Topology of Privacy: Lattice Structures and Information Bubbles for Inference and Obfuscation

Michael Erdmann*
Carnegie Mellon University
December 12, 2017

Abstract

Information has intrinsic geometric and topological structure, arising from relative relationships beyond absolute values or types. For instance, the fact that two people did or did not share a meal describes a relationship independent of the meal's ingredients. Multiple such relationships give rise to relations and their lattices. Lattices have topology. That topology informs the ways in which information may be observed, hidden, inferred, and dissembled. Privacy preservation may be understood as finding isotropic topologies, in which relations appear homogeneous. Moreover, the underlying lattice structure of those topologies has a temporal aspect, which reveals how isotropy may contract over time, thereby puncturing privacy.

Dowker's Theorem establishes a homotopy equivalence between two simplicial complexes derived from a relation. From a privacy perspective, one complex describes individuals with common attributes, the other describes attributes shared by individuals. The homotopy equivalence is an alignment of certain common cores of those complexes, effectively interpreting sets of individuals as sets of attributes, and vice-versa. That common core has a lattice structure. An element in the lattice consists of two components, one being a set of individuals, the other being an equivalent set of attributes. The lattice operations join and meet each amount to set intersection in one component and set union followed by a potentially privacy-puncturing inference in the other component.

One objective of this research has been to understand the topology of the Dowker complexes, from a privacy perspective. First, privacy loss appears as simplicial collapse of free faces. Such collapse is local, but the property of fully preserving both attribute and association privacy requires a global condition: a particular kind of spherical hole. Second, by looking at the link of an identifiable individual in its encompassing Dowker complex, one can characterize that individual's attribute privacy via another sphere condition. This characterization generalizes to certain groups' attribute privacy. Third, even when long-term attribute privacy is impossible, homology provides lower bounds on how an individual may defer identification, when that individual has control over how to reveal attributes. Intuitively, the idea is to first reveal information that could otherwise be inferred. This last result highlights privacy as a dynamic process. Privacy loss may be cast as gradient flow. Harmonic flow for privacy preservation may be fertile ground for future research.

*This report is based upon work supported in part by the Air Force Office of Scientific Research under award number FA9550-14-1-0012 and in part by the National Science Foundation under award number IIS-1409003. Any opinions, findings and conclusions or recommendations expressed in this report are those of the author and do not necessarily reflect the views of the Government, the U.S. Department of Defense, or the National Science Foundation.

Contents

1 Introduction

Privacy is the ability of an individual or entity to control how much that individual or entity reveals about itself to others. Fundamental research into privacy seeks to understand the limits of that ability.

A brief history of privacy should include the following:

- **The right** to privacy as a legal principle, appearing in an 1890 Harvard Law Review article [24]. The article was a reaction to the then modern technology of photography and the dissemination of gossip via print media.

- **A demonstration** linking supposedly anonymous public information with other more specific public data, thereby revealing sensitive attributes [21]. The demonstration employed zip code, gender, and birth date to link anonymous public insurance summaries with voter registration data. Doing so produced the health record of the governor of Massachusetts. This privacy failure suggested a first form of homogenization, called *k-anonymity*. Roughly, the idea was to structure databases in such a way that a database could respond to any query with an answer consisting of no fewer than k individuals matching the query parameters.

- **The discovery** that it is impossible to preserve the privacy of an individual for even a single attribute in the face of repeated statistical queries over a population [2], *unless* answers to those queries are purposefully perturbed with noise of magnitude on the order of at least $\sqrt{n}$. Here n is the size of the population. The significance of this discovery is to underscore how difficult it is to preserve privacy while retaining information utility.

- **Netflix Prize**. In 2006, Netflix offered a $1M prize for an algorithm that would predict viewer preferences better than Netflix's internal algorithm. Netflix made available some of its historical user preferences, in anonymized form, as a basis for the competition. Once again, it turned out that one could link this anonymized data with other publicly available databases, resulting in the potential (and in some cases actual) identification of Netflix viewers, thereby de-anonymizing their viewing history [17]. Whereas in the earlier health example, a few specific observables made linking possible (global coordinates, one might say, namely zip code, gender, birth date), in the Netflix example, the intrinsic geometric structure of the database facilitated linking via a wide variety of observables (local landmarks, one might say, namely movies that were characteristic for each individual). Key was sparsity of information: 8 movie ratings and dates were generally enough to uniquely characterize 99% of viewers in the Netflix Prize dataset, even with errors in the ratings and dates.

- **Differential Privacy** [5, 4] seeks to avoid the previous privacy failures by focusing on local rather than absolute privacy guarantees. The underlying approach in differential privacy is for a database to answer statistical queries with a particular stochastic blurring. Specifically, the probability that an interrogator of the database will make any particular inference should depend only in a very small way on whether any one individual does or does not have a particular attribute (such as even being in the database). We might call this *stochastic homogeneity*.

- **Randomized Response**. Differential privacy is further significant because it makes explicit the dynamic nature of privacy; there may be no enduring privacy guarantees but there are differential guarantees. A particular form is *randomized response*, a technique used in the social sciences to elicit reliable aggregate answers to sensitive questions, asking the question of many people, but perturbing individual answers stochastically so as not to learn much about any one individual from any single response [23]. A version has been employed by Google to find malware [8].

Privacy has both a combinatorial component and a statistical component. Prior research has largely focused on statistical techniques, both to preserve privacy and to puncture privacy. One of the goals of this research is to understand the combinatorial component of privacy, leading naturally to methods from combinatorial topology.

A desire to understand the geometry and topology of the types of inferences revealed by the Netflix Prize formed the specific motivation for our research initially. Subsequently, we realized that the lattice structure found in that geometry had broader applicability, providing an ability to model the dynamics of privacy more generally.

2 Outline

The remaining sections and appendices present the following material:

Main Narrative:

3: Toy examples illustrating how a relation may lead to privacy loss in the presence of background information. The section introduces the *doubly-labeled poset* associated with a relation, to model such inferences. The elements of the poset are ordered pairs, each a set of individuals and a set of attributes.

This section also states and discusses assumptions that hold throughout the report.

4: Formal description of the *Galois connection* associated with a relation. The section first defines, for any relation, two simplicial complexes called *Dowker complexes*. One complex represents sets of individuals with shared attributes, the other represents sets of attributes shared by individuals. The Galois connection then establishes a homotopy equivalence between the Dowker complexes, thereby generating the relation's doubly-labeled poset. The homotopy equivalence gives rise to closure operators, with "closure" in the poset modeling inference of unobserved attributes from observed attributes (or unobserved individuals from observed individuals).

This section also defines *attribute privacy* and *association privacy.*

5: A characterization of privacy in terms of the absence of free faces in the relevant Dowker complex. This section observes as well that the only connected relations able to preserve both attribute and association privacy must look either like linear cycles or like boundary complexes. In particular, the number of individuals and attributes must be the same.

6: Conditional relations, as models for simplicial links. A conditional relation is much like a conditional probability distribution. It might, for instance, represent the possible arrangement of remaining attributes among individuals, after some attributes have already been observed.

7: A characterization of individual and group attribute privacy in terms of spherical and boundary complexes for the relation that models the individual's or group's link in its Dowker complex.

8: A brief exploration of holes in relations, focusing on attribute spaces generated by bits.

9: A small example exploring the possibility of increasing privacy by change-of-coordinate transformations.

10: A lengthy exploration of how someone can delay identification, by releasing attributes selectively in a particular order. This idea leads to the notion of *informative attribute release sequences*, how to find such sequences in the *Galois lattice*, and the use of homology as a lower bound for the number and length of such sequences.

11: Computation of the homology and maximal informative attribute release sequences present in two relations found on the world wide web. One relation describes Olympic athletes and their medals, the other describes jazz musicians and their bands.

12: A more general perspective of inference as motion in lattices, not necessarily directly derived from a relation. This perspective suggests connections to randomized response techniques.

13: An examination of the ability to obfuscate strategies and/or goals in graphs where motions may be nondeterministic or stochastic.

14: A possible category for representing relations, along with an analysis of morphism properties. The morphisms between relations in this category induce simplicial and therefore continuous maps between the relations' corresponding Dowker complexes.

This section further shows by example how a morphism of relations, when it is surjective at the set level, generates the full lattice of the codomain's relation, via closure under lattice operations. (A general proof appears in Appendix I.)

15: Some thoughts for the future, including an example that connects stochastic sensing to the Galois lattice.

Appendices:

A: A summary of the basic notation and definitions used in this report.

B: A summary of the basic tools used in this report, establishing the homotopy equivalences and closure operators mentioned previously.

C: Construction of links and deletions, and examination of the privacy properties each inherits from its encompassing relation. This appendix explores the significance of free faces in the Dowker complexes. The appendix further proves that a relation with more attributes than individuals cannot preserve attribute privacy for every individual.

D: Proof that the problem of finding a minimal set of attributes from which another attribute may be inferred is *NP*-complete. This stands in contrast to the observation that the problem of finding *some* set of attributes from which another may be inferred (or reporting that no such set exists) is computable in polynomial time.

E: Detailed proofs of the results claimed in Section 7. Also a detailed proof of the assertion from Section 5 regarding relations that preserve both attribute and association privacy.

F: Detailed proofs of the connection between maximal chains in a relation's Galois lattice and informative attribute release sequences. When such sequences are order-independent they correspond to spherical holes, leading to the concept of an *isotropic* sequence.

G: Detailed proof that homology establishes a lower bound for the number and length of maximal chains in a relation's Galois lattice, and thus for the number and length of informative attribute release sequences that may be used to delay identification.

H: An application of the previous results with the aim of obfuscating the identification of strategies for attaining goals in graphs with uncertain transitions.

I: Detailed proofs of the assertions of Section 14 regarding morphisms.

J: Some additional examples:

1. Dunce Hat: modeled as a relation for which the Dowker attribute complex is contractible but has no free attribute faces, meaning the relation preserves attribute privacy.
2. Disinformation: An example that glues together two copies of the Möbius strip, thereby removing free faces and creating a form of homogeneity that preserves attribute privacy yet retains the utility of identifiability.
3. Insufficient Representation: If there are insufficiently many individuals in a relation generated by bits, attribute inference is possible.
4. A Matching Example: When many individuals are being observed, cardinality constraints allow for inferences beyond those discussed in this report.

List of Primary Symbols

Symbol	Typical Meaning	Page(s)
X	discrete space of individuals	13, 88
Y	discrete space of attributes	13, 88
R	relation on $X \times Y$	13, 88
X_y	individuals with attribute y (usually in the context of relation R)	13, 88
Y_x	attributes of individual x (usually in the context of relation R)	13, 88
Q	another relation, often representing a link in a simplicial complex	24, 94, 40
Σ, Γ	generic simplicial complexes (sometimes merely sets)	84
Ψ_R	complex; simplices are sets of individuals with a common attribute	13, 89
Φ_R	complex; simplices are sets of attributes shared by some individual	13, 88
σ	usually a simplex representing individuals in Ψ_R	
γ	usually a simplex representing attributes in Φ_R	
ϕ_R	homotopy equivalence from sets of individuals to shared attributes	14, 89
ψ_R	homotopy equivalence from sets of attributes to sharing individuals	14, 89
P	partially ordered set (poset)	86
$\mathfrak{F}(\Sigma)$	face poset of the simplicial complex Σ	14, 86
$\Delta(P)$	order complex of the poset P	15, 86
P_R	doubly-labeled poset associated with relation R	11, 17, 89
L	(inference) lattice	(58) 87
P_R^+	Galois lattice formed from P_R	35, 89
$\{(\sigma_k, \gamma_k) < \cdots < (\sigma_0, \gamma_0)\}$	chain of length k in the lattice P_R^+	41, 118, 86
$y_1, \ldots, y_k$	informative attribute release sequence (iars) of length k (for relation R)	38
V	set of vertices in a simplicial complex or states in a graph	
$\partial(V)$	simplicial boundary complex with vertices V	21, 85
$\mathbb{S}^{-1}$	sphere of dimension -1, modeling the empty complex $\{\emptyset\}$	84
$\mathbb{S}^1$	circle	21
$\mathbb{S}^{n-2}$	sphere of dimension $n-2$	21, 85
$C_k(\Sigma; \mathbb{Z})$	group of simplicial k-chains over Σ, with integer coefficients	84
$\widetilde{\partial}$	(family of) reduced boundary map(s) $C_k(\Sigma; \mathbb{Z}) \to C_{k-1}(\Sigma; \mathbb{Z})$	85
$\widetilde{H}_k(\Sigma; \mathbb{Z})$	reduced k-dimensional homology group of Σ, with integer coefficients	85
G	a graph, generally with nondeterministic and/or stochastic actions	62, 64
Δ_G	strategy complex of a graph	63, 64
$\overline{\Delta}_G$	source complex of a graph	128
$\simeq$	homotopy equivalence	86
$*$	simplicial join	86
$\vee$	either topological wedge sum or lattice join	86, 87
$\wedge$	lattice meet	87

3 Privacy: Relations and Partially Ordered Sets

Our investigation of privacy in this report will be in terms of relations. As we will see in this section and the next, relations give rise to simplicial complexes, which give rise to partially ordered sets, which expose an underlying lattice structure. That lattice structure makes explicit how privacy may be preserved or lost through so-called *background knowledge*. As we will see in Section 10, the lattice structure also makes explicit how identification may be delayed by careful release of information.

3.1 A Toy Example: Health Data and Attribute Privacy

Consider the following relation H, describing the results of a hypothetical health study for four patients and three attributes. The patients have been anonymized and are represented simply by the set of numbers $\{1, 2, 3, 4\}$. The three attributes are drawn from the set $\{\text{SMOKES}, \text{HAS\_CANCER}, \text{DRINKS\_SODA}\}$.

One can describe a relation equivalently either as a matrix or as a set of ordered pairs:

Relation H as a matrix:

H	SMOKES	HAS_CANCER	DRINKS_SODA
1	•	•	
2		•	•
3			•
4			•

Relation H as a set of ordered pairs:

$$\big\{(1, \text{SMOKES}),\ (1, \text{HAS\_CANCER}),\ (2, \text{HAS\_CANCER}),\ (2, \text{DRINKS\_SODA}),$$
$$(3, \text{DRINKS\_SODA}),\ (4, \text{DRINKS\_SODA})\big\}.$$

Assumptions

Before discussing privacy further, we make some assumptions that hold throughout the report:

Assumption of Relational Completeness: We assume that any given relation *is not missing any observable elements*, relative to some external (unspecified) ground truth.

For example, if we observe that someone drinks soda and has cancer in relation H, then we would conclude that we are observing individual #2. We would be surprised to see that individual smoke. If for some reason we ever do see the individual smoke, then we would deem our observations to be *inconsistent* with relation H. — The meaning of inconsistency depends on context. At top-level, an inconsistency may mean that the relation or observation is errorful. When making conditional observations, an inconsistency may actually supply useful information, as we will see in Lemma 12 on page 26.

Comment: A relation may contain extra elements, as may be useful for disinformation. A relation could even be missing elements that represent valid ordered pairs, so long as those elements are deemed to be unobservable for that relation. For example, one may have a time series of relations in which some attributes only become observable at later times. In such a setting, one may never know whether a particular individual had a particular attribute at an earlier time.

In the example, it could be that individual #1 drinks soda, but that it is impossible to observe this fact. In that case, relation H would still satisfy the assumption of relational completeness, even though H contains no entry[1] indicating that individual #1 drinks soda.

Assumption of Observational Monotonicity: Even though we assume *relations* are complete, we do *not* assume that *observations* are complete. Instead, we assume: *The observation of a particular attribute for an individual is meaningful; lack of such an observation does not necessarily imply that the individual fails to have the unobserved attribute.* The motivation for this assumption is that one may yet discover that the individual has the attribute. For example, suppose we observe someone (whom we know to be part of relation H) drinking soda. Even if that is all we observe, we do *not* conclude that the individual is cancer free. It could be that we might yet observe the individual to have cancer.

If absence of an attribute is significant *and* that absence is observable, then both the attribute and its negation could and perhaps should appear explicitly in the relation as distinct mutually exclusive attributes. For instance, Prime versus Composite might be such a pair of attributes for integers greater than 1.

Assumption of Observational Accuracy: We assume that *observations are accurate.* For instance, if we observe an integer to be either Prime or Composite, then we do so correctly.

Comments: The three assumptions above are *desiderata* for how the mathematical abstractions of this report fit into the real world. Some comments are in order:

- In and of itself, a relation defines a particular kind of world, a bipartite graph, and there is no external ground truth.
- In such a world, the completeness, monotonicity, and accuracy assumptions describe a sensor and the meaning of observations made by the sensor.

The purpose of the assumptions in the real world is largely to ensure consistency between different relations and with possible observations.

- The monotonicity assumption is important because information generally aggregates asynchronously. Together with the other assumptions, this assumption means that one may view relations as monotone Boolean functions, and thus may leverage methods from combinatorial topology.
- One may incorporate some errors into the relational and observational models, for instance by blurring a relation. For very large integers, a relation might allow some integers to have *both* Prime and Composite as attributes. Although an integer is one

[1] Terminology: We often use the term `'entry'` to mean an element of a relation, as in a matrix, or in one of its rows or columns.

or the other, the relation admits to uncertainty by allowing both attributes at once. Indeed, some relations purposefully introduce such blurring to preserve privacy, as with randomized response [23]. In robotics, natural relational blurring arising from noisy but environment-compatible sensors can actually help establish the topology of a region, for instance by dualizing sensors and landmarks [11].

Privacy Implications

Making the health study H of page 7 publicly available has some privacy implications, including the following:

- Suppose someone named Bob tells his friend Alice that he was part of the study. Alice knows that Bob smokes everywhere he goes, so she can infer that he is Patient #1 and has cancer. (This is an example of inference in a relation using background knowledge.)
- Suppose Cindy is Patient #2. She has full attribute privacy as far as relation H is concerned. In particular, as we saw already, Cindy can tell her friends that she was part of the health study while drinking soda and those friends will not be able to conclude that she has cancer.
- Patients #3 and #4 are not only indistinguishable from each other but also from Cindy (patient #2), as far as relation H is concerned. This is a very strong form of anonymity. Even if one of them reveals that s/he drinks soda, s/he will remain indistinguishable from the other two patients who drink soda.

Caveat: In the last case, if Cindy reveals that she has cancer and is seen to be different from the other individuals, then one may be able to remove her from the relation, narrowing the focus and creating a new relation that may allow additional inferences. Similar caveats hold for the other bullets. Deletions are discussed further in Appendix C.

Modifying a Relation to Increase Privacy We can make a small change in relation H that enhances privacy. If we artificially give patient #3 the attribute SMOKES, then we obtain the following modified relation H':

H'	SMOKES	HAS_CANCER	DRINKS_SODA
1	•	•	
2		•	•
3	•		•
4			•

Now Bob may reveal to Alice that he was part of the health study without Alice being able to infer that he has cancer, even though she knows that everyone knows that he smokes. In fact, more generally, one can no longer infer cancer from smoking, within the relation.

Such an artificial entry in the relation is a form of *disinformation*. It certainly skews statistics and utility. It also increases privacy.

3.2 A Dual Perspective: Payroll Data and Association Privacy

The previous example examined a relation from the perspective of *attribute privacy*: we were interested in understanding how observation of some attribute(s) implied other attribute(s), possibly identifying an individual. A dual perspective is *association privacy*, in which one seeks to understand how some associations between individuals imply others.

The following hypothetical "salary" relation S has the same matrix structure as relation H did earlier, but with different semantics. This relation represents employees $\{\text{Bob}, \text{Mary}, \text{Frank}, \text{Julie}\}$ working on secret projects $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Now the employee names are visible so that a payroll clerk can disburse salaries correctly, but the actual projects are anonymous.

S	a	b	c
Bob	•	•	
Mary		•	•
Frank			•
Julie			•

The salary relation S has some implications for association privacy, including the following:

- If someone tells the payroll clerk that Julie is the lead of a very important project with valuable information, then the payroll clerk can infer that Mary and Frank have also been exposed to valuable information.
- In contrast, if someone tells the payroll clerk that Bob is running a very important project, then the payroll clerk does not have enough information to conclude that Mary is also working on an important project.

Regarding disinformation: Observe how adding the artificial entry $(\text{Julie}, \mathtt{a})$ prevents the payroll clerk from using the relation to infer that Mary and Frank have valuable information, even if the payroll clerk learns via background information that Julie is the lead of a very important project with such information:

S'	a	b	c
Bob	•	•	
Mary		•	•
Frank			•
Julie	•		•

3.3 Privacy Preservation and Loss: A Poset Model

R	a	b	c
1	•	•	
2		•	•
3			•
4			•

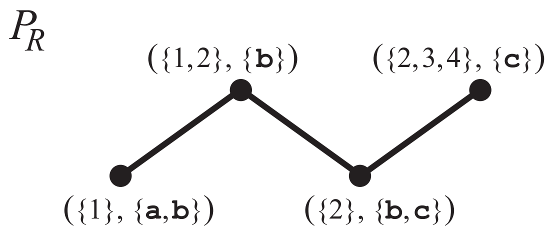


Figure 1: Relation R serves as a model for the two examples of Sections 3.1 and 3.2. The doubly-labeled poset P_R describes the inferences facilitated by R.

Figure 1 shows a relation R that serves as a model for both the health example of Section 3.1 and the payroll example of Section 3.2. The relation is identical to those given earlier, but with abstract labels in place of both individuals and attributes. The figure also depicts a partially ordered set (poset) P_R, designed to model the inferences discussed previously. We refer to that poset as the *doubly-labeled poset associated with* R. We next discuss the semantics of P_R. Section 4 discusses the construction of P_R. The underlying concepts are important throughout the report.

Semantics of the poset P_R:

- Each element in the poset consists of an ordered pair (σ, γ), with $\emptyset \neq \sigma \subseteq \{1, 2, 3, 4\}$ describing a set of individuals and $\emptyset \neq \gamma \subseteq \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ describing a set of attributes. We say that the poset element is *labeled with* σ *and* γ. The meaning of such a double-labeling (with respect to the information described by relation R) is:

  (a) All individuals in σ have all attributes in γ.

  (b) If (and only if) an individual has at least all the attributes in γ, then that individual must be in σ. For example, we see that individual #2, and only individual #2, has both attributes `b` and `c` in R.

  (c) If (and only if) an attribute is shared by at least all individuals in σ, then that attribute must be in γ. For example, individual #1 has both attributes `a` and `b`, so P_R cannot contain simply $(\{1\}, \{\mathtt{a}\})$, but must contain $(\{1\}, \{\mathtt{a}, \mathtt{b}\})$.

- The partial order for P_R is described by the edges in the figure. There is an edge between two elements (σ_1, γ_1) and (σ_2, γ_2) of P_R whenever the corresponding sets are subset comparable. In particular, $(\sigma_1, \gamma_1) \leq (\sigma_2, \gamma_2)$ in P_R precisely when $\sigma_1 \subseteq \sigma_2$ and $\gamma_1 \supseteq \gamma_2$. [Observe that the comparability ($\subseteq$ versus $\supseteq$) is opposite for σ versus γ.]

Using the poset P_R for attribute inference:

Suppose γ is *any* nonempty subset of attributes in $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Then one of (i) or (ii) holds:

(i) Perhaps no individual modeled by R has all the attributes γ. For example, no individual has attributes $\gamma = \{\mathtt{a}, \mathtt{c}\}$. We would not expect to see γ and so γ does not appear in the poset P_R.

(ii) Alternatively, γ is a subset of at least one set of attributes that does appear in the poset. In this case, one *may* be able to enlarge γ nontrivially, resulting in privacy loss.

For example, imagine we discover that a friend with attribute `a` is modeled by the given relation (e.g., Bob, who SMOKES, says he is part of the health study H).
Using $\gamma = \{\mathtt{a}\}$, the poset then allows us to infer that Bob must also have attribute `b` (that is, HAS_CANCER). Why? Because $\{\mathtt{a}, \mathtt{b}\}$ is a minimal set in P_R containing $\{\mathtt{a}\}$.

We can say yet more: The element labeled with $\{\mathtt{a}, \mathtt{b}\}$ is also labeled with $\{1\}$. So now we have *de-anonymized* individual #1 (identifying him to be Bob).

Regardless of whether Bob ever actually talks to us, the poset tells us that individual #1 *could* suffer privacy loss, and in fact, is uniquely identifiable in the context of relation R without needing to reveal everything about himself.

Similar reasoning is possible for **association inference**, as we saw earlier.

R'	a	b	c
1	•	•	
2		•	•
3	•		•
4			•

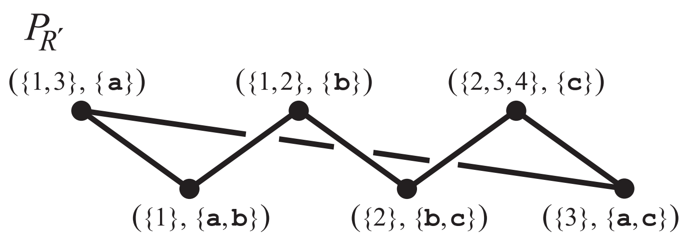


Figure 2: A relation R', along with its doubly-labeled poset $P_{R'}$. The relation preserves attribute privacy but allows a small amount of association inference: If ones sees individual #4 in some context `c`, then one can infer that individuals #2 and #3 are also present in that same context, without needing to observe them directly.

Disinformation Revisited: Figure 2 shows relation R', constructed from R by adding an entry of disinformation, much as we constructed H' from H earlier. The figure also shows the corresponding doubly-labeled poset $P_{R'}$. Observe that it is no longer possible to infer $\{\mathtt{a}, \mathtt{b}\}$ from $\{\mathtt{a}\}$, because $\{\mathtt{a}\}$ now appears directly in the poset. The added entry $(3, \mathtt{a})$ in R' has increased attribute privacy compared to R.

There is, however, still some opportunity for making association inferences. For instance, knowing that individual #4 (Julie, earlier) works on an important secret project still allows the inference that individuals #2 and #3 have valuable information. That is because the minimal set containing $\{4\}$ in the poset is $\{2, 3, 4\}$. Notice that no such association inference is possible if someone says that individual #3 works on an important secret project, though that would have been possible in the original relation R.

Comment: Artificial entries can potentially also produce inferences of disinformation. For instance, if, in our earlier relation H, the entry $(1, \text{HAS\_CANCER})$ is artificial, then inferring that Bob has cancer from his smoking, when in fact Bob is healthy, would be disinformation.

4 The Galois Connection for Modeling Privacy

Section 3 showed by example how a relation determines a partially ordered set (poset) useful for modeling privacy. The elements in the poset are ordered pairs — a set of attributes and a set of individuals — that are equivalent from the relation's perspective. Privacy loss occurs when an observer has data (for example, background knowledge) that is not directly in the poset but is a proper subset of some set of attributes or individuals in the poset. The observer may then infer some additional attributes or individuals. This section develops the connection between relations and posets more precisely, continuing to use the earlier examples for illustration. See also Appendices A and B for notation and additional material.

4.1 Dowker Complexes

Definition 1 (Dowker Complexes)**.** *Let X and Y be finite discrete spaces and let R be a relation on $X \times Y$. This means R is a set of ordered pairs (x, y), with $x \in X$ and $y \in Y$. We frequently view/depict R as a matrix of 0s and 1s, or as a matrix of blank and nonblank entries, with X indexing rows and Y indexing columns.*

(a) We often refer to elements of X as individuals *and to elements of Y as* attributes*.*

(b) For each $x \in X$, let $Y_x = \{y \in Y \mid (x, y) \in R\}$. Then Y_x consists of all attributes of individual x. We may view Y_x as a row *of R. We say that the* row is blank *if $Y_x = \emptyset$.*

(c) For each $y \in Y$, let $X_y = \{x \in X \mid (x, y) \in R\}$. Then X_y consists of all individuals who have attribute y. We may view X_y as a column *of R. The* column is blank *if $X_y = \emptyset$.*

(d) We next define two simplicial complexes Φ_R and Ψ_R (with some special cases below):

$$\Phi_R = \{\gamma \subseteq Y \mid \textit{there exists } x \in X \textit{ such that } (x, y) \in R \textit{ for all } y \in \gamma\},$$

$$\Psi_R = \{\sigma \subseteq X \mid \textit{there exists } y \in Y \textit{ such that } (x, y) \in R \textit{ for all } x \in \sigma\}.$$

Special cases: If $X = \emptyset$ and/or $Y = \emptyset$, then we say the relation is void*. In this case, with some exceptions discussed later (see Section 6, Section 10, and Appendix C), we let Φ_R and Ψ_R each be an instance of the void complex, containing no simplices. Otherwise, with X and Y both nonempty, each of Φ_R and Ψ_R contains at least the empty simplex $\emptyset$.*

We refer to Φ_R and Ψ_R as Dowker complexes*, after the author of upcoming Theorem 2. We say that each complex is the* Dowker dual *of the other, with respect to relation R.*

Interpretation: A nonempty set γ of attributes is a simplex in Φ_R precisely when at least one individual has at least all the attributes in γ. We refer to any such individual as a witness for γ*.*

Similarly, a nonempty set σ of individuals is a simplex in Ψ_R precisely when there is at least one attribute that is shared by at least all the individuals in σ. We refer to any such attribute as a witness for σ*.*

R	a	b	c
1	•	•	
2		•	•
3			•
4			•

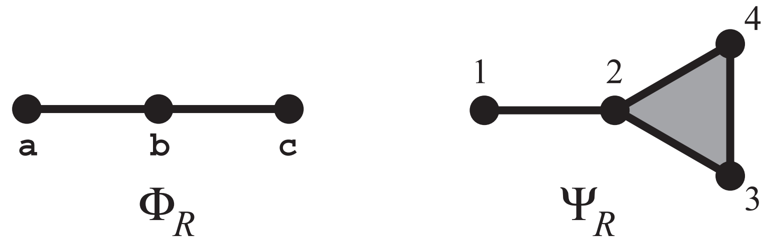


Figure 3: Dowker simplicial complexes Φ_R and Ψ_R determined by relation R.

Figure 3 shows the Dowker complexes for the relation R of Section 3.3.

Dowker's Theorem [3, 1] says that the two simplicial complexes Φ_R and Ψ_R have the same homotopy type. As we will see, the maps establishing that homotopy equivalence define the doubly-labeled poset P_R and describe how privacy may be lost.

Theorem 2 (Dowker Duality [3])**.** *Suppose R is a relation on $X \times Y$. Let Φ_R and Ψ_R be as in Definition 1. Then Φ_R and Ψ_R are homotopy equivalent.*

Every nonvoid simplicial complex Σ determines a partially ordered set $\mathfrak{F}(\Sigma)$ called the *face poset* of Σ. The elements of this poset are the *nonempty* simplices of Σ, partially ordered by set inclusion. (Recall that `'poset'` is short for `'partially ordered set'`.)

For the finite setting, the homotopy equivalence of Dowker's Theorem may be seen by explicit formulas for maps between the face posets of the two Dowker complexes. These maps describe what is known as a *Galois connection.* [This construction also appears as a core tool within the field of Formal Concept Analysis [25, 10].] Here are the formulas:

$$\phi_R \;:\; \mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Phi_R) \qquad\qquad \psi_R \;:\; \mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Psi_R)$$

$$\sigma \mapsto \bigcap_{x\in\sigma} Y_x \qquad\qquad \gamma \mapsto \bigcap_{y\in\gamma} X_y$$

These two maps are inverse homotopy equivalences. One sees this by considering the maps $\phi_R \circ \psi_R$ and $\psi_R \circ \phi_R$. These compositions turn out to be what are called *closure operators* on the face posets $\mathfrak{F}(\Phi_R)$ and $\mathfrak{F}(\Psi_R)$, respectively, implying that each is homotopic to an identity map, thereby establishing the desired homotopy equivalence. See Appendix B for detailed computations; see the next subsection for interpretation.

4.2 Inference from Closure Operators

An order-preserving poset map $f : P \to P$ is said to be a *closure operator* whenever $x \leq f(x)$ and $f(f(x)) = f(x)$ for all $x \in P$. If f is a closure operator, then it induces a homotopy equivalence between P and the image $f(P)$. See [1, 22, 19, 18] for more details.

One can think of a closure operator as "pushing elements up" in the poset. From a privacy perspective, "pushing up" amounts to inference. Specifically, $(\phi_R \circ \psi_R)(\gamma) \setminus \gamma$ consists of all additional attributes that may be inferred from observing attributes γ, while $(\psi_R \circ \phi_R)(\sigma) \setminus \sigma$ consists of all additional individuals that may be inferred from observing individuals σ.

Comment: The formulas for ϕ_R and ψ_R in Section 4.1 extend to the empty simplex and to the spaces X and Y, suggesting "inferences from nothing": Observe that $\psi_R(\emptyset) = X$, so $(\phi_R \circ \psi_R)(\emptyset)$ consists of all attributes that every individual in X has. If $(\phi_R \circ \psi_R)(\emptyset) \neq \emptyset$, then the attributes $(\phi_R \circ \psi_R)(\emptyset)$ are inferable "for free" from R, that is, without making any observations. Similarly, $(\psi_R \circ \phi_R)(\emptyset)$ consists of all individuals who have every attribute in Y.

Any poset P defines a simplicial complex $\Delta(P)$ called the *order complex* of P. The simplices of $\Delta(P)$ are given by the finite chains $\{p_0 < p_1 < \cdots < p_n\}$ in P. Suppose we start with a nonvoid simplicial complex Σ, construct its face poset $\mathfrak{F}(\Sigma)$, and then construct the order complex $\Delta(\mathfrak{F}(\Sigma))$. The result is isomorphic to the *first barycentric subdivision* of Σ [20, 22]. A convenient visualization of the face posets $\mathfrak{F}(\Phi_R)$ and $\mathfrak{F}(\Psi_R)$ therefore is to draw the first barycentric subdivisions of Φ_R and Ψ_R, respectively, as in Figure 4.

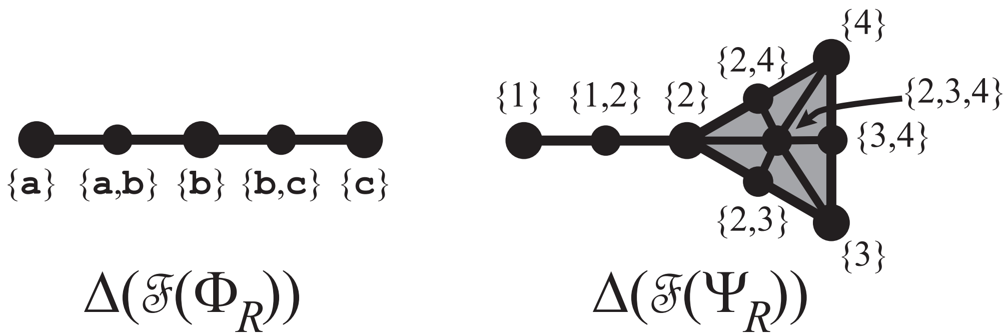


Figure 4: Order complexes of the face posets of the complexes Φ_R and Ψ_R shown in Figure 3.

Viewed in the order complexes, functions ψ_R and ϕ_R are easy to visualize. They are fully determined by their actions on vertices of the order complexes, as shown in Table 1. (Bear in mind that each element of $\mathfrak{F}(\Phi_R)$ represents a simplex in Φ_R but is a vertex in $\Delta(\mathfrak{F}(\Phi_R))$. Similarly, each element of $\mathfrak{F}(\Psi_R)$ represents a simplex in Ψ_R but is a vertex in $\Delta(\mathfrak{F}(\Psi_R))$.)

γ	$\psi_R(\gamma)$	$(\phi_R \circ \psi_R)(\gamma)$
{a}	{1}	{a, b}
{b}	{1, 2}	{b}
{c}	{2, 3, 4}	{c}
{a, b}	{1}	{a, b}
{b, c}	{2}	{b, c}

σ	$\phi_R(\sigma)$	$(\psi_R \circ \phi_R)(\sigma)$
{1}	{a, b}	{1}
{2}	{b, c}	{2}
{3}	{c}	{2, 3, 4}
{4}	{c}	{2, 3, 4}
{1, 2}	{b}	{1, 2}
{2, 3}	{c}	{2, 3, 4}
{3, 4}	{c}	{2, 3, 4}
{2, 4}	{c}	{2, 3, 4}
{2, 3, 4}	{c}	{2, 3, 4}

Table 1: The maps ψ_R and ϕ_R, and their compositions, for relation R of Figure 3.

Using Table 1, one can again see how privacy loss might occur via R.
For instance, the map $\phi_R \circ \psi_R$ gives rise to the closure (i.e., a "pushing up")

$$\{\mathtt{a}\} \xmapsto{\psi_R} \{1\} \xmapsto{\phi_R} \{\mathtt{a},\mathtt{b}\},$$

telling us how to infer unobserved attribute `b` from observed attribute `a` (in the health study example of Section 3.1, Alice could infer that Bob HAS_CANCER from knowing that he SMOKES). Similarly, for the map $\psi_R \circ \phi_R$,

$$\{4\} \xmapsto{\phi_R} \{\mathtt{c}\} \xmapsto{\psi_R} \{2,3,4\},$$

leading to association inference (in the payroll example from Section 3.2, the payroll clerk could infer Bob and Mary's exposure to valuable information after learning of Julie's work on an important project).

Figure 5 indicates the homotopy deformations produced by the maps $\phi_R \circ \psi_R$ and $\psi_R \circ \phi_R$, while Figure 6 shows the resulting image of each face poset.

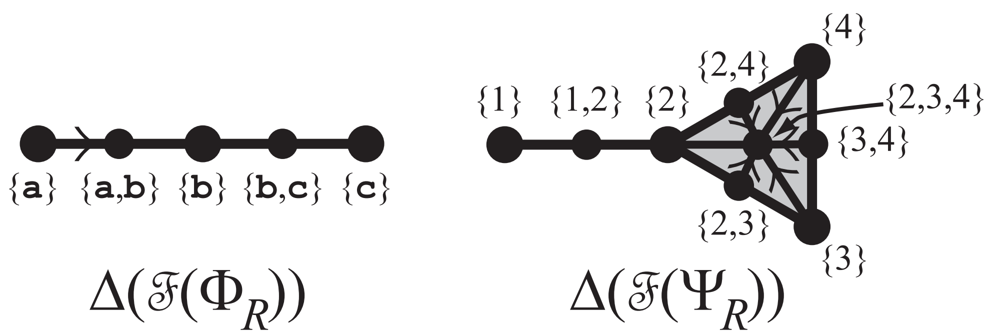


Figure 5: Closure operators $\phi_R \circ \psi_R$ and $\psi_R \circ \phi_R$ produce homotopy deformations, indicated by directed edges. In $\mathfrak{F}(\Phi_R)$, $\{\mathtt{a}\}$ closes up to $\{\mathtt{a},\mathtt{b}\}$. In $\mathfrak{F}(\Psi_R)$, most of the subsets of $\{2,3,4\}$ close up to $\{2,3,4\}$. The exception is subset $\{2\}$, which does not move.

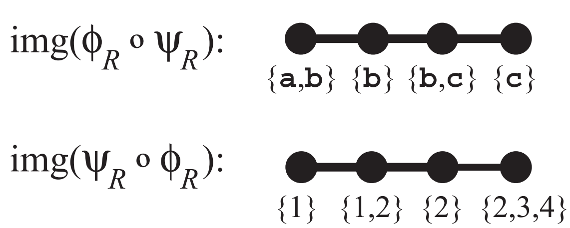


Figure 6: Images of the closure operators of Figure 5.

Observe that these two images are isomorphic. Matching up corresponding elements produces the poset P_R of Figure 1.

Summary: A relation R produces two simplicial complexes, Φ_R and Ψ_R, one modeling attributes shared by individuals, the other modeling individuals with common attributes. The complexes are related by two maps, ϕ_R and ψ_R, that are homotopy inverses. The compositions of these maps describe the attribute and association inferences possible via R, leveraging background information someone may have. These inferences are summarized by a poset P_R that pairs sets of individuals with sets of attributes. We may describe P_R as follows:

Definition 3 (Doubly-Labeled Poset)**.** *Let R be a relation with nonvoid Dowker complexes.*

The doubly-labeled poset P_R associated with R *consists of all ordered pairs of sets (σ, γ) such that $\emptyset \neq \sigma \in \Psi_R$, $\emptyset \neq \gamma \in \Phi_R$, $\sigma = \psi_R(\gamma)$, and $\gamma = \phi_R(\sigma)$.*

The partial order on P_R is defined by: $(\sigma_1, \gamma_1) \leq (\sigma_2, \gamma_2)$ if and only if $\sigma_1 \subseteq \sigma_2$ (and/or, equivalently, $\gamma_1 \supseteq \gamma_2$).

See Appendix A.4, specifically page 89, for some special cases.

(This definition agrees with the intuition that P_R is both the image $(\psi_R \circ \phi_R)(\mathfrak{F}(\Psi_R))$ and the image $(\phi_R \circ \psi_R)(\mathfrak{F}(\Phi_R))$, by Appendix B.)

4.3 Attribute and Association Privacy

Here are formal definitions for the intuition developed via the previous examples:

Definition 4 (Attribute Privacy)**.** *Let R be a relation with nonvoid Dowker complexes.*
We say that R preserves attribute privacy *precisely when*
$\phi_R \circ \psi_R$ is the identity operator on the poset $\mathfrak{F}(\Phi_R) \cup \{\emptyset\}$.

Definition 5 (Association Privacy)**.** *Let R be a relation with nonvoid Dowker complexes.*
We say that R preserves association privacy *precisely when*
$\psi_R \circ \phi_R$ is the identity operator on the poset $\mathfrak{F}(\Psi_R) \cup \{\emptyset\}$.

Comment: For notational simplicity, we frequently say simply that *$\phi_R \circ \psi_R$ is the identity on Φ_R* and/or that *$\psi_R \circ \phi_R$ is the identity on Ψ_R*.

4.4 Disinformation Example Re-Revisited

Recall the relation R' of Figure 2 on page 12, which is relation R of Figure 1 but with an added entry of disinformation. Figure 7 displays the resulting Dowker complexes and the actions of the closure operators. Figure 8 flattens out the poset $P_{R'}$ of Figure 2, so one sees its triangle structure and how it is the image of the Dowker complexes under the closure operators for R'.

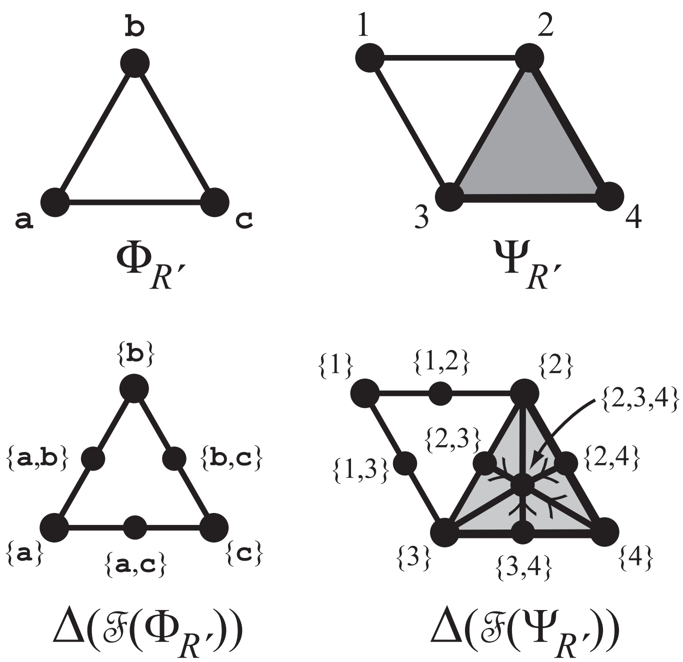


Figure 7: The Dowker complexes, as well as the order complexes of their face posets, for the relation R' of Figure 2 on page 12. The closure operator $\phi_{R'} \circ \psi_{R'}$ is the identity on $\mathfrak{F}(\Phi_{R'})$. The closure operator $\psi_{R'} \circ \phi_{R'}$ on $\mathfrak{F}(\Psi_{R'})$ closes many (but not all) subsets of $\{2, 3, 4\}$ up to $\{2, 3, 4\}$, as indicated by the directed arrows. The result is a poset isomorphic to the poset $P_{R'}$ of Figure 2, drawn again slightly differently in Figure 8. Also, $(\phi_{R'} \circ \psi_{R'})(\emptyset) = \emptyset$. Thus relation R' preserves attribute privacy but not association privacy.

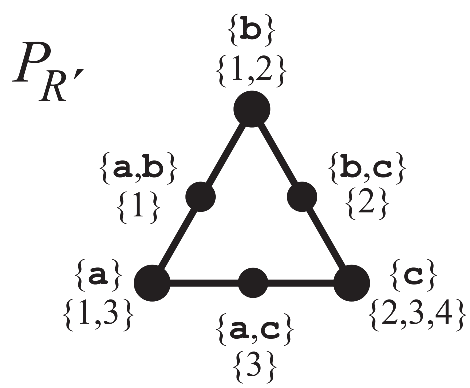


Figure 8: A flattened view of the doubly-labeled poset $P_{R'}$ from Figure 2. Combined with Figure 7, this perspective shows how $P_{R'}$ arises as the images of $\mathfrak{F}(\Phi_{R'})$ and $\mathfrak{F}(\Psi_{R'})$ under the closure operators $\phi_{R'} \circ \psi_{R'}$ and $\psi_{R'} \circ \phi_{R'}$, respectively. (The vertices drawn as bigger dots in the current figure were higher up in the poset of Figure 2 than those drawn as smaller dots.)

5 The Face Shape of Privacy

R	a	b	c
1	•	•	
2		•	•
3			•
4			•

R'	a	b	c
1	•	•	
2		•	•
3	•		•
4			•

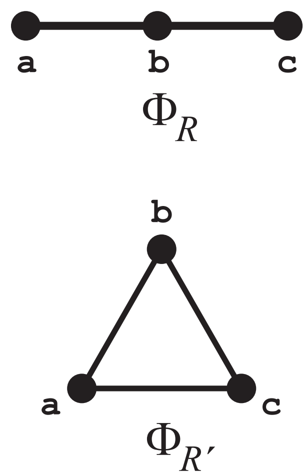


Figure 9: Relations R and R' of Section 3, along with their attribute complexes Φ_R and $\Phi_{R'}$.

5.1 Free Faces

Figure 9 recapitulates relation R and R' from the previous two sections, along with their Dowker attribute complexes, Φ_R and $\Phi_{R'}$, respectively. Recall that in R one could make the inference $\texttt{a} \Rightarrow \texttt{b}$, but no such inference was possible in R'.

The structure of Φ_R suggests that the inference $\texttt{a} \Rightarrow \texttt{b}$ *might* be possible in R. In contrast, the structure of $\Phi_{R'}$ makes clear that such an inference is *impossible* in R'. In particular, observe how vertex a has only one incident edge in Φ_R but has two incident edges in $\Phi_{R'}$. The fact that there are two edges in $\Phi_{R'}$, with those edges being maximal simplices, means, intuitively, that vertex a is being "pulled" in two different inference directions, so one cannot conclude anything additional from attribute a. In contrast, in Φ_R, vertex a is being "pulled" only toward b, so it is plausible that attribute a might imply attribute b.

The underlying geometry is that of a free face. A simplex γ of a simplicial complex Γ is said to be a *free face* of Γ if it is a proper subset of exactly one maximal simplex of Γ. That is true for $\{\texttt{a}\}$ in Φ_R but not for $\{\texttt{a}\}$ in $\Phi_{R'}$.

Of course, vertex $\{\texttt{c}\}$ also forms a free face in Φ_R, yet one cannot make any inferences upon observing just attribute c. What is going on? The difference is that c is also an attribute of individuals in R who have *only* c as an attribute (specifically, individuals #3 and #4). Even though $\{\texttt{c}\}$ is technically a free face of Φ_R, it is not really free to move under the closure operator $\phi_R \circ \psi_R$, whereas $\{\texttt{a}\}$ is.

Observe that individuals #2, #3, and #4 all have attribute c, but only individual #2 has additional attributes. This means that individuals #3 and #4 cannot ever be identified uniquely in the context of relation R; they have effectively "camouflaged" themselves with individual #2, as far as relation R is concerned. If one disallows or disregards such camouflage, then the idea of a free face and privacy loss are equivalent. The following definition is useful:

Definition 6 (Unique Identifiability)**.** *Let R be a relation on $X \times Y$ and suppose $x \in X$. We say that x is* uniquely identifiable via relation R *when $\psi_R(Y_x) = \{x\}$.*

Suppose R is a relation. Appendix C.3 proves that if Φ_R has no free faces, then R preserves attribute privacy. For the converse, Appendix C.3 further proves that if R preserves attribute privacy *and* if every individual is uniquely identifiable, then Φ_R has no free faces. (Dual statements hold for association privacy.)

5.2 Privacy versus Identifiability

Section 5.1 hinted at the difference between privacy and identifiability. In relation I below ("I" for "individuality" or "identity"), every individual has exactly one attribute and that attribute uniquely identifies the individual. Relation I *preserves privacy* fully (assuming $n > 1$). It is impossible to make any attribute inferences. If Bob reveals that he has attribute y_7, then Alice cannot infer any additional attributes for Bob. She now knows that Bob is individual x_7 but cannot infer any additional attributes. He has himself revealed everything about himself that there is to know, as far as relation I is concerned.

I	y_1	y_2	$\cdots$	y_n
x_1	$\bullet$			
x_2		$\bullet$		
$\vdots$			$\ddots$	
x_n				$\bullet$

In contrast, all individuals in relation C (for "conformism" or "confusion") have exactly the same set of attributes. As a result, there is *no privacy*: one can predict all the attributes of any individual in the relation without making any observations. On the other hand, no individual is uniquely identifiable (assuming $n > 1$).

C	y_1	y_2	$\cdots$	y_n
x_1	$\bullet$	$\bullet$	$\cdots$	$\bullet$
x_2	$\bullet$	$\bullet$	$\cdots$	$\bullet$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_n	$\bullet$	$\bullet$	$\cdots$	$\bullet$

Homogeneity: Relation C exhibits a form of homogeneity often sought by anonymization or other privacy techniques. As we have suggested before, the utility of relation C is essentially zero, unless one makes the entries stochastic, so that some utility is encoded in the distribution.

The discussion of free faces in Section 5.1 suggests an alternative approach to homogeneity: one may preserve privacy and retain utility by choosing the geometry of the relation appropriately, for instance, so the space Φ_R exhibits sphere-like homogeneity. There will be considerable discussion of the importance of spheres in the rest of the report.

5.3 Spheres and Privacy

The attribute complex $\Phi_{R'}$ of Figure 9 is equal to a *boundary complex*, namely the boundary of the full simplex consisting of the attributes $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. We will denote boundary complexes by $\partial(V)$, with V some nonempty set. The simplices of $\partial(V)$ are all proper subsets of V. Boundary complexes are homotopic to spheres, specifically $\partial(V) \simeq \mathbb{S}^{n-2}$, with $n = |V|$. For $\Phi_{R'}$ of Figure 9, we have that $\Phi_{R'} = \partial(\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}) \simeq \mathbb{S}^1$. (In English: The Dowker attribute complex is the boundary of a triangle, so homotopic to a circle.)

More generally, if for some relation R on $X \times Y$, $\Phi_R = \partial(Y)$, then Φ_R cannot have any free faces and so R preserves attribute privacy.

Privacy and Utility: An important observation is that boundary complexes exhibit homogeneity but still permit identifiability. If $\Phi_R = \partial(Y)$, with $|Y| > 1$, and if no individual's attributes are a subset of another's attributes, then one can and needs to specify $|Y| - 1$ attributes in order to identify an individual. The boundary structure ensures that one cannot infer any attributes by specifying fewer than $|Y|-1$ attributes, yet retains the ability to identify every individual.

Appendix J.1 gives an example of a contractible space that preserves attribute privacy. Observe, however, that the number of attributes needed to identify an individual in that example is considerably less than the total number of attributes in the space. For a boundary complex, it is just one less.

Preserving Attribute and Association Privacy: A consequence of these observations is that if one wishes to preserve both attribute and association privacy with a connected relation, then one requires both Dowker complexes to look like spheres. More specifically, either both Dowker complexes are linear cycles of the same length or both are boundary complexes of the same dimension. In the latter case, the relation is isomorphic to a relation of the following form, in which the diagonal $\{(x_i, y_i)\}$ is blank but all other entries are present:

$$
\begin{array}{c|cccccc}
R & y_1 & y_2 & \cdots & \cdots & y_{n-1} & y_n \\
\hline
x_1 & & \bullet & \bullet & \cdots & \bullet & \bullet \\
x_2 & \bullet & & \bullet & \cdots & \bullet & \bullet \\
\vdots & \bullet & \bullet & & \ddots & \vdots & \bullet \\
\vdots & \vdots & \vdots & \ddots & & \bullet & \vdots \\
x_{n-1} & \bullet & \bullet & \cdots & \bullet & & \bullet \\
x_n & \bullet & \bullet & \bullet & \cdots & \bullet &
\end{array}
$$

See Appendix E.3, starting on page 108, for further details.

5.4 A Spherical Non-Boundary Relation that Preserves Attribute Privacy

Consider relation R as in Figure 10. Relation R preserves attribute privacy, since Φ_R has no free faces. The relation does not preserve association privacy. In particular, the quadrilaterals drawn for Ψ_R in the figure are actually tetrahedra. This means that the diagonals of the quadrilaterals are free faces. For instance, one would expect to infer individuals #1 and #6 as

R	a	b	c	d	e
1	•	•		•	
2	•		•	•	
3		•	•	•	
4	•	•			•
5	•		•		•
6		•	•		•

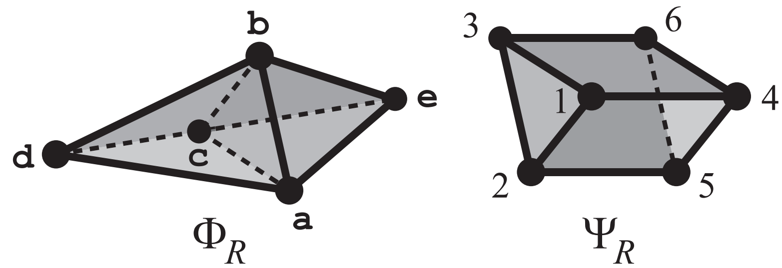


Figure 10: A relation R and its Dowker complexes Φ_R and Ψ_R, each homotopic to the two-dimensional sphere $\mathbb{S}^2$. (One may view Φ_R as two party hats glued together. One may view Ψ_R as a cylinder with a triangular cross-section and endcaps. However, the quadrilaterals drawn for the cylinder portion of Ψ_R are simply flattened sketches of what are actually solid tetrahedra.)

additional unobserved associates if one observes individuals #3 and #4. Indeed, computing using the closure operator $\psi_R \circ \phi_R$, we see that:

$$(\psi_R \circ \phi_R)(\{3,4\}) \;=\; \psi_R(\{\mathtt{b}\}) \;=\; \{1,3,4,6\}.$$

Relation R has another interesting feature. Even though Φ_R is not itself a boundary complex, it is the *simplicial join* (see page 86) of two boundary complexes:

$$\Phi_R \;=\; \partial(\{\mathtt{a},\mathtt{b},\mathtt{c}\}) * \partial(\{\mathtt{d},\mathtt{e}\}).$$

In fact, we can think of R as $R_1 \cup R_2$ and Φ_R as $\Phi_{R_1} * \Phi_{R_2}$, with R_1 the restriction of R to the attributes $\{\mathtt{a},\mathtt{b},\mathtt{c}\}$ and R_2 the restriction of R to the attributes $\{\mathtt{d},\mathtt{e}\}$. This join structure of Φ_R means that we can view every individual in R as being described by two *independent* attribute spaces. The attribute space $\{\mathtt{d},\mathtt{e}\}$ acts like a standard bit; every individual has exactly one of these two attributes. In contrast, the attribute space $\{\mathtt{a},\mathtt{b},\mathtt{c}\}$ is an "any 2 of 3" type of descriptor. Every individual has exactly two of these three attributes.

Figure 11 shows the relations R_1 and R_2 along with their Dowker attribute complexes.

R_1	a	b	c
1	•	•	
2	•		•
3		•	•
4	•	•	
5	•		•
6		•	•

R_2	d	e
1	•	
2	•	
3	•	
4		•
5		•
6		•

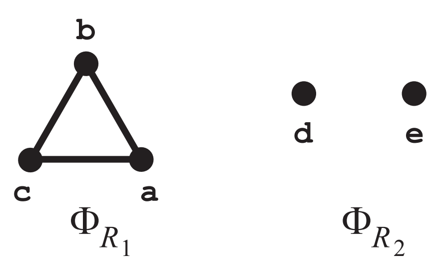


Figure 11: Relation R of Figure 10 decomposes into two disjoint relations R_1 and R_2 such that $\Phi_R = \Phi_{R_1} * \Phi_{R_2}$, with Φ_{R_1} the boundary complex of a triangle and Φ_{R_2} two isolated points. This means every individual in R has attributes that act like two independent coordinates: an "any 2 of 3" component and a bit.

6 Conditional Relations as Simplicial Links

The decomposition of Figures 10 and 11 is reminiscent of stochastic independence expressed as multiplication of probabilities. Similarly, there is a combinatorial analogue to the notion of a *conditional probability distribution*. It appears as the *link* of a simplex in a simplicial complex.

Given a relation R, suppose we have observed attributes γ for some unknown individual. The remaining possible combinations of attributes we might yet observe are described by the simplicial complex $\mathrm{Lk}(\Phi_R, \gamma) = \{\tau \in \Phi_R \mid \tau \cap \gamma = \emptyset \text{ and } \tau \cup \gamma \in \Phi_R\}$. Interpretation: $\tau \cap \gamma = \emptyset$ means that τ consists of as yet unobserved attributes, while $\tau \cup \gamma \in \Phi_R$ means that there is some individual who has the attributes τ in addition to the attributes γ that we have already observed.

Q	a	b	c
1	•	•	
2	•		•
3		•	•

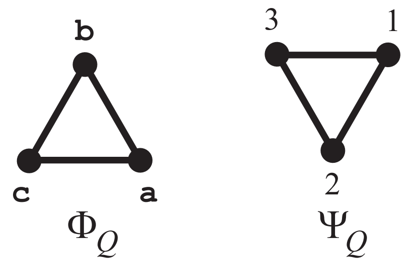


Figure 12: Relation Q describes the conditional relation resulting from R of Figure 10 upon observing attribute d. Note that $\Phi_Q = \mathrm{Lk}(\Phi_R, \{\mathtt{d}\})$.

For instance, after observing attribute d in relation R of Figure 10, we may conclude that we are observing one of the individuals in $\{1, 2, 3\}$ and that the remaining attributes we might yet observe are any two attributes drawn from $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. We can express these conclusions as yet another relation, namely the relation Q of Figure 12. Relation Q describes exactly which individuals could give rise to which attributes, consistent with the prior observation of attribute d. **Thus Φ_R plays a role much like a probability distribution, while Φ_Q plays the role of a conditional distribution.** For another example, suppose we have observed attribute b in R. Then the resulting conditional relation Q' is as in Figure 13.

Q'	a	c	d	e
1	•		•	
3		•	•	
4	•			•
6		•		•

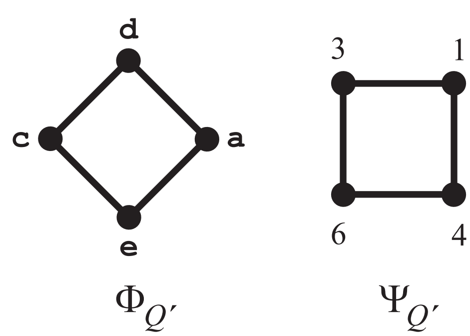


Figure 13: Relation Q' describes the conditional relation resulting from R of Figure 10 upon observing attribute b. Here $\Phi_{Q'} = \mathrm{Lk}(\Phi_R, \{\mathtt{b}\})$. Observe that the attribute space for Q' now factors into two independent bits: $\{\mathtt{a}, \mathtt{c}\}$ constitutes one bit, $\{\mathtt{d}, \mathtt{e}\}$ the other. This factoring is *conditional* on having observed b.

The formal constructions of conditional relations appear below. See also Appendix C.1.

Notation: A symbol of the form $R|_W$ means "restrict R to W". For instance, if R is a relation on $X \times Y$, and if $A \subseteq X$ and $B \subseteq Y$, then $R|_{A\times B} = R \cap (A \times B)$.

Definition 7 (Conditional Attribute Relations)**.** *Let R be a nonvoid relation on $X \times Y$ and suppose $\gamma \subseteq Y$. The following* relation Q models $\mathrm{Lk}(\Phi_R, \gamma)$*:*

$$Q \;=\; R\,|_{\sigma\times\overline{Y}}, \quad \textit{with} \quad \sigma = \psi_R(\gamma) \quad \textit{and} \quad \overline{Y} = \bigcup_{x\in\sigma} Y_x \setminus \gamma.$$

The Dowker complexes are defined in the standard way, except for this special case: If $\overline{Y} = \emptyset$ and $\sigma \neq \emptyset$, then we let Φ_Q and Ψ_Q be instances of the empty complex $\{\emptyset\}$.

Observe: $\mathrm{Lk}(\Phi_R, \gamma) = \Phi_Q$ (a proof appears in Appendix C.1, on page 94).

Comment: If $\gamma \notin \Phi_R$, then $\sigma = \emptyset$ and Q is void, and so Φ_Q is an instance of the void complex, consistent with the standard definition of $\mathrm{Lk}(\Phi_R, \gamma)$ being void in this situation. (See page 84 in Appendix A.1 for the definitions of *void simplicial complex* and *empty simplicial complex*, and page 88 in Appendix A.4 for the definition of *void relation.*)

There is a dual construction for links of individuals σ in the Dowker complex modeling associations:

Definition 8 (Conditional Association Relations)**.** *Let R be a nonvoid relation on $X \times Y$ and suppose $\sigma \subseteq X$. The following* relation Q models $\mathrm{Lk}(\Psi_R, \sigma)$*:*

$$Q \;=\; R\,|_{\overline{X}\times\gamma}, \quad \textit{with} \quad \gamma = \phi_R(\sigma) \quad \textit{and} \quad \overline{X} = \bigcup_{y\in\gamma} X_y \setminus \sigma.$$

The Dowker complexes are defined in the standard way, except for this special case: If $\overline{X} = \emptyset$ and $\gamma \neq \emptyset$, then we let Ψ_Q and Φ_Q be instances of the empty complex $\{\emptyset\}$.

Observe: $\mathrm{Lk}(\Psi_R, \sigma) = \Psi_Q$.

As we will see in Section 7, the complex $\mathrm{Lk}(\Psi_R, \{x\})$ is useful for characterizing individual x's attribute privacy. If that seems surprising, observe that $\mathrm{Lk}(\Psi_R, \{x\})$ describes other individuals in R who share attributes with x, with simplices modeling the extent of commonalities. These commonalities, or lack thereof, determine whether in Φ_Q, and thus back in Φ_R, there are attributes of x that are "free to move" under the closure operators.

7 Privacy Characterization via Boundary Complexes

Q	b	c	d
1	•		•
2		•	•
4	•		
5		•	
6	•	•	

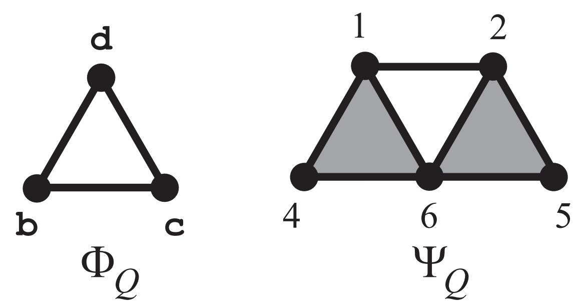


Figure 14: With R as in Figure 10, relation Q describes the conditional relation corresponding to $\text{Lk}(\Psi_R, \{3\})$. Also shown are the Dowker complexes of Q. By design, $\Psi_Q = \text{Lk}(\Psi_R, \{3\})$. Observe that Φ_Q is the boundary complex $\partial(\{\mathtt{b}, \mathtt{c}, \mathtt{d}\})$, with $\{\mathtt{b}, \mathtt{c}, \mathtt{d}\}$ being all of individual #3's attributes in relation R. That boundary condition characterizes attribute privacy for an identifiable individual. Here, it means that individual #3 has full attribute privacy.

We observed earlier that relation R of Figure 10 preserves attribute privacy. We came to that conclusion after observing that Φ_R has no free faces. In fact, one can focus on the privacy of any identifiable individual rather than look at the whole relation. Let us pick one such individual, say #3, and look at the conditional relation Q that models the link $\text{Lk}(\Psi_R, \{3\})$, as shown in Figure 14. (Observe that individual #3 is indeed uniquely identifiable via R.)

Individual #3 has attributes $\{\mathtt{b}, \mathtt{c}, \mathtt{d}\}$ in R. The attribute complex Φ_Q for Q is the boundary complex on exactly this set. Interpretation: for any nonempty proper subset of individual #3's attributes, some *other* individual in R has at least those attributes but not all of individual #3's attributes. Consequently, there is a different such individual for each proper subset of $\{\mathtt{b}, \mathtt{c}, \mathtt{d}\}$ that is missing exactly one of #3's attributes. That diversity of individuals ensures individual #3's attribute privacy.

The previous example suggests the following characterization: **An identifiable individual has full attribute privacy precisely when the attribute complex of the individual's link is the boundary complex of the individual's attributes**.

Observe that this characterization is local to the individual; it does not depend on other individuals having privacy. We now formalize this intuition. Proofs appear in Appendix E.

First, a definition to make precise the notion of individual privacy:

Definition 9 (Individual Privacy)**.** *Let R be a relation on $X \times Y$ and suppose $x \in X$.*
We say that <u>R preserves attribute privacy for x</u> *whenever $(\phi_R \circ \psi_R)(\gamma) = \gamma$ for all $\gamma \subseteq Y_x$.*
Informally, we may also say that <u>individual x has full attribute privacy</u>.

Recall also Definitions 4 and 6, from pages 17 and 20, respectively, formalizing the notions of (attribute) privacy preservation and unique identifiability. And recall the semantics of P_R, for instance from Definition 3 on page 17.

Here is the characterization of individual attribute privacy formalized:

Theorem 10 (Individual Attribute Privacy)**.** *Let R be a relation on $X \times Y$, with $|X| > 1$. Suppose $x \in X$ is uniquely identifiable via R. Let Q be the relation modeling* $\mathrm{Lk}(\Psi_R, x)$*. Then the following three conditions are equivalent:*

(a) R preserves attribute privacy for x.

(b) $\mathrm{Lk}(\Psi_R, x) \;\simeq\; \mathbb{S}^{k-2}$*, with $k = |Y_x|$.*

(c) $\Phi_Q \;=\; \partial(Y_x)$*.*

The previous theorem generalizes to sets of individuals for sets that are "stable" under the closure operators, i.e., that appear as the "set of individuals component" in an element of P_R:

Theorem 11 (Group Attribute Privacy)**.** *Let R be a relation on $X \times Y$. Suppose $(\sigma, \gamma) \in P_R$, with $\sigma \neq X$. Let Q be the relation modeling* $\mathrm{Lk}(\Psi_R, \sigma)$*. Then the following three conditions are equivalent:*

(a) $(\phi_R \circ \psi_R)(\gamma') \;=\; \gamma'$*, for every subset γ' of γ.*

(b) $\mathrm{Lk}(\Psi_R, \sigma) \;\simeq\; \mathbb{S}^{k-2}$*, with $k = |\gamma|$.*

(c) $\Phi_Q \;=\; \partial(\gamma)$*.*

The following lemma relates interpretation and inference in a link to the encompassing relation:

Lemma 12 (Interpreting Local Operators)**.** *Let R be a relation on $X \times Y$.*

Suppose $(\sigma, \gamma) \in P_R$, with $\sigma \neq X$.

Let Q be the relation on $\overline{X} \times \gamma$ that models $\mathrm{Lk}(\Psi_R, \sigma)$ *and suppose $\overline{X} \neq \emptyset$.*

Then, for every $\gamma' \subseteq \gamma$:

(i) If $\gamma' \notin \Phi_Q$, then $\psi_R(\gamma') = \sigma$.

(ii) If $\gamma' \in \Phi_Q$, then $\psi_R(\gamma') \supsetneq \sigma$.

Moreover, in this case:

For $\gamma' = \emptyset$, $(\phi_Q \circ \psi_Q)(\emptyset) \supseteq (\phi_R \circ \psi_R)(\emptyset)$.

If $\gamma' \neq \emptyset$, then $(\phi_Q \circ \psi_Q)(\gamma') = (\phi_R \circ \psi_R)(\gamma')$.

The lemma says that observations of attributes consistent in Q have as interpretation more individuals in R than just the individuals σ. However, if ever those observations become inconsistent in Q, then one has identified σ in R. Here "inconsistent in Q" means that the observed attributes are legitimate attributes for Q but do not constitute a simplex of Φ_Q. (Note: Such observed attributes necessarily constitute a simplex of Φ_R since they are a subset of $\gamma \in \Phi_R$).

Moreover, attribute inferences are identical in Q and R for nonempty simplices of Φ_Q.

8 The Meaning of Holes in Relations

We have seen how spheres characterize privacy. More generally, when working with topological spaces, holes are significant. One wonders what topological holes mean for relations.

- Some holes arise as a consequence of exclusion between attributes, as we saw in the decomposition of Figures 10 and 11.

  Sticking with binary exclusions, suppose a group of individuals are described by k bits. One can model those individuals via a relation containing $2k$ binary attributes (two such attributes per bit, one for each possible bit value). Every individual has exactly k of those $2k$ attributes. If all possible 2^k combinations of bit values are represented by individuals in the relation, then the two Dowker complexes are both homotopic to $\mathbb{S}^{k-1}$, the sphere of dimension $k-1$. In fact, Φ_R is the simplicial join of k copies of $\mathbb{S}^0$, while Ψ_R is visualizable as a hollow hypercube in k dimensions, in which solid $(k-1)$-dimensional subcubes represent $(2^{k-1}-1)$-dimensional simplices (flattened, when $k \geq 3$). Figures 15, 16, and 17 depict the cases $k = 1$, 2, and 3, respectively.

  In short, k bits means a hole of dimension $k-1$, *if* all possible individuals are actually present in the relation.

  (The lack of an expected hole may mean that the capacity of a relation has not been exhausted, hinting at possible inference. See Appendix J.3.)

S	a	¬a
1	•	
2		•

Figure 15: Relation S describes two individuals in terms of a single attribute and its negation. The topology of the Dowker complexes is $\mathbb{S}^0$.

Q	a	¬a	b	¬b
1	•		•	
2	•			•
3		•	•	
4		•		•

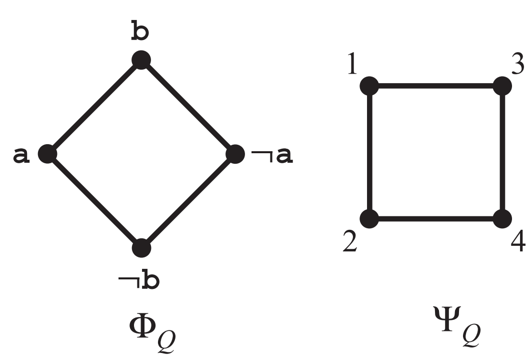


Figure 16: Relation Q describes four individuals in terms of two attributes and their negations. The topology of the Dowker complexes is $\mathbb{S}^1$.

R	`a`	`¬a`	`b`	`¬b`	`c`	`¬c`
1	•		•		•	
2	•		•			•
3	•			•	•	
4	•			•		•
5		•	•		•	
6		•	•			•
7		•		•	•	
8		•		•		•

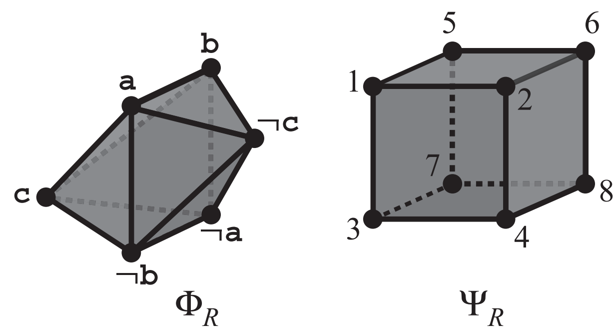


Figure 17: Relation R describes eight individuals in terms of three attributes and their negations. The topology of the Dowker complexes is $\mathbb{S}^2$. The cube faces are actually tetrahedra, flattened to parallelograms in the drawing.

- Suppose Σ is a simplicial complex with underlying vertex set X. A *minimal nonface* of Σ is a subset of X that is not itself a simplex but all of whose proper subsets are simplices in Σ. A minimal nonface may or may not be a topological hole. Regardless, a minimal nonface of size two or greater in a Dowker complex suggests restricting the relation to equal-numbered attributes and individuals for whom there is both attribute and association privacy, within the restricted relation. This observation dovetails with the following results (here we assume that each relation has no blank rows or columns):
  - A relation with more attributes than individuals cannot fully preserve attribute privacy.
  - A relation with more individuals than attributes cannot fully preserve association privacy.
  - A relation that preserves both attribute and association privacy must have the same number of attributes and individuals. Moreover, if the relation is connected, then both Dowker complexes are either linear cycles of the same length or boundary complexes of full simplices of the same dimension, as we indicated previously.

  See Appendices C and E for further details and proofs.

- Minimal nonfaces can have other context-dependent meanings. For instance, in a certain authorship relation, knowing that *each pair* of three individuals has written a paper together appears to be a good predictor that *all three* individuals will co-author a paper together [15]. This observation suggests the following: if one sees that such an authorship hole does *not* fill over time, then one likely can infer some kind of obstruction, perhaps an incompatibility in the group as a whole, or the death of an author, for instance.

- When designing relations or anonymizing relations, these results suggest transformations that create "bubbly spaces" of some sort, in order to retain identifiability but also reduce unwanted inference. Section 9 and Appendix J.2 discuss examples.

- Whatever topological holes there are in Φ_R and Ψ_R must also show up in the poset P_R, since that poset is formed by homotopy equivalences from Φ_R and Ψ_R. Interestingly, whereas one thinks of Φ_R and Ψ_R simply as spaces, one sees a partial order on P_R. Something can move, "up" or "down". The elements of P_R are inference-stable, by design. So, what is this possible motion? It is a dynamic process that describes how information acquisition changes interpretation. For instance, as an individual reveals information about him- or herself, an observer can attempt to identify the individual, by finding interpretations in P_R of the information revealed. As the individual reveals additional information, the observer's interpretation moves downward in P_R, narrowing the set of individuals.

  Topological holes in the spaces Φ_R and Ψ_R (and thus P_R) constrain how that interpretation moves downward in P_R. The greater a hole's dimension, the further a downward path has to move before identifying an individual. One can think of holes in a relation much like boulders in a stream. Eventually, the current of information sweeps past the hole, but it is forced to divert its motion, covering more distance. Moreover, there may be many paths around the hole, much like a leaf in a stream may divert around a boulder in different directions. The individual can force a particular path by choosing to reveal attributes in a particular order.

  Much of the rest of the report explores the implications of this stream analogy. The analogy merges with the realization that privacy is a dynamic process, certain to flow toward identification when attributes are static or persistent, yet subject to channeling (perhaps even turbulence in more fluid settings than those discussed in this report). See, in particular, Section 10 onward.

9 Change-of-Attribute Transformations

Free faces and holes in the Dowker complex Φ_R can sometimes suggest changes in attributes that preserve desired information but reduce inference. Consider the hypothetical "ice-cream cone" relation C of Figure 18 and the corresponding complexes shown in Figure 19. The relation describes four individuals in terms of the two-flavor two-scoop ice-cream cones each individual enjoys at a particular ice-cream parlor.

C	gc	gs	cs	cv	sv	gv
Bob	•	•	•			
Alice			•	•	•	
David		•			•	•
Cindy	•			•		•

g	=	ginger
c	=	chocolate
s	=	strawberry
v	=	vanilla

Figure 18: Four individuals and their preferences for ice-cream cones containing two scoops, with different flavors (each letter represents a flavor, as indicated). See Figure 19 for the Dowker complexes.

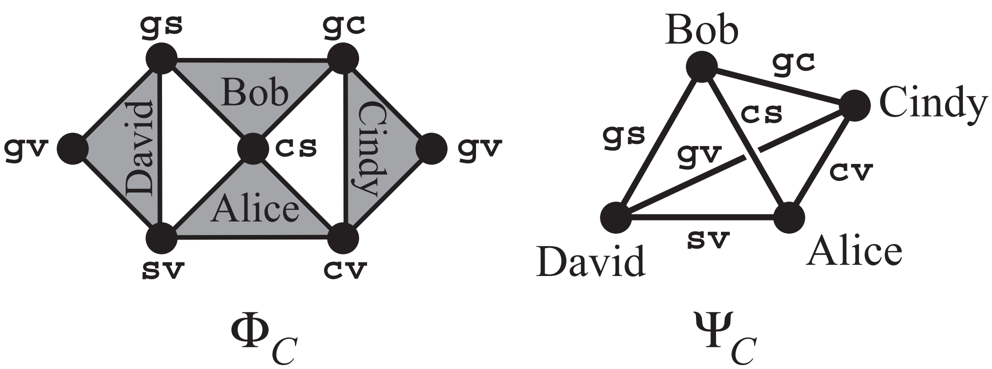


Figure 19: The Dowker complexes for the relation of Figure 18. Φ_C is a complex whose vertices are ice-cream cones (two flavors). (For visualization purposes, the complex is flattened, with the leftmost and rightmost vertices really representing the same ice-cream cone.) Each maximal simplex is a triangle, labeled with the individual who enjoys the three types of cones comprising the triangle. Ψ_C is a complex whose vertices are individuals. Each maximal simplex is an edge, representing a two-flavor two-scoop ice-cream cone that each of two individuals enjoys; the edge is labeled with the cone flavors. The homotopy type of each complex is $\mathbb{S}^1 \vee \mathbb{S}^1 \vee \mathbb{S}^1$.

Relation C is a typical "2-implies-3" relation: Any two different ice-cream cones uniquely identify an individual, thereby implying a third ice-cream cone, as can be seen from either Dowker complex: In Φ_C, every edge is a free face of its encompassing triangle. Moreover, the edge is not itself generated by any individual.[2] The closure operator $\phi_C \circ \psi_C$ must therefore map every edge to a triangle. Dually, in Ψ_C, any two edges intersecting at a vertex imply the third edge incident on that vertex.

[2]We say that an individual x of a relation R *generates* the simplex $Y_x \in \Phi_R$. Similarly, an attribute y generates the simplex $X_y \in \Psi_R$. Individuals generate triangles in Φ_C. Ice-cream cones generate edges in Ψ_C.

This type of relation models, in the small, inferences such as those reported in [21, 17]. For instance, [21] reported that zip code, gender, and birth date were likely sufficient in 1990 to identify 87% of individuals in the U.S. That is nearly a "3-implies-all" type of relation. Similarly, [17] reported that 8 movie ratings and dates were enough to uniquely identify 99% of viewers in the Netflix Prize dataset. That is essentially an "8-implies-all" type of relation.

Let us focus for a moment on Bob's neighborhood. That relation, let us call it B, and its complexes are depicted in Figure 20. (The relation models $\overline{\mathrm{St}}(\Psi_C, \{\mathrm{Bob}\})$; see Appendix A.1.)

B	gc	gs	cs
Bob	•	•	•
Alice			•
David		•	
Cindy	•		

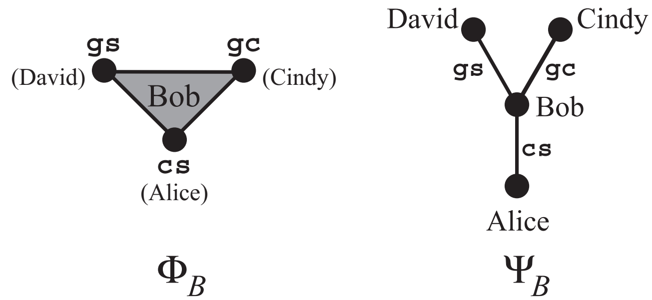


Figure 20: Relation B models Bob's neighborhood in the ice-cream relation of Figure 18. Each maximal simplex is labeled with its generator. Generators of nonmaximal simplices are indicated in parentheses.

As in C, seeing someone eat one ice-cream cone is not enough to identify anyone in B uniquely. Seeing someone (in this case, Bob) eat two *different* types of ice-cream cones is sufficient to infer the third type of ice-cream cone that individual prefers. How might we prevent this? We observe that the vertices of Φ_B are themselves generated by individuals while the edges are not. Homotopically, therefore, we want to expand the vertices of Φ_B into edges, and contract the edges of Φ_B into vertices. One possible way to accomplish this is the take logical ORs of the existing attributes. With $\oplus$ meaning Boolean OR, we define:

$$\alpha \;=\; \mathtt{gc} \,\oplus\, \mathtt{gs}, \qquad \beta \;=\; \mathtt{gc} \,\oplus\, \mathtt{cs}, \qquad \gamma \;=\; \mathtt{gs} \,\oplus\, \mathtt{cs}.$$

Then relation B becomes B' as in Figure 21. The result is that the free faces of $\Phi_{B'}$ now are generated by other individuals, so even though they are free, the closure operator does not move them. In fact, the closure operator $\phi_{B'} \circ \psi_{B'}$ is the identity on $\mathfrak{F}(\Phi_{B'}) \cup \{\emptyset\}$, meaning that no attribute inference is possible in B'.

B'	α	β	γ
Bob	•	•	•
Alice		•	•
David	•		•
Cindy	•	•	

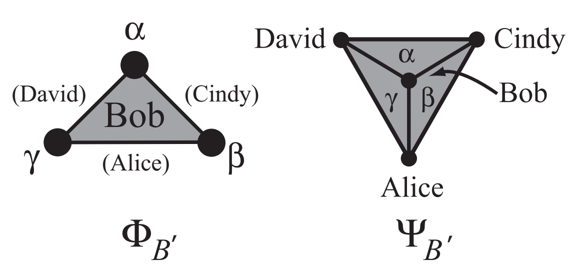


Figure 21: Relation B' represents relation B of Figure 20, now with a coordinate transformation for the attributes. Simplices are again labeled by generators.

Now imagine performing similar operations for all four individuals of relation C from Figure 18. One winds up constructing four logical ORs:

$$\texttt{gc} \oplus \texttt{gs} \oplus \texttt{gv}, \qquad \texttt{gc} \oplus \texttt{cs} \oplus \texttt{cv}, \qquad \texttt{gs} \oplus \texttt{cs} \oplus \texttt{sv}, \qquad \texttt{cv} \oplus \texttt{sv} \oplus \texttt{gv}.$$

Two observations:

1. Each OR describes three ice-cream cones that form a hole in the complex Φ_C of Fig. 19.
2. Each such hole may be interpreted as a single flavor, namely the flavor in common to the three ice-cream cones appearing in the OR. For instance, "ginger" (abbreviated as `g`) is the common flavor for the OR $\texttt{gc} \oplus \texttt{gs} \oplus \texttt{gv}$.

In order to describe the resulting relation, it is perhaps easiest to express those four new coordinates themselves via a relation S that describes the scoops present in an ice-cream cone:

S	g	c	s	v
gc	•	•		
gs	•		•	
cs		•	•	
cv		•		•
sv			•	•
gv	•			•

Finally, to perform the coordinate-transformation, one simply multiplies Boolean matrices, with addition being Boolean OR and multiplication being Boolean AND: $F = CS$. The relation F and its complexes appear in Figure 22.

F	g	c	s	v
Bob	•	•	•	
Alice		•	•	•
David	•		•	•
Cindy	•	•		•

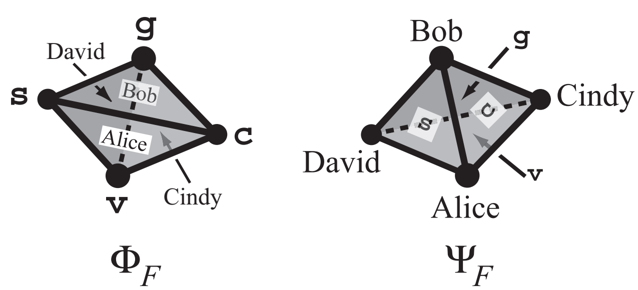


Figure 22: Relation F describes the ice-cream flavors each individual prefers. Φ_F is the boundary complex of a tetrahedron, with flavors as vertices. Ψ_F is the Dowker dual of Φ_F, with respect to relation F. Consequently, Ψ_F also is the boundary complex of a tetrahedron, now with the roles of flavors and individuals interchanged. For both Φ_F and Ψ_F, each maximal simplex is a triangle, labeled with its generator.

Relation F represents a description of the four individuals' preferences in terms of flavors not cones. The resulting complexes Φ_F and Ψ_F are now boundary complexes of full simplices, each homeomorphic to $\mathbb{S}^2$. These complexes have no free faces, so no inference is possible.

Observe further that Φ_F is homotopic to what one obtains from Φ_C by filling the $\mathbb{S}^1$-holes. Indeed, this idea implicitly motivated our construction, as a way to remove free faces. Similarly, Ψ_F is isomorphic to what one obtains from Ψ_C by filling its $\mathbb{S}^1$-holes.

One should ask how this approach might generalize. The answer is mixed. The idea of removing free faces is central. There are many ways to accomplish that, with relational composition being but one method. One issue with logical ORs is that it is very easy to obtain an OR that is always TRUE, at which point the resulting attribute is of little use.

Even with more general transformations, there remains the issue of whether the new attributes are grounded in what is actually observable. In the ice-cream example, it was fortunate that cones decomposed naturally into flavors. It is at least plausible that someone might merely observe the flavors a customer prefers, not the combinations of flavors as cones. If, however, only cones can be observed, then one is forced to deal with relation C as given.

10 Leveraging Lattices for Privacy Preservation

This section examines more carefully the lattice structure of a relation's poset, leading to the idea of *informative attribute release sequences*. Such a sequence consists of attributes that an individual releases in a particular order, so as to prevent inference of any attributes yet to be released via the sequence. The length of the lattice representing the individual's link relation then describes the extent to which that individual can defer identification. Homology provides lower bounds on that length.

10.1 Attribute Release Order

Relation G of Figure 23 describes hypothetical co-authorships among five authors in producing travel guides for five European cities. Each collaboration consists of three authors working together on one of the five travel guides.

G	Athens	Berlin	Caen	Dublin	Ełk
1 (`Alice`)	•	•			•
2 (`Ben`)	•	•	•		
3 (`Claire`)		•	•	•	
4 (`David`)			•	•	•
5 (`Eric`)	•			•	•

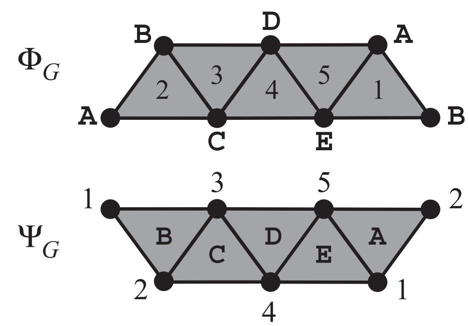


Figure 23: A relation G describing co-authorship of travel guides. The Dowker complexes are dual triangulations of the Möbius strip, with $\mathbb{S}^1$ homotopy type. (Notes: Integers indicate authors, letters indicate cities via first letter abbreviations. Some vertices and edges appear twice for ease of viewing. Each maximal simplex is labeled with its generating author or city.)

Suppose in casual conversation a person mentions that he/she worked on producing a travel guide for Berlin. In the context of relation G, that information means the author is one of {`Alice`, `Ben`, `Claire`}. If the author further mentions working on the travel guide for Dublin, then that identifies the author uniquely as `Claire`. Equivalently, the listener can infer that the author also helped write the travel guide for Caen. (This form of inference was a source of privacy problems for the Netflix Prize [17].)

`Claire` was a co-author on three travel guides, for Berlin, Caen, and Dublin. Now consider the different possible sequential ways in which `Claire` might reveal which books she helped co-author, along with the points at which her identity becomes known (see Figure 24).

Of the six possible ways, four do not uniquely identify `Claire` until she has revealed all three books that she co-authored. However, two of the possible six release sequences do allow a listener to identify the author and infer an additional book that she co-authored.

This example shows how inference may be a dynamic process. While a consumer of data may wish to identify `Claire` with as little information as possible, the author herself may wish to delay that identification for as long as possible (perhaps for reasons of public mystery in

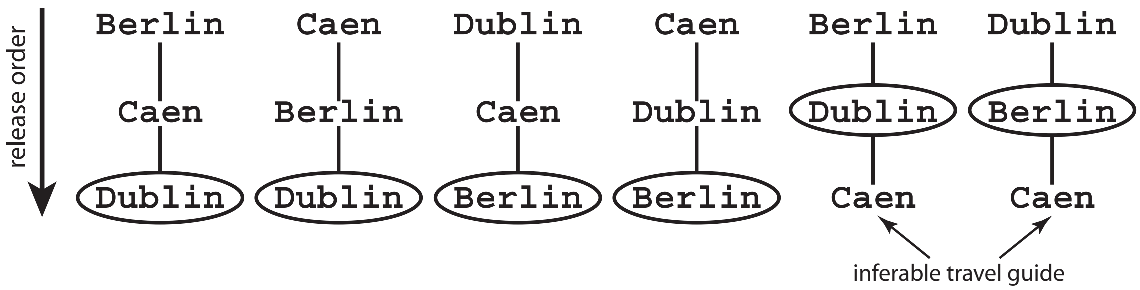


Figure 24: This figure shows the six possible sequential ways in which author #3 (`Claire`) of Figure 23 can mention the cities for which she co-authored travel guides. The point at which her identity becomes known in any such release sequence is circled. If `Claire` does not mention Caen, one can infer, via relation G of Figure 23, that she co-authored a travel guide for that city as soon as she mentions the other two cities, Berlin and Dublin, in either order.

selling books). In the example, the *minimal length* of an *identifying attribute release sequence* is two, while the *maximal length* is three. If `Claire` can control how information is released, then she can choose to reveal what might otherwise be inferred, namely that she co-authored a travel guide to Caen, thereby delaying her identification.

Finally, we observe that the order of attributes released may or may not matter. In the travel guide example, `Claire` should mention Caen before the end of her disclosures (if she wants to delay her identification), but the order of cities mentioned is otherwise irrelevant. The topology of the doubly-labeled poset P_G encodes this order (in)dependence, as we will see shortly. Indeed, much of the remainder of this report examines the connection between the topology of a relation's doubly-labeled poset and the length of attribute release sequences.

10.2 Inferences on a Lattice

The doubly-labeled poset of a relation produces a lattice [25], as follows:

Definition 13 (Galois Lattice)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Let P_R be the associated doubly-labeled poset.*

(Recall from Definition 3 on page 17 that an element of P_R is an ordered pair (σ, γ), with $\emptyset \neq \sigma = \psi_R(\gamma) \in \Psi_R$ and $\emptyset \neq \gamma = \phi_R(\sigma) \in \Phi_R$.

We previously defined a partial order on P_R by $(\sigma_1, \gamma_1) \leq (\sigma_2, \gamma_2)$ iff $\sigma_1 \subseteq \sigma_2$ (iff $\gamma_1 \supseteq \gamma_2$).)

P_R may already contain a unique bottom element of the form (σ, Y), with σ those individuals in X who have all the attributes in Y. If not, we adjoin $(\emptyset, Y)$ to the bottom of P_R.

P_R may already contain a unique top element of the form (X, γ), with γ those attributes in Y that every individual in X has. If not, we adjoin $(X, \emptyset)$ to the top of P_R.

We refer to the resulting poset as the <u>Galois lattice P_R^+</u>*. It has lattice operations $\vee$ and $\wedge$:*

$$\begin{aligned}(\sigma_1, \gamma_1) \ \vee \ (\sigma_2, \gamma_2) &= \big((\psi_R \circ \phi_R)(\sigma_1 \cup \sigma_2), \ \ \gamma_1 \cap \gamma_2\big), \\ (\sigma_1, \gamma_1) \ \wedge \ (\sigma_2, \gamma_2) &= \big(\sigma_1 \cap \sigma_2, \ \ (\phi_R \circ \psi_R)(\gamma_1 \cup \gamma_2)\big).\end{aligned}$$

We sometimes refer to the bottom element of P_R^+ by $\underline{\hat{0}_R}$ and to the top element by $\underline{\hat{1}_R}$.

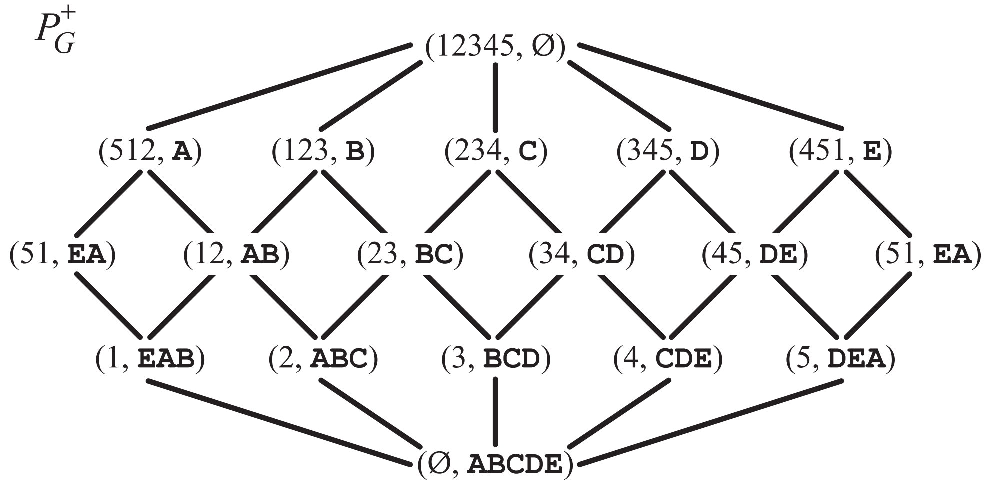


Figure 25: The lattice P_G^+ for the travel guide relation of Figure 23. Each element is an ordered pair of sets (σ, γ) such that $\sigma = \psi_G(\gamma)$ and $\gamma = \phi_G(\sigma)$. (We have elided commas and braces in sets, for ease of viewing.) The lattice operations model inferences possible from observations. For instance, $(123, \texttt{B}) \wedge (345, \texttt{D}) = (3, \texttt{BCD})$, meaning that observation of attributes `B` and `D` permits inference of additional attribute `C` and identification of author #3. (In Figure 23, attribute `C` is the travel guide for CAEN and author #3 is `Claire`.) The lattice wraps around, with element $(51, \texttt{EA})$ duplicated for ease of viewing. If one removes the top and bottom elements, the remaining poset P_G has $\mathbb{S}^1$ homotopy type, just like the Möbius strip.

Figure 25 shows the lattice P_G^+ for the travel guide relation of Figure 23. Observe how the lattice encodes attribute and association inferences (or lack thereof) via its lattice operations.

Special Cases: It can happen that the lattice consists of a single element. For example, with relation C as on page 20, $P_C^+ = P_C = \{(X, Y)\}$. In particular, $\hat{0}_C = \hat{1}_C$.

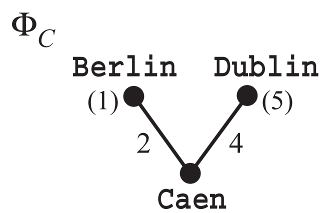
Definition 13 ignores the situation in which R is void. One possibility is to leave P_R undefined and let $P_R^+ = \emptyset$. See page 89 in Appendix A.4 for additional comments.

10.3 Preserving Attribute Privacy for Sets of Individuals

Theorem 10 on page 26 described the conditions under which an individual has full attribute privacy. For such an individual, the order in which that individual (or anyone) releases the individual's attributes is irrelevant. Any order is fine. Only once all attributes have been released, can an observer uniquely identify the individual. Theorem 11 described a similar result for certain sets of individuals, including sets of individuals with whom a given individual is confusable after only some of his/her attributes have been released.

Consider $\mathrm{Lk}(\Psi_G, 3)$, modeled by relation C as in Figure 26. This relation describes the authors with whom `Claire` has collaborated, via their co-authored books. The Dowker complexes are contractible, so by either Theorem 10 or Theorem 11, we know that some attribute inference is possible involving `Claire`. Lemma 12 on page 26 tells us to look for a proper subset of {BERLIN, CAEN, DUBLIN} that is *not* a simplex of Φ_C. As is apparent from

C		Berlin	Caen	Dublin
1	(Alice)	•		
2	(Ben)	•	•	
4	(David)		•	•
5	(Eric)			•



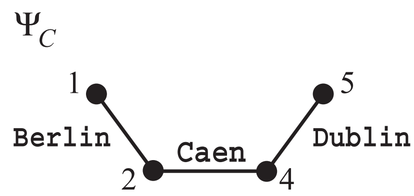


Figure 26: Relation C describes $\mathrm{Lk}(\Psi_G, 3)$, the link of `Claire` in the relation of Figure 23. (Each maximal simplex in any one complex is labeled with its generating attribute or individual from the other complex. Generators of nonmaximal simplices are indicated in parentheses.)

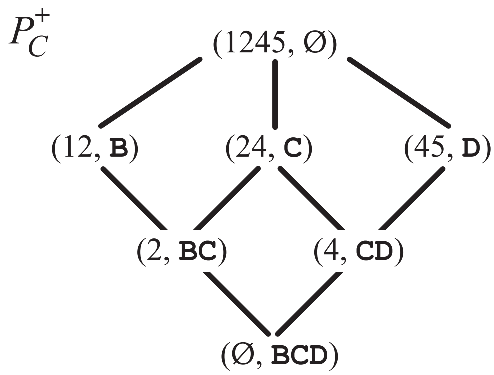


Figure 27: The lattice P_C^+ for the link of `Claire`, as given in Figure 26. (Here authors appear as integers and city names appear as first letter abbreviations.) Observe that P_C^+ may be viewed as a sublattice of P_G^+, consisting of all elements that include individual #3 there, but with that individual removed here. (See Figure 25 for P_G^+.)

Figure 26, the set {Berlin, Dublin} satisfies these conditions, consistent with our earlier observations. Alternatively, looking at P_C^+ in Figure 27, we see that $(12,\mathtt{B}) \wedge (45,\mathtt{D}) = (\emptyset,\mathtt{BCD})$, allowing us to draw the same conclusion. Consequently, `Claire` should be sure to mention her travel guide for Caen early on, not leave it for last, if she wants to delay identification.

Now let us take this reasoning one step further. Consider an element of P_G^+ corresponding to some state just prior to identification of `Claire`, for instance $(23,\mathtt{BC})$. This element corresponds to both of the first two release sequences of Figure 24: `Claire` has mentioned her work regarding the travel guides for Berlin and Caen, but has not yet mentioned Dublin. Thus there is still some ambiguity as to her identity (it is either author #2 or author #3). In terms of Theorem 11 on page 26, $\sigma = \{2, 3\}$, $\gamma = \{\text{Berlin}, \text{Caen}\}$, and $k = 2$.

Figure 28 shows the relation describing $\mathrm{Lk}(\Psi_G, \{2, 3\})$. The Dowker complexes have $\mathbb{S}^0$ homotopy type, thus satisfying the topological conditions of Theorem 11. Consequently, there is no attribute inference possible in the encompassing relation G based on attributes that appear in the link relation Q. That means the order in which `Claire` releases the two attributes Berlin and Caen is immaterial. This conclusion is consistent with the conclusion one draws upon explicitly enumerating all release sequences, as in Figure 24.

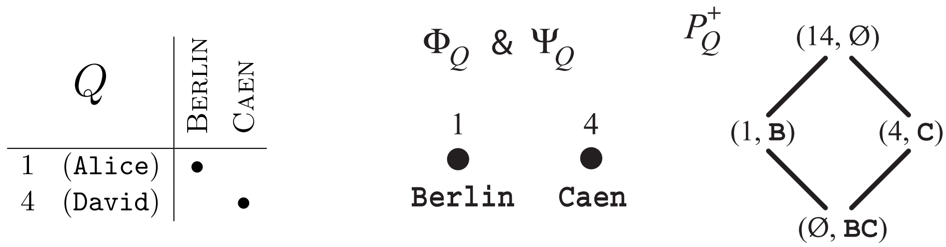


Figure 28: Relation Q describes $\mathrm{Lk}(\Psi_G, \{2,3\})$, the combined link of authors #2 and #3 (`Ben` and `Claire`) in the relation of Figure 23. These two authors have together collaborated with each of authors #1 and #4 (`Alice` and `David`) but have not both together collaborated with author #5 (`Eric`). The two Dowker complexes are each instances of $\mathbb{S}^0$, so essentially the same. The corresponding lattice P_Q^+ is also very simple.

10.4 Informative Attribute Release Sequences

This subsection defines more precisely the idea of controlled information release. These definitions will help us better understand topological holes in a relation's Dowker complexes. Subsequently, Section 11 will explore these insights with data from the world wide web.

Definition 14 (Attribute Release Sequence)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. An* attribute release sequence for R *is a nonempty set of attributes from Y released in a particular sequential order:*

$$y_1, y_2, \ldots, y_k, \quad \text{with } k \geq 1.$$

We say that the sequence has length *k.*

We say that an attribute release sequence is informative *if*

$$y_i \notin (\phi_R \circ \psi_R)(\{y_1, \ldots, y_{i-1}\}), \quad \text{for all } 1 \leq i \leq k.$$

(Note: for $i = 1$, the requirement states that $y_1 \notin (\phi_R \circ \psi_R)(\emptyset) = \phi_R(X)$.)

(We sometimes use the abbreviation `'iars'` *to mean either* `'informative attribute release sequence'` *or* `'informative attribute release sequences'`*.)*

Interpretation: When $i = 1$, the argument to $\phi_R \circ \psi_R$ is the empty set, so the condition requires that $y_1 \notin \phi_R(X)$. In other words, y_1 may not be any attribute that is shared by all individuals in X. Any such attribute could be inferred "for free" in the context of relation R, and thus would not be informative. Thereafter, the condition requires that any attribute to be released not be inferable from those already released.

We are interested in understanding the extent to which order of release matters:

Definition 15 (Isotropy)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty.*

Suppose $\emptyset \neq \gamma \subseteq Y$.

We say that γ is isotropic *if every possible ordering of all the elements in γ forms an informative attribute release sequence for R.*

We are interested in the minimal and maximal lengths of informative attribute release sequences:

Definition 16 (Identification and Minimal Identification)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty.*

We say that a set of attributes $\gamma \subseteq Y$ identifies a set of individuals $\sigma \subseteq X$ in R when $\psi_R(\gamma) = \sigma$. (We sometimes alternatively say that γ localizes to σ in R.)

We say that γ is minimally identifying (for σ) if both the following conditions hold:

(i) $\psi_R(\gamma) = \sigma$.

(ii) $\psi_R(\gamma') \supsetneq \sigma$ *for every* $\gamma' \subsetneq \gamma$.

Definition 17 (Identification Lengths)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $(\sigma, \gamma) \in P_R$. Define the fast and slow attribute release lengths for σ as:*

$$\mathrm{r_{fast}}(\sigma) = \min \left\{ |\chi| \;\middle|\; \chi \subseteq \gamma \text{ and } \psi_R(\chi) = \sigma \right\}.$$

$$\mathrm{r_{slow}}(\sigma) = \max \left\{ k \;\middle|\; y_1, \ldots, y_k \text{ is an iars for } R \text{ and } \psi_R\big(\{y_1, \ldots, y_k\}\big) = \sigma \right\}.$$

An argument similar to that in Appendix D shows that the following problem is *NP*-complete: Given R, σ, and k, is there some minimally identifying γ for σ with $|\gamma| \leq k$?

10.5 Isotropy, Minimal Identification, and Spheres

There is no requirement in Definition 14 that an informative attribute release sequence be a simplex in Φ_R. (Indeed, when working with links of individuals, it can be useful to create informative attribute release sequences that are not simplices in the link, thereby identifying the given individuals in the encompassing relation, as per Lemma 12 on page 26.) However, it is always the case that any inconsistency arises only with the last attribute released:

Lemma 18 (Almost a Simplex)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty.*
Suppose $\{y_1, \ldots, y_k\}$ is an informative attribute release sequence for R.
Then $\{y_1, \ldots, y_{k-1}\} \in \Phi_R$.

Proof. If $\{y_1, \ldots, y_{k-1}\} \notin \Phi_R$, then $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_{k-1}\}) = \phi_R(\emptyset) = Y$. Since $y_k \in Y$, this contradicts the requirement of Definition 14. □

Consequently, a nonempty set of attributes $\gamma \subseteq Y$, with $\gamma \notin \Phi_R$, is isotropic if and only if it is a minimal nonface of Φ_R. We can view such an isotropic γ as minimally identifying for $\emptyset$.

When a nonempty set of attributes γ *is* a simplex in Φ_R, then being isotropic is again equivalent to being minimally identifying, now for some nonempty set of individuals σ. Moreover, topologically, we can again characterize this isotropy as a sphere, appearing via a restricted link:

Q'	Berlin	Dublin
1 (Alice)	•	
2 (Ben)	•	
4 (David)		•
5 (Eric)		•

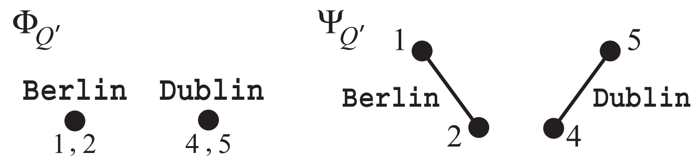


Figure 29: Relation $Q' = Q(\sigma, \gamma)$, for the book authorship example of Figure 23, with $\sigma = \{3\}$ and $\gamma = \{\text{Berlin}, \text{Dublin}\}$. Relation Q' describes the link of author #3 (`Claire`) restricted to the attribute set $\{\text{Berlin}, \text{Dublin}\}$. See Figure 26 for the whole link relation. (Each maximal simplex in the Dowker complexes is again labeled with its generating individuals or attribute.)

Definition 19 (Restricted Link)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\sigma \in \Psi_R$ and $\gamma \subseteq \phi_R(\sigma)$.*
Define relation $Q(\sigma, \gamma)$ as follows:

$$Q(\sigma, \gamma) \;=\; R\,|_{\overline{X} \times \gamma}, \quad \textit{with} \quad \overline{X} = \bigcup_{y \in \gamma} X_y \setminus \sigma.$$

The Dowker complexes are defined in the standard way, except for these special cases:
If $\sigma = X$, we let $\Psi_{Q(\sigma,\gamma)}$ and $\Phi_{Q(\sigma,\gamma)}$ be instances of the void complex $\emptyset$.
If $\sigma \neq X$ but $\overline{X} = \emptyset$, we let $\Psi_{Q(\sigma,\gamma)}$ and $\Phi_{Q(\sigma,\gamma)}$ be instances of the empty complex $\{\emptyset\}$.

We say that $Q(\sigma, \gamma)$ models the link of σ restricted to γ.

Comments: Although the previous definition looks similar to that for $\text{Lk}(\Psi_R, \sigma)$ on page 24, there are some differences: (a) Here, we require that σ be a simplex in Ψ_R. (b) Here, we do *not* assume $\gamma = \phi_R(\sigma)$, merely $\gamma \subseteq \phi_R(\sigma)$. (c) When $\sigma = X \in \Psi_R$, the current definition creates void complexes, whereas Definition 8 on page 24 creates empty complexes. (d) Finally, when $\sigma \neq X$ but $\gamma = \emptyset$, the current definition creates empty complexes rather than void complexes. Interpretation: When $\sigma \in \Psi_R$ and $\sigma \neq X$, $Q(\sigma, \gamma)$ models those simplices of $\text{Lk}(\Psi_R, \sigma)$ that are witnessed by attributes in γ, plus the empty simplex.

Theorem 20 (Isotropy = Minimal Identification = Sphere)**.** *Let R be a relation and suppose $\emptyset \neq \gamma \in \Phi_R$. Let $\sigma = \psi_R(\gamma)$. Then the following four conditions are equivalent:*

(a) γ is isotropic.

(b) γ is minimally identifying (for σ).

(c) $\Psi_{Q(\sigma,\gamma)} \;\simeq\; \mathbb{S}^{k-2}$, with $k = |\gamma|$.

(d) $\Phi_{Q(\sigma,\gamma)} \;=\; \partial(\gamma)$.

See Appendix F.3 for a proof.

Collaboration Example Revisited: To illustrate Theorem 20, consider again the example of Figure 23. Recall that together the travel guides for BERLIN and DUBLIN identify `Claire`. Indeed, {BERLIN, DUBLIN} is a minimally identifying set of books for `Claire`. It is isotropic, as Figure 24 shows. Figure 29 depicts the link of `Claire` restricted to {BERLIN, DUBLIN}, modeled by relation Q'. Observe that $\Phi_{Q'} = \partial(\{\text{BERLIN}, \text{DUBLIN}\})$ and that $\Psi_{Q'} \simeq \mathbb{S}^0$, as the theorem asserts.

10.6 Poset Lengths and Information Release

We have seen how minimal identification appears topologically via spheres. Spheres are isotropic so perhaps it is not surprising that they encode isotropic attribute release sequences. We cannot therefore expect a spherical characterization for the problem of finding a maximally long informative attribute release sequence. Instead, we find an answer in the combinatorial structure of the doubly-labeled poset P_R and its lattice P_R^+. We summarize the key results below. For proofs, see Appendix F.

Lemma 21 (Informative Attributes from Maximal Chains)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\{(\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1) < (\sigma_0, \gamma_0)\}$, with $k \geq 1$, is a maximal chain in P_R^+.*

Define $y_1, \ldots, y_k$ by selecting some $y_i \in \gamma_i \setminus \gamma_{i-1}$, for each $i = 1, \ldots, k$.

Then $y_1, \ldots, y_k$ is an informative attribute release sequence for R.

Moreover, $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_i\}) = \gamma_i$, for each $i = 0, 1, \ldots, k$.

(Notes: (a) For a maximal chain in P_R^+, $\gamma_k = Y$ and $\sigma_0 = X$. (b) The hypothesis $k \geq 1$ excludes any relation R for which $\hat{0}_R = \hat{1}_R$.)

Lemma 21 implies that every nontrivial maximal chain in the doubly-labeled poset associated with a relation gives rise to an informative attribute release sequence that tracks the chain. A partial converse holds as well:

Lemma 22 (Chains from Informative Attributes)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $y_1, \ldots, y_k$ is an informative attribute release sequence for R, with $k \geq 1$.*

Let $\gamma_i = (\phi_R \circ \psi_R)(\{y_1, \ldots, y_i\})$ and $\sigma_i = \psi_R(\gamma_i)$, for $i = 1, \ldots, k$.

Then $\{(\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1) < (X, \gamma_0)\}$ is a (not necessarily maximal) chain in P_R^+, with $\gamma_0 = \phi_R(X)$.

Consequently, one can obtain all informative attribute release sequences as subsequences of those constructed from maximal chains in P_R^+.

Comment about "length": The *length* $\ell(P)$ of a poset P is defined to be one less than the number of elements comprising a longest chain in the poset [22]. The *length* of an informative attribute release sequence $y_1, \ldots, y_k$ is k. These definitions match much like the dimension of a simplex is one less than the number of its elements. Consequently, one obtains:

Corollary 23 (Maximal Length)**.** *The maximum length of an informative attribute release sequence for a nonvoid relation R is $\ell(P_R^+)$. (If R has no iars, then the maximum length is 0.)*

Corollary 24 (Maximal Identification Length)**.** *Suppose R is a relation such that no attribute is shared by all individuals. For any $(\sigma, \gamma) \in P_R$, $\mathrm{r_{slow}}(\sigma) = \ell(P_{Q(\sigma,\gamma)}) + 2$.*

Collaboration Example Re-Revisited: Returning again to the travel guide example, observe in Figure 25 that $\ell(P_G^+) = 4$. This tells us, by Corollary 23, that a longest informative attribute release sequence for relation G contains four attributes. Indeed, we can pick three attributes to identify an individual, and then a fourth to form an inconsistency. How do we know that we can choose three attributes informatively to identify an individual? See, for example, $\mathrm{Lk}(\Psi_R, \texttt{Claire})$ in Figure 26, with associated lattice P_C^+ in Figure 27. In this case, $\ell(P_C) + 2 = \ell(P_C^+) = 3$. Moreover, by the construction of Lemma 21, one can read off four different such informative sequences, namely the first four sequences appearing in Figure 24.

We thus see that $\mathrm{r_{slow}}(\{\texttt{Claire}\}) = 3$, and as we have seen previously, $\mathrm{r_{fast}}(\{\texttt{Claire}\}) = 2$. In other words, if `Claire` has control over how to release information, she can draw out identification for three books, while the fastest anyone can identify her is via two books.

T	a	b	c	d
1	•	•	•	
2		•	•	•
3	•		•	•
4	•	•		•

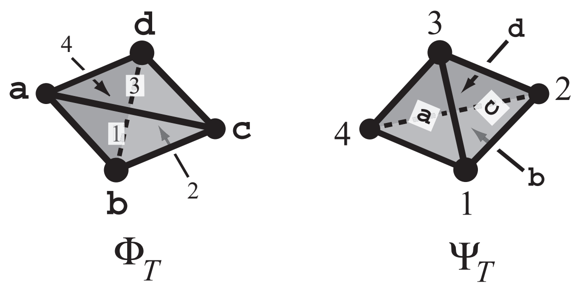


Figure 30: Relation T describes four individuals with four attributes, with Dowker complexes that are boundary complexes of tetrahedra, meaning they have homotopy type $\mathbb{S}^2$.

In contrast, consider the tetrahedral relation of Figure 30. The Dowker complexes are boundary complexes, so we know that no attribute or association inference is possible. This is evident from the lattice P_T^+ depicted in Figure 31 as well. It has length 4, just as did the travel guide lattice, but the inference structure is now different. For any $(\sigma, \gamma) \in P_T$, with $Q = Q(\sigma, \gamma)$ modeling $\mathrm{Lk}(\Psi_T, \sigma)$ on attributes γ, we see that $\Phi_Q = \partial(\gamma)$ and thus that $\ell(P_Q^+) = \ell(P_Q) + 2 = |\gamma|$. This tells us, by Theorem 20 and Corollary 24, that $\mathrm{r_{fast}}(\sigma) = \mathrm{r_{slow}}(\sigma) = |\gamma|$, as one would expect in an inference-free world. For a specific instance, Figure 32 depicts $Q = Q(\{3\}, \{\texttt{a}, \texttt{c}, \texttt{d}\})$ along with Q's Dowker complexes and the lattice P_Q^+.

10.7 Hidden Holes

We saw via Theorem 20 that whenever a nonempty set of attributes γ minimally identifies some set of individuals σ, then the link of σ, restricted to those simplices that are witnessed by attributes in γ, defines a sphere in both Dowker complexes. It is a topological hole.

All sets of individuals that are identifiable in some way, in other words, that appear in the doubly-labeled poset P_R of a relation, must be minimally identifiable in some way. That suggests there must be holes everywhere in a relation's Dowker complexes, and yet we do not see many holes. What is going on?

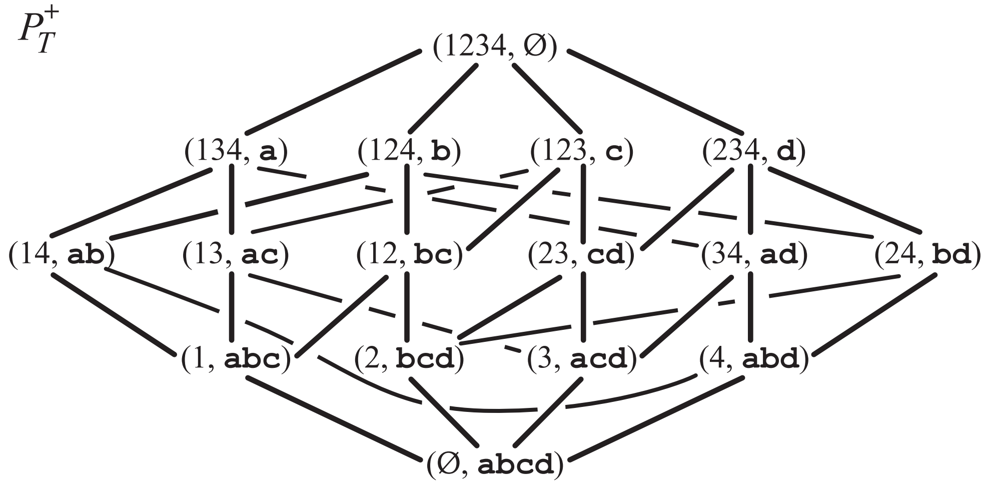


Figure 31: The lattice P_T^+ for the tetrahedral relation of Figure 30. Each element is an ordered pair of sets (σ, γ) such that $\sigma = \psi_T(\gamma)$ and $\gamma = \phi_T(\sigma)$. (We have elided commas and braces in sets, for ease of viewing.) This lattice is isomorphic to the Boolean lattice on four atoms, consistent with the fact that T preserves both attribute and association privacy. If one removes the top and bottom elements, the remaining poset P_T has $\mathbb{S}^2$ homotopy type.

The answer is that the restricted link construction $Q(\sigma, \gamma)$ focuses on a particular subrelation, thereby exposing/highlighting a potential hole. The hole could in fact be hidden, that is, filled-in by the encompassing relation. For instance, we saw that relation Q of Figure 32 defines an $\mathbb{S}^1$ hole. If Q happened to be a subrelation of relation R as in Figure 33, then Q would not appear as a hole when viewed in R, merely a boundary.

Notice that the lattice P_R^+ is isomorphic to the lattice P_Q^+. The difference is that for every lattice element (σ, γ), the set of individuals σ includes 3 in P_R^+ but not in P_Q^+. Consequently, the bottom element $(3, \texttt{acd})$ of P_R^+ is actually an element of the poset P_R, meaning $\Delta(P_R)$ is a cone, hence contractible. In contrast, the poset P_Q does not contain the bottom element $(\emptyset, \texttt{acd})$ of P_Q^+ and so $\Delta(P_Q)$ has $\mathbb{S}^1$ homotopy type.

Aside: Why not always focus on a relation's lattice rather than its doubly-labeled poset? Because the lattice is always contractible. Any informative topology lies in the poset. See [22].

Conclusion: Even though R is contractible, it offers the same choices for informative attribute release sequences as does Q. More generally, the analysis of this subsection suggests that one look for potential holes in *subrelations* of a given relation. Looking at links is one way to focus on subrelations. Removing individuals or attributes that represent cone apexes is another, as we just saw. More generally, any simplicial cycle may define a useful hole even though the hole appears to be filled-in. So long as one can remove any coboundary of that cycle, *by restricting the relation to a subrelation without destroying the cycle*, the cycle is informational. In particular, it offers opportunities for informative attribute release sequences, as the next subsection makes precise.

Q	a	c	d
1	•	•	
2		•	•
4	•		•

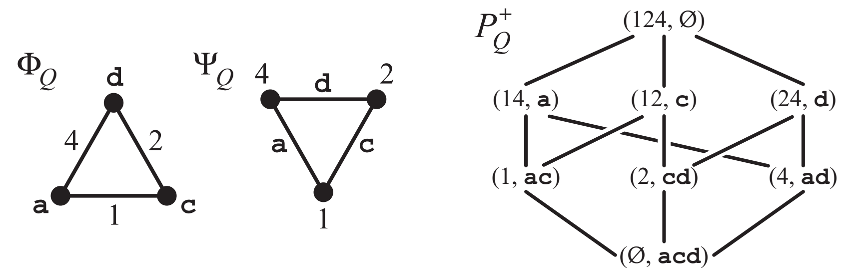


Figure 32: Relation Q models $\mathrm{Lk}(\Psi_T, 3)$, with T as in Figure 30.

R	a	c	d
1	•	•	
2		•	•
3	•	•	•
4	•		•

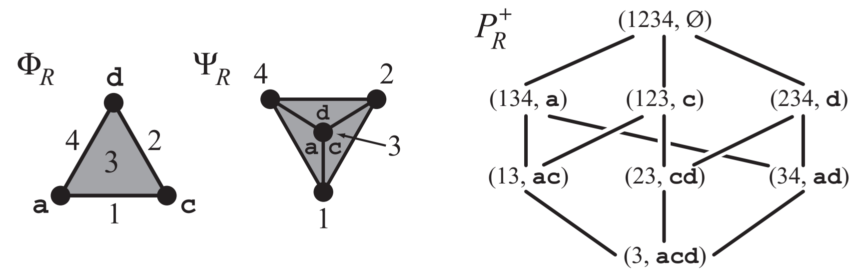


Figure 33: Relation R fills in the hole of relation Q from Figure 32. It is still true that Q models a link, namely $\mathrm{Lk}(\Psi_R, 3)$. Relations R and Q have the same lattice structure, but the bottom element of P_R^+ defines the set of individuals $\{3\}$, whereas the bottom element of P_Q^+ defines the empty set. Thus relation R defines a contractible poset for P_R, whereas relation Q defines an $\mathbb{S}^1$ hole for P_Q.

10.8 Bubbles are Lower Bounds for Privacy

We have seen minimal identifiability characterized by holes, via Theorem 20. The previous subsections make clear that the topological characterization of $\mathrm{r_{slow}}$ is not so direct. In this subsection we establish a sufficient condition. We will see that holes provide lower bounds for $\mathrm{r_{slow}}$. We will focus on a relation and its links, but these results apply more generally to any hidden holes made visible by focusing on subrelations, as suggested in the previous subsection.

The connection between a relation's poset P_R and its lattice P_R^+ suggests the following:

Definition 25 (Almost a Join-Based Lattice)**.** *Let P be a finite poset. We say that P is* <u>almost a join-based lattice</u> *if adjoining a new topmost element $\hat{1}$ means $P \cup \{\hat{1}\}$ is a join semi-lattice.*

Comments: (a) We adjoin a new $\hat{1}$ even if P already has a unique top (i.e., maximal) element. (b) Since P is finite, if P is almost a join-based lattice, then if we adjoin both a new topmost element $\hat{1}$ and a new bottommost element $\hat{0}$, the result will be a lattice. See also [22].

This definition leads to the following result (for a proof, see Appendix G):

Theorem 26 (Many Maximal Chains)**.** *Let P be almost a join-based lattice. Suppose P has reduced integral homology in dimension $k \geq 0$, that is, $\widetilde{H}_k(\Delta(P); \mathbb{Z}) \neq 0$.*

Then there are at least $(k+2)!$ maximal chains in P of length at least k.

Interpretation: The theorem says that a homology hole acts like a spherical hole, from the perspective of producing informative attribute release sequences. Consider again the tetrahedral relation of Figure 30. The Dowker complexes form two-dimensional spherical holes, so $k = 2$ and $(k+2)! = 24$. The poset P_T is the proper part of the lattice shown in Figure 31, that is, all the elements except the topmost and bottommost. There are indeed 24 different chains of length 2, i.e., containing three elements, in P_T.

These chains represent the 24 different ways in which one might start at a vertex of one of the Dowker complexes, walk from that vertex to the middle of an incident edge, then walk from the middle of that edge to the centroid of an encompassing triangle. For instance: the walk from the vertex $\{\mathtt{a}\}$ to the edge $\{\mathtt{a}, \mathtt{c}\}$ to the triangle $\{\mathtt{a}, \mathtt{c}, \mathtt{d}\}$ in Φ_T. One can think of this walk as sequential acquisition of attribute information about an individual in a particular order. The order may perhaps be determined by chance or perhaps by an individual purposefully releasing information in a particular order. Once (and only once) one has arrived at the centroid of the triangle, one has identified the individual uniquely (in this case, as individual #3).

With that observation, we finally see how the global geometry/topology of the Dowker complexes, as encoded in their doubly-labeled poset, affects inference, beyond the local simplicial collapses of the closure operators. We will presently formalize this insight via two corollaries to Theorem 26.

R	a	b	c
1	•	•	•
2	•	•	•
3	•	•	•

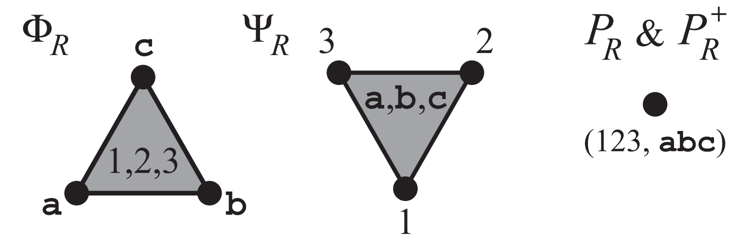


Figure 34: Relation R describes three individuals all of whom have the exact same three attributes. The Dowker complexes are both triangles, but the poset P_R is a single point. This single point captures the indistinguishability of the individuals and the attributes. In fact, $P_R^+ = P_R$, meaning one can infer everything from nothing (in the context of relation R).

We caution that the dimension of a simplex in a Dowker complex is not meaningful in and of itself, since the simplex may collapse under the closure operators. (Consider the example of Figure 34, in which the Dowker complexes are fully filled-in triangles, but the doubly-labeled poset is a single point.) Instead, the length of chains in a relation's poset is significant. Holes prevent these chains from being short, summarized as follows (proofs appear in Appendix G):

Corollary 27 (Holes Reduce Inference)**.** *Let R be a nonvoid relation. Suppose P_R has reduced integral homology in dimension $k \geq 0$. Then there are at least $(k+2)!$ maximal chains in P_R of length at least k.*

Corollary 28 (Holes Defer Recognition)**.** *Let R be a nonvoid relation and let $(\sigma, \gamma) \in P_R$.*

Define $Q = Q(\sigma, \gamma)$ as per Definition 19 and recall Definition 17, from pages 39–40.

Suppose P_Q is well-defined and has reduced integral homology in dimension $k \geq 0$.

Then there are at least $(k+2)!$ distinct informative attribute release sequences $y_1, \ldots, y_\ell$ for R, each with $\ell \geq k+2$, such that $\psi_R(\{y_1, \ldots, y_\ell\}) = \sigma$. Consequently, $\mathrm{r_{slow}}(\sigma) \geq k+2$.

Comment: Since $(\sigma, \gamma) \in P_R$ and by the assumptions about P_Q, relation $Q(\sigma, \gamma)$ models the link $\mathrm{Lk}(\Psi_R, \sigma)$.

Terminology: Here and elsewhere, the term `'distinct'` means `'different'` or `'distinguishable'`, as determined by the given context. For instance, the two sequences `a`, `b`, `c` and `a`, `c`, `b` are distinct sequences even though the underlying set is $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ in both cases.

Collaboration Example Once Again: The Dowker complexes for the travel guide example of Figure 23 have $\mathbb{S}^1$ homotopy type, meaning P_G has homology in dimension $k = 1$. Corollary 27 therefore says that there are at least 6 maximal informative attribute release sequences in P_G. Being maximal, each such sequence must identify some author, since each author is uniquely identifiable via relation G. In fact, we saw that there were 4 different maximal informative attribute release sequences for identifying any one author. Since there are 5 authors, P_G actually contains at least 20 distinct maximal informative attribute release sequences. Indeed, one can readily see, via P_G^+ in Figure 25 on page 36, that P_G contains exactly 20 maximal informative attribute release sequences.

Can we find these 20 sequences via our corollaries? Not by looking at individual authors, since, as we saw via Figure 26, the link of any one author is contractible, meaning that Corollary 28 does not help us directly.

There is more to be said, however: The proof of Theorem 26 actually establishes that, for certain representatives of a homology class, the maximal elements in the support of that representative each give rise to $(k+1)!$ many chains. In the collaboration example, by choosing the homology generator appropriately, this implies that for each author there are at least two informative attribute release sequences for identifying the author. That gives us 10 sequences overall for relation G. To find 20, we would likely want to examine links of pairs of co-authors. There are 10 such links, 5 of which[3] look similar to the one in Figure 28 on page 38. Each of those is an instance of $\mathbb{S}^0$, meaning each has two different iars for identifying the pair of co-authors. That therefore gives us 10 iars for identifying certain pairs of co-authors, and thus 20 iars for identifying individual authors (each author participates in two of the identifiable pairs).

Corollary 28 further allows us to conclude that the maximal length of an informative attribute release sequence for identifying an identifiable pair of co-authors is at least two. Consequently, the maximal length of an informative attribute release sequence that identifies a given individual author must be at least (and thus exactly) three.

[3]For the curious reader: Each of the remaining 5 links is a singleton. For instance, $\mathrm{Lk}(\Psi_G, \{2, 4\})$ is the simplicial complex consisting of the single vertex $\{3\}$. It is generated in the corresponding link relation by attribute Caen. These comments are another way of saying that the only author who has co-authored a book together with both `Ben` and `David` is `Claire`, producing the travel guide for Caen. Observe as well that the pair of co-authors $\{\mathtt{Ben}, \mathtt{David}\}$ does *not* appear as the σ component of an element (σ, γ) in P_G or P_G^+. This means one cannot identify just the pair of co-authors $\{\mathtt{Ben}, \mathtt{David}\}$, but invariably infers the full triple $\{\mathtt{Ben}, \mathtt{Claire}, \mathtt{David}\}$ of collaborators, given relation G.

11 Experiments

An individual may wish to reveal information about himself/herself while delaying full identification. We saw in Section 10.8 that homology provides a lower bound on the number and length of such informative attribute release sequences. The lower bound need not be tight. In order to explore these results experimentally, we examined two datasets of different character:

Medals: We obtained this dataset in August 2014 from

`http://www.tableausoftware.com/public/community/sample-data-sets`.

The dataset contained information about athletes who participated in the Olympics during the years 2000–2012. The attribute fields that we considered were:

`Age, Country, Year, Sport, Gold Medals, Silver Medals, Bronze Medals`

(The last three fields counted the number of medals won by an athlete.)

Every athlete therefore had exactly 7 attributes, with each attribute taking on one of a finite discrete set of pairwise exclusive values. We represented these 7 dimensions of multivalent attributes as a collection of 223 binary attributes.

There were 8613 individuals (we regarded the same physical person in different years as distinct athletes), who partitioned into 6955 equivalence classes (for team sports, athletes often were indistinguishable).

The result was a binary relation M with 6955 rows and 223 columns.

Jazz: We assembled this relation in June 2015 by examining the website

`http://www.redhotjazz.com`.

The website contained information about jazz musicians and bands, mainly from the early to late-mid 20th century.

We assembled a relation J whose rows were indexed by musicians and whose columns were indexed by bands, with $(m, b) \in J$ meaning that musician m played in band b.

The result was a binary relation J with 4896 rows and 990 columns.

Caution: We were somewhat but not particularly careful in determining whether similar names constituted different spellings of the same musician's actual name. For some bands, the website listed one or more bandmembers as "unknown". We ignored those bandmembers. We ignored bands for whom we could not determine any bandmembers. Since our goal was to examine and compare homology and informative attribute release sequences, merely constructing a relation was sufficient for our purposes. However, it is unlikely that the resulting relation satisfied the assumption of relational completeness stated on page 7, relative to data obtainable from other sources.

We encountered the jazz website because it was the source of data for a paper on collaboration networks [12] that explored the dual nature of individuals and attributes.

The paper constructed two graphs, one with musicians as vertices and bands as edges, the other with those roles reversed, then analyzed the information each representation highlighted. We may view those graphs as the 1-skeleta of our Dowker complexes.

11.1 Compare and Contrast

We review some key differences between the two relations M and J.

Identifiability: The original 8613 individuals in the Olympic Medals dataset were not all uniquely identifiable. For some athletes, even knowing an athlete's full set of 7 attributes left ambiguity as to the athlete's identity. This was true for 2810 of the athletes. Fortuitously, an athlete's ambiguity was fully symmetric, meaning that one could in fact partition the set of all athletes into equivalence classes. This symmetry was likely due to the fact that some competitions involved teams, with team members indistinguishable from each other. Each equivalence class then formed a uniquely identifiable "individual" in relation M.

For the Jazz relation, 863 of the 4896 musicians were uniquely identifiable, but 4033 were not. Unfortunately, this time the ambiguity was not fully symmetric. One could again partition the 4033 individuals into 1022 equivalence classes based on having identical rows in J. However, some rows remained subsets of other rows, giving a directionality to the ambiguity. For this reason, we did not pass to equivalence classes.

Attribute Size: In the medals relation M, every individual had exactly 7 binary attributes, describing one value for each of the 7 possible fields: `Age, Country, Year, Sport, Gold Medals, Silver Medals, Bronze Medals`. Consequently, there were also always exactly 7 binary attributes in each relation modeling the link of an athlete in Ψ_M.

In the Jazz dataset, there was no structural bound to the number of bands in which a musician might have played, so a musician's attributes could be many. The largest number of bands in which any one musician played was in fact 44. The average was a little over 2 and the median 1. Dually, the largest band had 288 musicians, with an average of 10.4 and a median of 7.

Link Size: For M, the number of other athletes in any given athlete's link was always close to the entire set of possible athletes. With only 7 attribute fields and few distinct values, any two athletes shared almost certainly some attribute value (for instance, winning zero gold medals).

In contrast, for the 767 musicians in J for whom we computed links (described further in Section 11.4), the number of other musicians in any given musician's link was relatively small. The average was 55.3, the median 37, with a maximum of 301. With musicians generally playing in few bands, each collaborated artistically on average with only a few score fellow musicians of the 4895 musicians in the database.

11.2 Homology Computations

For each of the link relations discussed below, we computed homology of the Dowker complex Φ_Q, with relation Q modeling the link.[4] Since our goal was to find lower bounds for informative attribute release sequences, we modified Φ_Q slightly, as suggested by Section 10.7. Specifically, whenever Φ_Q was a cone with more than one maximal simplex, we removed all its cone apexes.

Comment: The homology lower bound results of Section 10 and Appendix G do not depend directly on the chain coefficients being integers (of course, the actual homology observed may depend on the type of coefficients). We therefore computed homology with $\mathbb{Z}_2$ coefficients, using the `Perseus` software previously written at the University of Pennsylvania. We downloaded an executable version in 2014 from `http://www.sas.upenn.edu/~vnanda/perseus/`.

11.3 Homology and Release Sequences in the Olympic Medals Dataset

Overall Homology: A collection of k attributes, each taking on one of a finite discrete set of pairwise exclusive values, produces Dowker complexes with homotopy types that are wedge sums of $\mathbb{S}^{k-1}$s, assuming all possible combinations of attributes are represented by individuals.

Consequently, with every individual having exactly 7 attributes, one might expect to see some homology in dimension 6. But of course, not every combination is possible. For instance, no one athlete is going to simultaneously win the gold, silver, and bronze medals in the same event. *From this perspective, real-world constraints show up as absence of potential homology.* In fact, relation M had the Betti numbers described in Table 2, computed using $\mathbb{Z}_2$ coefficients:

d	0	1	2	3	4
β_d	1	0	23	757	503

Table 2: Betti numbers for the topology of the Olympic Medals relation M.

The table does suggest that there could be quite a few informative attribute release sequences of length at least 5 for identifying athletes ($\beta_4 \neq 0$ in P_M implies length 5 iars).

Link Homology: We computed the link of each athlete in M (or more precisely, of each equivalence class), and determined homology for the resulting relation, with the modifications mentioned before. Specifically, we removed all cone apexes from an athlete's Dowker complex Φ_Q (assuming it contained more than one maximal simplex) before computing homology, with Q being the link relation. Of the 6955 links, 3822 contained attribute cone apexes in Φ_Q.

Table 3 summarizes the results. One may conclude more strongly now that (at least) 2198 athletes could find (at least) 120 different ways of releasing (at least) 5 of their 7 attributes without identifying themselves uniquely prior to having released all 5 attributes ($\beta_3 \neq 0$ in P_Q minimally implies 5! many iars of length 5 for relation M, by Corollary 28 on page 45).

Informative Attribute Release Sequences: We computed a maximal length informative attribute release sequence for each link relation. One can find such a sequence by searching for a least-cost path from $\hat{1}_Q$ to $\hat{0}_Q$ in P_Q^+, picking attributes along the way as per the construction of Lemma 21 on page 41, with cost being the number of attributes inferred as one traverses the path. Here Q is again the link relation. Of the 6955 athletes, 6229 actually had a maximal

[4] Formally, the link is equal to Ψ_Q. By Dowker's Theorem, Ψ_Q and Φ_Q have the same homology.

d	0	1	2	3	4
# of athletes	229	1355	2773	2198	57
$\max_{\text{athletes}} \beta_d$	2	4	7	4	2

Table 3: Histogram indexed by dimension d, describing athletes whose links $\text{Lk}(\Psi_M, \text{athlete})$ had reduced homology in dimension d (after removal of attribute cone apexes from the dual complexes), for the 6955 athletes in the Olympic Medals relation M. (525 of the 6955 links had no reduced homology; they do not appear in the histogram.) Also shown are the maximum Betti numbers seen in each dimension, with the maximum taken over all possible athletes.

informative attribute release sequence of length 7. Each such athlete could order his/her attributes in such a way that his/her identity would not become fully known until s/he had released all 7 attributes. Of the remaining athletes, 719 had a maximal informative attribute release sequence of length 6, and 7 had a maximal length of 5.

Of course, Corollary 28 on page 45 makes a stronger claim, suggesting possible permutability of some attributes. Consequently, we computed for each link relation all possible isotropic sets of attributes (see again Definition 15 on page 38, now with Q in place of R). Table 4 summarizes the results:

$\lvert\kappa\rvert$	2	3	4	5	6
# of athletes	6955	6955	6955	5568	171
$\max_{\text{athletes}} \lvert\{\kappa\}\rvert$	21	35	35	21	5

Table 4: Histogram indexed by size $|\kappa|$, describing athletes whose link relations contained isotropic attribute sets κ. An athlete could have several distinct (possibly overlapping) such sets for any given size. Also shown therefore are the maximum numbers of such sets, with the maximum taken over all possible athletes. For example: 171 athletes had at least one isotropic set of size 6, and the maximum number of isotropic sets of size 6 any one athlete had was 5.

Scatterplot: Finally, we computed for each link a pair of numbers (h, i), with h representing a measure of link homology and i representing a measure of informative attribute release sequences for the link relation. The resulting scatterplot appears in Figure 35.

The exact formulas for h and i are not that significant, but we mention them here for completeness. To obtain a measure of homology, we assembled for each link a vector with the Betti numbers computed earlier: $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$. We determined maximum values for each component (as given in Table 3). We could then think of any such vector as defining, in reverse order, a varying-radix numeral. We converted that numeral to an integer. For example, a link with Betti vector $(2, 3, 6, 0, 0)$ would have h value $2 + 3{\cdot}(2 + 1) + 6{\cdot}(4 + 1){\cdot}(2 + 1) = 101$. A link relation that remains contractible after removal of attribute cone apexes would have h value 1. In order to graph the scatterplot nicely, we scaled the h-axis by taking a fourth root.

We computed a link's i value similarly, now from the following vector of data: $(\ell_{\max}, c_2, c_3, c_4, c_5, c_6)$. Here $\ell_{\max}$ is the largest ℓ in an informative attribute release sequence $y_1, \ldots, y_\ell$ for the link relation, while c_k is the number of different isotropic attribute sets κ in the link relation such that $|\kappa| = k$. We scaled the i-axis by taking a logarithm.

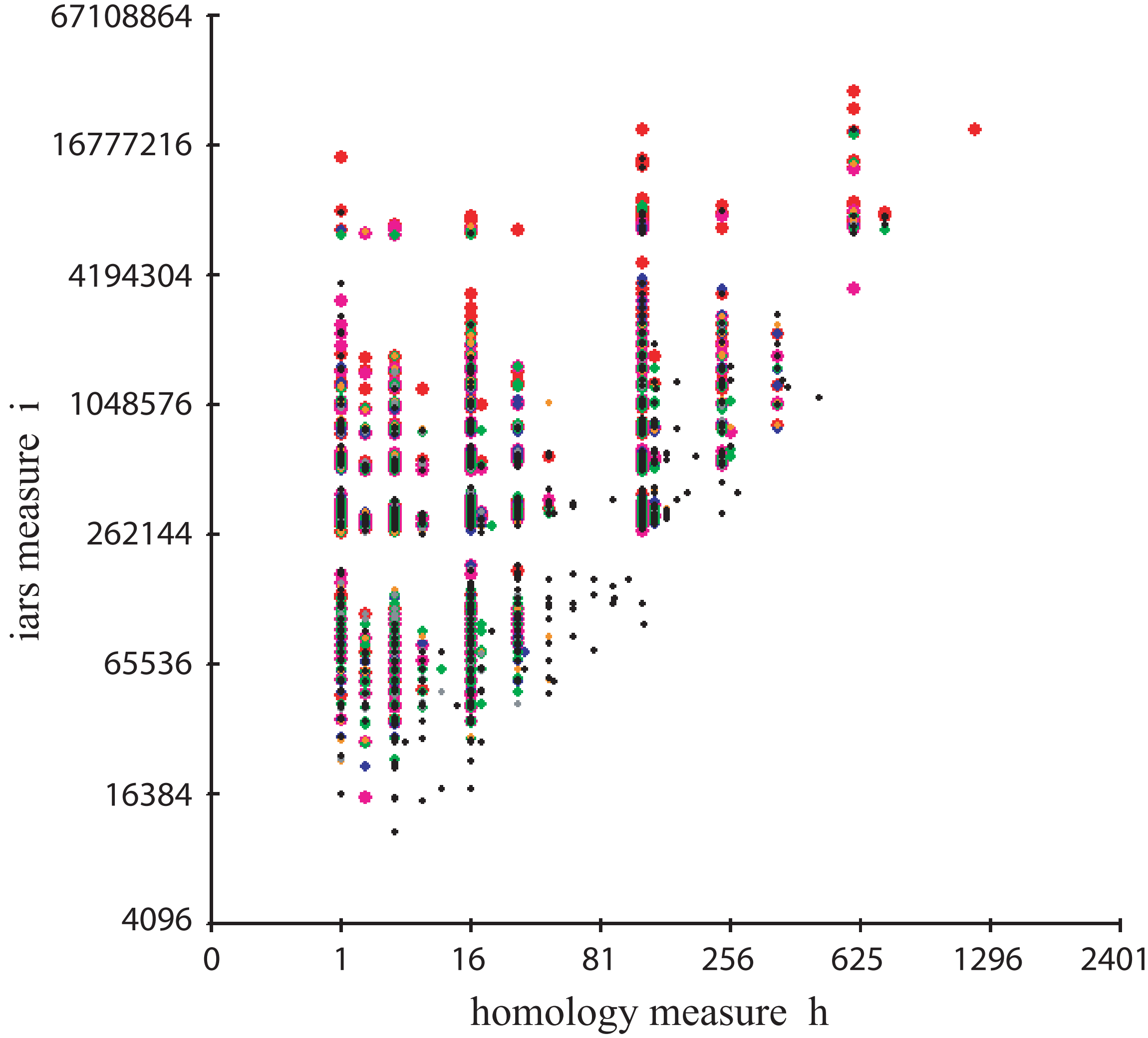


Figure 35: Scatterplot describing each athlete's link in the medals relation M. The scatterplot shows for each link a point (h, i), with h a measure of the link's homology (after removal of attribute cone apexes) and i a measure of how many significant informative attribute release sequences exist for the link relation. The scatterplot suggests that the homology measure h serves as a loose lower bound for the iars measure i. See also Corollary 28 on page 45.

(The colors and radii indicate the numbers of athletes in the links. The color ordering and size boundaries are:

BLACK$-_{6821}-$SILVER$-_{6831}-$ORANGE$-_{6851}-$GREEN$-_{6859}-$BLUE$-_{6865}-$MAGENTA$-_{6872}-$RED.

In this figure, the boundaries between colors were chosen so that each bucket would hold roughly 1000 links. As one can see, the number of athletes in a link was generally large.)

11.4 Homology and Release Sequences in the Jazz Dataset

Overall Homology: Given the large number of bands in which some musicians played, and given memory constraints of our machines, we did not compute homology for the whole Jazz relation J. Instead, we computed homology for restricted relations consisting of musicians who played in fewer than 20 bands. This covered 4856 of the 4896 musicians in the overall relation. Since we did not see any homology above dimension 2 after considering several of these restricted cases, we used the 4-skeleton (all simplices of dimension 4 or less) of Φ_J as proxy for the topology of the whole relation J and computed its homology. Table 5 summarizes the results. Using a graph algorithm, we verified that the whole relation J did indeed have 107 components, as indicated by β_0 for the 4-skeleton $\Phi_J^{(4)}$. Given the low dimension of homology for the restricted relations, conceivably even J might not tell us much about the length of informative attribute release sequences for the various musicians, suggesting we look at links.

b	Σ	m	β_0	β_1	β_2	β_3	β_4
14	$\Phi_{J\|b}$	4819	111	613	20	0	0
15	$\Phi_{J\|b}$	4831	111	613	32	0	0
16	$\Phi_{J\|b}$	4838	111	605	42	0	0
17	$\Phi_{J\|b}$	4848	110	603	58	0	0
18	$\Phi_{J\|b}$	4851	110	603	65	0	0
19	$\Phi_{J\|b}$	4856	109	596	75	0	0
∞	$\Phi_J^{(4)}$	4896	107	550	93	10	–
15	$\Phi_{J'}$	767	18	595	32	0	0

Table 5: Betti numbers for subcomplexes Σ of Φ_J, with J being the Jazz relation. The first six rows correspond to restrictions of J to musicians who played in at most b bands. For each row, m indicates the number of musicians in the relation. The penultimate row describes the 4-skeleton of Φ_J. The last row refers to a relation J' described further in the text.

Link Homology: We computed the link of some of the musicians in J, and determined homology for the resulting relations (again after removal of attribute cone apexes, when appropriate). Table 6 summarizes the results. Given the inability to uniquely identify some musicians even knowing all their bands (as described in Section 11.1) and the difficulty of computing homology when musicians played in many bands, we computed links only for a subset of the musicians. We required each musician to be uniquely identifiable, to have played in at most 15 bands, and to have a nontrivial link. There were 767 such musicians. Betti numbers for the relation J' representing the restriction of J to these 767 musicians also appear in Table 5. (Note, however, that we computed the complete link $\mathrm{Lk}(\Psi_J, \text{musician})$ for each of the 767 musicians, not merely $\mathrm{Lk}(\Psi_{J'}, \text{musician})$.) We removed attribute cone apexes from the link relation for 106 of these 767 musicians.

These results suggest that the relationships to other musicians do indeed *not* have many high-dimensional holes in them. Recall, by Corollary 28 on page 45, one can assert the existence of at least $(k+2)!$ distinct informative attribute release sequences of length at least $k+2$ for any musician with a k-dimensional hole. For almost all musicians this lower bound means 2

d	0	1	2	3
# of musicians	604	145	20	1
$\max_{\text{musicians}} \beta_d$	7	6	3	1

Table 6: Histogram indexed by dimension d, describing musicians whose links $\text{Lk}(\Psi_J, \text{musician})$ had reduced homology in dimension d (after removal of attribute cone apexes from the dual complexes), for the 767 musicians who were uniquely identifiable in J, played in at most 15 bands, and had nontrivial link. (52 of the 767 links had no reduced homology; they do not appear in the histogram.) Also shown are the maximum Betti numbers seen in each dimension, with the maximum taken over the 767 possible musicians. For $d = 0$, this means that 604 of the 767 musicians had collaborations with other musicians that split into pairwise disjoint groups. The maximum number of such components for any one musician was 7.

sequences of length 2, for some it means 6 sequences of length 3, for a few it means 24 sequences of length 4, and for one musician it means 120 sequences of length 5. These implications are roughly consistent with the data for informative attribute release sequences described next, though, as expected for the theoretical reasons discussed earlier, they constitute lower bounds.

Informative Attribute Release Sequences: We computed a maximal length informative attribute release sequence for each link relation. Table 7 summarizes the results. We mention in passing: Any attribute release sequence that was informative for a musician's link relation was also informative for the encompassing relation J (by Lemma 12(ii) on page 26). For a few musicians, the maximal sequence found within the link relation Q could be further extended in the encompassing relation J, with a prefix of one attribute, namely an attribute shared by all members of the link, yet remain informative and identifying within J. This occurred for the 17 musicians whose maximum sequence length ℓ in the musician's link was 1.

We also computed for each link relation all possible isotropic sets of attributes. Table 8 summarizes those results.

ℓ	1	2	3	4	5	6	7	8	9	10	11
# of musicians	17	248	218	125	72	35	23	15	11	2	1

Table 7: Histogram of musicians, indexed by length ℓ of a longest informative attribute release sequence for the musician's link relation, for the 767 musicians described in the text.

$\lvert\kappa\rvert$	2	3	4	5
# of musicians	750	219	49	3
$\max_{\text{musicians}} \lvert\{\kappa\}\rvert$	105	202	40	2

Table 8: Histogram indexed by size $|\kappa|$, describing musicians whose link relations contained isotropic attribute sets κ. Also shown are the maximum numbers of such sets, with the maximum taken over the 767 possible musicians described in the text.

Scatterplot: We computed for each link a pair of numbers (h, i), with h representing a measure of homology and i representing a measure of the link's informative attribute release sequences, much as for the medals relation M of Section 11.3. Figure 36 depicts the scatterplot.

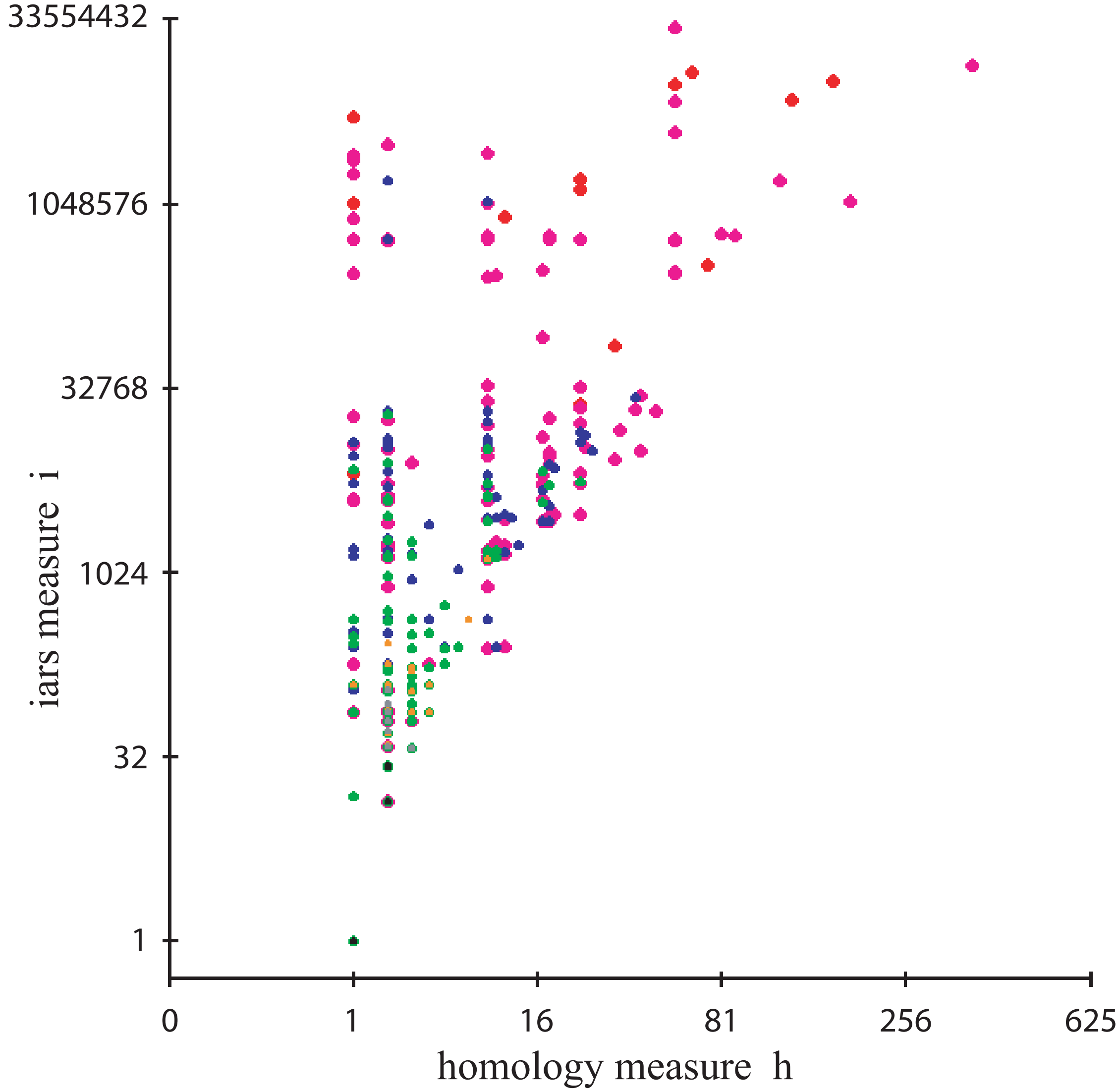


Figure 36: Scatterplot describing the links computed for 767 of the musicians in the Jazz relation J. The scatterplot shows for each link a point (h, i), with h a measure of the link's homology (after removal of attribute cone apexes) and i a measure of the link's informative attribute release sequences.

(The colors and radii indicate the numbers of musicians in the links. Link sizes were fairly small. The color ordering and size boundaries are:

BLACK$-_5-$SILVER$-_{10}-$ORANGE$-_{20}-$GREEN$-_{50}-$BLUE$-_{100}-$MAGENTA$-_{200}-$RED.

In this figure, the buckets could hold noticeably varying numbers of links.)

12 Inference in Sequence Lattices

We have seen how a relation gives rise to a lattice via the Galois connection, as per Definition 13 on page 35. The lattice structure describes the ways in which privacy may be preserved or lost. Consequently, when thinking about privacy, perhaps one can also start with lattices that do not necessarily arise initially from relations.

This section will look at inferences from sequences of observations. The next section examines strategy obfuscation in planning with uncertainty.

We should mention some equivalences: Lattices are particular kinds of partially ordered sets (posets). Posets and simplicial complexes are topologically identical; one can move back and forth between these representations while preserving homeomorphism type (see [22] and Appendix A). Furthermore, one may describe a finite simplicial complex by a relation in several different ways that preserve homotopy type, including ways in which one of the two resulting Dowker complexes is identical to the original simplicial complex. For instance, maximal simplices can play the role of individuals and vertices can play the role of attributes. In short, one has three different categories of structures with which to think about privacy: relations, simplicial complexes, and lattices. One may start with any one representation and build the other two from that.

12.1 Sequence Lattices for Dynamic Attribute Observations

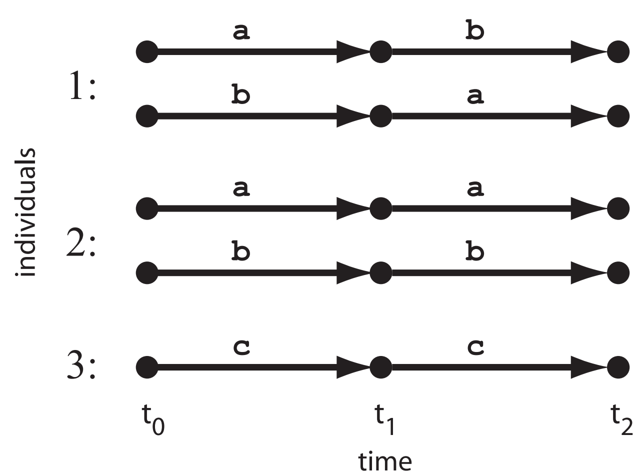


Figure 37: Three types of individuals and the attributes each might reveal in two successive time intervals.

Consider the dynamic process of Figure 37. The process models observations of individuals who reveal attributes over successive time steps. There are three possible individuals (or more generally, types of individuals). The first individual emits attributes "`a`" and "`b`" alternatingly at successive times, but one does not know which of those attributes one might see first. The second individual always emits the same attribute, either "`a`" or "`b`", but one does not know *a priori* which it is. The third individual always emits the same attribute "`c`".

S	a	b	c
1	•	•	
2	•	•	
3			•

T	aa	bb	ab	ba	cc
1			•	•	
2	•	•			
3					•

Figure 38: Relation S describes individuals and single attributes, while T describes individuals and sequences of two attributes.

A relation for these (types of) individuals that models the individuals in terms of single attributes appears as relation S in Figure 38. Individual #3 is distinguishable from the other two individuals, but the relation provides no means for distinguishing those two individuals from each other. The relation is homogeneous with regard to single attributes for individuals #1 and #2. Of course, we can see from the dynamic process of Figure 37, that distinguishing information appears via sequences of two attributes. Relation T of Figure 38 models such sequences. Now all three individuals are uniquely identifiable. Should one wish to model inferences based on both one and two observations, one could use the relation $S \cup T$.

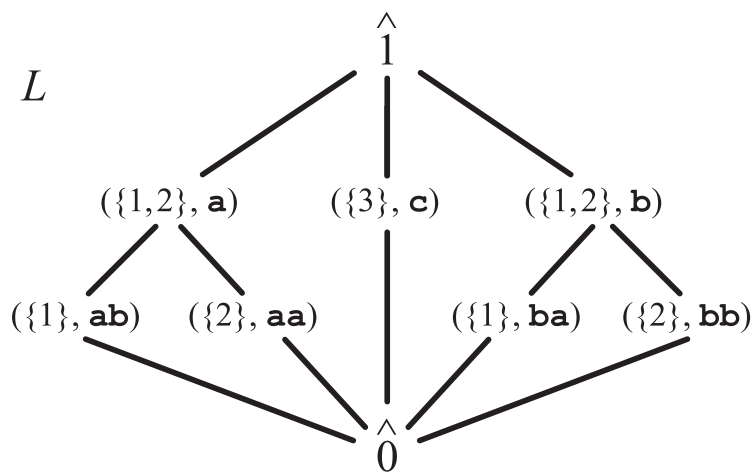


Figure 39: Lattice representing the dynamic process of Figure 37.

That jump from single to double attributes is useful, but where does it come from intrinsically? After all, without additional knowledge, we might simply consider infinitely long sequences, even though those would not add anything in this example. In fact, the dynamic process of Figure 37 gives us the information. It is itself basically a decision tree that amounts to the lattice of Figure 39. In that figure, we have annotated each internal node of the lattice with an ordered pair, consisting of a set of individuals and either a single attribute or a sequence of two attributes. This lattice differs from previous ones in this report in that a set of individuals (or attributes, more generally) is no longer constrained to appear in at most one node of the lattice. By allowing multiple nodes, we enhance our ability to encode state in the lattice. For example, observing attribute "a" carries different meaning depending on whether one has already seen attribute "a" or attribute "b" or no attribute at all. Also: While we could have included ({3}, cc) in the lattice, we did not need that element.

In the lattice of Figure 39 it is tempting to merge the two identifying nodes for individual #1 into one node and to merge the two identifying nodes for individual #2 into one node. There is apparently no harm in doing so, in that the decision process would still be correct. However, the resulting structure would no longer be a lattice but merely a poset. That may or may not be desirable in a given application. For instance, using homology to estimate lower bounds for how long one can delay identification suggests using almost a join-based lattice, if one wishes to fulfill the hypotheses of Theorem 26 on page 44.

If we did want to merge nodes as just described, while maintaining a lattice, then we would perhaps also merge the two nodes containing the set $\{1, 2\}$, giving us the lattice of Figure 40. This lattice is similar to the lattice $P^{+}_{S\cup T}$ that one would construct from the relation $S \cup T$, except that it does not include singleton attributes in the nodes identifying individuals #1 and #2 and it does not include the sequence "`cc`" in the node identifying individual #3.

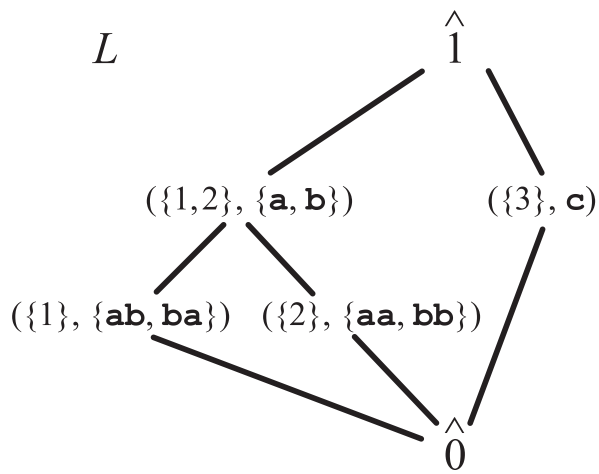


Figure 40: Modified lattice of Figure 39, after merging some nodes.

Regardless, the lattices of Figures 39 and 40 encode the inferences possible for the dynamic process of Figure 37. In particular, if we observe either attribute "`a`" or attribute "`b`", then we know the set of possible individuals is $\{1, 2\}$; we have excluded individual #3. Moreover, if we observe any two-attribute sequence, with attributes drawn from $\{\texttt{a}, \texttt{b}\}$, then we can identify the observed individual uniquely as either #1 or #2. Thus the required sequences come directly from the dynamic process, not requiring an explicit intermediate representation as a relation. (One might argue, however, that a relation is implicit in our reasoning.)

12.2 Lattices of Stochastic Observations

The dynamic sequence perspective incorporates repeated randomized response within the lattice framework. Instead of arising via a (non)deterministic process as in Figure 37, the attributes "`a`" and "`b`" for two (types of) individuals could flow from a stochastic process. One obtains an infinite lattice determined by increasingly longer sequences of observations. Depending on the confidence intervals one wishes to set, one obtains stochastic decision regions such as those sketched in Figure 41, with a central region of ambiguity, bounded by regions of exclusion, for identifying individuals.

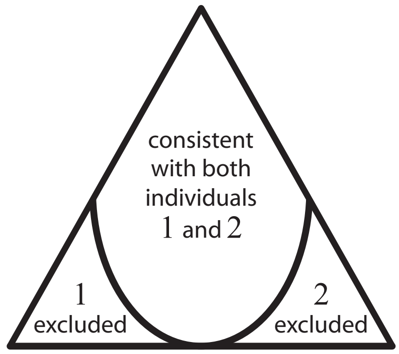


Figure 41: Sketch of an inference lattice for sequences of randomized response queries.

12.3 General Inference Lattices

Lattices are useful tools for inference. Rather than work with completely arbitrary lattices, we give here a definition that makes explicit the existence of two underlying structures over which we wish to perform inferences. However, we no longer assume a pair of underlying discrete spaces X and Y for individuals and attributes, but instead posit posets P and Q. The connection to our earlier relational perspective is that P would be the powerset of X and Q the powerset of Y. By allowing potentially different posets P and Q for a given lattice L, one can in some instances obtain different "views" of that lattice, thereby increasing flexibility in the interpretation process. For instance, Q might consist of all sequences up to a specified length or it might consist of *sets* of such sequences.

Definition 29 (Inference Lattice)**.** *Let P and Q be finite posets.*

An inference lattice L with respect to P and Q *is a bounded lattice whose proper part $\overline{L}$ consists of ordered pairs (p, q), with $p \in P$ and $q \in Q$, satisfying the following conditions:*

For all (p_1, q_1) and (p_2, q_2) in $\overline{L}$:

(i) $(p_1, q_1) \leq_L (p_2, q_2)$ *if and only if* $p_1 \leq_P p_2$ *and* $q_1 \geq_Q q_2$*;*

(ii) $(p_1, q_1) \vee_L (p_2, q_2)$ *is either* $\hat{1}_L$ *or a pair* $(p, q) \in \overline{L}$ *such that p is an upper bound for both p_1 and p_2 in P and q is a lower bound for both q_1 and q_2 in Q;*

(iii) $(p_1, q_1) \wedge_L (p_2, q_2)$ *is either* $\hat{0}_L$ *or a pair* $(p, q) \in \overline{L}$ *such that p is a lower bound for both p_1 and p_2 in P and q is an upper bound for both q_1 and q_2 in Q.*

(Note that $\hat{0}_L <_L (p, q) <_L \hat{1}_L$ *for every* $(p, q) \in \overline{L}$*, when* $\overline{L} \neq \emptyset$*.*
Also, be aware that $\overline{L}$ need not*, and generally will not, contain all possible pairs (p, q) of $P \times Q$.)*

Inference Protocol: Suppose we have observed some $q \in Q$. How should we interpret that observation in terms of the lattice L? Here is a possible protocol:
(In terms of our earlier relational model, one may view this protocol as inferring sets of individuals from sets of attributes.)

- Let $\Gamma = \{(p', q') \in \overline{L} \mid q \ \leq_Q \ q'\}$.
- If $\Gamma = \emptyset$, then we view q as inconsistent, producing interpretation $\hat{0}_L \in L$.
- Otherwise, let $\Gamma_{\max}$ consist of all the maximal elements of Γ (maximal with respect to the partial order on L). We view q as implying this set of elements in L. One can project each of those elements onto its P coordinate, if that is useful.

There is a dual protocol for interpreting an observation $p \in P$:
(In terms of our earlier relational model, one may view this protocol as inferring sets of attributes from sets of individuals.)

- Let $\Sigma = \{(p', q') \in \overline{L} \mid p \ \leq_P \ p'\}$.
- If $\Sigma = \emptyset$, then we view p as inconsistent, producing interpretation $\hat{1}_L \in L$.
- Otherwise, let $\Sigma_{\min}$ consist of all the minimal elements of Σ (minimal with respect to the partial order on L). We view p as implying this set of elements in L. Again, one can project each of those elements onto its Q coordinate, if that is useful.

Comments: (1) In our previous relational setting, the structure of Galois lattices ensured that, for nonempty observations, each of $\Gamma_{\max}$ and $\Sigma_{\min}$ never contained more than one element. That need not be true for general inference lattices. (2) One may augment the previous protocols, so as to regard some element(s) of Q much like the empty attribute simplex, giving interpretation $\hat{1}_L \in L$. Similarly, some element(s) of P might have interpretation $\hat{0}_L \in L$.

Example: Consider Figure 42. Poset P models subsets drawn from the set of two individuals $\{1, 2\}$, while poset Q models sequential observations of "`a`" and "`b`", of lengths one and two, as in our earlier example of Figure 37. (For presentational simplicity, P and Q ignore individual #3 and attribute "`c`", instead focusing on individuals $\{1, 2\}$ and attributes $\{$`a`, `b`$\}$.) Let lattice L be as in Figure 39. Assume the interpretation of $\emptyset \in P$ is $\hat{0}_L$ in L, and that of $\hat{0} \in Q$ is $\hat{1}_L$.

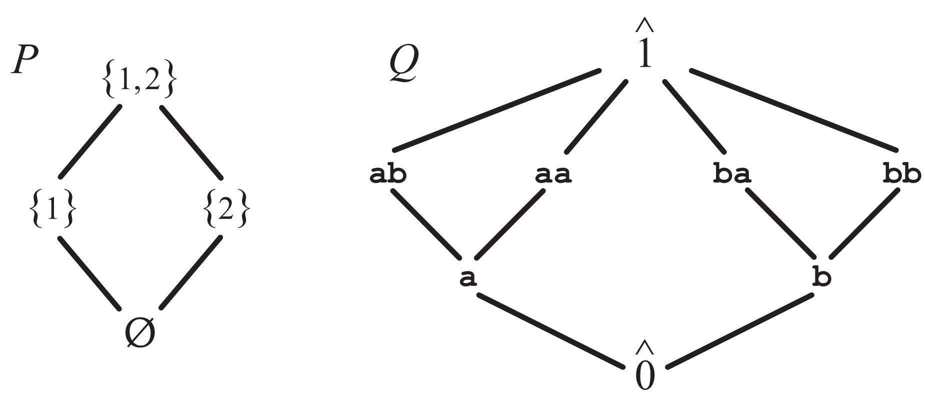
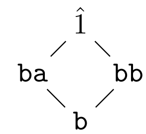
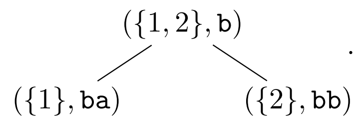


Figure 42: Poset P models some sets of individuals; poset Q models some sequences of attributes.

Observing an attribute: Suppose we have observed attribute "b", i.e., $q = \mathtt{b}$. What can we infer from q in P via L? Let us follow the protocol given earlier:

- The subposet of Q consisting of elements q' greater than or equal to q is: 

.

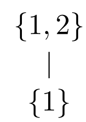
- Consequently, Γ is the following subposet of L: 

.

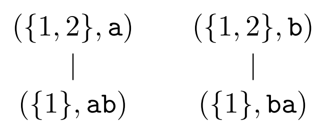
- There is one maximal element in Γ, so $\Gamma_{\max} = \{(\{1,2\},\mathtt{b})\}$.

Projecting onto the P component tells us how to interpret q: The observation "b" must have come from either individual #1 or individual #2, as one would hope. (This conclusion would hold as well if P had modeled individual #3 and if Q had modeled attribute "c".)

Observing an individual: Suppose we have observed individual #1, i.e., $p = \{1\}$. What can we infer from p in Q via L? Again, let us follow the inference protocol given earlier:

- The subposet of P consisting of elements p' greater than or equal to p is: 

.

- Consequently, Σ is the following subposet of L: 

.

- The minimal elements of Σ give us $\Sigma_{\min} = \{(\{1\},\mathtt{ab}),\ (\{1\},\mathtt{ba})\}$.

Projecting onto the Q component tells us how to interpret p: The individual observed can or did reveal one of the two-attribute sequences "ab" or "ba".

Comment: The poset Q of Figure 42 would not be relevant for inferences in the lattice of Figure 40, since that lattice now models attribute observations involving "a" and/or "b" as sets of sequences rather than merely as sequences. We would instead probably want Q to be something like the poset of Figure 43. So even though L has become simpler than in Figure 39, Q has become more complicated. On the other hand, the new (L, P, Q) triple means that one can infer $(\{1,2\},\{\mathtt{a},\mathtt{b}\})$ from the observation "b". As before, that says the observation "b" must have come from individual #1 or #2, but it also says directly that the individual could alternatively have produced attribute "a". In summary, by altering the triple (L, P, Q), one changes the possible inferences.

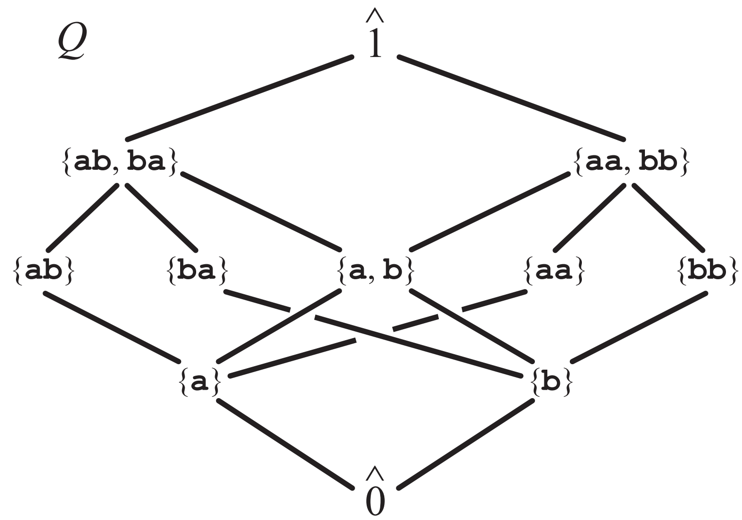


Figure 43: Poset Q modeling *sets* of attribute sequences, for inferences in the lattice of Figure 40.

Aside: The poset Q of Figure 43 is a conveniently chosen finite subposet of a particular infinite poset modeling sets of sequences. In that model, each set is required to be finite and *prefix-free*, meaning that if two distinct sequences appear in an element of $\overline{Q}$, neither may be a prefix of the other. The partial order on $\overline{Q}$ is defined by: $q_1 \leq_Q q_2$ precisely when every sequence in q_1 is a prefix of (possibly equal to) some sequence in q_2. (Notation: $\overline{Q}$ is the proper part of Q, that is, $\overline{Q} = Q \setminus \{\hat{0}, \hat{1}\}$, and $\hat{0} < q < \hat{1}$ for every $q \in \overline{Q}$.)

13 Lattices for Strategy Obfuscation

In Section 12, we saw sublattices of powerset lattices, those being prototypical examples of Boolean lattices. A related example is given by *strategy complexes* [6, 7], which may be viewed as lattices of (stochastic) partial orders formed from potentially nondeterministic or stochastic transitions in a graph. The basic elements in such a lattice are *strategies* for attaining various goals. Our work on privacy now raises the question of strategy obfuscation: How can someone reveal the *actions* of a strategy in a fashion that delays identification of the strategy?

13.1 Strategies for Nondeterministic Graphs

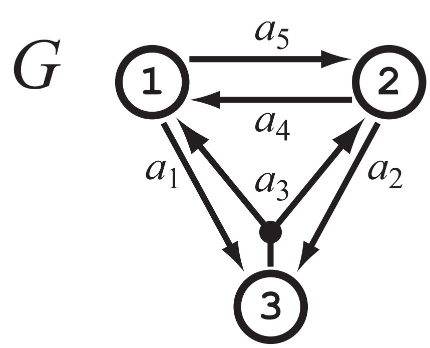


Figure 44: A graph G with three states, four deterministic actions, and one nondeterministic action (a_3).

For a very simple example, consider the graph of Figure 44. We might think of this graph as modeling some kind of dynamic system, for instance, a person driving between three shopping malls or a robot moving among clutter in a warehouse or an intruder in a server network.

There are three states in the graph, along with five actions. Each action has a *source* state and one or more *target* states, indicated in the figure by arrows. An action may be *executed* when the system is at the source state of the action, causing the system to move from the action's source state to one of its target states.

Four of the actions, $\{a_1, a_2, a_4, a_5\}$, are standard deterministic directed edges, leading for certain from one state to another. The remaining action, a_3, is nondeterministic. Nondeterminism of a_3 means that if the system is at state 3 and executes action a_3, then the precise outcome is uncertain: The system might move either to state 1 or to state 2. Nondeterminism is potentially adversarial: The precise target state attained is unpredictable and could vary nonstochastically on different executions of the action, perhaps determined by an adversary outside the graph. One may generalize this idea to include stochastic actions along with deterministic and nondeterministic actions, thus modeling adversarial combinations of Markov chains [6, 7].

In the nondeterministic setting, a *strategy* is a set of actions whose underlying directed edge set contains no directed cycles. The semantics of a strategy are: If the system is at the source state of an action in the strategy, then the system executes that action. If the strategy contains multiple actions with that same source state, then the actual action executed is again determined nondeterministically. For instance, in the example, if actions a_1 and a_5 both appear in a strategy, then the strategy is indifferent as to whether the system will transition to state `2` or to state `3` from state `1`. One or the other will occur. If a strategy does not contain any action with a given source state, then the system will stop moving if it is ever in that state.

The lattice operations for strategies are set union and set intersection, with one proviso: Suppose σ_1 and σ_2 are two strategies. Each strategy is a set of actions with no directed cycles in its underlying directed edge set. If the union of the two strategies, $\sigma_1 \cup \sigma_2$, contains a directed cycle in its underlying directed edge set, then the lattice operation becomes $\sigma_1 \vee \sigma_2 = \hat{1}$, with $\hat{1}$ the top element of the lattice. That top element represents cyclicity. The bottom element $\hat{0}$ of the lattice is equivalent to the empty strategy $\emptyset$, amounting to no motion.

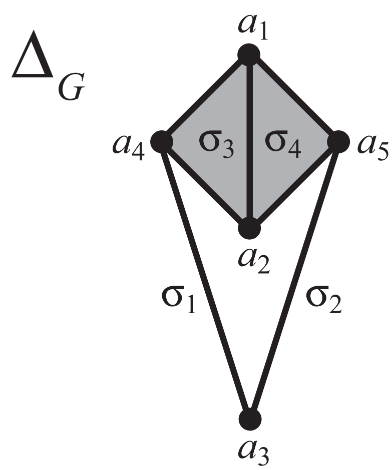


Figure 45: The strategy complex for the graph of Figure 44. Each vertex represents an action, as indicated by the labels. Each maximal simplex also has a label, for the purposes of Figure 46.

Rather than draw a lattice of strategies L, it is more convenient to draw an equivalent simplicial complex whose vertices are the (acyclic) actions of the graph. This simplicial complex is denoted by Δ_G and is called the *strategy complex* of G. The connection is that the proper part of the lattice is the face poset of the simplicial complex, that is $L \setminus \{\hat{0}, \hat{1}\} = \mathfrak{F}(\Delta_G)$. Figure 45 shows the strategy complex for the graph of Figure 44. The constituent simplices of the strategy complex are strategies, that is, all sets of actions whose underlying directed edge sets are acyclic.

Now that we have a simplicial complex, we can form a relation, whose "individuals" are all maximal strategies of the complex and whose "attributes" are the underlying actions, as shown in Figure 46. The figure also shows each maximal strategy's *goal*, that is, the state at which the strategy would stop moving. (In general, a strategy, even a maximal strategy, may have a multi-state goal set, but in this example the goals of all maximal strategies are singleton states.) Of course, a system could employ nonmaximal strategies, but for identifiability purposes it is natural to consider maximal strategies.

We make the following observations:

- There is at least one strategy for attaining each state in the graph, meaning it is possible to move from every state to every other state, despite uncertainty in the outcome of one

A	a_1	a_2	a_3	a_4	a_5	Goal
σ_1			•	•		1
σ_2			•		•	2
σ_3	•	•		•		3
σ_4	•	•			•	3

Figure 46: Relation A describes the strategy complex of Figure 45 in terms of its maximal simplices and their constituent actions. The rightmost column shows each maximal strategy's goal, i.e., that state at which motion ceases.

of the actions. Such graphs are called *fully controllable* in [6, 7], and have properties similar to those of strongly connected directed graphs.

- Each maximal strategy contains *two* informative attribute (i.e., action) release sequences, with each sequence consisting of two actions that together identify the strategy. For instance, for σ_1, one could reveal actions a_3 and a_4 in either order, identifying σ_1 only after revealing both actions. For σ_3, one could reveal actions a_1 and a_4 in either order, now identifying σ_3 only after revealing both actions.

- Some actions reveal the goal even though they do not identify the maximal strategy. In particular, actions a_1 and a_2 each individually reveal the goal to be 3. (The two actions are in fact equivalent in A, in that either one implies the other.) For instance, if one knows that a_1 is in a maximal strategy σ, then one knows that the strategy cannot also contain a_3, as adding a_3 would create a directed cycle in the underlying directed edge set. Action a_2 must therefore also be in the strategy, since the strategy is maximal. Consequently, the goal is state 3 and σ is either σ_3 or σ_4. The difference between these two maximal strategies is a choice between a_4 and a_5. That choice does not affect the final goal, but could affect intermediate motions and the time to reach the goal. A rough analogy is knowing that a car on a freeway must continue on the freeway until at least the next exit but has a choice between lanes enroute.

- Each maximal strategy contains at least *one* informative action release sequence consisting of two actions that do not reveal the goal until the second action has been released. For instance, for σ_3, one could first release a_4, leaving open the possibility of either state 1 or state 3 being the goal, then subsequently release either a_1 or a_2.

The rest of this section and Appendix H explore these observations more generally.

13.2 Connecting the Topologies of Strategy Complexes and Privacy

Notation:

- $G = (V, \mathfrak{A})$ denotes a graph with states V and actions $\mathfrak{A}$. An action may be deterministic, nondeterministic, or stochastic. (For simplicity, we assume here that $V \neq \emptyset$ and $\mathfrak{A} \neq \emptyset$.)

- Δ_G denotes the strategy complex of G; it includes the empty strategy $\emptyset$.

Lemma 30. *Let $G = (V, \mathfrak{A})$ be a graph as above and $\mathfrak{M}$ the set of maximal simplices of Δ_G.*

Define relation A on $\mathfrak{M} \times \mathfrak{A}$ by $A = \{(\sigma, a) \mid a \in \sigma \in \mathfrak{M}\}$. Then $\Phi_A = \Delta_G$. In other words, the Dowker complex over the set of actions is the same as the graph's strategy complex.

(The lemma holds more generally for simplicial complexes. The proof is nearly definitional.)

(The "A" stands for "Action" and we refer to relation A as G's *action relation.*)

One of the fundamental results from [6, 7] is that a graph is fully controllable if and only if its strategy complex is homotopic to a sphere of dimension two less than the number of states in the graph: (Recall that "$\simeq$" denotes a homotopy equivalence.)

Theorem 31. *A graph $G = (V, \mathfrak{A})$ is fully controllable if and only if $\Delta_G \simeq \mathbb{S}^{n-2}$, with $n = |V|$.*

Now recall our fundamental privacy result, Corollary 27 from page 45. That corollary, along with Theorem 31, tells us that if a graph $G = (V, \mathfrak{A})$ is fully controllable, then the poset P_A, formed from relation A of Lemma 30, must contain at least $n!$ maximal chains, each consisting of at least $n - 1$ elements, with $n = |V|$ (recall that the number of elements in a chain is one more than its length).

We actually want a stronger result, speaking to individual strategies and we can get that by looking into the details of the proof of Theorem 26. The proof is an induction that recursively considers links, giving us the following (see Appendices G and H):

Theorem 32 (Delaying Strategy Identification)**.** *Let $G = (V, \mathfrak{A})$ be a fully controllable graph, with $n = |V| > 1$. Let A be the relation constructed as in Lemma 30 and let P_A be its associated doubly-labeled poset. Then:*

For each $v \in V$, there exists a maximal strategy $\sigma_v \in \Delta_G$ for attaining singleton goal state v such that P_A contains at least $(n-1)!$ distinct maximal chains for identifying σ_v, with each chain consisting of at least $n - 1$ elements.

Clarifying Observation: Each maximal chain for identifying σ_v specifies, via the construction of Lemma 21 on page 41, at least $n - 1$ actions and an order for releasing them, such that no action is implied by those previously released. In particular, the sequence of actions does not identify σ_v until all actions have been released.

Comments: Theorem 32 does *not* assert that *every* maximal strategy in Δ_G has $(n-1)!$ many "long" identifying chains, merely that, for every possible singleton goal v, there is *some* strategy for attaining v with $(n-1)!$ many "long" identifying chains. It is not hard to construct examples for which some maximal strategy has fewer than $(n - 1)!$ identifying chains (see Section 13.3). This fact raises further questions (G is assumed fully controllable throughout):

- Given an arbitrary maximal strategy σ_v for attaining a singleton goal state v, can we find at least *one* chain in P_A that identifies σ_v but requires release of at least $n - 1$ actions before doing so? The answer in general is "no", but "yes" for certain kinds of graphs.

  One can construct counterexamples, in which the strategy σ_v is always inferable before $n - 1$ of its actions have been revealed, regardless of the order in which one reveals the actions. Appendix H.5 describes one such example, containing a mix of stochastic and nondeterministic actions. Nonetheless, even for such mixed graphs one can describe situations in which the answer is "yes". This occurs for instance when the graph contains a Hamiltonian cycle consisting of directed edges that arise from deterministic or stochastic actions (see Lemma 98 on page 133, in Appendix H.4). Leveraging that insight, one can prove that, for *pure* nondeterministic or stochastic graphs (defined on page 136 in Appendix H.6), every maximal strategy (even one with a multi-state goal) has an informative action release sequence of length at least $n-1$.

- Given a singleton goal state v, can we find at least one maximal strategy τ_v and at least one chain in P_A that eventually identifies τ_v, but does not reveal the goal v before releasing at least $n-1$ actions? The answer to this question is "yes". The proof operates by repeatedly creating quotient graphs. In forming a quotient graph, the proof regards as equivalent a certain set of states that are connected by a cycle of directed edges, with each edge coming from some deterministic or stochastic action. For instance, in the graph of Figure 44, the proof would regard states 1 and 2 as equivalent. The resulting quotient graph would then consist of two states with deterministic actions between them, since action a_3 becomes a deterministic transition in the quotient graph. Inductively, one therefore sees that an entity can hide its true goal until at least two actions in the original graph G have been revealed. (See Appendix H.3 for further details.)

A comment/caution regarding the availability of many chains: The $(n-1)!$ chains mentioned above may come from all possible permutations of the same underling set of $n-1$ actions. Alternatively, these $(n-1)!$ chains may involve creative sequencing of more than $n-1$ actions. The precise makeup of the chains depends on the underlying homology generators. However, even if the chains are merely reordering the same $n-1$ actions, there is good reason to take advantage of that capability, rather than pick one particular sequence via a deterministic algorithm. The reason is that knowledge of how an algorithm releases actions may leak information to an adversary. Such leakage may be understood as changing the effective relation. For instance, despite thinking one is working with relation A, a particular release protocol may simply be focusing on some proper subset of A or some proper subset of the poset P_A, possibly resulting in very different inference characteristics. A good release strategy may be to choose randomly from among the $(n-1)!$ possible chains. In that way, one is taking good advantage of the spherical homogeneity suggested by homology.

13.3 Example: Multi-State Goals and Multi-Strategy Singleton Goals

Figure 47 shows a fully controllable nondeterministic graph on four states. The graph contains four deterministic actions and three nondeterministic actions.

Three of the deterministic actions form a directed cycle: $1 \xrightarrow{e_1} 2 \xrightarrow{e_2} 3 \xrightarrow{e_3} 1$. The remaining deterministic action, a_2, moves from state 2 to state 4.

Actions a_1, a_3, and b_4 are nondeterministic. Action a_1 moves nondeterministically from state 1 to either state 3 or state 4, while action a_3 moves nondeterministically from state 3 to either state 2 or state 4. Finally, action b_4 moves nondeterministically back from state 4 to any of the other three states.

Figure 48 shows relation A for the graph of Figure 47, with A as defined in Lemma 30 on page 65. As indicated in the figure, some maximal strategies have two-state goals. In addition, two of the maximal strategies converge to the same singleton goal, namely state 4.

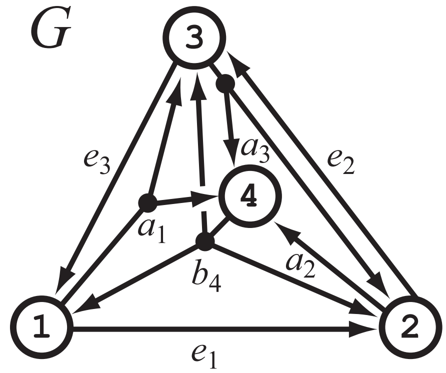


Figure 47: A graph with four states $\{1, 2, 3, 4\}$, four deterministic actions $\{e_1, e_2, e_3, a_2\}$, and three nondeterministic actions $\{a_1, a_3, b_4\}$.

A	e_1	e_2	e_3	a_1	a_2	a_3	b_4	Goal
σ_1		•	•				•	1
σ_2	•		•				•	2
σ_3	•	•					•	3
σ_4	•		•		•	•		4
σ_5	•			•	•	•		4
σ_{14}		•	•		•			$\{1, 4\}$
σ_{34}	•	•		•	•			$\{3, 4\}$

Figure 48: Relation A describes the strategy complex for the graph of Figure 47 in terms of its maximal strategies and their constituent actions. The rightmost column further shows each maximal strategy's goal. Observe that some strategies converge to multi-state goals.

Strategies σ_1, σ_2, σ_3, and σ_4 each have (at least) six different informative action release sequences of length (at least) 3, as guaranteed to exist in Δ_G by Theorem 32. Strategy σ_5 also has informative action release sequences of length at least 3 (in fact, length 4), but it only has four such sequences, consistent with the comments on page 65. We can see this by constructing the Galois lattice P_Q^+, with Q modeling $\mathrm{Lk}(\Psi_A, \sigma_5)$. Figures 49 and 50 depict Q and P_Q^+, respectively. There are indeed four downward paths of length 4 from $\hat{1}_Q$ to $\hat{0}_Q$.

In order to delay identification of σ_5 as long as possible, the lattice P_Q^+ further tells us that one should reveal action a_2 either right away or right after first revealing action e_1. However, as soon as one has revealed action a_2, an observer knows that the goal is either state 4 or a two-state set containing state 4. If the observer has adversarial control over the outcome of nondeterministic actions in G, then as soon as one has revealed two of σ_5's actions, the observer-adversary can lie in wait for the system at state 4.

Of course, as indicated on page 66, one can delay goal recognition for 3 steps if one is free to choose any strategy for that goal. For goal state 4, one should choose strategy σ_4 rather than σ_5, and reveal actions e_1, e_3, a_2, either in that order or in the order e_3, e_1, a_2.

Q	e_1	a_1	a_2	a_3	Goal
σ_2	•				2
σ_3	•				3
σ_4	•		•	•	4
σ_{14}			•		$\{1, 4\}$
σ_{34}	•	•	•		$\{3, 4\}$

Figure 49: Relation Q models $\mathrm{Lk}(\Psi_A, \sigma_5)$, with A as in Figure 48.

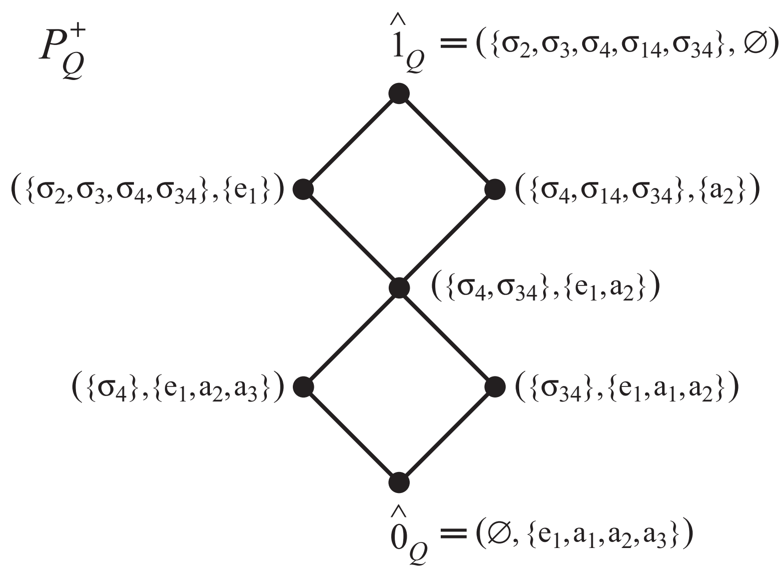


Figure 50: The Galois lattice P_Q^+ has length 4, with Q as in Figure 49.

13.4 Randomization

Suppose $G = (V, \mathfrak{A})$ is a fully controllable graph with $|V| = |\mathfrak{A}| = n > 1$. These conditions imply that there is exactly one action at each state and that the maximal simplices of Δ_G consist of all subsets of $\mathfrak{A}$ of size $n - 1$. Consequently, Δ_G is a boundary complex, specifically $\Delta_G = \partial(\mathfrak{A})$. As we have seen, such complexes preserve attribute privacy, meaning it is impossible to infer any additional actions from actions already revealed. The existence of $(n-1)!$ different informative action release sequences of length $n - 1$ for any given maximal strategy here simply means that one may reveal the actions of that strategy in any order.

The technical complications discussed previously arise when there are multiple actions at some or all states of V. One may circumvent such complications by creating a single "effective action" at each state, for instance by choosing stochastically among the given actions available at a state when the system is in that state. The precise probabilities are not significant from the perspective of combinatorial strategy obfuscation, so long as the probability of choosing any given original action is greater than 0, and all such probabilities sum to 1. (Of course, the actual probabilities will affect the expected time to attain the goal.)

A related issue concerns execution order versus release order. The privacy results in this report assume that a system can control the order in which it reveals actions. If instead actions are revealed as they are executed, then an observer may be able to infer the underlying strategy more quickly than desired. In order to obfuscate the strategy, the system may need to be willing to ignore early arrival at the goal and instead continue moving. The precise criteria determining whether the system stops or continues could be stochastic, or could reflect a protocol determined by the desired action release sequence.

14 Relations as a Category

We have discussed disinformation, obfuscation, and other manipulation of relations. The goal of such transformations has been to preserve privacy by removing or hiding free faces. We have not yet discussed such transformations formally. For instance, the coordinate transformations of Section 9 raise the question:

How should one think about maps between relations?

14.1 Relationship-Preserving Morphisms

Traditionally, relations are themselves morphisms between sets (with functions a special case). In thinking about privacy, it is useful to define a category in which relations are the objects. We have some choices in defining morphisms for this category. Bearing in mind our Dowker constructions (see again Definition 1 on page 13), we adopt the following standard definition:

Notation: (1) We frequently will be working with two relations: R is a relation on $X^R \times Y^R$ and Q is a relation on $X^Q \times Y^Q$ (the superscripts are just indices to indicate the underlying relation). In order to distinguish rows and columns between the two relations, we will also use notation of the form X^R_y, Y^R_x, X^Q_y, and Y^Q_x. (2) By a *set map* we mean a function between two sets.

Definition 33 (Morphism)**.** *Let R be a relation on $X^R \times Y^R$ and let Q be a relation on $X^Q \times Y^Q$. A* morphism of relations *$f : R \to Q$ is a pair of set maps:*

$$\begin{aligned} f_X &: X^R \to X^Q \\ f_Y &: Y^R \to Y^Q \end{aligned}$$

such that $\big(f_X(x),\, f_Y(y)\big) \in Q$ whenever $(x, y) \in R$.

In other words, a morphism of relations maps individuals to individuals and attributes to attributes in a way that preserves relationships.

The following lemma follows from the definitions (a proof appears in Appendix I.1):

Lemma 34 (Induced Simplicial Maps)**.** *A morphism $f : R \to Q$ between nonvoid relations induces simplicial maps between the Dowker complexes:*

$$\begin{aligned} f_X &: \Psi_R \to \Psi_Q \\ f_Y &: \Phi_R \to \Phi_Q \end{aligned}$$

Notational comment: The symbols f_X and f_Y are overloaded intentionally. The simplicial map f_X is precisely the set map f_X applied to the vertices of any simplex: If $\sigma = \{x_0, \ldots, x_k\} \in \Psi_R$, then $f_X(\sigma) = \{f_X(x_0), \ldots, f_X(x_k)\} \in \Psi_Q$. Similarly for f_Y.

Connectivity Implication: Intuitively, one cannot partition the individuals of a connected relation into two or more pairwise disjoint classes without misclassifying some individuals or ignoring some relationships. A graph connectivity argument provides a possible proof. Lemma 34 provides another, with additional insight. Let us look at some examples:

Two Bits onto One: Consider again the relations S and Q of Figures 15 and 16, respectively, on page 27. Relation S models a one-bit relation — an attribute and its negation. Relation Q models a two-bit relation — two attributes and their negations. The Dowker complexes for S have $\mathbb{S}^0$ homotopy type, while those for Q have $\mathbb{S}^1$ homotopy type. We can think of S as a classification, splitting individuals into those that have some attribute `a` and those that do not.

By Lemma 34, a morphism $f : Q \to S$ induces simplicial (hence continuous) maps between the corresponding Dowker complexes of S and Q. Since $\mathbb{S}^1$ is connected but $\mathbb{S}^0$ is not, there is no surjective continuous function from $\mathbb{S}^1$ to $\mathbb{S}^0$. Consequently, no morphism $f : Q \to S$ can truly be a classification: f_Y can map all four attributes $\{\mathtt{a}, \neg\mathtt{a}, \mathtt{b}, \neg\mathtt{b}\}$ of Q to the single attribute `a` or all four attributes to ¬`a`, but f_Y cannot map to both `a` and ¬`a`.

Q'	`a`	¬`a`
1	•	
2	•	
3		•
4		•

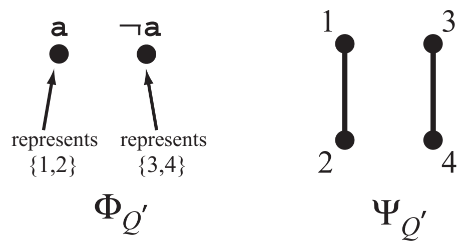


Figure 51: Relation Q' obtained from relation Q of Fig. 16 by discarding attributes `b` and ¬`b`.

This impossibility may at first seem paradoxical. After all, one can simply cut relation Q down the middle and throw away the columns involving attributes `b` and ¬`b`, as shown in Figure 51. After that, a surjective morphism $f' : Q' \to S$ is immediate. Indeed, that is possible. However, in so doing, one has discarded some relationships, perhaps purposefully, perhaps accidentally. In particular, the relationship between individuals #1 and #3 of Q via attribute `b` is lost, as is the relationship between individuals #2 and #4 via attribute ¬`b`. This reasoning simply underscores the fact that morphisms of relations preserve relationships. Lack of continuity in a function therefore is a sign that one is discarding some relationships. Whether such discard is desirable depends on one's goals in a particular application.

Three Bits onto Two: Recall as well Figure 17 on page 28, which depicts a three-bit relation R — three attributes and their negations, capable of distinguishing between eight individuals. The homotopy type of the Dowker complexes is $\mathbb{S}^2$. With Q as above, the following question arises naturally when trying to reduce complexity of data yet preserve information:

Does there exist a surjective morphism $f : R \to Q$?

Unlike the previous example, there do exist continuous maps from $\mathbb{S}^2$ onto $\mathbb{S}^1$, so perhaps one can find a surjective morphism $f : R \to Q$. In fact, one can not. Intuitively, the issue is that the two-dimensional relationships of R try to fill the one-dimensional hole of relation Q. Here is a simplex-based argument:

- Suppose surjective $f : R \to Q$ exists. As will be discussed later (see page 74), this means the component functions $f_X : X^R \to X^Q$ and $f_Y : Y^R \to Y^Q$ are surjective as set maps.

- One may therefore assume without loss of generality that $f_Y(\texttt{a}) = \texttt{a}$ and $f_Y(\texttt{b}) = \texttt{b}$.
- The triangles $\{\texttt{a}, \texttt{b}, \texttt{c}\}$ and $\{\texttt{a}, \texttt{b}, \neg\texttt{c}\}$ are both simplices in Φ_R. The maximal simplices of Φ_Q are edges.
- By Lemma 34, this means that $f_Y(\texttt{c})$ and $f_Y(\neg\texttt{c})$ are both elements of $\{\texttt{a}, \texttt{b}\}$ in Φ_Q.
- Again by surjectivity, we therefore see that $\{f_Y(\neg\texttt{a}), f_Y(\neg\texttt{b})\} = \{\neg\texttt{a}, \neg\texttt{b}\}$.
- Another triangle-versus-edge argument then says that $f_Y(\texttt{c})$ and $f_Y(\neg\texttt{c})$ are both elements of $\{\neg\texttt{a}, \neg\texttt{b}\}$, giving us a contradiction.

Of course, as in constructing Q' of Figure 51, if we are willing to tolerate discontinuities, we could discard one attribute and its negation to obtain Q from R. As before, discontinuity means losing awareness of some relationship(s). For instance, if we omit attribute `c`, we would become unaware in Q of the relationship that exists in R among the set of individuals $\{1, 3, 5, 7\}$.

14.2 Privacy-Establishing Morphisms

M	a	b	c	d	e
1	•	•			•
2	•	•	•		
3		•	•	•	
4			•	•	•
5	•			•	•

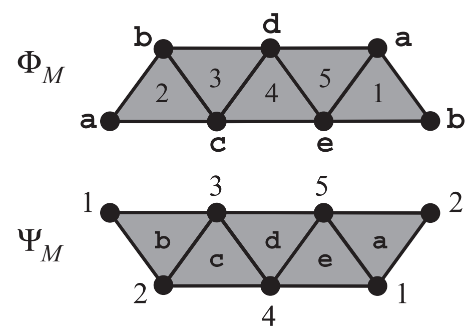


Figure 52: Relation M is isomorphic to relation G of Figure 23 on page 34, now without the author-book semantics. The Dowker complexes are dual triangulations of the Möbius strip, with $\mathbb{S}^1$ homotopy type.

Relations involve two spaces. Looking at just Φ_R or just Ψ_R may hide some interesting properties. For instance, consider the Möbius strip relation M of Figure 52. We encountered this relation previously, in Section 10.

We might wish to remove some of the inferences discussed in Section 10 by reshaping the underlying relation without discarding any relationships. Doing so leads to the following question:

> Does there exist a surjective morphism $f : M \to T$, with T a relation that preserves both attribute and association privacy ?

If such a morphism f exists, then, as we mentioned in Sections 5 and 8, relation T must have the topology of either a linear cycle or a spherical boundary complex. It turns out that the answer to the question above is "yes", with T being a relation whose Dowker complexes are boundaries of tetrahedra (see Figure 30 on page 42).

This construction is not immediately obvious from the complexes Φ_M and Ψ_M. Although those simplicial complexes are 2-dimensional, suggesting that their triangles can be wrapped around a tetrahedron, doing so actually collapses two of the five triangles to edges. Indeed, the component functions for one such surjective morphism $f : M \to T$ are:

$$\begin{array}{rcl} f_X \ : \ X^M & \to & X^T \\ 1 & \mapsto & 4 \\ 2 & \mapsto & 1 \\ 3 & \mapsto & 2 \\ 4 & \mapsto & 3 \\ 5 & \mapsto & 4 \end{array} \qquad\qquad \begin{array}{rcl} f_Y \ : \ Y^M & \to & Y^T \\ \mathtt{a} & \mapsto & \mathtt{a} \\ \mathtt{b} & \mapsto & \mathtt{b} \\ \mathtt{c} & \mapsto & \mathtt{c} \\ \mathtt{d} & \mapsto & \mathtt{d} \\ \mathtt{e} & \mapsto & \mathtt{a} \end{array}$$

The induced simplicial maps act on the five maximal simplices of Ψ_M and Φ_M as follows:

$$\begin{array}{rcl} f_X \ : \ \Psi_M & \to & \Psi_T \\ \{1,2,3\} & \mapsto & \{1,2,4\} \\ \{2,3,4\} & \mapsto & \{1,2,3\} \\ \{3,4,5\} & \mapsto & \{2,3,4\} \\ \{1,4,5\} & \mapsto & \{3,4\} \\ \{1,2,5\} & \mapsto & \{1,4\} \end{array} \qquad\qquad \begin{array}{rcl} f_Y \ : \ \Phi_M & \to & \Phi_T \\ \{\mathtt{a},\mathtt{b},\mathtt{c}\} & \mapsto & \{\mathtt{a},\mathtt{b},\mathtt{c}\} \\ \{\mathtt{b},\mathtt{c},\mathtt{d}\} & \mapsto & \{\mathtt{b},\mathtt{c},\mathtt{d}\} \\ \{\mathtt{c},\mathtt{d},\mathtt{e}\} & \mapsto & \{\mathtt{a},\mathtt{c},\mathtt{d}\} \\ \{\mathtt{a},\mathtt{d},\mathtt{e}\} & \mapsto & \{\mathtt{a},\mathtt{d}\} \\ \{\mathtt{a},\mathtt{b},\mathtt{e}\} & \mapsto & \{\mathtt{a},\mathtt{b}\} \end{array}$$

Even though f_X and f_Y are surjective as set maps on the vertices of the Dowker complexes, they are *not* surjective as simplicial maps on the complexes themselves. Each only covers three of the four triangles comprising the tetrahedron in its codomain. At first glance it may therefore seem that the morphism $f : M \to T$ resulting from f_X and f_Y does not achieve the desired privacy preservation. A closer look, however, reveals that f is actually surjective as a map of relations: it maps the elements of M *onto* the elements of T. Therefore, it does represent a transformation that achieves privacy preservation.

In order to understand this paradox, imagine again that M represents an authorship database. Think of the maps f_X and f_Y as quotient maps, in this case equating authors 1 and 5 and books `a` and `e`. The equivalencing of authors might constitute a recognition of pseudonyms. The equivalencing of books might represent a generalization from titles to genres. Such changes of resolution, carefully chosen, perhaps based on external structure, can preserve relationships while reducing recognition and inference granularity.

14.3 Summary of Morphism Properties

Definition 33 defines a morphism of relations $f : R \to Q$ in terms of underlying set functions $f_X : X^R \to X^Q$ and $f_Y : Y^R \to Y^Q$. These set functions further induce simplicial maps $f_X : \Psi_R \to \Psi_Q$ and $f_Y : \Phi_R \to \Phi_Q$. The previous subsections spoke of surjectivity in varying contexts. Similarly, one could speak of maps as being injective in varying contexts. Finally, one also speaks of morphisms as being epimorphisms and monomorphisms. This subsection summarizes how these properties relate for the various maps. See Appendix I.1 for proofs.

First, some definitional context and reminders:

- A morphism of relations $f : R \to Q$ is also a set map between the set of pairs comprising R and the set of pairs comprising Q. Specifically, $f(x, y) = (f_X(x), f_Y(y))$ for all $(x, y) \in R$.

  One may speak of f as being *surjective* and/or *injective*, meaning as a set map.

- We say that two morphisms of relations $g, h : R \to Q$ are *equal*, written $g = h$, when they are equal as set maps of ordered pairs, meaning $g(x, y) = h(x, y)$ for all $(x, y) \in R$.

  (Note: If R contains blank rows and/or columns, then $g = h$ is possible even though $g_X \neq h_X$ and/or $g_Y \neq h_Y$, as set maps. This will not cause us problems; one could pass to equivalence classes in Definition 33.)

- The functions $f_X : X^R \to X^Q$ and $f_Y : Y^R \to Y^Q$ are set maps. One may speak of them as being surjective and/or injective.

- One may also ask whether the induced simplicial maps $f_X : \Psi_R \to \Psi_Q$ and $f_Y : \Phi_R \to \Phi_Q$ are surjective and/or injective as maps between simplicial complexes viewed as sets.

- Suppose $f : R \to Q$ is a morphism of relations. Recall from category theory that f is an *epimorphism* if, for any pair of morphisms $g, h : Q \to S$, $g \circ f = h \circ f$ implies $g = h$.

  Recall further that a morphism $f : R \to Q$ is a *monomorphism* if, for any pair of morphisms $g, h : S \to R$, $f \circ g = f \circ h$ implies $g = h$.

Lemma 35 (Morphism Properties). *Assume the notation from above and that all relevant relations are nonvoid. Let $f : R \to Q$ be a morphism of relations (as per Definition 33). Then:*

(i) f_X and f_Y *are injective set maps* $\Longrightarrow$ f *is injective* $\Longleftrightarrow$ f *is a monomorphism.*

(ii) f *surjective* $\Longrightarrow$ f *epimorphism* $\Longleftrightarrow$ f_X and f_Y *are surjective set maps.*

(Additional conditions for that last $\Longleftrightarrow$: The $\Longrightarrow$ direction assumes that Q has no blank rows or columns, while the $\Longleftarrow$ direction assumes that R has no blank rows or columns.)

The two uni-directional implications $\Longrightarrow$ above are strict.

(iii) If $f_X : \Psi_R \to \Psi_Q$ is surjective and Q has no blank rows, then $f_X : X^R \to X^Q$ is surjective.

Similarly for f_Y, now assuming that Q has no blank columns.

The converses need not hold. Indeed, f itself can be surjective but the maps of simplicial complexes need not be (as we saw with the maps of page 73).

(iv) If $f_X : X^R \to X^Q$ is injective, then $f_X : \Psi_R \to \Psi_Q$ is injective. The converse holds if R has no blank rows.

Similarly for f_Y, now assuming that R has no blank columns for the converse.

14.4 G-Morphisms

Since a relation R defines a poset P_R, rather than merely create morphisms from set maps between individuals and attributes as in Definition 33, we may broaden the definition by considering maps between posets:

Definition 36 (G-Morphism)**.** *Let R and Q be nonvoid relations.*
A <u>G-morphism</u> $f : R \to Q$ *is any poset map* $f : P_R \to P_Q$.

Comments: The "G" stands for "Galois". We might have insisted that a G-morphism $R \to Q$ be a lattice morphism $P_R^+ \to P_Q^+$ rather than merely a poset map $P_R \to P_Q$, but that might be too restrictive. Instead, as subsequent lemmas will describe, we view a G-morphism as providing homotopy flexibility. In particular, a morphism between relations as per Definition 33 induces two homotopic G-morphisms. The lattice structure of the codomain is relevant, in that it allows one to fill in elements not directly in the image of any one G-morphism, as will become apparent in Theorem 41.

$$\begin{array}{ccc} \mathfrak{F}(\Psi_R) & \xrightarrow{\quad f_X \quad} & \mathfrak{F}(\Psi_Q) \\ \phi_R \Big\downarrow \Big\uparrow \psi_R & & \phi_Q \Big\downarrow \Big\uparrow \psi_Q \\ \mathfrak{F}(\Phi_R) & \xrightarrow{\quad f_Y \quad} & \mathfrak{F}(\Phi_Q) \end{array}$$

Figure 53: Diagram showing the poset maps f_X and f_Y induced by a morphism $f : R \to Q$, along with the homotopy equivalences between each relation's face posets. (The diagram need not be commutative, but is almost so; see Lemma 37.)

Recall that a morphism $f : R \to Q$ as per Definition 33 is built from two set maps f_X and f_Y and that these set maps induce simplicial maps between the Dowker complexes, as per Lemma 34. We may therefore further regard f_X and f_Y as order-preserving poset maps between the face posets of the Dowker complexes: $f_X : \mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Psi_Q)$ and $f_Y : \mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Phi_Q)$. Consequently, we have a diagram of maps as in Figure 53. The diagram need not be commutative, but the following containments hold:

Lemma 37 (Witness Containment)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. Then:*

(a) $(f_Y \circ \phi_R)(\sigma) \subseteq (\phi_Q \circ f_X)(\sigma)$, *for every* $\sigma \in \Psi_R$,

(b) $(f_X \circ \psi_R)(\gamma) \subseteq (\psi_Q \circ f_Y)(\gamma)$, *for every* $\gamma \in \Phi_R$.

(See Appendix I.2 for a proof of the previous lemma and its upcoming corollaries.)

As a corollary, we see that the diagram of Figure 53 describes two pairs of homotopic maps:

Corollary 38 (Homotopic Face Maps)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. Then:*

(a) f_X and $\psi_Q \circ f_Y \circ \phi_R$ are homotopic poset maps $\mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Psi_Q)$,

(b) f_Y and $\phi_Q \circ f_X \circ \psi_R$ are homotopic poset maps $\mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Phi_Q)$.

The images of the compositions that appear in Corollary 38 may be regarded as lying in P_Q. We may further restrict the domain of these maps to be P_R, giving us the following G-morphisms:

Definition 39 (Induced G-Morphisms)**.** *A morphism of nonvoid relations $f : R \to Q$ induces two G-morphisms $R \to Q$, defined by the following poset maps $P_R \to P_Q$:*

$$f_X^g \;=\; (\psi_Q \circ f_Y \circ \phi_R)|_{P_R} \qquad\qquad f_Y^g \;=\; (\phi_Q \circ f_X \circ \psi_R)|_{P_R}.$$

(The "g" superscript stands for "Galois" while the vertical bar "|" means "restricted to". See also Appendix I.2.)

Corollary 40 (Homotopic G-Morphisms)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. The induced G-morphisms given by the poset maps $f_X^g, f_Y^g : P_R \to P_Q$ are homotopic.*

The proof of Corollary 40 on page 142 says that we may view the underlying maps f_X and f_Y of a morphism f as mapping any inference-closed set (viewed either as a set of individuals or as a set of attributes) from the domain of f to an interval (in the poset sense) of inference-closed sets in the codomain of f.

R	`a`
1	•

Q	`a`	`b`
1	•	•
2	•	

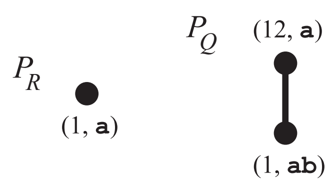


Figure 54: Relation R is a subrelation of Q. How should one embed P_R into P_Q? There are two possible embeddings, related by a homotopy.

For a simple example, see Figure 54. One may regard relation R as a subrelation of Q, then define $f : R \to Q$ to be inclusion. For instance, maybe R and Q represent individuals #1 and #2 at two parties `a` and `b`, with R representing known parties and party-attendees at some time and Q representing an update of that information at a later time. Observe that:

$$f_X^g((1,\texttt{a})) \;=\; (\psi_Q \circ f_Y \circ \phi_R)(\{1\}) \;=\; (\psi_Q \circ f_Y)(\{\texttt{a}\}) \;=\; \psi_Q(\{\texttt{a}\}) \;=\; \{1,2\} \text{ “=” } (12,\texttt{a}),$$

$$f_Y^g((1,\texttt{a})) \;=\; (\phi_Q \circ f_X \circ \psi_R)(\{\texttt{a}\}) \;=\; (\phi_Q \circ f_X)(\{1\}) \;=\; \phi_Q(\{1\}) \;=\; \{\texttt{a},\texttt{b}\} \text{ “=” } (1,\texttt{ab}).$$

The last equality in each row indicates how to view the image element on the left of the "=" as an element of the poset P_Q.

Both f_X^g and f_Y^g tell us how to update inference-closed sets from P_R into inference-closed sets within P_Q:

- The map f_X^g updates associations while holding observed attributes fixed. In this example, based on initial information (relation R), we know that person #1 attended party `a`. Once we update that information (relation Q) we can conclude that person #2 also attended a party at which person #1 was present.

- Similarly, the map f_Y^g updates attributes while holding observed individuals fixed. In this example, updated information allows us to conclude that person #1 attended not only party `a` but also party `b`.

In general, for any fixed element of P_R, the two maps may give different results, but those results are comparable in P_Q. Here f was inclusion, so we could speak of holding attributes or individuals "fixed". More generally, "fixed" is replaced by whatever f does.

14.5 Surjectivity Revisited

A paradox: We saw on page 73 a surjective morphism f, from the Möbius strip relation of Figure 52 to the tetrahedral relation of Figure 30, whose induced simplicial maps $f_X : \Psi_M \to \Psi_T$ and $f_Y : \Phi_M \to \Phi_T$ were not surjective. This raises some questions:

1. Are the induced poset maps $f_X^g, f_Y^g : P_M \to P_T$ surjective?

2. If not, how can one speak of a surjective morphism?

(Note that P_M^+ is isomorphic to P_G^+ as shown in Figure 25 on page 36. A rendering would be identical, except for lowercase letters in place of uppercase ones. The lattice P_T^+ appears in Figure 31 on page 43.)

The answer to Question 1 is that the two poset maps are *not* surjective. Observe in Table 9 on page 78, for instance, that the image of f_X^g does not include $(4, \texttt{abd})$. Similarly, the image of f_Y^g does not include $(134, \texttt{a})$.

These missing elements *are* in the image of both maps *together*, viewed as a pair of homotopic maps, as per Corollary 40. Unfortunately, that explanation is not a full answer to Question 2. For instance, neither map's image includes the element $(13, \texttt{ac})$ of P_T, nor does that element appear in any interval $[f_Y^g(p), f_X^g(p)]$ as p varies throughout P_M.

To answer question 2, the lattice structure of P_T is useful. In the example, the image of f_X^g includes all elements of P_T that correspond to maximal simplices of Ψ_T. Similarly, the image of f_Y^g includes all elements of P_T that correspond to maximal simplices of Φ_T. Intuitively, we therefore expect that the lattice operations (which correspond to intersection in either Ψ_T or Φ_T) will generate all the elements of P_T. In that sense, the surjectivity of f appears as surjectivity of each of f_X^g and f_Y^g, once one *completes* their images under lattice operations.

p	$f^g_X(p)$	$f^g_Y(p)$
(12 , ab)	(14 , ab)	(14 , ab)
(2 , abc)	(1 , abc)	(1 , abc)
(123 , b)	(124 , b)	(124 , b)
(23 , bc)	(12 , bc)	(12 , bc)
(3 , bcd)	(2 , bcd)	(2 , bcd)
(234 , c)	(123 , c)	(123 , c)
(34 , cd)	(23 , cd)	(23 , cd)
(4 , cde)	(3 , acd)	(3 , acd)
(345 , d)	(234 , d)	(234 , d)
(45 , de)	(34 , ad)	(34 , ad)
(5 , ade)	(34 , ad)	(4 , abd)
(145 , e)	(134 , a)	(34 , ad)
(15 , ae)	(134 , a)	(4 , abd)
(1 , abe)	(14 , ab)	(4 , abd)
(125 , a)	(134 , a)	(14 , ab)

Table 9: Each p is of the form $(\sigma, \gamma) \in P_M$. The elements $f^g_X(p)$ and $f^g_Y(p)$ lie in P_T. See also Figures 25 and 31, on pages 36 and 43, respectively. (As in those figures, the table elides commas and braces from set notation. For Figure 25, recall that M and G are isomorphic relations.)

The following theorem summarizes the intuition of the previous pages:

Theorem 41 (Lattice Surjectivity)**.** *Let R and Q be nonvoid relations with no blank rows or columns. Suppose $f : R \to Q$ is a surjective morphism (in the sense of Definition 33). For any $q \in P_Q$:*

$$q = \bigwedge_j \bigvee_i q_{ji}, \quad \text{with each } q_{ji} \text{ in the image of } f^g_X : P_R \to P_Q,$$

$$q = \bigvee_k \bigwedge_\ell q'_{k\ell}, \quad \text{with each } q'_{k\ell} \text{ in the image of } f^g_Y : P_R \to P_Q.$$

(Here, $\bigvee$ and $\bigwedge$ are the lattice operations of P^+_Q.)

See Appendix I.3 for a proof.

15 Future Thoughts

Throughout this report, one senses the inevitability of privacy loss, that the topology of relations necessarily converts attribute information into revealing gradient flow. Gradient flow appears both in the collapse of free faces [9] and in the lattice structure of information acquisition: Collapse of free faces infers unobserved attributes from observed attributes. The meet operation of a relation's Galois lattice propels observed attributes into downward motion, toward minima of identification. Still, alternatives exist.

15.1 Relaxing Assumptions

Gradient flow is a natural consequence of the assumptions stated in Section 3.1. Let us discuss briefly how to relax those assumptions, while leaving detailed explorations for the future.

1. We could drop the assumption of relational completeness. We might then observe a set of attributes γ inconsistent with the given relation R, meaning $\gamma \notin \Phi_R$. One possibility is that some individual in R has attributes γ, but the relation does not capture this fact. There is another possibility, that γ represents the attributes of some individual external to R. For instance, recall that in Lemma 12 a set of attributes inconsistent with a link's relation identifies the linking set of individuals. Deciding between these two scenarios (new attributes for given individuals versus wholly new individuals) requires additional information, not so unlike the decisions faced in mapping unknown environments.

2. We could drop the assumption of observational monotonicity. We do wish to retain the ability to observe attributes asynchronously. However, we might be able to place algebraic structure on some attributes, so that certain newly observed attributes can cancel previously observed ones. Spending a dollar versus earning a dollar for instance. Such an algebraic structure would then permit upward motion in a relation's Galois lattice. (This is not always possible, e.g., if a relation encodes history by time-indexing.)

3. We could drop the assumption of observational accuracy. Attributes frequently are measured by noisy sensors, whether based on physical instruments or errorful databases. Existing privacy work has frequently assumed noise intrinsically (e.g., identification in [17] was successful despite database errors). This report ignored noise in order to focus on the combinatorial structure of privacy. Presently, we will sketch a possible noise model, in which a sensor reports attributes stochastically. A sensor is a physical device and an interpretation algorithm, producing observed attributes $\gamma(t)$ as functions of time. Thus, as t varies, $\gamma(t)$ may move either up or down in a relation's Galois lattice, not just down.

   A caution: Moving from a purely combinatorial system to a stochastic system need not turn gradient flow into harmonic flow. Reasonable but noisy sensors create a (stochastic) gradient flow, by the Central Limit Theorem. Rather, a noisy sensor model in the observation of attributes facilitates the connection to other privacy work. For instance, one view of Differential Privacy [5] is that it injects noise into a sensor, with the noise magnitude chosen in part as a function of the time interval allotted for observations, thus preventing gradient flow from reaching a minimum. Moreover, adversarial control over $\gamma(t)$, perhaps by sensor disinformation, may be able to create more general flows.

15.2 Sensing Attributes Stochastically

We briefly explore a model for stochastic sensors within combinatorial relations, via a simple example. Consider relation R of Figure 55. This relation produces Dowker complexes with $\mathbb{S}^1$ homotopy type, as indicated in the figure. The relation's Galois lattice P_R^+ appears in Figure 56.

R	a	b	c
1	•	•	
2		•	•
3	•		•

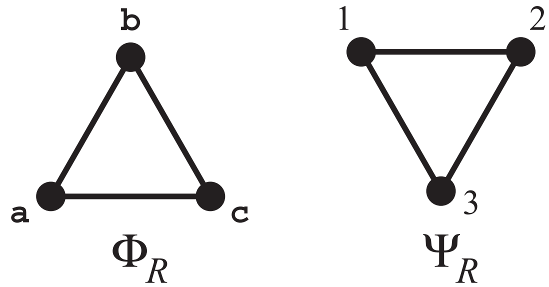


Figure 55: A relation whose Dowker complexes are dual triangulations of the circle. See Figure 56 for the associated Galois lattice.

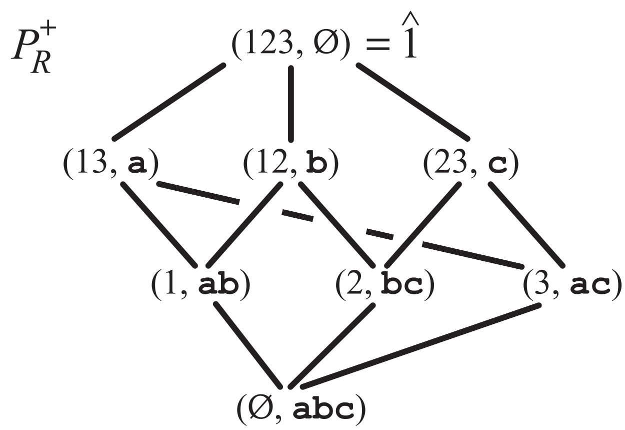


Figure 56: The lattice P_R^+ for relation R of Figure 55. (We have elided commas and braces in sets.) — For later reference: The poset $P_R \cup \{\hat{1}\}$ consists of all elements in P_R^+ except for the bottom element, $(\emptyset, \mathtt{abc})$.

Relation R's space of attributes is $Y = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Suppose that a sensor reports these attributes by observing the world and performing some computation. The report may be inaccurate. Such inaccuracy could be either adversarial or stochastic. We focus here on the stochastic case, and on one particular model: The sensor computes three probabilities, $p_{\mathtt{a}}$, $p_{\mathtt{b}}$, and $p_{\mathtt{c}}$, with p_y being the probability that the sensor's observation came from actual attribute $y \in Y$. These three probabilities constitute a point $\mathbf{p} = (p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}})$ in the full simplex $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, with the point's barycentric coordinates being the three probabilities. Subsequently, the sensor reports an attribute by interpreting $\mathbf{p}$, perhaps by maximum probability. In order to reduce false positives, the sensor sets a confidence threshold below which it interprets $\mathbf{p}$ as too ambiguous. This thresholding carves the simplex $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ into four regions, one for each attribute, plus a *zone of indecision.* A sketch of this process appears in Figure 57.

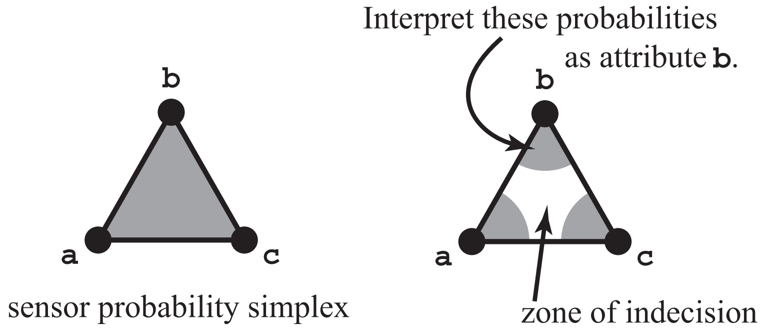


Figure 57: A sensor computes probabilities over a set of attributes $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Left panel: A simplex whose vertices are those attributes models the possible distributions, with barycentric coordinates being probabilities. Right panel: An interpretation of probabilities as attributes, along with a zone of indecision, partitions the simplex into four regions.

Individuals in relation R of Figure 55 have two attributes. One could model two observations as a probability distribution over $Y \times Y$. However, since the first observation may lead to an indecision, it is convenient to model the second observation as conditional on the first having produced a particular attribute. The second observation may: (i) repeat the just seen attribute, (ii) report an as yet unseen attribute, or (iii) fail to report anything. The net effect for (i) and (iii) is observation of one attribute. Consequently, the space of two observations may be interpreted as the first barycentric subdivision of the boundary of the attribute simplex, i.e., as $\mathrm{sd}(\partial(Y))$, with the empty simplex modeling a zone of indecision. This process appears in Figure 58. One obtains a map from a stochastic sensor's observations to the poset $P_R \cup \{\hat{1}\}$, with $\hat{1}$ representing an inability to decipher any attribute. These calculations suggest that existing stochastic results fit naturally into this report's combinatorial framework.

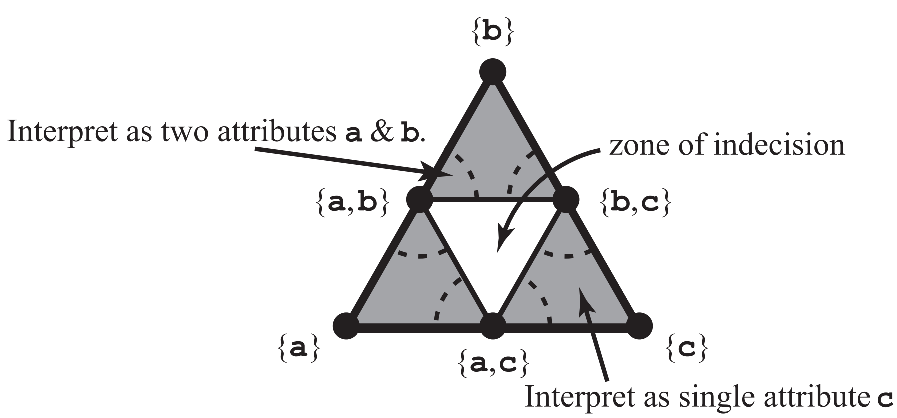


Figure 58: Two stochastic sensor observations may be modeled as a first observation followed by a conditional second observation. The resulting decision space has a representation isomorphic to the first barycentric subdivision of the boundary of the original probability simplex, with the empty simplex representing a zone of indecision. (The two observations may be understood as two points, **p** and **q**, with **p** in the encompassing triangle and **q** in a shaded subtriangle.)

Acknowledgments

We are grateful to Professor Robert Ghrist and his students and post-docs, particularly Paweł Dłotko, Vidit Nanda, and Greg Henselman, for their topological advice and suggestions, as well as their discussions related to this project.

We are also grateful for the publicly available `Perseus` software previously written at the University of Pennsylvania, which we used for our homology computations: `http://www.sas.upenn.edu/~vnanda/perseus/`.

References

[1] A. Björner. Topological methods. In R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, volume II, pages 1819–1872. Elsevier, Amsterdam, 1995.

[2] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*, pages 202–210, 2003.

[3] C. H. Dowker. Homology groups of relations. *Annals of Mathematics*, 56(1):84–95, 1952.

[4] C. Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, pages 1–19, 2008.

[5] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[6] M. A. Erdmann. On the topology of discrete strategies. *International Journal of Robotics Research*, 29(7):855–896, 2010.

[7] M. A. Erdmann. On the topology of discrete planning with uncertainty. In A. Zomorodian, editor, *Advances in Applied and Computational Topology*, pages 147–193. AMS, 2012.

[8] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security*, pages 1054–1067, 2014.

[9] R. Forman. A user's guide to discrete Morse theory. *Séminaire Lotharingien de Combinatoire*, 48:B48c, 2002.

[10] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin, 1999.

[11] R. Ghrist, D. Lipsky, J. Derenick, and A. Speranzon. Topological landmark-based navigation and mapping. Preprint, 2012.

[12] P. M. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573, 2003.

[13] F. Harary and R. Read. Is the null-graph a pointless concept? In *Proceedings of the Capital Conference on Graph Theory and Combinatorics*, pages 37–44, 1973.

[14] A. Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, 2002.

[15] M. Hoang, R. Ramanathan, and A. Singh. Structure and evolution of missed collaborations in large networks. In *The Sixth IEEE International Workshop on Network Science for Communication Networks, in conjunction with IEEE Infocom*, pages 849–854, 2014.

[16] J. R. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, Menlo Park, CA, 1984.

[17] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[18] D. Quillen. Higher algebraic K-theory: I. In *Lecture Notes in Mathematics*, volume 341, pages 85–147. Springer-Verlag, Berlin, 1973.

[19] D. Quillen. Homotopy properties of the poset of nontrivial p-subgroups of a group. *Advances in Mathematics*, 28(2):101–128, 1978.

[20] J. J. Rotman. *An Introduction to Algebraic Topology.* Springer-Verlag, Berlin, 1988.

[21] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[22] M. L. Wachs. *Poset Topology: Tools and Applications.* IAS/Park City Mathematics Institute, Summer 2004. Also available here: `http://arxiv.org/abs/math/0602226`.

[23] S. L. Warner. Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[24] S. D. Warren and L. D. Brandeis. The right to privacy. *Harvard Law Review*, IV(5):193–220, 1890.

[25] R. Wille. Concept lattices and conceptual knowledge systems. *Computers & Mathematics with Applications*, 23(6–9):493–515, 1992.

A Preliminaries

Assumption: All simplicial complexes, relations, posets, and lattices in this report are finite.

A.1 Simplicial Complexes

We largely follow the notation and definitions in [16] and [1].

- An *(abstract) simplicial complex* Σ *with underlying vertex set* X is a collection of finite subsets of X, such that if σ is in Σ, then so is every subset of σ. An element of Σ is called a *simplex*. We allow the empty set $\emptyset$ to be a simplex in Σ, for combinatorial reasons. An element of a simplex is called a *vertex*. It is also convenient to refer indistinguishably to any singleton simplex as a *vertex*. Not all elements of X need to be vertices of Σ.

- The *dimension* of a simplex σ is one less than its cardinality. The empty simplex $\emptyset$ has dimension -1. If a simplex has dimension k we sometimes call it a *k-simplex*.

- If Σ is a simplicial complex with underlying vertex set X, we let $\text{verts}(\Sigma)$ denote the set of elements in X that actually appear as vertices in Σ. Viewed as a set of 0-simplices, this set is called the *zero-skeleton* of Σ. [The standard notation for the zero-skeleton is $\Sigma^{(0)}$ but that conflicts with some iterative notation in the proof of Theorem 26.]

- The *void complex* $\emptyset$ has no simplices in it. We view it as a *degenerate* space. The *empty complex* $\{\emptyset\}$ consists solely of the empty simplex. The empty complex represents the empty topological space. It is also the sphere of dimension -1, written $\mathbb{S}^{-1}$. (There could be be different instances of the void or empty complex, depending on the underlying vertex set X, though frequently one takes X to be empty in these situations.)

- A simplex σ of a simplicial complex Σ is a *free face* of Σ if it is a proper subset of exactly one maximal simplex τ of Σ. (The empty simplex $\emptyset$ can sometimes be a free face.)

- Suppose Σ is a simplicial complex. Then $C_k(\Sigma;\mathbb{Z})$ denotes the group of simplicial *k-chains* over Σ, with integer coefficients. A k-chain $c \in C_k(\Sigma;\mathbb{Z})$ is a function that assigns to each oriented k-simplex τ of Σ an integer. When $k > 0$, one further requires that $c(-\tau) = -c(\tau)$. Here $-\tau$ refers to the same combinatorial set as τ but with opposite orientation. (When $k = 0$ or $k = -1$, each k-simplex has only one possible orientation.)

  Caution: We later also use the word "chain" in the poset sense; there should be no ambiguity given context.

- Suppose Σ is a simplicial complex and $c \in C_k(\Sigma;\mathbb{Z})$. Assume all simplices have been assigned an orientation in Σ. One can write $c = \sum_i n_i\tau_i$ uniquely, for some subcollection $\{\tau_i\}$ of the k-simplices in Σ, such that $n_i \neq 0$ for each i. (Any k-simplex τ of Σ may appear at most once in the sum, with its assigned orientation.) This means $c(\tau_i) = n_i$ for each τ_i that appears in the sum and, if $k > 0$, $c(-\tau_i) = -n_i$. For all other oriented k-simplices τ of Σ, $c(\tau) = 0$.

  We define the *support* of c as $\|c\| = \cup_i\tau_i$. [This is not standard notation.] The support is the set of all vertices that appear in any of the simplices τ for which $c(\tau)$ is nonzero.

- We let ∂ and $\widetilde{\partial}$ stand for "boundary". There are two contexts:

  1. When V is a nonempty finite set, then $\partial(V)$ means the simplicial complex whose underlying vertex set is V and whose simplices are all the proper subsets of V. We refer to this complex as the *boundary complex of the full simplex on vertex set* V. It has the homotopy type of a sphere, specifically $\mathbb{S}^{n-2}$, with $n = |V|$, for all $n \geq 1$.

  2. We also designate the *simplicial boundary operator* by ∂ and the *reduced boundary operator* by $\widetilde{\partial}$. These operators are families of maps, describing for each dimension k a group homomorphism $C_k(\Sigma;\mathbb{Z}) \to C_{k-1}(\Sigma;\mathbb{Z})$, defined on basis elements by:

     When $\sigma = \{x_0, \ldots, x_k\}$ is an oriented k-simplex, with $k \geq 1$, $\widetilde{\partial}_k(\sigma) = \partial_k(\sigma) = \sum_{i=0}^{k}(-1)^i \tau_i$, where τ_i is the oriented $(k-1)$-simplex formed from σ by removing vertex x_i and using the induced orientation of σ on τ_i. (See [16, 14] for details.)

     For $k = 0$, $\partial_0 : C_0(\Sigma;\mathbb{Z}) \to 0$, while $\widetilde{\partial}_0 : C_0(\Sigma;\mathbb{Z}) \to C_{-1}(\Sigma;\mathbb{Z})$, with $\widetilde{\partial}_0(\{v\}) = \mathbb{1}$, for each vertex $\{v\} \in \Sigma$. (Here $\mathbb{1}$ represents the generator of $C_{-1}(\Sigma;\mathbb{Z})$ when Σ is nonvoid. If Σ is void, then $\widetilde{\partial}_0 = 0$.) There is also a map $\widetilde{\partial}_{-1} : C_{-1}(\Sigma;\mathbb{Z}) \to 0$.

     We are mainly interested in the reduced boundary operator $\widetilde{\partial}$.
     We may write $\widetilde{\partial}$ in place of $\widetilde{\partial}_k$ when the dimensional context k is clear.

     Elements of the subgroup $\ker(\widetilde{\partial}_k)$ are called *reduced k-cycles.*
     Elements of the subgroup $\mathrm{img}(\widetilde{\partial}_{k+1})$ are called *reduced k-boundaries.*

- Given a simplicial complex Σ, $\widetilde{H}_k(\Sigma;\mathbb{Z})$ is the *reduced homology group in dimension* k based on simplicial chains over Σ with integer coefficients. It is a quotient group, measuring the reduced k-cycles that are not reduced k-boundaries.

  Formally, $\widetilde{H}_k(\Sigma;\mathbb{Z}) = \ker(\widetilde{\partial}_k)/\mathrm{img}(\widetilde{\partial}_{k+1})$. (That makes sense since $\widetilde{\partial}_k \circ \widetilde{\partial}_{k+1} = 0$.)

- Given a simplicial complex Σ and a set σ, we define the following three simplicial subcomplexes of Σ in the standard way:

  – The *link* of σ in Σ: $\mathrm{Lk}(\Sigma,\sigma) = \{\tau \in \Sigma \mid \tau \cap \sigma = \emptyset \text{ and } \tau \cup \sigma \in \Sigma\}$.
  – The *deletion* of σ in Σ: $\mathrm{dl}(\Sigma,\sigma) = \{\tau \in \Sigma \mid \tau \cap \sigma = \emptyset\}$.
  – The *closed star* of σ in Σ: $\overline{\mathrm{St}}(\Sigma,\sigma) = \{\tau \in \Sigma \mid \tau \cup \sigma \in \Sigma\}$.

  The definitions make sense even when σ is not itself a simplex in Σ, though in that case both $\mathrm{Lk}(\Sigma,\sigma)$ and $\overline{\mathrm{St}}(\Sigma,\sigma)$ are instances of the void complex $\emptyset$.

  Observe that $\mathrm{dl}(\Sigma,\sigma) \cap \overline{\mathrm{St}}(\Sigma,\sigma) = \mathrm{Lk}(\Sigma,\sigma)$ and $\overline{\mathrm{St}}(\Sigma,\sigma) = \mathrm{Lk}(\Sigma,\sigma) * {<}\sigma{>}$.
  Here $*$ means simplicial join (described on page 86) and ${<}\sigma{>}$ is the *simplicial complex generated by* σ (defined to be the collection of all subsets of σ).

  When σ consists of a single element v, i.e., $\sigma = \{v\}$, we tend simply to write $\mathrm{Lk}(\Sigma,v)$, $\mathrm{dl}(\Sigma,v)$, $\overline{\mathrm{St}}(\Sigma,v)$. Aside: For a singleton v, it is further true that $\mathrm{dl}(\Sigma,v) \cup \overline{\mathrm{St}}(\Sigma,v) = \Sigma$.

- One may associate a *geometric realization* to a finite nonvoid abstract simplicial complex Σ by embedding Σ into a finite-dimensional Euclidean space. One may therefore think of Σ as a topological space in a well-defined way [16, 1].

- Suppose Σ and Γ are two simplicial complexes with underlying vertex sets X and Y, respectively. A set function $f : X \to Y$ is said to be a *simplicial map* if it satisfies the following condition: If $\sigma \in \Sigma$, then $f(\sigma) \in \Gamma$.

  In that case, one may view f as a map of simplicial complexes, $f : \Sigma \to \Gamma$.

  A simplicial map may further be viewed as a continuous function between the geometric realizations of Σ and Γ [16].

- When X_1 and X_2 are topological spaces, the notation $X_1 \simeq X_2$ means that X_1 and X_2 have the same *homotopy type* [1, 14]. One may also say that X_1 and X_2 are *homotopic* or *homotopy equivalent*. A topological space homotopic to a point is said to be *contractible*.

- When X_1 and X_2 are topological spaces, $X_1 \vee X_2$ means a *wedge sum* of X_1 and X_2 [14].

- Suppose $\mathcal{U}$ is a nonempty collection of (not necessarily distinct) topological subspaces of some nonempty ambient topological space. One may define a simplicial complex $\mathcal{N}(\mathcal{U})$, called the *nerve of* $\mathcal{U}$, whose simplices are the finite subcollections $\mathcal{W}$ of $\mathcal{U}$ for which $\bigcap_{W \in \mathcal{W}} W$ is not the empty space. If $\bigcap_{W \in \mathcal{W}} W$ is contractible for each nonempty simplex $\mathcal{W}$ of $\mathcal{N}(\mathcal{U})$, then, under a variety of additional finiteness conditions [1, 14], the nerve has the same homotopy type as the union of all the spaces in $\mathcal{U}$: $\mathcal{N}(\mathcal{U}) \simeq \bigcup_{U \in \mathcal{U}} U$.

- Suppose Σ and Γ are simplicial complexes with disjoint underlying vertex sets. The *simplicial join* [22] of Σ and Γ is the simplicial complex

  $$\Sigma * \Gamma \;=\; \{\sigma \cup \gamma \mid \sigma \in \Sigma \text{ and } \gamma \in \Gamma\}.$$

  The underlying vertex set of $\Sigma * \Gamma$ is the union of the underlying vertex sets of Σ and Γ.

A.2 Partially Ordered Sets (Posets)

We largely follow the notation of [22].

- A *poset* P is a set of elements with a partial order, sometimes written simply as "$\leq$" other times as "$\leq_P$". The symbols "$\geq$", "$<$", "$>$" and "$=$" are defined accordingly.

- A *chain* c in a poset P is a totally ordered subset of P, which we often write as $c = \{p_0 < p_1 < \cdots < p_\ell\}$. The *length* $\ell(c)$ of chain c is ℓ, one less than the number of elements in the chain (analogous to simplex dimension). The length of the empty chain is -1. The length $\ell(P)$ of a poset P is the maximum length of any chain in P.

- The *face poset* $\mathfrak{F}(\Sigma)$ of a nonvoid simplicial complex Σ consists of all nonempty simplices of Σ, partially ordered by set inclusion. (If Σ is void, we leave $\mathfrak{F}(\Sigma)$ *undefined*.)

- The *order complex* $\Delta(P)$ of a poset P is the simplicial complex whose simplices are given by all finite chains $\{p_0 < p_1 < \cdots < p_\ell\}$ in P. (If $P = \emptyset$, then $\Delta(P) = \{\emptyset\}$.)

- One may speak of the *topology of a poset*: One says that a poset P has a topological property when its order complex $\Delta(P)$ has that property and the property is an invariant of homeomorphism type. For instance, to say that a poset is contractible means that its order complex is contractible. To say that two posets P and Q are homotopic means that $\Delta(P)$ and $\Delta(Q)$ have the same homotopy type. Etc.

- For nonvoid Σ, it is a fact that $\Delta(\mathfrak{F}(\Sigma))$ is homeomorphic to Σ. Indeed, $\Delta(\mathfrak{F}(\Sigma))$ may be viewed as the *first barycentric subdivision of* Σ, which we write as $\mathrm{sd}(\Sigma)$. See [20, 22].

- A set function $\theta : P \to Q$ between two posets P and Q is said to be a *poset map* if it is either *order-preserving* or *order-reversing*. That means:

$$\begin{aligned} &\text{order-preserving:} \quad \text{For all } x, y \in P, \quad \text{if } \ x \leq_P y, \ \text{ then } \ \theta(x) \leq_Q \theta(y). \\ &\text{order-reversing:} \quad \text{For all } x, y \in P, \quad \text{if } \ x \leq_P y, \ \text{ then } \ \theta(x) \geq_Q \theta(y). \end{aligned}$$

- A poset map $\theta : P \to Q$ between two posets P and Q induces a simplicial map between the associated order complexes $\theta : \Delta(P) \to \Delta(Q)$.

- An order-preserving poset self-map $\theta : P \to P$ is said to be a *closure operator* when $x \leq_P \theta(x)$, for all $x \in P$, and $\theta \circ \theta = \theta$. A closure operator θ induces a homotopy equivalence between P and the image $\theta(P)$. See [1, 22, 19, 18].

A.3 Semi-Lattices and Lattices

We largely follow the development in [22] and [1]. Let L be a partially ordered set:

- Suppose $p, q \in L$. If p and q have a unique least upper bound, then one writes $p \vee q$ to mean that least upper bound. (One may also write $p \vee_L q$.) If every pair of elements in L has a unique least upper bound in L, then one refers to L as a *join semi-lattice*.

- Suppose $p, q \in L$. If p and q have a unique greatest lower bound, then one writes $p \wedge q$ (or possibly $p \wedge_L q$) to mean that greatest lower bound. If every pair of elements in L has a unique greatest lower bound in L, then one refers to L as a *meet semi-lattice*.

- A poset that is both a join semi-lattice and a meet semi-lattice is known as a *lattice*.

- If L has a unique top (i.e., maximal) element, we may designate that element by $\hat{1}$ or $\hat{1}_L$.

- If L has a unique bottom (minimal) element, we may designate that element by $\hat{0}$ or $\hat{0}_L$.

- If L is a finite join semi-lattice with a unique bottom element, then L is a lattice. Similarly, if L is a finite meet semi-lattice with a unique top element, then L is a lattice.

- A lattice L is called *bounded* if it has a unique top element $\hat{1}$ and a unique bottom element $\hat{0}$. (These are same element if L is a singleton.)

- When L is a bounded lattice, the *proper part of* L is the poset $\overline{L} = L \setminus \{\hat{0}, \hat{1}\}$.

- Suppose L is a bounded lattice and $p \in L$. Then the *complements of* p are given by the set $\mathfrak{C}(p) = \{q \in L \mid q \vee p = \hat{1} \text{ and } q \wedge p = \hat{0}\}$.

- A bounded lattice L is said to be *noncomplemented* if $\mathfrak{C}(p) = \emptyset$ for at least one $p \in L$. If L is a noncomplemented bounded lattice with $\overline{L} \neq \emptyset$, then $\overline{L}$ is contractible [1].

- Suppose L is a bounded lattice with $\overline{L} \neq \emptyset$. The elements of L immediately below $\hat{1}$ are called *co-atoms*. These are the maximal elements of $\overline{L}$. The elements immediately above $\hat{0}$ are called *atoms*. These are the minimal elements of $\overline{L}$.

A.4 Relations

Let R be a relation on $X \times Y$, with X and Y finite discrete spaces.
We use the following notation and conventions (see also page 13):

- R is a set of ordered pairs, namely a subset of the cross product $X \times Y$. It is convenient sometimes to view R as a matrix of 0s and 1s, perhaps drawn as a matrix of blank and nonblank entries, representing the characteristic function of this set of ordered pairs.

- Even if $X \neq \emptyset$ and $Y \neq \emptyset$, it is possible that $R = \emptyset$, in which case we say that R is an *empty relation*.

- If $X = \emptyset$ and/or $Y = \emptyset$, then we say that R is a *void relation*.

  On some occasions, we may treat a void relation R much like an empty relation, in the sense that we will let the *Dowker complexes* defined below (and on page 13) be empty rather than void. That view will sometimes be convenient when R is derived from some encompassing relation as a link or deletion in a simplicial complex.

- We often refer to elements of X as *individuals* and elements of Y as *attributes*.

- For each $x \in X$, Y_x is the set of attributes that individual x has (in relation R). Viewing R as a matrix, one may think of Y_x as the row of R indexed by x. We say that *the row is blank* when $Y_x = \emptyset$.

- For each $y \in Y$, X_y is the set of individuals who have attribute y (in relation R). Viewing R as a matrix, one may think of X_y as the column of R indexed by y. We say that *the column is blank* when $X_y = \emptyset$.

- Φ_R is the *Dowker attribute complex* determined by R. It is a simplicial complex with underlying vertex set Y. A nonempty subset γ of Y is a simplex in Φ_R precisely when there exists $x \in X$ such that $(x, y) \in R$ for all $y \in \gamma$. We refer to x as a *witness for* γ.

  When R is void, we let Φ_R be void as well, except as otherwise indicated in the text.

  When R is nonvoid, Φ_R contains at least the empty simplex. We then may view Φ_R as generated by the rows of R, so $\Phi_R = \bigcup_{x \in X} <\!Y_x\!>$. Thus, $Y_x \in \Phi_R$ for each $x \in X$.

- Ψ_R is the *Dowker association complex* determined by R. It is a simplicial complex with underlying vertex set X. A nonempty subset σ of X is a simplex in Ψ_R precisely when there exists $y \in Y$ such that $(x, y) \in R$ for all $x \in \sigma$. We refer to y as a *witness for* σ.

  When R is void, we let Ψ_R be void as well, except as otherwise indicated in the text.

  When R is nonvoid, Ψ_R contains at least the empty simplex. We then may view Ψ_R as generated by the columns of R, so $\Psi_R = \bigcup_{y \in Y} <X_y>$. Thus, $X_y \in \Psi_R$ for each $y \in Y$.

- There exist homotopy equivalences $\phi_R : \Psi_R \to \Phi_R$ and $\psi_R : \Phi_R \to \Psi_R$.

  Viewed as (order-reversing) poset maps $\phi_R : \mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Phi_R)$ and $\psi_R : \mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Psi_R)$, one obtains explicit formulas, sending nonempty simplices to nonempty simplices:

  $$\phi_R(\sigma) = \bigcap_{x \in \sigma} Y_x \quad \text{and} \quad \psi_R(\gamma) = \bigcap_{y \in \gamma} X_y.$$

  Suppose $X \neq \emptyset$ and $Y \neq \emptyset$. Then the intersections appearing in the previous formulas comprise the witnesses for the respective simplex arguments. Consequently, one may use the formulas more generally as tests for membership in the Dowker complexes:

  – For any $\sigma \subseteq X$, $\sigma \in \Psi_R$ if and only if $\phi_R(\sigma) \neq \emptyset$.
  – For any $\gamma \subseteq Y$, $\gamma \in \Phi_R$ if and only if $\psi_R(\gamma) \neq \emptyset$.

  These tests also make sense for the empty set, that is, when $\sigma = \emptyset$ or $\gamma = \emptyset$. In particular, $\phi_R(\emptyset) = Y$ and $\psi_R(\emptyset) = X$.

- Composing ϕ_R and ψ_R as $\psi_R \circ \phi_R : \mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Psi_R)$ and $\phi_R \circ \psi_R : \mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Phi_R)$ produces closure operators. See Appendix B for further details.

- P_R is the *doubly-labeled poset* associated with R as per Definition 3 on page 17. Each element in P_R is of the form (σ, γ), with $\sigma \neq \emptyset$ and $\gamma \neq \emptyset$, such that $\sigma = \psi_R(\gamma)$ and $\gamma = \phi_R(\sigma)$.

  One may view P_R either as the image $(\psi_R \circ \phi_R)(\mathfrak{F}(\Psi_R))$ or as the image $(\phi_R \circ \psi_R)(\mathfrak{F}(\Phi_R))$.

  We mention some special cases:

  – If Ψ_R and Φ_R are instances of the empty complex $\{\emptyset\}$, then $P_R = \emptyset$. This occurs when R is an empty relation, or when R is void but we let $\Psi_R = \{\emptyset\}$ and $\Phi_R = \{\emptyset\}$.
  – If Ψ_R and Φ_R are instances of the void complex $\emptyset$, then P_R is *undefined.*

- P_R^+ is the *Galois lattice* formed from P_R as per Definition 13 on page 35.

  If R is an empty relation, then $P_R = \emptyset$ and so P_R^+ consists simply of $\hat{0}_R$ and $\hat{1}_R$.

  Definition 13 assumes that the underlying spaces X and Y of R are both nonempty. One could imagine extending the definition, perhaps as follows: (i) When the Dowker complexes are void, leave P_R undefined and let $P_R^+ = \emptyset$. (ii) When R is technically void but the Dowker complexes are artificially empty, with one of X or Y empty, let $P_R = \emptyset$ and $P_R^+ = \{(X, Y)\}$. Fortunately, we will not need these boundary cases.

(Different perspectives often suggest conflicting interpretations in null situations [13]. This report chooses to preserve the validity of Dowker's Theorem, meaning $\Psi_R \simeq \Phi_R$.)

- We sometimes view P_R as "almost a join-based lattice", as per Definition 25 on page 44. That amounts to adjoining a single new element $\hat{1}$ above P_R, then inducing a join operation on $P_R \cup \{\hat{1}\}$ from the join operation on P_R^+. Thus $P_R \cup \{\hat{1}\}$ is a join semi-lattice. If we further adjoin a new bottom element $\hat{0}$, then $P_R \cup \{\hat{0}, \hat{1}\}$ is a lattice.

- One may speak of the *topology of a relation (modulo homotopy equivalence)*: One says that a relation R has a topological property when any and all of Φ_R, Ψ_R, and $\Delta(P_R)$ have that property and the property is an invariant of homotopy type. (This convention makes sense by Dowker's Theorem on page 14 and the nature of P_R.) Connectivity is an example of such a property.

B Basic Tools

This appendix reviews some basic facts about relations, their Dowker complexes, and the Galois connection. Recall the formulas for ϕ_R and ψ_R from page 89.

Although we do not always say so explicitly, there are dual statements for the lemmas and corollaries in this appendix, for each of the two perspectives offered by Dowker's Theorem, by inverting the roles of individuals and attributes.

Lemma 42. *Let R be a relation on $X \times Y$. Then ϕ_R is inclusion-reversing.*

Proof. Let $\sigma' \subseteq \sigma \subseteq X$. Then: $$\phi_R(\sigma') \;=\; \bigcap_{x\in\sigma'} Y_x \;\supseteq\; \bigcap_{x\in\sigma} Y_x \;=\; \phi_R(\sigma).$$

Just to be careful: if $\sigma' = \emptyset$, then $\phi_R(\sigma') = Y$, which does indeed contain $\phi_R(\sigma)$. □

Each of ϕ_R and ψ_R is inclusion-reversing, so $\phi_R \circ \psi_R$ is inclusion-preserving. Lemmas 43 and 45 establish that $\phi_R \circ \psi_R$ is a closure operator when viewed as a poset map $\mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Phi_R)$:

Lemma 43. *Let R be a relation on $X \times Y$. For all $\gamma \subseteq Y$, $\gamma \subseteq (\phi_R \circ \psi_R)(\gamma)$.*

Proof.
$$(\phi_R \circ \psi_R)(\gamma) = \bigcap_{x\in\sigma} Y_x, \quad \text{with } \sigma = \bigcap_{y\in\gamma} X_y.$$

The assertion is clear if $\gamma = \emptyset$ or $\sigma = \emptyset$. Otherwise, let $y \in \gamma$ and $x \in \sigma$. Then $x \in X_y$, so $y \in Y_x$. Since x is arbitrary in σ, we see that $y \in (\phi_R \circ \psi_R)(\gamma)$ and thus $\gamma \subseteq (\phi_R \circ \psi_R)(\gamma)$. □

Corollary 44. *Let R be a relation on $X \times Y$, with both X and Y nonempty.*
If γ is a maximal simplex of Φ_R, then $(\phi_R \circ \psi_R)(\gamma) = \gamma$.

Proof. When $\gamma \neq \emptyset$, this assertion follows from Lemma 43 and maximality of γ. Otherwise, apparently $\Phi_R = \{\emptyset\}$ and so $(\phi_R \circ \psi_R)(\emptyset) = \phi_R(X) = \emptyset$ (since ϕ_R must map X into Φ_R). □

Lemma 45. *Let R be a relation on $X \times Y$.*
For all $\gamma \subseteq Y$, $\big((\phi_R \circ \psi_R) \circ (\phi_R \circ \psi_R)\big)(\gamma) = (\phi_R \circ \psi_R)(\gamma)$.

Proof. Consider: $\gamma \xmapsto{\psi_R} \sigma \xmapsto{\phi_R} \gamma' \xmapsto{\psi_R} \sigma' \xmapsto{\phi_R} \gamma''$.

We need to show that $\gamma' = \gamma''$.
By Lemma 43 and its dualization, $\gamma \subseteq \gamma' \subseteq \gamma''$ and $\sigma \subseteq \sigma'$.
By Lemma 42, ϕ_R is inclusion-reversing, so $\sigma \subseteq \sigma'$ implies $\gamma' \supseteq \gamma''$, and thus $\gamma' = \gamma''$.
Comment: By the dual of Lemma 42, ψ_R is inclusion-reversing, so in fact also $\sigma = \sigma'$. □

Corollary 46. *Let R be a relation on $X \times Y$. For all $\sigma \subseteq X$, $(\phi_R \circ \psi_R)(\phi_R(\sigma)) = \phi_R(\sigma)$.*

Proof. This follows from a dual version of the comment at the end of the proof of Lemma 45. □

Corollary 47. *Let R be a relation on $X \times Y$. For all $x \in X$, $(\phi_R \circ \psi_R)(Y_x) = Y_x$.*

Proof. The assertion follows from Corollary 46, with $\sigma = \{x\}$.
(This includes the case $Y_x = \emptyset$.) □

Lemma 48. *Let R be a relation on $X \times Y$ and suppose $\eta \subseteq Y$. Then the following two conditions are equivalent:*

(a) $(\phi_R \circ \psi_R)(\chi) = \chi$, *for every proper subset* χ *of* η.

(b) $(\phi_R \circ \psi_R)(\gamma) = \gamma$, *for all* γ *of the form* $\gamma = \eta \setminus \{y\}$ *with* $y \in \eta$.

Proof. Certainly (a) implies (b). Suppose (b) holds, but there is some $\chi \subsetneq \eta$ such that $\chi \subsetneq (\phi_R \circ \psi_R)(\chi)$. Since (b) holds, $(\phi_R \circ \psi_R)(\chi) \subseteq \eta$. Let $y \in (\phi_R \circ \psi_R)(\chi) \setminus \chi$ and consider $\gamma = \eta \setminus \{y\}$.

Observe that $\chi \subseteq \gamma$, so $y \in (\phi_R \circ \psi_R)(\chi) \subseteq (\phi_R \circ \psi_R)(\gamma)$. Consequently,

$$\eta \;=\; \gamma \cup \{y\} \;\subseteq\; (\phi_R \circ \psi_R)(\gamma) \;=\; \gamma \;\subsetneq\; \eta, \qquad \text{which is a contradiction.}$$ □

Definition 49 (Connected)**.** *A relation R on $X \times Y$ is* connected *if R is connected when viewed as an undirected bipartite graph on the vertex sets X and Y. (This definition regards X and Y as disjoint.)*

Definition 50 (Tight)**.** *A relation R on $X \times Y$ is* tight *if it has no blank rows or columns.*

Comment: As mentioned on page 86, one can view an abstract simplicial complex as a topological space, via its geometric realization. In particular, one may ask whether a simplicial complex is *path-connected.*

Lemma 51 (Connectedness)**.** *Let R be a tight relation on $X \times Y$, with both X and Y nonempty. Then the following three conditions are equivalent:*

(a) R *is connected.*

(b) Ψ_R *is path-connected.*

(c) Φ_R *is path-connected.*

Proof. We will show that (a) and (b) are equivalent. The proof for (a) and (c) is similar, or one can simply invoke Dowker duality.

I. Suppose R is connected. Consider two vertices x_0 and x_f of Ψ_R. Since R is connected as a bipartite graph, there exists a path $x_0, y_1, x_1, y_2, \ldots, y_n, x_n = x_f$ in this graph. Observe that each y_i is a witness for the simplex $\{x_{i-1}, x_i\} \in \Psi_R$. We can assume without loss of generality that $x_{i-1} \neq x_i$, for each relevant i. So in Ψ_R there exist edges $\{x_0, x_1\}, \ldots, \{x_{n-1}, x_n\}$. Since Ψ_R is a simplicial complex, we see that it is path-connected.

II. Suppose Ψ_R is path-connected. Since R is tight, each $y \in Y$ appears as the vertex of an edge (x, y) in the bipartite graph R. To show that R is connected, it therefore is enough to show that any two elements x_0 and x_f of X may be connected by a path in the bipartite graph. Since R is tight, x_0 and x_f are each vertices of Ψ_R. Since Ψ_R is path-connected, there exists a path between x_0 and x_f in Ψ_R. Since Ψ_R is a finite simplicial complex, we can deform that path so that it consists of finitely many edges $\{x_0, x_1\}, \ldots, \{x_{n-1}, x_n\}$, with each x_i a vertex of Ψ_R and $x_n = x_f$. Each edge $\{x_{i-1}, x_i\}$ has some witness $y_i \in Y$. So $x_0, y_1, x_1, y_2, \ldots, y_n, x_f$ is a path connecting x_0 and x_f in the bipartite graph R. □

Lemma 52 (Components). *Let R be a tight relation on $X \times Y$, with both X and Y nonempty. Suppose $R = R_1 \cup \cdots \cup R_\ell$, with the $\{R_i\}$ pairwise disjoint and each R_i a connected component of R viewed as a bipartite graph on X and Y. Then X, Y, Ψ_R, and Φ_R decompose as follows:*

(a) $X = X_1 \cup \cdots \cup X_\ell$, with the $\{X_i\}$ pairwise disjoint and each X_i not empty.

(b) $Y = Y_1 \cup \cdots \cup Y_\ell$, with the $\{Y_i\}$ pairwise disjoint and each Y_i not empty.

(c) R_i is the restriction of R to $X_i \times Y_i$, and is tight, for $i = 1, \ldots, \ell$.

(d) $\Psi_R = \Psi_{R_1} \cup \cdots \cup \Psi_{R_\ell}$, with pairwise disjoint face posets and each Ψ_{R_i} path-connected.

(e) $\Phi_R = \Phi_{R_1} \cup \cdots \cup \Phi_{R_\ell}$, with pairwise disjoint face posets and each Φ_{R_i} path-connected.

Proof. Let $X_i = \{x \mid (x, y) \in R_i \text{ for some } y \in Y\}$ and $Y_i = \{y \mid (x, y) \in R_i \text{ for some } x \in X\}$, for $i = 1, \ldots, \ell$. These sets are nonempty since the components of R are necessarily nonempty.

To see that $X_i \cap X_j = \emptyset$ unless $i = j$, suppose $x \in X_i \cap X_j$. Then $(x, y) \in R_i$ for some $y \in Y$ and $(x, y') \in R_j$ for some $y' \in Y$. Since R_i and R_j are connected components of R, $i = j$. Next observe that each x of X must appear in some X_i since R has no blank rows. Point (a) follows. Point (b) is similar.

For (c), observe that if $(x, y) \in R_i \subseteq R$ then $x \in X_i$ and $y \in Y_i$, so (x, y) is in the restriction of R to $X_i \times Y_i$. Conversely, if $(x, y) \in R$ with $x \in X_i$ and $y \in Y_i$, then $(x, y) \in R_j$ for some j. By the previous reasoning, $i = j$. Tightness follows by definition of X_i and Y_i.

For (d), $\Psi_{R_i} \subseteq \Psi_R$ since $R_i \subseteq R$, for each $i = 1, \ldots, \ell$. Now let $\emptyset \neq \sigma \in \Psi_R$. Then there exists $y \in Y$ such that $(x, y) \in R$ for every $x \in \sigma$. For some i, $y \in Y_i$. Since R_i is a connected component of R, $(x, y) \in R_i$ for every $x \in \sigma$, so $\sigma \in \Psi_{R_i}$. The collections $\{\mathfrak{F}(\Psi_{R_i})\}$ are pairwise disjoint since the underlying vertex sets $\{X_i\}$ are pairwise disjoint. Path-connectedness follows from Lemma 51, since each R_i is tight and connected. Point (e) is similar. □

Corollary 53 (Component Maps). *Assume the hypotheses and constructions as in Lemma 52 and its proof. Then:*

$$\begin{aligned} \psi_{R_i}(\gamma) &= \psi_R(\gamma), \quad \textit{for each } \emptyset \neq \gamma \in \Phi_{R_i}, \\ \phi_{R_i}(\sigma) &= \phi_R(\sigma), \quad \textit{for each } \emptyset \neq \sigma \in \Psi_{R_i}, \quad i = 1, \ldots, \ell. \end{aligned}$$

Proof. By direct computation (be aware, the subscripts in X_y and X_i have different meanings):

$$\psi_{R_i}(\gamma) = \bigcap_{y \in \gamma} (X_y \cap X_i) = \bigcap_{y \in \gamma} X_y = \psi_R(\gamma).$$

The second equality comes from the fact that each X_y can touch only X_i, since R_i is a connected component of R. The argument for the $\phi_{\ldots}$ maps is similar. □

Corollary 54 (Component Privacy). *Assume the hypotheses and constructions as in Lemma 52 and its proof. Let $i \in \{1, \ldots, \ell\}$.*

If $\psi_R \circ \phi_R$ is the identity on Ψ_R and $Y_i \notin \Phi_{R_i}$, then $\psi_{R_i} \circ \phi_{R_i}$ is the identity on Ψ_{R_i}.

If $\phi_R \circ \psi_R$ is the identity on Φ_R and $X_i \notin \Psi_{R_i}$, then $\phi_{R_i} \circ \psi_{R_i}$ is the identity on Φ_{R_i}.

Proof. Suppose $\emptyset \neq \sigma \in \Psi_{R_i}$. Then $\emptyset \neq \phi_{R_i}(\sigma) \in \Phi_{R_i}$, so by Corollary 53, $(\psi_{R_i} \circ \phi_{R_i})(\sigma) = (\psi_R \circ \phi_R)(\sigma) = \sigma$. And $(\psi_{R_i} \circ \phi_{R_i})(\emptyset) = \psi_{R_i}(Y_i) = \emptyset$, since $Y_i \notin \Phi_{R_i}$.

The argument for $\phi_{R_i} \circ \psi_{R_i}$ is similar. □

C Links, Deletions, and Inference

This appendix provides some technical tools for modeling inference, particularly in links, ending with some instances in which inference is unavoidable.

C.1 Links, Deletions, and Induced Maps

Intuition: The link $\mathrm{Lk}(\Phi_R, \gamma)$ of a set of attributes γ in the Dowker complex Φ_R can be understood as a description of what may yet be observed or inferred, *conditional* on having already observed γ.

Lemma 55. *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\gamma \in \Phi_R$. Define relation Q as a restriction of R by*

$$Q \;=\; R\,\big|_{\sigma\times\overline{Y}}, \quad \textit{with} \quad \sigma = \psi_R(\gamma) \quad \textit{and} \quad \overline{Y} = \bigcup_{x\in\sigma} Y_x \setminus \gamma.$$

(See the comments below for the case in which $\overline{Y} = \emptyset$.)

Then $\mathrm{Lk}(\Phi_R, \gamma) = \Phi_Q$, *as collections of simplices (i.e., ignoring underlying vertex sets).*

Comments: (a) Observe that $\sigma \neq \emptyset$. (b) If $\overline{Y} = \emptyset$, then technically Q is void, but it is convenient to let both Φ_Q and Ψ_Q be instances of the empty complex $\{\emptyset\}$, as in Definition 7 on page 24. (c) In a standard link, one might define $\overline{Y} = Y \setminus \gamma$. With $\overline{Y}$ as above, Q always discards blank columns of R, even when $\gamma = \emptyset$.

Proof. Observe that $\gamma \subseteq Y_x$ if and only if $x \in \sigma$.

We discuss the case $\overline{Y} = \emptyset$ separately, for clarity. We need to show that $\mathrm{Lk}(\Phi_R, \gamma) = \{\emptyset\}$. If $\mathrm{Lk}(\Phi_R, \gamma) \neq \{\emptyset\}$, then there exists some $\overline{y} \in \mathrm{verts}(\mathrm{Lk}(\Phi_R, \gamma))$. By definition of link, $\overline{y} \notin \gamma$ and there exists $\overline{x} \in X$ such that $(\overline{x}, y) \in R$ for all $y \in \gamma \cup \{\overline{y}\}$. That means $\overline{x} \in \sigma$, so $\overline{y} \in \overline{Y}$, a contradiction.

The converse is true as well: If $\mathrm{Lk}(\Phi_R, \gamma) = \{\emptyset\}$, then $\overline{Y} = \emptyset$. For if some $x \in \sigma$ has an attribute y in addition to all those in γ, then y would be a vertex in the link.

Now suppose $\overline{Y} \neq \emptyset$:

I. If $\xi \in \mathrm{Lk}(\Phi_R, \gamma)$, then $\xi \cap \gamma = \emptyset$ and there exists $x \in X$ such that $(x, y) \in R$ for every $y \in \xi \cup \gamma$. So $\xi \subseteq Y_x \setminus \gamma$ and $x \in \psi_R(\gamma) = \sigma$. Thus $(x, y) \in Q$ for every $y \in \xi$, meaning $\xi \in \Phi_Q$.

II. Conversely, if $\xi \in \Phi_Q$, then there exists $x \in \sigma$ such that $(x, y) \in Q \subseteq R$ for every $y \in \xi$. By definition of σ, $(x, y) \in R$ for every $y \in \gamma$. Combining these two assertions, we see that $(x, y) \in R$ for every $y \in \xi \cup \gamma$. And $\xi \cap \gamma = \emptyset$, since $\xi \subseteq \overline{Y}$. So $\xi \in \mathrm{Lk}(\Phi_R, \gamma)$. $\square$

Additional Comment: There is a dual version of this lemma for links of individuals σ, modeling $\mathrm{Lk}(\Psi_R, \sigma)$ by Ψ_Q, for an appropriate relation Q. We see instances of that construction in Theorems 10 and 11, as well as in Lemma 12, on pages 104–107 (previously stated on page 26), including the case in which σ consists of a single individual x.

Link Witness Formulas. With notation and construction as in Lemma 55, the following formulas hold, assuming $\overline{Y} \neq \emptyset$:

- Suppose $\xi \subseteq \overline{Y}$ and define $\tau = \xi \cup \gamma$. Then

$$\psi_Q(\xi) \;=\; \bigcap_{y\in\xi}(X_y \cap \sigma) \;=\; \Big(\bigcap_{y\in\xi} X_y\Big) \bigcap \Big(\bigcap_{y\in\gamma} X_y\Big) \;=\; \bigcap_{y\in(\xi\cup\gamma)} X_y \;=\; \psi_R(\tau).$$

  Notes: We allow $\xi = \emptyset$, since $\psi_Q(\emptyset) = \sigma = \psi_R(\gamma)$. We do not require $\xi \in \Phi_Q$. The equalities hold regardless. Of course, $\xi \in \Phi_Q$ if and only if $\psi_Q(\xi) \neq \emptyset$.

- Suppose $\emptyset \neq \kappa \subseteq \sigma$. Then

$$\phi_Q(\kappa) \;=\; \bigcap_{x\in\kappa}(Y_x \cap \overline{Y}) \;=\; \Big(\bigcap_{x\in\kappa} Y_x\Big) \setminus \gamma \;=\; \phi_R(\kappa) \setminus \gamma.$$

  And thus also $\phi_R(\kappa) = \phi_Q(\kappa) \cup \gamma$, since $\gamma \subseteq Y_x$ for all $x \in \sigma$.

  Notes: Here we do *not* allow $\kappa = \emptyset$, since $\phi_Q(\emptyset) = \overline{Y}$ whereas $\phi_R(\emptyset) = Y$. It need not be true that $Y = \overline{Y} \cup \gamma$. Again, $\kappa \in \Psi_Q$ if and only if $\phi_Q(\kappa) \neq \emptyset$, this valid also for $\kappa = \emptyset$.

Comment: If $\overline{Y} = \emptyset$, the previous formulas still hold, albeit trivially. However, testing for membership in Ψ_Q via the question "Is $\phi_Q(\kappa)$ nonempty?" no longer makes sense.

Lemma 56. *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\gamma \subseteq Y$. Then* $\mathrm{dl}(\Phi_R, \gamma) = \Phi_{Q'}$*, with Q' formed from R by removing the columns corresponding to γ, that is, $Q' = R|_{X\times(Y\setminus\gamma)}$. (Here we let $\Psi_{Q'}$ and $\Phi_{Q'}$ each be an empty complex if $\gamma = Y$.)*

Proof. An individual $x \in X$ is a witness to a set of attributes $\xi \subseteq Y \setminus \gamma$ in R if and only if x is a witness to ξ in Q'. (If $\gamma = Y$, then $\mathrm{dl}(\Phi_R, \gamma) = \{\emptyset\} = \Phi_{Q'}$.) □

Deletion Witness Formulas. With notation and construction as in Lemma 56, the following formulas hold, assuming $\gamma \neq Y$:

- If $\xi \subseteq (Y \setminus \gamma)$, then $\psi_{Q'}(\xi) = \bigcap_{y\in\xi} X_y = \psi_R(\xi)$.

- If $\kappa \subseteq X$, then $\phi_{Q'}(\kappa) = \bigcap_{x\in\kappa}(Y_x \setminus \gamma) = \phi_R(\kappa) \setminus \gamma$.

  Caution: It need *not* be true that $\phi_R(\kappa) = \phi_{Q'}(\kappa) \cup \gamma$.

Comments: (1) The first formula holds for $\xi = \emptyset$ and the second formula holds for $\kappa = \emptyset$. (2) The simplex tests hold: For $\xi \subseteq (Y \setminus \gamma)$, $\xi \in \Phi_{Q'}$ if and only if $\psi_{Q'}(\xi) \neq \emptyset$; and, for $\kappa \subseteq X$, $\kappa \in \Psi_{Q'}$ if and only if $\phi_{Q'}(\kappa) \neq \emptyset$. (3) If $\gamma = Y$, the formulas still hold, but testing for membership in $\Psi_{Q'}$ via the question "Is $\phi_{Q'}(\kappa)$ nonempty?" no longer makes sense.

C.2 Privacy Preservation in Links and Deletions

Recall: A relation R preserves attribute privacy when the closure operator $\phi_R \circ \psi_R$ is the identity on Φ_R and it preserves association privacy when the closure operator $\psi_R \circ \phi_R$ is the identity on Ψ_R (see page 17).

Lemma 57. *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\gamma \in \Phi_R$.*
If $\phi_R \circ \psi_R$ is the identity on Φ_R, then the corresponding closure operators for the relations modeling $\mathrm{Lk}(\Phi_R, \gamma)$ and $\mathrm{dl}(\Phi_R, \gamma)$ are also identities.
(The assertion for $\mathrm{dl}(\Phi_R, \gamma)$ holds even if γ is merely a subset of Y.)

Technical reminder: The operators are formally defined as self-maps on the face posets of the simplicial complexes mentioned in the lemma, but we can extend each operator to the empty simplex and therefore think of it as a self-map on a simplicial complex viewed as a collection of simplices. See again pages 14–17 and page 89.

Proof. Define Q as in Lemma 55. That lemma tells us $\Phi_Q = \mathrm{Lk}(\Phi_R, \gamma)$.
Given $\xi \in \Phi_Q$, let $\tau = \xi \cup \gamma$ and calculate:

$$(\phi_Q \circ \psi_Q)(\xi) \;=\; \phi_Q(\psi_R(\tau)) \;=\; \phi_R(\psi_R(\tau)) \setminus \gamma \;=\; \tau \setminus \gamma \;=\; \xi.$$

Define Q' as in Lemma 56. That lemma tells us $\Phi_{Q'} = \mathrm{dl}(\Phi_R, \gamma)$.
Given $\xi \in \Phi_{Q'}$, calculate:

$$(\phi_{Q'} \circ \psi_{Q'})(\xi) \;=\; \phi_{Q'}(\psi_R(\xi)) \;=\; \phi_R(\psi_R(\xi)) \setminus \gamma \;=\; \xi \setminus \gamma \;=\; \xi. \qquad \square$$

Here is a variation, in which one again computes a link of attributes, but then considers the closure operator on the dual association complex, modeling individuals consistent with the attributes:

Lemma 58. *Let R be a tight relation on $X \times Y$, with both X and Y nonempty. Let $\gamma \in \Phi_R$.*
Define Q, σ, and $\overline{Y}$ as in the construction of Lemma 55. Assume $|\sigma| > 1$ and $\overline{Y} \neq \emptyset$.
If $\psi_R \circ \phi_R$ is the identity on Ψ_R, then $\psi_Q \circ \phi_Q$ is the identity on Ψ_Q.

Proof. Suppose $\emptyset \neq \kappa \in \Psi_Q$. Observe that $\gamma \subseteq \phi_R(\kappa)$ and calculate:

$$(\psi_Q \circ \phi_Q)(\kappa) \;=\; \psi_Q(\phi_R(\kappa) \setminus \gamma) \;=\; \psi_R(\phi_R(\kappa)) \;=\; \kappa.$$

Additionally,

$$(\psi_Q \circ \phi_Q)(\emptyset) \;=\; \psi_Q(\overline{Y}) \;=\; \psi_R(\overline{Y} \cup \gamma) \;=\; \psi_R\Big(\bigcup_{x \in \sigma} Y_x\Big) \;=$$

$$=\; \bigcap_{x \in \sigma} \psi_R(Y_x) \;=\; \bigcap_{x \in \sigma} (\psi_R \circ \phi_R)(\{x\}) \;=\; \bigcap_{x \in \sigma} \{x\} \;=\; \emptyset.$$

The last equality holds since $|\sigma| > 1$. The equality before that holds since $\psi_R \circ \phi_R$ is the identity on Ψ_R and since R has no blank rows.
So we see that $(\psi_Q \circ \phi_Q)(\kappa) = \kappa$ for all $\kappa \in \Psi_Q$. $\square$

Comment: Assume the setting of the previous two lemmas, but suppose $\overline{Y} = \emptyset$. We would then take Φ_Q and Ψ_Q to be instances of the empty simplicial complex $\{\emptyset\}$. It is sensible to say that $\phi_Q \circ \psi_Q$ is the identity on Φ_Q, since $\phi_Q(\psi_Q(\emptyset)) = \phi_Q(\sigma) = \emptyset$. It could be confusing to say that $\psi_Q \circ \phi_Q$ is the identity on Ψ_Q, since $\psi_Q(\phi_Q(\emptyset)) = \psi_Q(\overline{Y}) = \psi_Q(\emptyset) = \sigma$. On the other hand, one could argue that one may nonetheless say that there is no association inference in Q, since there are no attributes that could witness associations.

Corollary 59. *Let R be a tight relation on $X \times Y$, with both X and Y nonempty. Let $\gamma \in \Phi_R$. Define Q and $\overline{Y}$ as in the construction of Lemma 55. Assume $\overline{Y} \neq \emptyset$.*
If R preserves both attribute and association privacy, then so does Q.

Proof. Relation Q preserves attribute privacy by Lemma 57. Let $\sigma = \psi_R(\gamma)$. If we can show that $|\sigma| > 1$, then Q preserves association privacy by Lemma 58.

Observe that $|\sigma| > 0$, since $\gamma \in \Phi_R$. If $\psi_R(\gamma)$ consists of a single individual $x \in X$, then

$$\gamma \;=\; (\phi_R \circ \psi_R)(\gamma) \;=\; \phi_R(\sigma) \;=\; Y_x \;=\; \overline{Y} \cup \gamma.$$

That is impossible for nonempty $\overline{Y}$, since $\overline{Y} \cap \gamma = \emptyset$. □

The following lemma formalizes the intuition that a set of attributes γ implies another attribute y precisely when the columns corresponding to γ have nonempty intersection and that intersection is a subset of the column corresponding to y.

Lemma 60. *Let R be a relation on $X \times Y$, with both X and Y nonempty.*
R preserves attribute privacy if and only if the following condition is true:
For all $\gamma \in \Phi_R$ and all $y \in Y$, if $\psi_R(\gamma) \subseteq \psi_R(\{y\})$ then $y \in \gamma$.

Proof. I. Suppose there exist $\gamma \in \Phi_R$ and $y \in Y$ such that $\psi_R(\gamma) \subseteq \psi_R(\{y\})$ but $y \notin \gamma$. Since $\phi_R \circ \psi_R$ is a closure operator, $y \in (\phi_R \circ \psi_R)(\{y\})$ and $\gamma \subseteq (\phi_R \circ \psi_R)(\gamma)$. Now observe that $(\phi_R \circ \psi_R)(\{y\}) \subseteq (\phi_R \circ \psi_R)(\gamma)$ by supposition and because ϕ_R is inclusion-reversing. Consequently, $(\phi_R \circ \psi_R)(\gamma)$ must be a proper superset of γ, telling us there is attribute inference.

II. If there is attribute inference, then for some $\gamma \in \Phi_R$, $\gamma \subsetneq (\phi_R \circ \psi_R)(\gamma)$. Pick some $y \in (\phi_R \circ \psi_R)(\gamma) \setminus \gamma$. Then $y \notin \gamma$ but

$$\psi_R(\gamma) \;=\; \psi_R\big((\phi_R \circ \psi_R)(\gamma)\big) \;\subseteq\; \psi_R\big(\,(\phi_R \circ \psi_R)(\gamma) \setminus \gamma\,\big) \;\subseteq\; \psi_R(\{y\}).$$

(The equality holds by associativity of $\circ$ and the dual version of Corollary 46 on page 91. The two subset relations hold by inclusion-reversal of ψ_R.)

(Technical comment: In both parts above, $\gamma = \emptyset$ is permissible.) □

C.3 Unique Identifiability, Free Faces, and Privacy Preservation

Recall the following definition:

Definition 6 (Unique Identifiability)**.** *Let R be a relation on $X \times Y$ and suppose $x \in X$. We say that x is* uniquely identifiable via relation R *when $\psi_R(Y_x) = \{x\}$.*

Comment: It is entirely possible that one or more proper subsets γ of Y_x already *identifies* x, meaning $\psi_R(\gamma) = \{x\}$. Certainly x is uniquely identifiable in that case. Moreover, the attributes $Y_x \setminus \gamma$ can be inferred from γ.

Lemma 61. *Let R be a relation on $X \times Y$ that preserves attribute privacy. Let $x \in X$. Then no proper subset of Y_x identifies x.*

Proof. Suppose, for some $\gamma \subsetneq Y_x$, $\psi_R(\gamma) = \{x\}$. We obtain a contradiction as follows:

$$\gamma \;\subsetneq\; Y_x \;=\; \phi_R(\{x\}) \;=\; (\phi_R \circ \psi_R)(\gamma) \;=\; \gamma.$$

□

We turn now to proving the assertions of Section 5 regarding free faces.

Lemma 62. *Let R be a relation on $X \times Y$, with both X and Y nonempty. If Φ_R contains no free faces, then R preserves attribute privacy.*

Proof. We will show that $\phi_R \circ \psi_R$ is the identity on Φ_R.

To build intuition, we treat the empty simplex separately. As usual, $(\phi_R \circ \psi_R)(\emptyset) = \phi_R(X)$. Therefore, we will show that $\phi_R(X) = \emptyset$. Observe that every maximal simplex of Φ_R contains $\phi_R(X)$, since any witness for such a simplex must have all the attributes in $\phi_R(X)$. Pick some maximal simplex η of Φ_R and consider $\gamma = \eta \setminus \phi_R(X)$. Let η' be any maximal simplex of Φ_R containing γ. Then

$$\eta \;=\; \gamma \cup \phi_R(X) \;\subseteq\; \eta' \cup \phi_R(X) \;=\; \eta'.$$

So $\eta = \eta'$ by maximality. Since Φ_R has no free faces, γ cannot be a proper subset of η, meaning $\phi_R(X) = \emptyset$, as desired.

Now consider $\emptyset \neq \gamma \in \Phi_R$. Suppose γ is a proper subset of $(\phi_R \circ \psi_R)(\gamma)$. By Corollary 44 and Lemma 48 on pages 91 and 92, respectively, we can assume without loss of generality that $\gamma = \eta \setminus \{y\}$ for some maximal η of Φ_R and some $y \in \eta$. Observe that

$$\eta \setminus \{y\} = \gamma \subsetneq (\phi_R \circ \psi_R)(\gamma) \subseteq (\phi_R \circ \psi_R)(\eta) = \eta,$$

so $\eta = (\phi_R \circ \psi_R)(\gamma)$. Now let η' be any maximal simplex of Φ_R containing γ. Then

$$\eta \;=\; (\phi_R \circ \psi_R)(\gamma) \;\subseteq\; (\phi_R \circ \psi_R)(\eta') \;=\; \eta'.$$

(Note: The last equality in each of the lines of comparisons above follows from Corollary 44 by maximality.)

So $\eta = \eta'$ by maximality. That says γ is a free face of Φ_R, a contradiction. □

The converse of Lemma 62 need not hold if there exists an individual who can hide, with attributes that form a strict subset of some other individual's attributes. However:

Lemma 63. *Let R be a relation on $X \times Y$, with both X and Y nonempty. If R preserves attribute privacy and if every $x \in X$ is uniquely identifiable via R, then Φ_R contains no free faces.*

Proof. Suppose that γ is a free face of Φ_R. We can assume without loss of generality that $\gamma = \eta \setminus \{y\}$ for some maximal $\eta \in \Phi_R$ and $y \in \eta$. Since a Dowker attribute complex is generated by the rows of the underlying relation, it must be that $\eta = Y_x$ for at least one $x \in X$. By Lemma 61, there is at least one x' besides x in $\psi_R(\gamma)$. Then

$$\gamma \;=\; (\phi_R \circ \psi_R)(\gamma) \;\subseteq\; \phi_R(\{x, x'\}) \;=\; Y_x \cap Y_{x'}.$$

Since we have assumed that γ is free and Y_x is maximal, we see that $Y_{x'}$ must be a subset of Y_x. That means x' is not uniquely identifiable, a contradiction.

(Technical comment: $\gamma = \emptyset$ is permissible throughout this argument.) □

The following lemma will help us later in Appendix E, to establish the assertions of Sections 5 and 8 regarding relations that preserve both attribute and association privacy:

Lemma 64. *Let R be a relation on $X \times Y$ such that $|X| = |Y| > 1$. If R has no blank columns and preserves attribute privacy, then every $x \in X$ is uniquely identifiable via R.*

Proof. The proof is by induction on $n = |X| = |Y|$.

I. The base case $n = 2$ implies that R is isomorphic to

R	y_1	y_2
x_1	•	
x_2		•

(Any other type of 2×2 relation without blank columns would allow for attribute inference.)

Each x_i is uniquely identifiable in R above.

II. For the induction step, assume that, for some $n > 2$, the lemma holds for all relations with X and Y spaces of size strictly less than n (and bigger than 1). We need to establish the lemma for all relations with X and Y spaces of size n.

Subclaim: R has no blank rows.

To see this, suppose that $Y_{\tilde{x}} = \emptyset$ for some $\tilde{x} \in X$. Let Q be the restriction of R to $X' \times Y$, with $X' = X \setminus \{\tilde{x}\}$. There is no significant difference between R and Q; in particular, Q also preserves attribute privacy.

(Perhaps the empty simplex is slightly tricky: $(\phi_Q \circ \psi_Q)(\emptyset) = \bigcap_{x \in X'} Y_x$. If this intersection is nonempty, it contains some $y_1 \in Y$. Pick $y_2 \in Y$ with $y_2 \neq y_1$; this is possible since $|Y| > 2$. Note that $\emptyset \neq X_{y_2} \subseteq X'$, so $y_1 \in \bigcap_{x \in X'} Y_x \subseteq \bigcap_{x \in X_{y_2}} Y_x = (\phi_R \circ \psi_R)(\{y_2\}) = \{y_2\}$, a contradiction. So $(\phi_Q \circ \psi_Q)(\emptyset) = \emptyset$.)

Now let Q' be the further restriction of R to $X' \times Y'$, where $Y' = Y \setminus \{\tilde{y}\}$, with $\tilde{y}$ *any* attribute in Y. By Lemma 57 on page 96, Q' preserves attribute privacy. The underlying spaces X' and Y' of Q' each have size $n-1$ and Q' has no blank columns. The induction hypothesis therefore tells us that every individual in X' is uniquely identifiable via Q'. Bearing in mind that $\tilde{x}$ does not appear in any X_y, one sees that for each $x \in X'$, there is some $\gamma \subseteq Y'$ such that $\bigcap_{y \in \gamma} X_y = \{x\}$. That intersection is a column vector all of whose entries are 0 (blank) except for the entry indexed by x. Since R preserves attribute privacy and x is arbitrary in X', Lemma 60 implies that in fact $X_{\tilde{y}} = \emptyset$, contradicting the assumption that R has no blank columns.

Next, pick $\overline{x} \in X$. We will show that $\overline{x}$ is uniquely identifiable via R. Without loss of generality, write R as in Figure 59 (the figure indicates blank entries by "0"s):

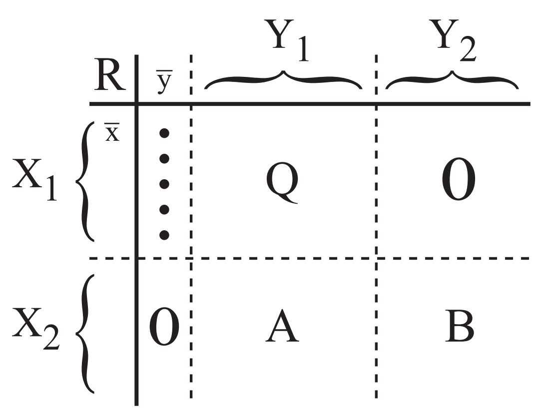


Figure 59: Relation R decomposed into blocks for the proof of Lemma 64, as described below.

Specifically, pick some $\overline{y} \in Y$ such that $(\overline{x}, \overline{y}) \in R$. This is possible since R has no blank rows. Then decompose $X = X_1 \cup X_2$, with $X_1 = X_{\overline{y}}$ and $X_2 = X \setminus X_1$. Since R preserves attribute privacy, $X_2 \neq \emptyset$.

Let Q model $\mathrm{Lk}(\Phi_R, \overline{y})$. So Q is R restricted to $X_1 \times Y_1$, with $Y_1 = \bigcup_{x \in X_1} Y_x \setminus \{\overline{y}\}$. If $Y_1 \neq \emptyset$, then Q preserves attribute privacy, by Lemma 57, and Q has no blank columns.

Now write Y as the disjoint union $Y = \{\overline{y}\} \cup Y_1 \cup Y_2$, with $Y_2 = Y \setminus (Y_1 \cup \{\overline{y}\})$.

Observe that no individual in X_2 has attribute $\overline{y}$. Observe further that every individual in X_1 has attribute $\overline{y}$ but has no attributes in Y_2, by construction.

Let A be the restriction of R to $X_2 \times Y_1$ and let B be the restriction of R to $X_2 \times Y_2$.

If $Y_2 \neq \emptyset$, then B has no blank columns and $\Phi_B = \text{dl}(\Phi_R, Y_1 \cup \{\overline{y}\})$. If $|Y_2| \geq 2$, then the blank rows indexed by X_1 that remain after deleting from R the columns indexed by $Y_1 \cup \{\overline{y}\}$ are irrelevant and so B preserves attribute privacy (by Lemma 57 and by an argument similar to that appearing in the proof of the Subclaim on page 100).

Let us look at some cases:

- $\underline{|Y_2| \geq |X_2| = 1}$: Then any attribute of Y_2 identifies the one individual in X_2. Since R preserves attribute privacy, this implies both that $|Y_2| = 1$ and that relation A is blank. Consequently, every attribute in Y_1 implies $\overline{y}$ in R. Since R preserves attribute privacy, we conclude that $Y_1 = \emptyset$. That means we are actually in the base case, with $n = 2$.

- $\underline{|Y_2| > |X_2| \geq 2}$: By removing some columns of B, we obtain a square relation to which we can apply the induction hypothesis. That means every $x \in X_2$ is uniquely identifiable by the remaining columns. Since B preserves attribute privacy that means the columns removed must have been blank, a contradiction.

- $\underline{|Y_2| = |X_2| \geq 2}$: We can apply the induction hypothesis directly to B. That again tells us that every $x \in X_2$ is uniquely identifiable by Y_2-indexed columns, both in B and in R. We conclude that relation A must be blank and so $Y_1 = \emptyset$, arguing as above. Thus $|X_2| = |Y_2| = n - 1$, implying $|X_1| = 1$. So $\overline{y}$ uniquely identifies $\overline{x}$, as desired.

- $\underline{|Y_2| < |X_2|}$: This means $|Y_1| \geq |X_1|$. Additionally, $|X_1| \geq 2$, as otherwise $\overline{y}$ implies all the attributes Y_1. If actually $|Y_1| > |X_1|$, then we could argue as above to see that some columns of Q are blank, contrary to the construction of Q. So we have that $n > |Y_1| = |X_1| \geq 2$ and the induction hypothesis applies. Consequently, $\overline{x}$ is uniquely identifiable via Q, say as $\{\overline{x}\} = \psi_Q(\gamma)$, for some $\gamma \subseteq Y_1$. If we adjoin $\overline{y}$, we get that $\psi_R(\gamma \cup \{\overline{y}\}) = \psi_Q(\gamma) = \{\overline{x}\}$, as desired. □

Theorem 65 (Too Many Attributes)**.** *Let R be a relation on $X \times Y$ with no blank columns. Suppose $|Y| > |X| \geq 1$. Then R does not preserve attribute privacy.*

Proof. The proof is a corollary to Lemma 64:

If $|Y| > |X| = 1$, then all attributes in Y are inferable from nothing (in the context of R).

Otherwise, suppose R preserves attribute privacy. We have $|Y| > |X| > 1$, so we can delete some columns of R and apply Lemmas 57 and 64 to a resulting square relation. Every individual in X is therefore uniquely identifiable via the columns retained. Consequently, either there is attribute inference in R or the discarded columns were blank, a contradiction. □

Comment: One implication of this result and those in Appendix E is the old detective show mantra "eliminate suspects": Reduce the number of relevant individuals sufficiently, and some attribute inference is assured. This amounts to moving from relation R to a subrelation Q representing $\text{dl}(\Psi_R, \sigma)$, with σ a *set* of "eliminated suspects".

An additional conclusion: The proof of Lemma 64 suggests that perhaps R also preserves association privacy. Indeed, we will see that to be true in Appendix E.4.

D Inference Hardness

Recall the following definition:

Definition 9 (Individual Privacy)**.** *Let R be a relation on $X \times Y$ and suppose $x \in X$.*
We say that R <u>preserves attribute privacy for x</u> *whenever $(\phi_R \circ \psi_R)(\gamma) = \gamma$ for all $\gamma \subseteq Y_x$.*

We have seen the following basic result within the proofs of other lemmas:

Lemma 66. *Let R be a relation on $X \times Y$, with both X and Y nonempty. Let $x \in X$. Then:*

R preserves attribute privacy for x
if and only if
$(\phi_R \circ \psi_R)(\gamma) = \gamma$, for all γ of the form $\gamma = Y_x \setminus \{y\}$, with $y \in Y_x$.

Proof. I. If R preserves attribute privacy for x, then the condition is satisfied by definition.

II. Suppose R does not preserve attribute privacy for x. Then for some $\gamma \subseteq Y_x$, $\gamma \subsetneq (\phi_R \circ \psi_R)(\gamma)$. We know $(\phi_R \circ \psi_R)(Y_x) = Y_x$ by Corollary 47 on page 91, so by Lemma 48 on page 92 we can assume that $\gamma = Y_x \setminus \{y\}$, for some $y \in Y_x$. □

Lemma 66 tells us that it is fairly easy to check whether an individual's attribute privacy is preserved. One merely needs to check whether any one attribute is implied by all the remaining attributes. That may be done quickly since the maps ϕ_R and ψ_R amount to set intersections. Harder is finding a smallest set of attributes that implies another of the individual's attributes.

Influenced by Lemma 60 on page 97, we formulate the following problem:

Definition 67 (Minimal Inference)**.** <u>MinInf</u> *is the following decision problem:*
Given relation R on $X \times Y$, $x \in X$, $y \in Y$, and $k \geq 0$, is there a simplex $\gamma \in \Phi_R$, with $\gamma \subseteq Y_x \setminus \{y\}$, such that $|\gamma| \leq k$ and $\psi_R(\gamma) \subseteq \psi_R(\{y\})$?

Lemma 68. MinInf *is NP-complete*

Proof. (A) Observe that the problem lies in *NP*: Given some γ, one can verify the stated conditions in polynomial time. The verifications amount to set intersection, cardinality, and subset computations, drawn from the columns and one row of R.

(B) We will establish *NP*-hardness by a reduction from Set Cover. Recall: Given a collection of sets $\{S_1, \ldots, S_m\}$, Set Cover asks whether there is some subcollection of size at most k such that the union of the subcollection is the overall union (often called the *universe*).

Given an instance of the Set Cover problem, we define the following relation:

- $X = \{x_0\} \cup \bigcup_{i=1}^{m} S_i$, with x_0 a new element distinct from any elements in the sets S_i.
- $Y = \{0, 1, \ldots, m\}$.
- $R = \big(\{x_0\} \times Y\big) \cup \bigcup_{i=1}^{m} \big\{(x, i) \in X \times Y \mid x \in X \setminus S_i \big\}$.

In words: The 0^{th} column of R is the singleton set $\{x_0\}$ and the i^{th} column of R, for $i = 1, \ldots, m$, is $X \setminus S_i$, i.e., the complement of S_i in the original set cover universe, but now with x_0 added. The row for x_0 has entries for all possible attributes. All other rows have no entry in column 0.

Reduction: Given an instance of SET COVER, we transform it into an instance of MININF using the relation R given above and by letting $x = x_0$ and $y = 0$. The parameter k is the same for both problems. Observe that $Y_x \setminus \{y\} = \{1, \ldots, m\}$.

Observe further that $|X| = |\bigcup_{i=1}^m S_i| + 1 = n + 1$ and $|Y| = m + 1$, with n the number of elements in the set cover universe and m the number of subsets specified for the set cover problem. The reduction can therefore be computed in polynomial time.

To avoid trivialities, we assume that $n > 0$ and $m > 0$.

To complete the proof, we will establish the following:

Claim: The answer to SET COVER is "yes" if and only if the answer to MININF is "yes".

I. A "yes" answer to SET COVER means that there is some set of indices $\gamma \subseteq \{1, \ldots, m\}$, with $|\gamma| \leq k$, such that $\bigcup_{j\in\gamma} S_j = \bigcup_{i=1}^m S_i$. Therefore, since $0 \notin \gamma$,

$$\psi_R(\gamma) = \bigcap_{j\in\gamma} X_j = \bigcap_{j\in\gamma}(X \setminus S_j) = X \setminus (\bigcup_{j\in\gamma} S_j) = X \setminus (\bigcup_{i=1}^{m} S_i) = \{x_0\} = \psi_R(\{0\}) = \psi_R(\{y\}).$$

Consequently, $\emptyset \neq \psi_R(\gamma) \subseteq \psi_R(\{y\})$ with $\gamma \subseteq Y_x \setminus \{y\}$ and $|\gamma| \leq k$, meaning that the answer to MININF is "yes" as well.

II. A "yes" answer to MININF means there is some $\gamma \subseteq \{1, \ldots, m\}$ such that $|\gamma| \leq k$ and $\emptyset \neq \psi_R(\gamma) \subseteq \psi_R(\{y\})$. Observe that $\psi_R(\{y\}) = \psi_R(\{0\}) = \{x_0\}$ and that

$$\psi_R(\gamma) \;=\; \bigcap_{j\in\gamma} X_j \;=\; \bigcap_{j\in\gamma}(X \setminus S_j) \;=\; X \setminus (\bigcup_{j\in\gamma} S_j).$$

The middle equality holds as before because $0 \notin \gamma$.

So we see that $x_0 \in X \setminus (\bigcup_{j\in\gamma} S_j) \subseteq \{x_0\}$, telling us

$$\bigcup_{j\in\gamma} S_j \;=\; X \setminus \{x_0\} \;=\; \bigcup_{i=1}^{m} S_i.$$

That means γ describes a set of indices sought for by SET COVER, with $|\gamma| \leq k$, so the answer to SET COVER is also "yes". □

E Privacy Spheres

The aim of this appendix is to characterize privacy and inference in terms of spheres. Spheres exhibit homogeneity, which is good for privacy, while still admitting a coordinate system for identifiability.

We first prove a theorem characterizing individual attribute privacy, then a generalization that holds for arbitrary elements of a relation's doubly-labeled poset, and finally a characterization of those relations that preserve both attribute and association privacy.

E.1 Individual Attribute Privacy

We first state a lemma as a tool. Recall also Definitions 6 and 9 (see pages 98 and 102).

Lemma 69. *Let R be a relation on $X \times Y$, with both X and Y nonempty. Let $x \in X$ be uniquely identifiable via R. Then:*

$$\Big(\bigcap_{y \in Y_x} X_y\Big) \setminus \{x\} \;=\; \emptyset.$$

Moreover, R preserves attribute privacy for x if and only if

$$\Big(\bigcap_{y \in \gamma} X_y\Big) \setminus \{x\} \;\neq\; \emptyset, \quad \textit{for all } \gamma \subsetneq Y_x.$$

Proof. The first statement follows from the definition of unique identifiability: $\bigcap_{y \in Y_x} X_y = \psi_R(Y_x) = \{x\}$.

For the second statement:

I. Assume that R preserves attribute privacy for x. Let $\gamma \subsetneq Y_x$. If $\big(\bigcap_{y\in\gamma} X_y\big) \setminus \{x\} = \emptyset$, then $\psi_R(\gamma) = \bigcap_{y\in\gamma} X_y = \{x\}$, since $x \in X_y$ whenever $y \in \gamma \subseteq Y_x$ (when $\gamma = \emptyset$, the vacuous intersection is all of X, containing x). That says a proper subset of Y_x identifies x, contradicting the proof of Lemma 61 on page 98.

II. Assume $\big(\bigcap_{y\in\gamma} X_y\big) \setminus \{x\} \neq \emptyset$ for all proper subsets γ of Y_x. If R fails to preserve attribute privacy for x, then by Lemma 66 on page 102 there is some γ of the form $Y_x \setminus \{y\}$, with $y \in Y_x$, such that $\gamma \subsetneq (\phi_R \circ \psi_R)(\gamma) = Y_x$. Applying ψ_R to both sides of that last equality gives $\psi_R(\gamma) = \psi_R(Y_x) = \{x\}$, by unique identifiability. That is a contradiction, since $\psi_R(\gamma) = \bigcap_{y\in\gamma} X_y$. □

We now address our characterization of individual privacy, proving a theorem stated previously:

Theorem 10 (Individual Attribute Privacy)**.** *Let R be a relation on $X \times Y$, with $|X| > 1$. Suppose $x \in X$ is uniquely identifiable via R. Let Q be the relation modeling* $\mathrm{Lk}(\Psi_R, x)$*. Then the following three conditions are equivalent:*

(a) R preserves attribute privacy for x.

(b) $\mathrm{Lk}(\Psi_R, x) \;\simeq\; \mathbb{S}^{k-2}$*, with $k = |Y_x|$.*

(c) $\Phi_Q \;=\; \partial(Y_x)$.

Proof. The hypotheses ensure that $Y_x \neq \emptyset$ (and so also $Y \neq \emptyset$). They also ensure that x is a vertex of Ψ_R, so the link is not void. It could be an empty complex $\{\emptyset\}$, of course.

Observe that Q is the restriction of R to $\overline{X} \times Y_x$, with $\overline{X} = \bigcup_{y \in Y_x} X_y \setminus \{x\}$.

If $\overline{X} = \emptyset$, then, reasoning as in the proof of Lemma 55 on page 94, we see that $\mathrm{Lk}(\Psi_R, x) = \{\emptyset\} = \mathbb{S}^{-1}$. Furthermore, x does not share any of its attributes with any other individuals in X. By convention, $\Phi_Q = \{\emptyset\}$ as well. If $k = |Y_x| = 1$, meaning x has a single attribute, then R preserves attribute privacy for x, since $|X| > 1$. Also, $\mathbb{S}^{k-2} = \mathbb{S}^{-1} = \{\emptyset\} = \partial(Y_x)$. So conditions (a), (b), (c) all hold. If $k = |Y_x| \geq 2$, then any one attribute of Y_x implies all the others, so condition (a) does not hold. Moreover, conditions (b) and (c) also do not hold. In short, the theorem holds when $\overline{X} = \emptyset$.

We now assume that $\overline{X} \neq \emptyset$. We then know that $\mathrm{Lk}(\Psi_R, x) = \Psi_Q \simeq \Phi_Q$ by a dual version of Lemma 55 and by Dowker duality. Definitionally, $\partial(Y_x) \simeq \mathbb{S}^{k-2}$, with $k = |Y_x| > 0$. We therefore see that (c) implies (b). To see that (b) implies (c), observe that the underlying vertex set of Φ_Q is Y_x, so $\Phi_Q \simeq \mathbb{S}^{k-2}$ means $\Phi_Q = \partial(Y_x)$, since no proper subset of a sphere can be homotopic to that same sphere. To prove the theorem we therefore only need to establish that conditions (a) and (c) are equivalent.

Recall the formulas relating ϕ_Q and ϕ_R from page 95 and dualize them here. We see that:

$$\psi_Q(\chi) \;=\; \psi_R(\chi) \setminus \{x\} \;=\; \Big(\bigcap_{y \in \chi} X_y\Big) \setminus \{x\}, \quad \text{for all } \emptyset \neq \chi \subseteq Y_x.$$

I. Assume that R preserves attribute privacy for x. By Lemma 69 and the formula above we see that $\psi_Q(\chi) \neq \emptyset$ for all nonempty proper subsets χ of Y_x and that $\psi_Q(Y_x) = \emptyset$, since x is uniquely identifiable via R. Consequently, Φ_Q contains every nonempty proper subset of Y_x as a simplex, but does not contain Y_x. (Also, Φ_Q contains the empty simplex since the complex is not void.) Thus $\Phi_Q = \partial(Y_x)$.

II. Assume that $\Phi_Q = \partial(Y_x)$. Then $\psi_Q(\chi) \neq \emptyset$ for every nonempty proper subset χ of Y_x. By the formula above, $\big(\bigcap_{y \in \chi} X_y\big) \setminus \{x\} \neq \emptyset$, for each such χ. Now suppose $\chi = \emptyset \subsetneq Y_x$. Then:

$$\emptyset \;\neq\; \overline{X} \;=\; \psi_Q(\emptyset) \;\subseteq\; X \setminus \{x\} \;=\; \Big(\bigcap_{y \in \emptyset} X_y\Big) \setminus \{x\}.$$

So we see that $\big(\bigcap_{y \in \chi} X_y\big) \setminus \{x\} \neq \emptyset$ for every proper subset χ of Y_x, implying that R preserves attribute privacy for x, by Lemma 69. □

Comment: It is impossible to satisfy the following three conditions simultaneously:
(1) x is uniquely identifiable, (2) $|Y_x| = 1$, (3) $\overline{X} \neq \emptyset$.

E.2 Group Attribute Privacy

We now generalize the previous theorem to arbitrary elements (σ, γ) of the doubly-labeled poset P_R associated with a relation R. We stated the generalized theorem previously in the report, as Theorem 11, and replicate that below. One may view this generalized theorem as a characterization of the conditions under which a set σ of individuals (i.e., a *group* of individuals, in the non-mathematical sense) has its attribute privacy preserved, as a whole, not necessarily individually. Theorem 10 is a special case of Theorem 11, with the "group" a single individual x, since $(\{x\}, Y_x) \in P_R$ whenever x is uniquely identifiable via R and $Y_x \neq \emptyset$.

Theorem 11 (Group Attribute Privacy)**.** *Let R be a relation on $X \times Y$.*
Suppose $(\sigma, \gamma) \in P_R$, with $\sigma \neq X$. Let Q be the relation modeling $\mathrm{Lk}(\Psi_R, \sigma)$.
Then the following three conditions are equivalent:

(a) $(\phi_R \circ \psi_R)(\gamma') \;=\; \gamma'$, *for every subset γ' of γ.*

(b) $\mathrm{Lk}(\Psi_R, \sigma) \;\simeq\; \mathbb{S}^{k-2}$, *with $k = |\gamma|$.*

(c) $\Phi_Q \;=\; \partial(\gamma)$.

Proof. Reminder: Since $(\sigma, \gamma) \in P_R$, $\emptyset \neq \sigma \in \Psi_R$, $\emptyset \neq \gamma \in \Phi_R$, $\phi_R(\sigma) = \gamma$, and $\psi_R(\gamma) = \sigma$.

Thus also $(\phi_R \circ \psi_R)(\gamma) = \gamma$, meaning we can focus on proper subsets of γ for part (a).

Recall also that Q is the restriction of R to $\overline{X} \times \gamma$, with $\overline{X} = \bigcup_{y \in \gamma} X_y \setminus \sigma$.

If $\overline{X} = \emptyset$, then $\mathrm{Lk}(\Psi_R, \sigma) = \{\emptyset\} = \mathbb{S}^{-1}$. By convention, $\Phi_Q = \{\emptyset\}$ as well. If $k = |\gamma| = 1$, then $\mathbb{S}^{k-2} = \mathbb{S}^{-1} = \{\emptyset\} = \partial(\gamma)$. The only proper subset of γ in this case is $\gamma' = \emptyset$, and $(\phi_R \circ \psi_R)(\emptyset) = \phi_R(X) = \emptyset$. (Reason: If $y \in \phi_R(X)$, then $y \in \gamma$, so $\gamma = \{y\}$, implying $\sigma = X$, which is disallowed.) Thus conditions (a), (b), (c) all hold. If $k = |\gamma| \geq 2$, then conditions (b) and (c) cannot hold. Also, condition (a) does not hold since $(\phi_R \circ \psi_R)(\{y\}) = \gamma$ for each $y \in \gamma$, bearing in mind that $\overline{X} = \emptyset$ means $X_y = \sigma$ for each $y \in \gamma$. In short, the theorem holds when $\overline{X} = \emptyset$.

We now assume that $\overline{X} \neq \emptyset$. As in the proof of Theorem 10, we see readily that conditions (b) and (c) are equivalent, so we will prove that conditions (a) and (c) are equivalent. And, as in the previous proof, dualizing a formula from page 95 gives this formula:

$$\psi_Q(\chi) \;=\; \psi_R(\chi) \setminus \sigma, \quad \text{for all } \emptyset \neq \chi \subseteq \gamma.$$

I. Assume that $(\phi_R \circ \psi_R)(\gamma') \;=\; \gamma'$, for every subset γ' of γ.
We will establish that Φ_Q contains all proper subsets of γ but not γ, telling us $\Phi_Q = \partial(\gamma)$.

Since Φ_Q is not void, it contains the empty simplex.

Pick some $\emptyset \neq \gamma' \subsetneq \gamma$. Since $(\phi_R \circ \psi_R)(\gamma') \;=\; \gamma'$, $\psi_R(\gamma') \supsetneq \sigma$.

The formula above therefore says $\psi_Q(\gamma') \neq \emptyset$, telling us $\gamma' \in \Phi_Q$.

Similarly, $\psi_Q(\gamma) \;=\; \psi_R(\gamma) \setminus \sigma \;=\; \sigma \setminus \sigma \;=\; \emptyset$, so $\gamma \notin \Phi_Q$.

II. Assume that $\Phi_Q = \partial(\gamma)$.

Recall that $k = |\gamma| > 0$. We look at two cases based on the value of k:

$k = 1$: In this case, $\gamma = \{y\}$, for some $y \in Y$, so $\sigma = X_y$ and $\overline{X} = \emptyset$, which we discussed above.

$k > 1$: Suppose, for the sake of contradiction, that $\gamma' \subsetneq (\phi_R \circ \psi_R)(\gamma')$, for some $\gamma' \subsetneq \gamma$. By Lemma 48 on page 92, we can assume $\gamma' = \gamma \setminus \{y\}$, for some $y \in \gamma$. Consequently, $(\phi_R \circ \psi_R)(\gamma') = \gamma$, which implies $\psi_R(\gamma') = \sigma$. The formula on the previous page then says $\psi_Q(\gamma') = \emptyset$, whereas $\gamma' \in \Phi_Q$ means $\psi_Q(\gamma') \neq \emptyset$, a contradiction. □

The following lemma, previously stated on page 26, relates privacy preservation in a link to privacy preservation in the encompassing relation.

Lemma 12 (Interpreting Local Operators)**.** *Let R be a relation on $X \times Y$.*

Suppose $(\sigma, \gamma) \in P_R$, with $\sigma \neq X$.

Let Q be the relation on $\overline{X} \times \gamma$ that models $\mathrm{Lk}(\Psi_R, \sigma)$ *and suppose $\overline{X} \neq \emptyset$.*

Then, for every $\gamma' \subseteq \gamma$:

(i) If $\gamma' \notin \Phi_Q$, then $\psi_R(\gamma') = \sigma$.

(ii) If $\gamma' \in \Phi_Q$, then $\psi_R(\gamma') \supsetneq \sigma$.

Moreover, in this case:

For $\gamma' = \emptyset$, $(\phi_Q \circ \psi_Q)(\emptyset) \supseteq (\phi_R \circ \psi_R)(\emptyset)$.

If $\gamma' \neq \emptyset$, then $(\phi_Q \circ \psi_Q)(\gamma') = (\phi_R \circ \psi_R)(\gamma')$.

Proof. Observe that for every $\gamma' \subseteq \gamma$, one has $\gamma' \in \Phi_R$ and $\psi_R(\gamma') \supseteq \psi_R(\gamma) = \sigma$.

By formulas from page 95 dualized, $\psi_Q(\gamma') = \psi_R(\gamma') \setminus \sigma$ and $\psi_R(\gamma') = \psi_Q(\gamma') \cup \sigma$, when $\emptyset \neq \gamma' \subseteq \gamma$.

(i) Suppose $\gamma' \notin \Phi_Q$. Then $\gamma' \neq \emptyset$, since $\emptyset \in \Phi_Q$. Also, $\psi_Q(\gamma') = \emptyset$, so by the second formula above, $\psi_R(\gamma') = \sigma$.

(ii) Suppose $\gamma' \in \Phi_Q$. If $\gamma' = \emptyset$, then $\psi_R(\emptyset) = X \supsetneq \sigma$, by hypothesis. If $\gamma' \neq \emptyset$, then $\psi_Q(\gamma') \neq \emptyset$, so by the formulas above, $\psi_R(\gamma') \supsetneq \sigma$.

Turning to the "Moreover":

If $y \in (\phi_R \circ \psi_R)(\emptyset)$, then y is an attribute for all individuals in X, so $y \in \gamma$ and $y \in \phi_Q(\overline{X}) = (\phi_Q \circ \psi_Q)(\emptyset)$.

Let $\emptyset \neq \gamma' \in \Phi_Q$. By another formula on page 95 dualized, if $\kappa \subseteq \overline{X}$, then $\phi_Q(\kappa) = \phi_R(\kappa \cup \sigma)$.

Therefore, using the first formula above: $(\phi_Q \circ \psi_Q)(\gamma') = \phi_Q(\psi_R(\gamma') \setminus \sigma) = (\phi_R \circ \psi_R)(\gamma')$. □

Comment: Also, $(\phi_Q \circ \psi_Q)(\emptyset) = \phi_Q(\overline{X}) = \phi_R\big(\bigcup_{y \in \gamma} X_y\big) = \bigcap_{y \in \gamma} (\phi_R \circ \psi_R)(\{y\})$.

E.3 Preserving Attribute and Association Privacy

In this subsection, we are interested in understanding relations that preserve *both* attribute and association privacy. We will discover that this requirement is severely limiting. As we already see from Theorem 65 on page 101, if R is a nonvoid tight relation on $X \times Y$ that preserves both attribute and association privacy, then $|X| = |Y| = n$. What are the possibilities?

$n = 0$: Not relevant; this is a void relation.

$n = 1$: Not possible; such a relation does not preserve privacy; one can infer the single individual or single attribute "for free" (e.g., merely by knowing someone is covered by the relation).

$n = 2$: As we have seen before, such a relation must be isomorphic to the following relation:

R	y_1	y_2
x_1	•	
x_2		•

Then both Ψ_R and Φ_R are instances of the 0-sphere $\mathbb{S}^0$.

$n \geq 3$: Now there are several possibilities:

- The relation could be isomorphic to a *cyclic staircase relation*:

R	y_1	y_2	$\cdots$	$\cdots$	y_{n-1}	y_n
x_1	•	•				
x_2		•	•			
$\vdots$			$\ddots$	$\ddots$		
$\vdots$				$\ddots$	•	
x_{n-1}					•	•
x_n	•					•

Then both Ψ_R and Φ_R are homotopic to the 1-sphere $\mathbb{S}^1$. Each is simply a linear cycle of edges, with vertices in one complex dualizing to edges in the other.

- The relation could be isomorphic to a *spherical boundary relation*, in which every entry is present except that a diagonal is blank. For example, in the following relation all entries are present except those of the form (x_i, y_{n-i+1}), $i = 1, \ldots, n$:

R	y_1	y_2	$\cdots$	$\cdots$	y_{n-1}	y_n
x_1	•	•	•	$\cdots$	•	
x_2	•	•	$\cdots$	•		•
$\vdots$	$\vdots$	$\vdots$	$\iddots$		•	$\vdots$
$\vdots$	•	•		$\iddots$	$\vdots$	•
x_{n-1}	•		•	$\cdots$	•	•
x_n		•	•	$\cdots$	•	•

Then Ψ_R and Φ_R are each boundary complexes, namely $\Psi_R = \partial(X)$ and $\Phi_R = \partial(Y)$. Thus both are homotopic to the $(n-2)$-sphere $\mathbb{S}^{n-2}$.

- Finally, R could have multiple components, each of which is isomorphic to one of the following: A singleton, a cyclic staircase relation, or a spherical boundary relation, all as above. (Observe that even though a nonblank 1×1 relation in and of itself preserves no privacy, a relation containing a nonblank 1×1 subrelation can preserve privacy when that subrelation is one of several components.)

(Comment: the staircase and spherical relations are isomorphic when $n = 3$.)

The aim of this subsection is to prove that these are the only possibilities.

Lemma 70. *Let R be a connected tight relation on $X \times Y$, with $|X| = |Y| \geq 3$, that preserves both attribute and association privacy.*

Let $x \in X$ and define Q to be the relation on $\overline{X} \times Y_x$ that models $\mathrm{Lk}(\Psi_R, x)$*.*

Then $\Psi_Q = \partial(\overline{X})$ and $\Phi_Q = \partial(Y_x)$, with $|\overline{X}| = |Y_x|$.

Proof. Observe that $Y_x \neq \emptyset$ since R is tight. Recall that $\overline{X} = \bigcup_{y \in Y_x} X_y \setminus \{x\}$, which is nonempty since R is connected and X contains not just x.

By Lemma 64 on page 99, x is uniquely identifiable via R, so Theorem 10 on page 104 says that $\Psi_Q \simeq \mathbb{S}^{k-2}$ and $\Phi_Q = \partial(Y_x)$, with $k = |Y_x|$. If we can show that $|\overline{X}| = k$, then we can conclude that $\Psi_Q = \partial(\overline{X})$. (We also see that $k \geq 2$, since $\overline{X} \neq \emptyset$.)

The vertices of Ψ_Q generate the maximal simplices of Φ_Q. In particular, there exist distinct $x_1, \ldots, x_k \in \overline{X}$ such that $\overline{Y}_1, \ldots, \overline{Y}_k$ are the maximal simplices of Φ_Q, with $\overline{Y}_i = Y_{x_i} \cap Y_x$, and $|\overline{Y}_i| = k - 1$, for $i = 1, \ldots, k$.

Let $\tilde{x} \in \overline{X}$. Then $Y_{\tilde{x}} \cap Y_x \subseteq \overline{Y}_i \subseteq Y_{x_i}$, for some $i \in \{1, \ldots, k\}$.

That says $\emptyset \neq \phi_R(\{\tilde{x}, x\}) \subseteq \phi_R(\{x_i\})$.

Since R preserves association privacy, the dualization of Lemma 60 on page 97 implies $\tilde{x} = x_i$. Thus $|\overline{X}| = k$. □

Comment: Where did we use the assumption that each of X and Y has at least three elements? In fact, for much of the proof it is enough to assume that $|X| = |Y| \geq 2$. However, there is no connected tight relation that preserves privacy when $|X| = |Y| = 2$.

Corollary 71. *Let R be a connected tight relation on $X \times Y$, with $|X| = |Y|$, that preserves both attribute and association privacy.*

Let $y \in Y$ and suppose $|X_y| \geq 4$.

Then $\mathrm{Lk}(\Phi_R, y)$ *is not a linear cycle. (In other words, the relation Q that models* $\mathrm{Lk}(\Phi_R, y)$ *is <u>not</u> isomorphic to a cyclic staircase relation.)*

Proof. Arguing as in the proof of Lemma 70, now in dual form, we see that $\mathrm{Lk}(\Phi_R, y) \simeq \mathbb{S}^{k-2}$, with $k = |X_y|$. Since $k - 2 \geq 2$, $\mathrm{Lk}(\Phi_R, y)$ is not a linear cycle. □

Corollary 72. *Let R be a connected tight relation on $X \times Y$, with $|X| = |Y| \geq 3$, that preserves both attribute and association privacy.*

Suppose $\{x, x'\}$, with $x \neq x'$, is an edge (1-simplex) in Ψ_R.

Then $|Y_x| = |Y_{x'}|$.

Proof. Let $k = |Y_x|$ and $k' = |Y_{x'}|$.

Observe that x' is a vertex of $\mathrm{Lk}(\Psi_R, x)$ and x is a vertex of $\mathrm{Lk}(\Psi_R, x')$.

By the proof of Lemma 70, each of x' and x generates a maximal simplex in the attribute complex associated with the other's link. That simplex is $Y_x \cap Y_{x'}$ in both complexes.

So $k - 1 = |Y_x \cap Y_{x'}| = k' - 1$, hence $k = k'$. □

Corollary 73. *Let R be a connected tight relation on $X \times Y$, with $|X| = |Y| \geq 3$, that preserves both attribute and association privacy.*

Then all rows and columns have the same number of nonblank entries.

Proof. By Lemma 51 on page 92 and Corollary 72 above, all rows have the same number, k_r, of nonblank entries. Dualizing, one sees that all columns have the same number, k_c, of nonblank entries. We claim that $k_c = k_r$. This assertion follows from Lemma 70 and its proof as follows:

Pick some $x \in X$ and let Q be the relation modeling $\mathrm{Lk}(\Psi_R, x)$. By Lemma 70, Ψ_Q and Φ_Q are each boundary complexes, with $k_r = |Y_x|$ vertices. Moreover, each attribute $y \in Y_x$ generates a maximal simplex $X_y \cap \overline{X}$ in Ψ_Q, which must have size $k_r - 1$. The column X_y contains one additional individual, namely x. So $k_c = |X_y| = (k_r - 1) + 1 = k_r$. □

Theorem 74 (Privacy as Sphere)**.** *Let R be a nonvoid connected tight relation on $X \times Y$ that preserves both attribute and association privacy.*

Then $|X| = |Y| \geq 3$ and R is isomorphic to either a cyclic staircase relation or a spherical boundary relation (each described on page 108).

Proof. As we commented previously, Theorem 65 on page 101 implies that $|X| = |Y| = n$, for some $n \geq 2$. Connectedness further means that $n \geq 3$.

By Corollary 73, all rows and columns in R have the same number of nonblank entries. In other words, $|X_y| = |Y_x| = k$, for all $x \in X$ and all $y \in Y$, for some fixed k. By connectedness, $k \geq 2$.

By Lemma 64 on page 99, each $x \in X$ is uniquely identifiable via R. Dualized, each $y \in Y$ is uniquely identifiable via R as well.

If $k = 2$, then Ψ_R and Φ_R contain vertices and edges but no higher-dimensional simplices. By duality, each vertex therefore has at most two incident edges. By unique identifiability, each vertex has exactly two incident edges. Thus, by connectedness, each complex is a linear cycle. So R is isomorphic to a cyclic staircase relation.

Now assume that $k \geq 3$.

Pick a $\overline{y} \in Y$ and consider the decomposition of Figure 60, similar to the one we saw in the proof of Lemma 64.

Let $X_1 = X_{\overline{y}}$ and write $X = X_1 \cup X_2$ with $X_2 = X \setminus X_1$. $X_1 \neq \emptyset$ since every column of R has k nonblank entries and $X_2 \neq \emptyset$ since R preserves attribute privacy.

Let Q model $\mathrm{Lk}(\Phi_R, \overline{y})$. So Q is R restricted to $X_1 \times Y_1$, with $Y_1 = \bigcup_{x \in X_1} Y_x \setminus \{\overline{y}\}$. $Y_1 \neq \emptyset$ because every row of R has k nonblank entries. In particular, there are exactly $k - 1$ entries in each row of Q, so at least two entries in each row.

Now write Y as the disjoint union $Y = \{\overline{y}\} \cup Y_1 \cup Y_2$, with $Y_2 = Y \setminus (Y_1 \cup \{\overline{y}\})$. Observe that every individual in X_1 has attribute $\overline{y}$ but has no attributes in Y_2, by construction.

By the dual to Lemma 70, we know that $\Psi_Q = \partial(X_1)$ and $\Phi_Q = \partial(Y_1)$, with $k = |X_1| = |Y_1|$. Therefore, for each each $y \in Y_1$, column X_y of R has $k - 1$ entries that lie in X_1 and one entry

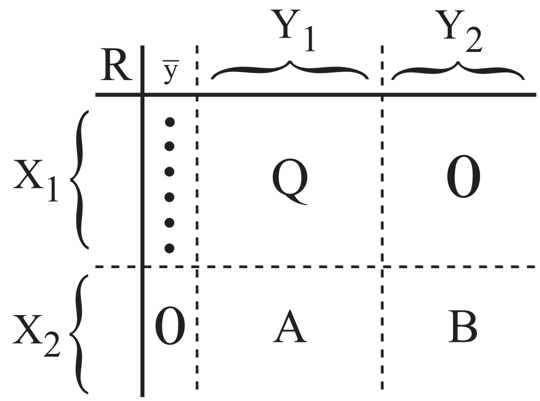


Figure 60: Relation R decomposed into blocks for the proof of Theorem 74.

that lies in X_2. We claim that the X_2 entry is the same across all columns X_y as y varies over Y_1. For otherwise, at least two such columns would have an intersection (nonempty, since $k-2 \geq 1$) contained wholly within $X_{\overline{y}}$, implying that R permits attribute inference after all, by Lemma 60 on page 97. Call that common individual $\overline{x}$. Observe that $Y_{\overline{x}} = Y_1$ since every row of R has exactly k attributes. Consequently, the block diagram for R becomes as in Figure 61. (The figures now indicate blank entries either by blanks or by explicit "0"s.)

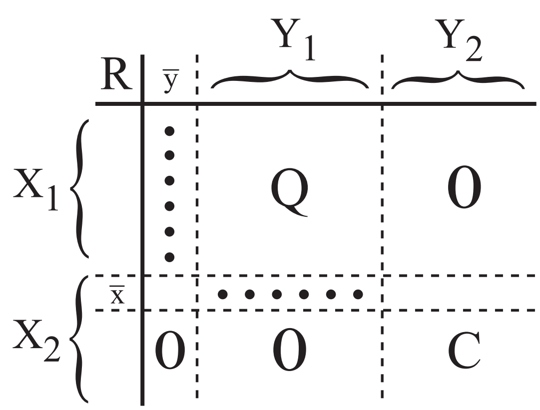


Figure 61: Relation R decomposed further.

Observe that no individual of $X_1 \cup \{\overline{x}\}$ has any attributes in Y_2 and that no individual of $X_2 \setminus \{\overline{x}\}$ has any attributes in $Y_1 \cup \{\overline{y}\}$, by the row and column cardinality constraints. That means relation C, which is the restriction of R to $(X_2 \setminus \{\overline{x}\}) \times Y_2$, would be disconnected from the rest of R, if C were to exist. We conclude that $Y_2 = \emptyset$ and that $X_2 = \{\overline{x}\}$. Thus, finally, R must decompose as in Figure 62. As we have seen, Q is nearly a full relation, missing only a diagonal. We now see that R is also nearly a full relation, missing only a diagonal. Thus $\Psi_R = \partial(X)$ and $\Phi_R = \partial(Y)$, meaning R is isomorphic to a spherical boundary relation, as claimed. □

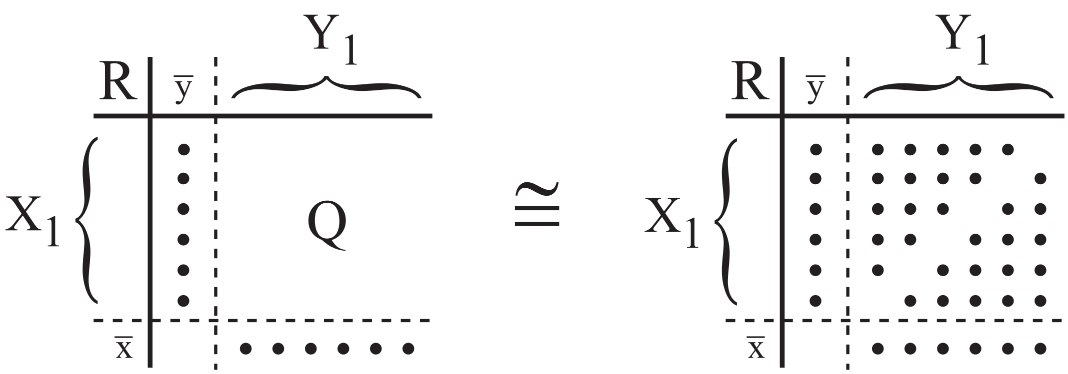


Figure 62: Relation R decomposes diagonally.

Corollary 75. *Let R be a nonvoid tight relation that preserves both attribute and association privacy. Decompose R into its connected components as $R = R_1 \cup \cdots \cup R_\ell$, with each R_i a nonvoid tight relation on $X_i \times Y_i$, as per the proof of Lemma 52 on page 93. Then, for each $i \in \{1, \ldots, \ell\}$, R_i is isomorphic to a singleton or a cyclic staircase relation or a spherical boundary relation, and $|X_i| = |Y_i|$.*

Comment: When $\ell = 2$ and each of R_1 and R_2 is a singleton, then the Dowker complexes of R itself, Ψ_R and Φ_R, are each an instance of $\mathbb{S}^0$.

Proof. Consider R_i, for some $i \in \{1, \ldots, \ell\}$.

Suppose that $X_i \in \Psi_{R_i}$. Then some attribute $y \in Y_i$ is shared by all individuals in X_i. If there were any other attributes in Y_i, then each of those would individually imply y in R. Since R preserves attribute privacy, $|Y_i| = 1$. Consequently, since R also preserves association privacy, $|X_i| = 1$, so R_i is a singleton.

If R_i is not a singleton, then $X_i \notin \Psi_{R_i}$ and similarly $Y_i \notin \Phi_{R_i}$.

Consequently, Lemma 52 and Corollary 54 on page 93 tell us that R_i is a nonvoid connected tight relation that preserves both attribute and association privacy. Theorem 74 completes the proof. □

Comment: The development leading to Corollary 75 used the language of relations, privacy, and inference as proof tools, in part to build intuition. One can take an alternate, more directly simplicial and combinatorial approach. For instance, by counting vertices, maximal simplices, and free faces that are just one vertex shy of being maximal simplices, one can obtain an alternate proof of Theorem 65 on page 101.

E.4 Square Relations Preserve Privacy Symmetrically

At the end of Appendix C.3, we observed that one could perhaps strengthen the conclusions of Lemma 64 on page 99. According to the lemma, if a square relation with no blank columns preserves attribute privacy, then each individual is uniquely identifiable via the relation. The proof of the lemma further established that the relation necessarily has no blank rows. In fact, we will now prove that the relation also preserves association privacy.

Summary: By Theorem 74 and Corollary 75, any nonvoid tight relation preserving both attribute and association privacy must be a square relation whose components are isomorphic to singletons, cyclic staircase relations, or spherical boundary relations. Complementing this statement, by upcoming Theorem 76 and its dual form, any nonvoid tight square relation must either preserve *both* attribute and association privacy or fail to preserve both, that is, allow some attribute inference *and* some association inference.

We now state the theorem, but will need to develop some tools before proving it.

Theorem 76 (Privacy in Square Relations)**.** *Let R be a relation on $X \times Y$ with $|X| = |Y| > 1$. If R has no blank columns and preserves attribute privacy, then these three conditions hold:*

(i) R has no blank rows.

(ii) Every $x \in X$ is uniquely identifiable via R.

(iii) R preserves association privacy.

We now develop the tools:

Definition 77 (Individuals with Maximal Attributes)**.** *Let R be a relation on $X \times Y$. The* restriction of R to its maximally attributed individuals *is the relation $Q_{max} = R|_{\overline{X}\times\overline{Y}}$, with*

$$\begin{aligned} k_{max} &= \max_{x\in X} |Y_x|, \\ \overline{X} &= \left\{ x \in X \;\middle|\; |Y_x| = k_{max} \right\}, \\ \text{and} \quad \overline{Y} &= \bigcup_{x\in\overline{X}} Y_x. \end{aligned}$$

Lemma 78 (Privacy Preservation in Q_{max})**.** *Let R be a relation on $X \times Y$, with $|X| \geq |Y| > 1$. Suppose that R is tight, that R preserves attribute privacy, and that every $x \in X$ is uniquely identifiable via R. Let k_{max} and Q_{max} be as in Definition 77.*

Then $k_{max} \geq 1$, Q_{max} is tight, Q_{max} preserves attribute privacy, and every individual $\overline{x} \in \overline{X}$ is uniquely identifiable via Q_{max}.

Proof. Since neither X nor Y is empty and since R is tight, $k_{max} \geq 1$. Consequently, neither $\overline{X}$ nor $\overline{Y}$ is empty in the definition of Q_{max}, from which it follows that Q_{max} is tight by construction.

Suppose $k_{max} = 1$. Since R has no blank rows, every individual in X has a single attribute in Y. By unique identifiability, distinct individuals have distinct attributes. Consequently, $|X| = |Y|$. So R is isomorphic to a square diagonal relation, and $Q_{max} = R$. The lemma's assertions therefore hold.

Henceforth, assume that $k_{max} > 1$. Let $\overline{x} \in \overline{X} \subseteq X$.

Observe that $\overline{x}$ is uniquely identifiable via Q_{max}, since $\overline{x}$ is uniquely identifiable via R, $Y_{\overline{x}} \subseteq \overline{Y}$, and $\psi_{Q_{max}}(Y_{\overline{x}}) = \psi_R(Y_{\overline{x}}) \cap \overline{X} = \{\overline{x}\} \cap \overline{X} = \{\overline{x}\}$.

By assumption, R preserves attribute privacy, every $x \in X$ is uniquely identifiable via R, and $|X| > 1$. Consequently, Theorem 10 on page 104 says that $\Phi_Q = \partial(Y_{\overline{x}})$, with Q modeling $\mathrm{Lk}(\Psi_R, \overline{x})$. As in the proof of Lemma 70 on page 109, this means there exist distinct vertices $x_1, \ldots, x_{k_{max}}$ in Ψ_Q such that $\overline{Y}_1, \ldots, \overline{Y}_{k_{max}}$ are the maximal simplices of Φ_Q, with $\overline{Y}_i = Y_{x_i} \cap Y_{\overline{x}}$, and $|\overline{Y}_i| = k_{max} - 1$, for $i = 1, \ldots, k_{max}$. (Aside: Here, Ψ_Q could contain additional vertices.)

Since each x_i is uniquely identifiable via R, $Y_{x_i} \not\subseteq Y_{\overline{x}}$. Bearing in mind the definition of k_{max}, this means each Y_{x_i} contains exactly one attribute in $Y \setminus Y_{\overline{x}}$. Consequently, $|Y_{x_i}| = k_{max}$ and each x_i is an individual in $\overline{X}$.

We therefore see that each x_i is a vertex as well of $\Psi_{Q'}$ and that $\Phi_{Q'} = \partial(Y_{\overline{x}})$, with Q' now modeling $\mathrm{Lk}(\Psi_{Q_{max}}, \overline{x})$. We also see that $\overline{X}$ must contain at least $k_{max} + 1$ individuals, so $|\overline{X}| > 1$. Theorem 10 then says that Q_{max} preserves attribute privacy for $\overline{x}$. Since $\overline{x}$ is arbitrary in $\overline{X}$, that means Q_{max} preserves attribute privacy generally. □

Lemma 79 (Square Uniform Relations)**.** *Let R be a relation on $X \times Y$, with $|X| = |Y| > 1$.*
Suppose that R is tight and that R preserves attribute privacy.
Suppose further that every row has exactly k nonblank entries, with $k \geq 1$.
Then every column has exactly k nonblank entries.

Proof. By Lemma 64 on page 99, every $x \in X$ is uniquely identifiable via R.

We first claim that $|X_y| \geq k$ for every $y \in Y$. To see this, let $y \in Y$ be arbitrary. Pick some $x \in X$ such that $(x, y) \in R$. Such an x exists since R has no blank columns. Since R preserves attribute privacy for x, we can argue as in the proof of Lemma 78, concluding that at least $k - 1$ other individuals in X must share attribute y with x. So $|X_y| \geq k$.

Counting the total number of nonblank entries in R in two ways, we obtain:

$$n\,k \;=\; \sum_{x \in X} |Y_x| \;=\; \sum_{y \in Y} |X_y| \;\geq\; \sum_{y \in Y} k \;=\; n\,k, \qquad \text{with } n = |X| = |Y|.$$

Thus $|X_y| = k$ for every $y \in Y$. □

Corollary 80. *Assume the hypotheses of Lemma 79 and that $k = 1$.*
Then R is isomorphic to a square diagonal relation.

Proof. As we saw in the proof of Lemma 79, every $x \in X$ is uniquely identifiable via R. The argument in the proof of Lemma 78, for $k_{max} = 1$, therefore establishes this corollary. □

Corollary 81. *Assume the hypotheses of Lemma 79.*
Suppose further that R is connected and that $k = 2$.
Then R is isomorphic to a cyclic staircase relation.

Proof. As we saw in the proof of Lemma 79, every $x \in X$ is uniquely identifiable via R.

Observe as well that, in a dual sense, each $y \in Y$ is uniquely identifiable via R, since R has no blank columns and preserves attribute privacy: $\phi_R(X_y) = (\phi_R \circ \psi_R)(\{y\}) = \{y\}$.

By Lemma 79, all rows and columns of R have exactly two nonblank entries.

Consequently, the argument in the proof of Theorem 74 on page 110, for $k = 2$, establishes this corollary. □

Corollary 82. *Assume the hypotheses of Lemma 79.*
Suppose further that R is connected and that $k \geq 3$. Let $n = |X| = |Y|$.
Then $k = n - 1$ and R is isomorphic to a spherical boundary relation.

Proof. As we saw in the proof of Lemma 79, every $x \in X$ is uniquely identifiable via R.

Pick some such x and let Q model $\mathrm{Lk}(\Psi_R, x)$. We can again argue as we did in the proof of Lemma 78 (and elsewhere), that there exist distinct individuals (vertices) $x_1, \ldots, x_k$ in Ψ_Q such that $\overline{Y}_1, \ldots, \overline{Y}_k$ are the maximal simplices of Φ_Q, with $\overline{Y}_i = Y_{x_i} \cap Y_x$, and $|\overline{Y}_i| = k - 1$, for $i = 1, \ldots, k$. (This time there are exactly k individuals in Ψ_Q, since every column of R has exactly k nonblank entries, by Lemma 79.)

Since every row of R has exactly k nonblank entries, each x_i has one additional attribute in $Y \setminus Y_x$. We claim that this additional attribute is the same y for all x_i. Given that claim and the general row and column cardinality constraints, R must be isomorphic to the decomposition shown in Figure 63. (This figure and the next indicate blank entries either by blanks or by explicit "0"s.)

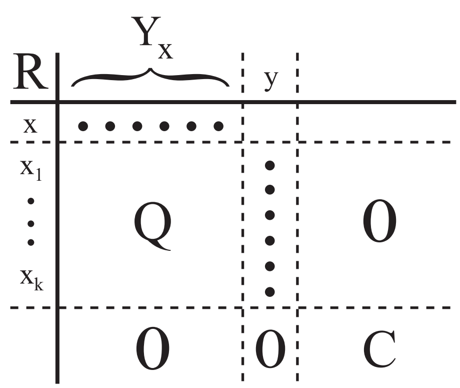


Figure 63: Relation R decomposed as for the proof of Corollary 82.

Since R is connected, C cannot exist and the corollary follows, somewhat as in the proof of Theorem 74 as shown in Figure 62 on page 112.

In order to establish the claim, suppose that $Y_x = \{y_1, \ldots, y_k\}$. Suppose further that for each $i = 1, \ldots, k$, individual x_i has all the attributes of Y_x except for y_i. Now let y be x_2's attribute outside Y_x. We can assume without loss of generality that x_1 does not have this attribute, and then derive a contradiction, as follows:

Let $X_{\text{imply}} = X_{y_3} \cap X_y$. The intersection is well-defined since $k \geq 3$. Moreover, $x_2 \in X_{\text{imply}}$, since x_2 has attributes y_3 and y. However, $x_1 \notin X_{\text{imply}}$, since x_1 does not have attribute y.

Observe that $X_{y_1} = \{x, x_2, x_3, \ldots, x_k\}$ and $X_{y_3} = \{x, x_1, x_2, \ldots, x_k\} \setminus \{x_3\}$.

Thus $\emptyset \neq X_{\text{imply}} = X_{y_3} \cap X_y \subseteq X_{y_1}$. In other words, attributes y_3 and y imply attribute y_1, contradicting the assumption that R preserves attribute privacy. □

We turn now to the proof of Theorem 76, which we had stated previously on page 113:

Proof. Part (i) follows from the Subclaim on page 100 and part (ii) follows from Lemma 64 on page 99. We therefore focus on proving part (iii), assuming parts (i) and (ii) hold:

The proof is by induction on $n = |X| = |Y|$.

I. The base case $n = 2$ means R is isomorphic to a standard two element diagonal relation as on page 99, which preserves association privacy.

II. For the induction step, assume that, for some $n > 2$, part (iii) of the theorem holds for all relations with X and Y spaces of size strictly less than n (and bigger than 1). We need to establish part (iii) for all relations with X and Y spaces of size n.

As we observed in the proof of Lemma 78, if $k_{max} = 1$ in Definition 77, then every individual has exactly one attribute and R is isomorphic to a square diagonal relation, hence preserves association privacy. We therefore assume that $k_{max} > 1$ for the rest of the proof.

Let Q_{max} be as in Definition 77 and consider the decomposition of R as in Figure 64.

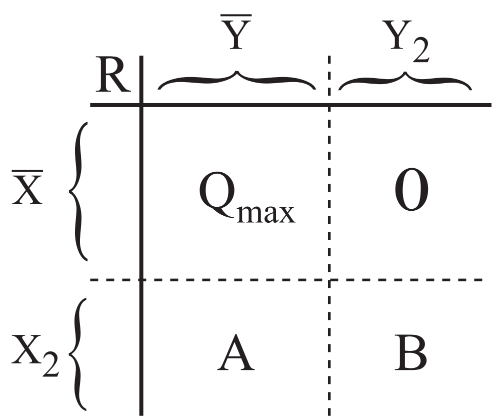


Figure 64: Relation R decomposed into blocks by Q_{max}.

Here $X_2 = X \setminus \overline{X}$ and $Y_2 = Y \setminus \overline{Y}$, with $\overline{X}$ and $\overline{Y}$ as in Definition 77. Then A is the restriction of R to $X_2 \times \overline{Y}$ and B is the restriction of R to $X_2 \times Y_2$.

Given parts (i) and (ii), Lemma 78 tells us that Q_{max} is tight and preserves attribute privacy. By Theorem 65 on page 101, we see that $|\overline{X}| \geq |\overline{Y}|$ and therefore that $|X_2| \leq |Y_2|$, since $|X| = |Y|$.

Let us look at some cases:

- $|Y_2| = |X_2| = 1$: Then B is a singleton, so $A = \emptyset$, since R preserves attribute privacy. The induction hypothesis applies to Q_{max}, telling us that Q_{max} preserves association privacy. Since R is the disjoint union of Q_{max} and B, both nonvoid, we see that R must also preserve association privacy.

- $|Y_2| = |X_2| > 1$: Arguing as on page 101, we see that B preserves attribute privacy. Lemmas 64 and 60, on pages 99 and 97, respectively, then imply that $A = \emptyset$. The induction hypothesis applies to each of Q_{max} and B, since neither is now a singleton (since $k_{max} > 1$). Again, R is a disjoint union of these two relations, so we see that R preserves association privacy. (One can formalize that argument by using the dual version of Lemma 60 on page 97.)

- $|Y_2| > |X_2| \geq 1$: This case cannot occur, since B would preserve attribute privacy but have more attributes than individuals (see again also Theorem 65 on page 101).

- $|X_2| = 0$: Then $Y_2 = \emptyset$ and $Q_{max} = R$.

  If R has more than one connected component, then one can apply the induction hypothesis to each component separately. (In order to apply the induction hypothesis, one should first make a small argument that each component is a tight square relation, contains more than one entry, and preserves attribute privacy. This is straightforward.) One concludes that each component preserves association privacy and therefore that R preserves association privacy.

  Otherwise, R is square, tight, connected, and preserves attribute privacy. Furthermore, every row of R has exactly k_{max} entries. By assumption, $k_{max} > 1$ in this part of the proof. So Corollaries 81 and 82 tell us that R is isomorphic to either a cyclic staircase relation or a spherical boundary relation. Thus R preserves association privacy. □

F Poset Chains

Recall Definition 13, on page 35, of the Galois lattice P_R^+ associated with a relation R, and Definition 14, on page 38, defining informative attribute release sequences. In this appendix we will explore connections between these two concepts.

F.1 Maximal Chains and Informative Attribute Release Sequences

Let R be a relation on $X \times Y$, with both X and Y nonempty:

Suppose $\{(\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1) < (\sigma_0, \gamma_0)\}$, with $k \geq 1$, is a maximal chain in P_R^+.
Then, for $1 \leq i \leq k$, $\sigma_i \subsetneq \sigma_{i-1}$ and $\gamma_i \supsetneq \gamma_{i-1}$.
Also, $\sigma_0 = X$ and $\gamma_k = Y$, so $\gamma_0 = \phi_R(X)$ and $\sigma_k = \psi_R(Y)$.
Consequently, $\gamma_0 \neq \emptyset$ if and only if $X \in \Psi_R$, and $\sigma_k \neq \emptyset$ if and only if $Y \in \Phi_R$.

We sometimes speak of a *maximal chain at and above* (σ, γ), by which we mean a chain $\{(\sigma, \gamma) < \cdots < (\sigma_1, \gamma_1) < (\sigma_0, \gamma_0)\}$ in P_R^+ that is maximal among all chains in P_R^+ containing (σ, γ) as least element. Such a chain is a prefix of a full maximal chain in P_R^+ ("prefix" with respect to our subscript ordering, which starts at the top of a poset and moves downward).

Recall the following lemma, previously stated on page 41 in Section 10.6:

Lemma 21 (Informative Attributes from Maximal Chains)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $\{(\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1) < (\sigma_0, \gamma_0)\}$, with $k \geq 1$, is a maximal chain in P_R^+.*
Define $y_1, \ldots, y_k$ by selecting some $y_i \in \gamma_i \setminus \gamma_{i-1}$, for each $i = 1, \ldots, k$.
Then $y_1, \ldots, y_k$ is an informative attribute release sequence for R.
Moreover, $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_i\}) = \gamma_i$, for each $i = 0, 1, \ldots, k$.

Proof. Establishing the "Moreover" also establishes the "iars" assertion.
The proof is by induction on i.
For the base case, $i = 0$ and we need to show that $(\phi_R \circ \psi_R)(\emptyset) = \gamma_0$.
Calculating, $(\phi_R \circ \psi_R)(\emptyset) = \phi_R(X) = \gamma_0$, by our earlier comments about maximal chains.

For the induction step, we assume that, for some $1 \leq i \leq k$, the assertion holds for indices smaller than i and we need to show the assertion holds for i. First, observe:

$$\psi_R(\{y_1, \ldots, y_i\}) \;=\; \psi_R(\{y_1, \ldots, y_{i-1}\}) \,\cap\, X_{y_i} \;=\; \psi_R(\gamma_{i-1}) \,\cap\, X_{y_i} \;=\; \psi_R(\gamma_{i-1} \cup \{y_i\}).$$

(The middle equality follows from the induction hypothesis and a dual version of Corollary 46 from page 91, specifically because $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_{i-1}\}) = \gamma_{i-1}$ and $\psi_R \circ \phi_R \circ \psi_R = \psi_R$.)

Since $\gamma_{i-1} \subsetneq \gamma_{i-1} \cup \{y_i\} \subseteq \gamma_i$,

$$\gamma_{i-1} \;=\; (\phi_R \circ \psi_R)(\gamma_{i-1}) \;\subsetneq\; (\phi_R \circ \psi_R)(\gamma_{i-1} \cup \{y_i\}) \;\subseteq\; (\phi_R \circ \psi_R)(\gamma_i) \;=\; \gamma_i.$$

By maximality of the original chain and the nature of elements in P_R^+, we see that $(\phi_R \circ \psi_R)(\gamma_{i-1} \cup \{y_i\}) = \gamma_i$, so $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_i\}) = (\phi_R \circ \psi_R)(\gamma_{i-1} \cup \{y_i\}) = \gamma_i$. ☐

Here is a partial converse (also previously stated in Section 10.6):

Lemma 22 (Chains from Informative Attributes)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty. Suppose $y_1, \ldots, y_k$ is an informative attribute release sequence for R, with $k \geq 1$.*

Let $\gamma_i = (\phi_R \circ \psi_R)(\{y_1, \ldots, y_i\})$ and $\sigma_i = \psi_R(\gamma_i)$, for $i = 1, \ldots, k$.

Let $\gamma_0 = \phi_R(X)$. Then $\{(\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1) < (X, \gamma_0)\}$ is a chain in P_R^+.

Comment: The resulting chain need not be maximal.

Proof. Observe that each $(\sigma_i, \gamma_i) \in P_R^+$ by construction, so we need to establish the total ordering. Letting $\sigma_0 = X$, we need to show that $\sigma_i \subsetneq \sigma_{i-1}$, for each $i = 1, \ldots, k$.

Since $\{y_1, \ldots, y_i\} \supseteq \{y_1, \ldots, y_{i-1}\}$, we see that $\sigma_i \subseteq \sigma_{i-1}$. If $\sigma_i = \sigma_{i-1}$, then also $\gamma_i = \gamma_{i-1}$, contradicting the fact that $y_i \in \gamma_i \setminus \gamma_{i-1}$ (which is true by the nature of informative attribute release sequences). □

As a corollary to Lemmas 21 and 22, one sees that every informative attribute release sequence (iars) for R is a subsequence of an iars derived from a maximal chain in P_R^+. (Technically, one needs to show that any nonempty subsequence of an iars is itself an iars. And one needs to show that extending any chain obtained via Lemma 22 to a maximal chain retains the original iars as a subsequence of one subsequently obtainable via Lemma 21. All that is straightforward.)

F.2 Chains and Links

We are interested in understanding how chains and informative attribute release sequences behave as one passes to links. (Small caution: whereas we were looking at chains in P_R^+ before, we focus here on P_R (and P_Q).)

Lemma 83 (Chains in Links)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty, and suppose $(\sigma, \gamma) \in P_R$. Let Q be the relation modeling* $\mathrm{Lk}(\Psi_R, \sigma)$*. Then*

$$P_Q \;=\; \bigl\{(\sigma' \setminus \sigma,\, \gamma') \;\big|\; (\sigma, \gamma) < (\sigma', \gamma') \in P_R \bigr\}.$$

Comments:

- Q is the restriction of R to $\overline{X} \times \gamma$, with $\overline{X} = \bigcup_{y \in \gamma} X_y \setminus \sigma$, as per Definition 8 on page 24.
- P_Q could be empty. This occurs precisely when (σ, γ) is a maximal element of P_R, which occurs precisely when $\mathrm{Lk}(\Psi_R, \sigma) = \{\emptyset\}$, which occurs precisely when $\overline{X} = \emptyset$.
- If $\sigma = X$, then $\mathrm{Lk}(\Psi_R, \sigma) = \{\emptyset\}$ and so $P_Q = \emptyset$, given Definition 8 on page 24. (For future reference, observe that $P_{Q(\sigma,\gamma)}$ is undefined when $\sigma = X$ and $Q(\sigma, \gamma)$ is given by Definition 19 on page 40. See also page 89 for comments about the doubly-labeled poset.)
- P_Q never contains the element $\hat{0}_Q$ of P_Q^+. Indeed, $\hat{0}_Q = (\emptyset, \gamma)$, corresponding to (σ, γ) in P_R. That value is consistent with the idea of Lemma 12 on page 26 that one has "localized to σ upon observing γ". (See also Definition 16 on page 39.)

- P_Q could contain the element $\hat{1}_Q = (\overline{X}, \chi)$ of P_Q^+, for some $\chi \subsetneq \gamma$. That happens precisely when $\overline{X} \neq \emptyset$ and all individuals in $\overline{X}$ share an attribute of γ, in which case $\chi \neq \emptyset$.

Proof. The proof relies on dual versions of the formulas appearing in the top half of page 95.

I. Suppose $(\kappa, \eta) \in P_Q$. So $\kappa \neq \emptyset$ and $\eta \neq \emptyset$. Also, $\Psi_Q = \mathrm{Lk}(\Psi_R, \sigma)$, so $\kappa \cap \sigma = \emptyset$ and $\kappa \cup \sigma \in \Psi_R$. Let $\sigma' = \kappa \cup \sigma$. So $\sigma \subsetneq \sigma'$. We can take γ' to be η since $\eta = \phi_Q(\kappa) = \phi_R(\sigma')$. Note that $\psi_R(\gamma') = \psi_Q(\eta) \cup \sigma = \kappa \cup \sigma = \sigma'$. We have shown that $(\sigma', \gamma') \in P_R$ and $(\sigma, \gamma) < (\sigma', \gamma')$.

II. Suppose $(\sigma', \gamma') \in P_R$ and $(\sigma, \gamma) < (\sigma', \gamma')$. So $\sigma \subsetneq \sigma'$ and $\gamma \supsetneq \gamma'$. Let $\kappa = \sigma' \setminus \sigma$. Note that $\kappa \neq \emptyset$ and $\gamma' \neq \emptyset$. Moreover, $\kappa \in \mathrm{Lk}(\Psi_R, \sigma)$, so $\overline{X} \neq \emptyset$.

Verifying correspondence: $\phi_Q(\kappa) = \phi_R(\sigma') = \gamma'$ and $\psi_Q(\gamma') = \psi_R(\gamma') \setminus \sigma = \sigma' \setminus \sigma = \kappa$.

We have shown that $(\sigma' \setminus \sigma,\, \gamma') \in P_Q$. □

Corollary 84 (Order Preservation)**.** *Let R and Q be as in Lemma 83, with $(\sigma, \gamma) \in P_R$. Then $(\sigma, \gamma) < (\sigma_1, \gamma_1) < (\sigma_2, \gamma_2)$ in P_R if and only if $(\sigma_1 \setminus \sigma,\, \gamma_1) < (\sigma_2 \setminus \sigma,\, \gamma_2)$ in P_Q.*

Proof. By Lemma 83 and because:

(a) $\sigma \subsetneq \sigma_1 \subsetneq \sigma_2$ implies $\emptyset \neq \sigma_1 \setminus \sigma \subsetneq \sigma_2 \setminus \sigma$;

(b) $\emptyset \neq \kappa_1 \subsetneq \kappa_2$ and $\kappa_2 \cap \sigma = \emptyset$ implies $\sigma \subsetneq (\kappa_1 \cup \sigma) \subsetneq (\kappa_2 \cup \sigma)$. □

Corollary 85 (Maximal Chain Preservation)**.** *Let R and Q be as in Lemma 83, with $(\sigma, \gamma) \in P_R$. Then $\{(\sigma, \gamma) < (\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1)\}$ is a maximal chain at and above (σ, γ) in P_R if and only if $\{(\sigma_k \setminus \sigma,\, \gamma_k) < \cdots < (\sigma_1 \setminus \sigma,\, \gamma_1)\}$ is a maximal chain in P_Q.*

Proof. By Lemma 83 and Corollary 84, we know that $\{(\sigma, \gamma) < (\sigma_k, \gamma_k) < \cdots < (\sigma_1, \gamma_1)\}$ is a chain extending upward from (σ, γ) in P_R if and only if $\{(\sigma_k \setminus \sigma,\, \gamma_k) < \cdots < (\sigma_1 \setminus \sigma,\, \gamma_1)\}$ is a chain in P_Q.

Maximality follows for the same reason: Refine or extend a chain in one poset and one can refine or extend the corresponding chain in the other poset as well. □

Comment about "length": Recall that the length of a chain in a poset is one less than the number of elements in the chain. We also speak of the *length* of an informative attribute release sequence $y_1, \ldots, y_k$, which is k, the actual number of attributes in the sequence.

In the context of Lemmas 21 and 22, there is a happy alignment of definitions: The length k of a longest iars for R is the length $\ell(P_R^+)$.

In thinking about poset lengths, bear in mind that $\ell(P_R^+)$ may be any of $\ell(P_R)$, $\ell(P_R)+1$, or $\ell(P_R)+2$, depending on whether the top and/or bottom elements of P_R^+ already lie in P_R.

Corollary 86 (Longest Localization Sequences)**.** *Let R be a relation on $X \times Y$, with both X and Y nonempty, and suppose $(\sigma, \gamma) \in P_R$. Let Q be the relation modeling* $\mathrm{Lk}(\Psi_R, \sigma)$.

If $X \notin \Psi_R$, then the length of a longest informative attribute release sequence for localizing to σ in R is $\ell(P_Q) + 2$. If $X \in \Psi_R$ and $\sigma \neq X$, then that length is $\ell(P_Q) + 1$.

(Note: If $\sigma = X \in \Psi_R$, then the length is 0; one can localize to X in R without observation.)

Comment: If P_Q does not contain the top element $\hat{1}_Q$ of P_Q^+, then $\ell(P_Q) + 2 = \ell(P_Q^+)$, since P_Q never contains the bottom element $\hat{0}_Q$. This occurs precisely when no attribute is shared by all the individuals in the link. Also, if $\sigma \subsetneq X \in \Psi_R$, then $\ell(P_Q) + 1 = \ell(P_Q^+)$.

Proof. Let us address one special case first, namely when $\mathrm{Lk}(\Psi_R, \sigma)$ is an empty complex. We only care about the situation in which σ is not all of X, which implies $X \notin \Psi_R$. Observe that P_Q is empty, so $\ell(P_Q) = -1$ and $\ell(P_Q) + 2 = 1$. Observe further that any $y \in \gamma$ identifies σ, as otherwise $\overline{X}$ in the definition of Q would not be empty. So the Corollary holds in this case.

Suppose $\mathrm{Lk}(\Psi_R, \sigma)$ is not an empty complex and that $X \notin \Psi_R$. Lemmas 21 and 22 imply that a longest informative attribute release sequence for localizing to σ comes from a longest maximal chain in P_R^+ at and above (σ, γ). Thus, by Corollary 85, this sequence arises from a maximal chain in P_Q, augmented by considering also $\hat{0}_Q$ and $\hat{1}_R$. The length of the chain in P_Q is two shorter than that in P_R^+. Why? Because $(\sigma, \gamma) \in P_R^+$ becomes $\hat{0}_Q \in P_Q^+$, which is not present in P_Q, and because the top element $\hat{1}_R = (X, \emptyset) \in P_R^+$ disappears altogether ($\hat{1}_Q$ may or may not be in P_Q). So $\ell(P_Q) + 2$ gives the correct length of the iars for R.

Suppose $\mathrm{Lk}(\Psi_R, \sigma)$ is not an empty complex but that $\sigma \subsetneq X \in \Psi_R$. The argument proceeds as before except that now the top element of P_R^+ looks like $\hat{1}_R = (X, \gamma_0)$, with $\gamma_0 \neq \emptyset$. It appears in P_R. Consequently, $\hat{1}_Q = (X \setminus \sigma, \gamma_0)$ and so $\hat{1}_Q$ also appears in P_Q. So a maximal chain in P_Q is now only one shorter than a corresponding maximal chain in P_R^+ at and above (σ, γ), meaning $\ell(P_Q) + 1$ gives the correct length of a longest iars. □

F.3 Isotropy

We turn now to the proof of our isotropy sphere theorem, with the theorem replicated here from earlier in the report. Recall also Definitions 14, 15, 16, and 19 from pages 38–40.

Theorem 20 (Isotropy = Minimal Identification = Sphere)**.** *Let R be a relation and suppose $\emptyset \neq \gamma \in \Phi_R$. Let $\sigma = \psi_R(\gamma)$. Then the following four conditions are equivalent:*

(a) γ is isotropic.

(b) γ is minimally identifying (for σ).

(c) $\Psi_{Q(\sigma,\gamma)} \simeq \mathbb{S}^{k-2}$, with $k = |\gamma|$.

(d) $\Phi_{Q(\sigma,\gamma)} = \partial(\gamma)$.

Proof. Observe that $\sigma \in \Psi_R$ and $\gamma \subseteq (\phi_R \circ \psi_R)(\gamma) = \phi_R(\sigma)$, so constructing $Q(\sigma, \gamma)$ is valid. Also, (a)–(d) each imply $\sigma \neq X$. Finally, observe that $\gamma \notin \Phi_{Q(\sigma,\gamma)}$. For if there were some $x \in \overline{X}$ such that $(x, y) \in Q(\sigma, \gamma) \subseteq R$ for every $y \in \gamma$, then $x \in \sigma$, but σ is disjoint from $\overline{X}$.

If $|\gamma| = 1$, then $\mathbb{S}^{k-2} = \mathbb{S}^{-1} = \{\emptyset\} = \partial(\gamma)$. Write $\gamma = \{y\}$. Then γ is isotropic if and only if y constitutes an informative attribute release sequence, if and only if $y \notin \phi_R(X)$. If $y \in \phi_R(X)$, then $\sigma = X$, so our conventions say $\Psi_{Q(\sigma,\gamma)} = \emptyset \not\simeq \{\emptyset\}$ and $\Phi_{Q(\sigma,\gamma)} = \emptyset \neq \{\emptyset\}$. Moreover, $\psi_R(\emptyset) = \sigma$, so γ is not minimally identifying. If $y \notin \phi_R(X)$, then $\sigma = X_y \subsetneq X$ and $\overline{X} = \emptyset$, so both $\Psi_{Q(\sigma,\gamma)}$ are $\Phi_{Q(\sigma,\gamma)}$ are instances of $\{\emptyset\}$, by our conventions. Moreover, $X = \psi_R(\emptyset) \supsetneq \sigma$. So we see that (a), (b), (c), (d) are all equivalent when $|\gamma| = 1$.

Henceforth assume that $|\gamma| > 1$. It will be convenient to write $\gamma = \{y_1, \ldots, y_k\}$, with $k > 1$, and with the attribute indexing chosen arbitrarily.

As we have observed elsewhere, (c) and (d) are equivalent by Dowker duality and the fact that only a boundary complex can produce $\mathbb{S}^{k-2}$ homotopy type when the underlying vertex set has size k.

We will first show that (a) implies (d) and (b):

Suppose that γ is isotropic.

We wish to show that all proper subsets of γ are simplices in $\Phi_{Q(\sigma,\gamma)}$. Without loss of generality, consider $\{y_1, \ldots, y_{k-1}\}$. If we can show that $\psi_R(\{y_1, \ldots, y_{k-1}\}) \setminus \sigma \neq \emptyset$, then that provides an $x \in \overline{X}$ such that $(x, y_i) \in R$ for $i = 1, \ldots, k-1$, thereby establishing that $\{y_1, \ldots, y_{k-1}\} \in \Phi_{Q(\sigma,\gamma)}$. It also establishes that $\psi_R(\{y_1, \ldots, y_{k-1}\}) \supsetneq \sigma$. Since the "missing attribute" y_k is arbitrary in γ, we see that $\Phi_{Q(\sigma,\gamma)} = \partial(\gamma)$ and that γ is minimally identifying.

Suppose otherwise: $\psi_R(\{y_1, \ldots, y_{k-1}\}) = \sigma = \psi_R(\gamma)$, so also $(\phi_R \circ \psi_R)(\{y_1, \ldots, y_{k-1}\}) = (\phi_R \circ \psi_R)(\gamma) \supseteq \gamma$. That says $y_k \in (\phi_R \circ \psi_R)(\{y_1, \ldots, y_{k-1}\})$, violating the assumption that any ordering of γ is an informative attribute release sequence.

We will now show that (d) implies (a):

Suppose that $\Phi_{Q(\sigma,\gamma)} = \partial(\gamma)$.

If some ordering of γ is not an informative attribute release sequence, then we can rearrange the sequence further to establish that the last attribute is implied by all the others, i.e., that $y_k \in (\phi_R \circ \psi_R)(\{y_1, \ldots, y_{k-1}\})$. Arguing as we did in the proof of Lemma 21 on page 118, we obtain:

$$\begin{aligned}
\psi_R\big(\{y_1, \ldots, y_{k-1}\}\big) &= (\psi_R \circ \phi_R)\Big(\psi_R\big(\{y_1, \ldots, y_{k-1}\}\big)\Big) \\
&= \psi_R\Big((\phi_R \circ \psi_R)\big(\{y_1, \ldots, y_{k-1}\}\big)\Big) \\
&= \psi_R\Big(\{y_k\} \cup (\phi_R \circ \psi_R)\big(\{y_1, \ldots, y_{k-1}\}\big)\Big) \\
&= X_{y_k} \cap \psi_R\Big((\phi_R \circ \psi_R)\big(\{y_1, \ldots, y_{k-1}\}\big)\Big) \\
&= X_{y_k} \cap \psi_R\big(\{y_1, \ldots, y_{k-1}\}\big) \\
&= \psi_R\big(\{y_1, \ldots, y_k\}\big) \\
&= \psi_R(\gamma) \\
&= \sigma.
\end{aligned}$$

On the other hand, since $\{y_1, \ldots, y_{k-1}\} \in \Phi_{Q(\sigma,\gamma)}$, there is a witness $x \in \overline{X}$, meaning $x \in \psi_R(\{y_1, \ldots, y_{k-1}\})$, which contradicts $\overline{X} \cap \sigma = \emptyset$.

Finally, we will show that (b) implies (d):

Suppose that γ is minimally identifying.

Observe that $\psi_R(\{y_1, \ldots, y_{k-1}\}) \supsetneq \sigma$. As above, this establishes $\{y_1, \ldots, y_{k-1}\} \in \Phi_{Q(\sigma,\gamma)}$, from which we conclude that $\Phi_{Q(\sigma,\gamma)} = \partial(\gamma)$, since the missing attribute y_k was arbitrary. $\square$

G Many Long Chains

This appendix provides a proof of Theorem 26 from page 44.

First, we need some tools:

Recall what it means for a poset to be almost a join-based lattice from Definition 25 on page 44.

Definition 87 (Join Completion). *Suppose P is almost a join-based lattice. Let S be a subset of P. The* bounded join-completion of S in P *is the set $S^\vee$ defined by:*

$$S^\vee = \{p \in P \mid p \leq s, \text{ some } s \in S, \text{ and } p = s_1 \vee \cdots \vee s_m, \text{ with each } s_i \in S, \text{ and } m \geq 1\}.$$

Here and in the rest of this appendix, "$\leq$" and "$<$" refer to the partial order on P, while "$\vee$" denotes the resulting join operation on $P \cup \{\hat{1}\}$. S and $S^\vee$ inherit this partial order.

We also define $S_{\max}$ to consist of all the maximal elements of S relative to the partial order inherited from P.

The following facts will be useful. Assume $S \subseteq P$, with P almost a join-based lattice. Then:

1. $S^\vee$ is almost a join-based lattice. The join operation for elements $p, q \in S^\vee$ is given by:

$$p \vee_{S^\vee} q = \begin{cases} p \vee q, & \text{if } p \vee q \leq s, \text{ for some } s \in S; \\ \hat{1}, & \text{otherwise.} \end{cases}$$

2. $S \subseteq S^\vee$ and $S_{\max} = (S^\vee)_{\max}$.

3. $(S^\vee)^\vee = S^\vee$.

4. If $T \subseteq S$, then $T^\vee \subseteq S^\vee$.

5. If $T \subseteq S^\vee$ such that $S_{\max} \setminus T \neq \emptyset$, then $T^\vee \subsetneq S^\vee$.

6. Let $\emptyset \neq T \subseteq S$. Then the poset

$$S_T = \left\{p \in S^\vee \mid p \leq t, \text{ for all } t \in T\right\}$$

is almost a join-based lattice. The join operation for elements $p, q \in S_T$ is given by:

$$p \vee_{S_T} q = \begin{cases} p \vee q, & \text{if } p \vee q \leq t, \text{ for } \textit{all } t \in T; \\ \hat{1}, & \text{otherwise.} \end{cases}$$

7. Fact 6 holds as well for the poset $S'_T = \{p \in S^\vee \mid p < t, \text{ for all } t \in T\}$, now using "$<$" in place of "$\leq$" throughout.

Lemma 88 (Contractibility of Closed Semi-Intervals)**.** *Suppose $\emptyset \neq T \subseteq S \subseteq P$, with P almost a join-based lattice. Define the poset S_T as in Fact 6 on page 123.*

If $S_T \neq \emptyset$, then S_T is contractible.

Proof. Suppose p and q are arbitrary elements of S_T. Every element of T is an upper bound for both p and q. Since T is not empty, this means $p \vee q$ exists in P and $p \vee q \leq t$ for all $t \in T$. Since $t \in S$, we have that $p \vee q \in S^{\vee}$ and thus $p \vee q \in S_T$ as well. Consequently, the lattice $S_T \cup \{\hat{0}, \hat{1}\}$ is noncomplemented, implying that S_T is contractible, by a fact on page 88. □

Intuitively: $\Delta(S_T)$ is a cone with apex $\bigwedge T$, the meet in $S^{\vee}$ of all the upper bounds T.

Caution: The lemma need *not* hold for S'_T as defined in Fact 7 on page 123.

We now specialize a topological tool to our current setting. We refer to the lemma as "cycle tightening" because we will apply the lemma with $p \in S_{\max}$ and with z a reduced homology generator for $\Delta(P)$. The lemma will allow us to move that generator downward in P.

Lemma 89 (Cycle Tightening)**.** *Let P be almost a join-based lattice. Suppose z is a nontrivial reduced k-cycle for $\Delta(P)$, i.e., $0 \neq z \in C_k(\Delta(P); \mathbb{Z})$ and $\widetilde{\partial} z = 0$, for some $k \geq 0$.*

Define $S = \|z\|$ and $K = \{\tau \in \Delta(P) \mid \tau \subseteq S^{\vee}\}$.

Let $p \in S$.

If $\widetilde{H}_{k-1}(\mathrm{Lk}(K,p); \mathbb{Z}) = 0$, then there exists $\eta \in C_{k+1}(\overline{\mathrm{St}}(K,p); \mathbb{Z})$ such that $p \notin \|z + \widetilde{\partial}\eta\|$, now viewing $\eta \in C_{k+1}(\Delta(P); \mathbb{Z})$.

Proof. Let $W = \overline{\mathrm{St}}(K,p)$ and $A = \mathrm{Lk}(K,p)$. Note that A is not an empty complex (that observation follows from the reduced homology assumption when $k = 0$ and the fact that p is part of a simplex containing at least one other element when $k > 0$).

The long exact sequence for a pair [16, 14] therefore gives us the following exact sequence:

$$0 = \widetilde{H}_k(W; \mathbb{Z}) \longrightarrow \widetilde{H}_k(W, A;\, \mathbb{Z}) \longrightarrow \widetilde{H}_{k-1}(A; \mathbb{Z}) = 0.$$

The left 0 comes from W being a cone and the right 0 comes from the lemma's hypotheses. Consequently, $\widetilde{H}_k(W, A;\, \mathbb{Z}) = 0$.

Suppose $z = \sum_i n_i \tau_i$, for some collection $\{\tau_i\}$ of (oriented) k-simplices such that $n_i \neq 0$ for each i. Let z_S consist of the part of z that lies within W, so:

$$z_S = \sum_{\tau_i \in W} n_i \tau_i \qquad \text{(with each } n_i \text{ and } \tau_i \text{ as in } z).$$

Since z is a reduced k-cycle with support in verts(K), z_S is a reduced relative k-cycle for the pair (W, A). Since $\widetilde{H}_k(W, A;\, \mathbb{Z}) = 0$, z_S must be a reduced relative boundary, so there exists $\kappa \in C_{k+1}(W; \mathbb{Z})$ such that $z_S = \widetilde{\partial}\kappa + \gamma$, with $\gamma \in C_k(A; \mathbb{Z})$.

Now let $\eta = -\kappa$ and view $\eta \in C_{k+1}(\Delta(P); \mathbb{Z})$.

Observe that $\|z_S + \widetilde{\partial}\eta\| \subseteq \mathrm{verts}(A) \subseteq \mathrm{verts}(\mathrm{dl}(K,p))$. Consequently, $p \notin \|z + \widetilde{\partial}\eta\|$. □

Lemma 90 (Maximal Element Cardinality)**.** *Let P be almost a join-based lattice. Suppose P has reduced integral homology in dimension $k \geq 0$, that is, $\widetilde{H}_k(\Delta(P); \mathbb{Z}) \neq 0$.*

Let $S = \|z\|$, with $z \in C_k(\Delta(P); \mathbb{Z})$ a reduced homology generator for $\widetilde{H}_k(\Delta(P); \mathbb{Z})$.

Then $|S_{\max}| \geq k + 2$.

Proof. Since $S \subseteq S^\vee$, we can view $z \in C_k(\Delta(S^\vee); \mathbb{Z})$. If there were to exist $\eta \in C_{k+1}(\Delta(S^\vee); \mathbb{Z})$ such that $\widetilde{\partial}\eta = z$, then z would also be a reduced boundary in $\Delta(P)$. So, $\widetilde{H}_k(\Delta(S^\vee); \mathbb{Z}) \neq 0$ and z is a reduced homology generator for $\Delta(S^\vee)$.

Recall the notation S_T in Fact 6 on page 123. Observe that

$$\bigcup_{t \in S_{\max}} \Delta(S_{\{t\}}) \;=\; \Delta(S^\vee).$$

To see this, first observe that the empty simplex $\emptyset$ appears in both these sets. Then:

I. Suppose $\emptyset \neq \sigma \in \Delta(S_{\{t\}})$ for some $t \in S_{\max}$. Being a chain in $S_{\{t\}}$, we can write σ as $\{p_0 < p_1 < \cdots < p_\ell\}$, for some $\ell \geq 0$, with each $p_i \in S^\vee$ (and $p_\ell \leq t \in S_{\max} \subseteq S \subseteq S^\vee$).

   Consequently, $\sigma \in \Delta(S^\vee)$ as well.

II. Suppose $\emptyset \neq \sigma \in \Delta(S^\vee)$. Then $\sigma = \{p_0 < p_1 < \cdots < p_\ell\}$, for some $\ell \geq 0$, with each $p_i \in S^\vee$. By definition of $S^\vee$ and $S_{\max}$, $p_\ell \leq s \leq t$, for some $s \in S$ and $t \in S_{\max}$.

   Consequently, $\sigma \in \Delta(S_{\{t\}})$ as well, for that t.

Similarly, one sees that, for any $\emptyset \neq T \subseteq S$,

$$\bigcap_{t \in T} \Delta(S_{\{t\}}) \;=\; \Delta(S_T).$$

The complex on the right is either an empty complex or it is contractible, by Lemma 88.

A variation of the Nerve Lemma now implies that $\Delta(S^\vee)$ and the nerve of the simplicial complexes $\{\Delta(S_{\{t\}})\}_{t \in S_{\max}}$ have the same homotopy type (see Theorem 10.6(i) in [1]).

Since $\Delta(S^\vee)$ has reduced homology in dimension k, so does the nerve of $\{\Delta(S_{\{t\}})\}_{t \in S_{\max}}$.

The nerve of $\{\Delta(S_{\{t\}})\}_{t \in S_{\max}}$ is isomorphic to a simplicial complex with underlying vertex set $S_{\max}$. In order for a simplicial complex to have reduced homology in dimension k, with $k \geq 0$, the complex must have at least $k + 2$ vertices. Thus $|S_{\max}| \geq k + 2$. □

We now turn to the proof of the main theorem, the statement of which is replicated here:

Theorem 26 (Many Maximal Chains)**.** *Let P be almost a join-based lattice. Suppose P has reduced integral homology in dimension $k \geq 0$, that is, $\widetilde{H}_k(\Delta(P); \mathbb{Z}) \neq 0$.*

Then there are at least $(k + 2)!$ maximal chains in P of length at least k.

Proof. The proof is by induction on k.

I. For the base case, $k = 0$, observe that $\Delta(P)$ must have at least two vertices that are incomparable in P, as otherwise $\Delta(P)$ would be either empty or contractible. Each vertex sits inside a maximal chain of P. The chains are different since the vertices are incomparable.

II. For the induction step, assume that, for some $k \geq 1$, the theorem holds for all relevant P with reduced homology in dimension $k-1$. We need to establish the theorem for all relevant P with reduced homology in dimension k.

Let $z = \sum_i n_i \tau_i$ be a reduced homology generator for $\widetilde{H}_k(\Delta(P);\mathbb{Z})$, with $n_i \neq 0$ for each i.

Define S and K by $S = \|z\|$ and $K = \{\tau \in \Delta(P) \mid \tau \subseteq S^{\vee}\}$. Interpretation: S is the support of the reduced homology generator z and K is the subcomplex of $\Delta(P)$ formed by restricting to the bounded join-completion of z's support.

We now have an inner induction, which we will describe as an iterative loop:
(Notation: superscript (j) indicates the j^{th} iteration.)

1. Initialize with $z^{(0)} = z$, $S^{(0)} = S$, and $K^{(0)} = K$.

2. Suppose $z^{(j)}$, $S^{(j)}$, and $K^{(j)}$ have been defined, with $z^{(j)}$ a reduced homology generator for $\widetilde{H}_k(\Delta(P);\mathbb{Z})$, and with $S^{(j)}$ and $K^{(j)}$ similar in meaning to S and K, now based on $z^{(j)}$. In particular, $z^{(j)}$ has support $S^{(j)}$ and all of $K^{(j)}$'s vertices lie in $(S^{(j)})^{\vee}$.

   Pick some $p \in (S^{(j)})_{\max}$ such that $\widetilde{H}_{k-1}(\text{Lk}(K^{(j)}, p);\mathbb{Z}) = 0$.

   If no such p exists, then the loop ends.

3. Otherwise, invoke Lemma 89 to find an $\eta \in C_{k+1}(\overline{\text{St}}(K^{(j)}, p);\mathbb{Z})$ such that $p \notin \|z^{(j)} + \widetilde{\partial}\eta\|$. Let $z^{(j+1)} = z^{(j)} + \widetilde{\partial}\eta$, so $z^{(j+1)}$ is again a generator of reduced homology in dimension k. Further, let $S^{(j+1)} = \|z^{(j+1)}\|$ and $K^{(j+1)} = \Big\{\tau \in \Delta(P) \;\Big|\; \tau \subseteq (S^{(j+1)})^{\vee}\Big\}$.

Observe that $S^{(j+1)} \subseteq \|z^{(j)}\| \cup \|\widetilde{\partial}\eta\| \subseteq (S^{(j)})^{\vee}$.
On the other hand, $p \in (S^{(j)})_{\max} \setminus S^{(j+1)}$. So by Fact 5 on page 123, $(S^{(j+1)})^{\vee} \subsetneq (S^{(j)})^{\vee}$.
In other words, the possible vertex set for the simplicial complex shrinks with each iteration, and so the loop must eventually end, P being finite.

Given this iterative algorithm, we can now assume without loss of generality that $\widetilde{H}_{k-1}(\text{Lk}(K, p);\mathbb{Z}) \neq 0$ for each p that is a maximal element in the support S of the given reduced homology generator z.
Observe that $\text{Lk}(K, p) = \{\tau \in \Delta(P) \mid \tau \subseteq S^{\vee} \text{ and } s < p \text{ for every } s \in \tau\}$, when $p \in S_{\max}$.

Consequently, $\text{Lk}(K, p) = \Delta(Q_p)$, where Q_p is the subposet of P given by

$$Q_p \;=\; \big\{s \in S^{\vee} \;\mid s < p\big\}.$$

By Fact 7 on page 123, Q_p is itself almost a join-based lattice.

Q_p has reduced integral homology in dimension $k-1$, so by the induction hypothesis, there are at least $(k+1)!$ maximal chains in Q_p of length at least $k-1$. As the description of Q_p makes clear, we can extend each of these chains in P by adding p as a top element, then further refine and/or extend each chain as needed into a maximal chain in P. Distinct chains remain distinct after this augmentation since the process only adds elements of P that lie outside Q_p.

Consequently, we obtain, for each $p \in S_{\max}$, at least $(k+1)!$ distinct maximal chains in P of length at least k, each touching p. A maximal chain in P cannot contain more than one element of $S_{\max}$, since such elements are necessarily incomparable. Letting p vary over $S_{\max}$ therefore produces at least $|S_{\max}| \cdot (k+1)!$ distinct maximal chains in P of length at least k.

By Lemma 90, $|S_{\max}| \geq k+2$. So P contains at least $(k+2)!$ distinct maximal chains of length at least k. □

Here are two corollaries, previously stated on page 45 in Section 10.8:

Corollary 27 (Holes Reduce Inference)**.** *Let R be a nonvoid relation. Suppose P_R has reduced integral homology in dimension $k \geq 0$. Then there are at least $(k+2)!$ maximal chains in P_R of length at least k.*

Proof. The assertion follows from Theorem 26, since P_R is almost a join-based lattice.

(The join operation is exactly that of P_R^+. In particular, the top element $\hat{1}_R$ of P_R^+ is not already in P_R, since P_R has homology, so we may adjoin that as the upper bound $\hat{1}$ for P_R.) □

Recall informative attribute release sequences from Section 10.4 and Appendix F.

Corollary 28 (Holes Defer Recognition)**.** *Let R be a nonvoid relation and let $(\sigma, \gamma) \in P_R$.*

Define $Q = Q(\sigma, \gamma)$ as per Definition 19 and recall Definition 17, from pages 39–40.

Suppose P_Q is well-defined and has reduced integral homology in dimension $k \geq 0$.

Then there are at least $(k+2)!$ distinct informative attribute release sequences $y_1, \ldots, y_\ell$ for R, each with $\ell \geq k+2$, such that $\psi_R(\{y_1, \ldots, y_\ell\}) = \sigma$. Consequently, $\mathrm{r_{slow}}(\sigma) \geq k+2$.

Proof. By Corollary 27, P_Q contains at least $(k+2)!$ maximal chains of length at least k.

The rest of the argument is much like that in the proof of Corollary 86 from page 120:

- Each maximal chain in P_Q gives rise to a maximal chain in P_R^+ at or above (σ, γ).
- Distinctness in P_Q carries over to P_R^+.
- In moving from P_Q to P_R^+ one adds two elements:
  1. One adds (σ, γ), corresponding to $\hat{0}_Q$ in P_Q^+.
  2. P_Q has reduced homology, so no attribute is shared by all individuals, either in Q or in R. One thus also adds the top element $\hat{1}_R$ of P_R^+.

Summarizing the length argument: Each distinct maximal chain in P_Q of length at least k gives rise to a distinct maximal chain at or above (σ, γ) in P_R^+ of length at least $k+2$, and therefore a distinct informative attribute release sequence for R of length at least $k+2$. By construction of Q and by Lemma 21 on page 118, each such iars identifies σ in R. So by Definition 17, $\mathrm{r_{slow}}(\sigma) \geq k+2$.

(How do we know that distinct maximal chains produce distinct iars? Because if two iars are the same, the chains must be the same, by the "Moreover" of Lemma 21. It is true that one may be able to obtain different iars from the same maximal chain, but our counting is over maximal chains, so provides a lower bound for the number of distinct iars.) □

Comment: Since P_Q is well-defined and has reduced homology in nonnegative dimension, Q's Dowker complexes are neither void nor empty. Thus $\sigma \neq X$. Along with the assumption $(\sigma, \gamma) \in P_R$, that means relation $Q(\sigma, \gamma)$ models the link $\mathrm{Lk}(\Psi_R, \sigma)$.

H Obfuscating Strategies

Recall the discussion and terminology of Section 13.

The primary goal of this appendix is to provide a proof of Theorem 32 appearing on page 65. In addition, this appendix provides proof of some of the assertions in the bullets on pages 65–66.

Once again, we first need to develop some tools:

H.1 Source Complex

Subsection 13.1 introduced the strategy complex Δ_G of a graph $G = (V, \mathfrak{A})$. Recall that every action $a \in \mathfrak{A}$ has a unique source state in V. Given a set of actions $\mathcal{A} \subseteq \mathfrak{A}$, we define the *start region* of $\mathcal{A}$, denoted by $\mathrm{src}(\mathcal{A})$, as

$$\mathrm{src}(\mathcal{A}) \;=\; \{v \in V \mid v \text{ is the source state of some action } a \in \mathcal{A}\}.$$

One obtains another simplicial complex from G via src, now on underlying vertex set V:

$$\overline{\Delta}_G \;= \{\mathrm{src}(\sigma) \mid \sigma \in \Delta_G\}.$$

We refer to this complex as G's *source complex.*

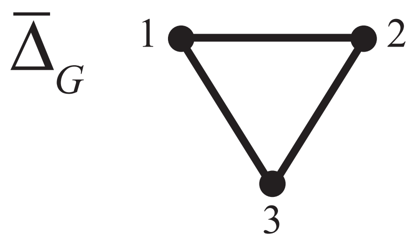


Figure 65: Source complex for the graph of Figure 44 on page 62.

The map $\mathrm{src} : \mathfrak{F}(\Delta_G) \to \mathfrak{F}(\overline{\Delta}_G)$ is a homotopy equivalence, so $\Delta_G \simeq \overline{\Delta}_G$ [6, 7]. Consequently, the source complex of a fully controllable graph G is equal to the boundary complex of the full simplex on the graph's state space V, that is, $\overline{\Delta}_G = \partial(V)$. For the graph of Figure 44 on page 62, the source complex is the boundary of a triangle, as shown in Figure 65.

B	1	2	3	Goal
σ_1		•	•	1
σ_2	•		•	2
σ_3	•	•		3
σ_4	•	•		3

Figure 66: Relation B describes the source complex $\overline{\Delta}_G$ of the graph of Figure 44. Each row describes the start region of a maximal simplex of Δ_G, with Δ_G as in Figure 45 on page 63. The rightmost column again shows each maximal strategy's goal. (See also Figure 46 on page 64.)

In Lemma 30 on page 65 we saw that $\Delta_G = \Phi_A$ for the action relation A defined there. We now see that $\overline{\Delta}_G = \Phi_B$, with relation B as defined in the next lemma. As example, Figure 66 shows relation B for the graph of Figure 44 from page 62.

Lemma 91. *Let $G = (V, \mathfrak{A})$ be a graph as discussed in Section 13. Let $\mathfrak{M}$ be the set of maximal simplices of Δ_G. Define relation B on $\mathfrak{M} \times V$ by $B = \{(\sigma, v) \mid v \in \mathrm{src}(\sigma) \text{ and } \sigma \in \mathfrak{M}\}$.*

Then $\Phi_B = \overline{\Delta}_G$.

(Again, the proof is nearly definitional, so we omit it.)

(The "B" stands for "Beginning" — while "S" for "source" might be desirable, we have already used S to mean "support" elsewhere.)

How should we interpret the remaining Dowker complexes, Ψ_A and Ψ_B, for relations A and B? To answer this question, let us look at the semantics of simplices in these complexes. Suppose Δ_G is not void or empty. A nonempty simplex in Ψ_A represents a *collection* of maximal simplices of Δ_G, namely maximal simplices that have at least one action in common. A nonempty simplex in Ψ_B again represents a collection of maximal simplices of Δ_G, now with at least one source state in common. Thus $\Psi_A \subseteq \Psi_B$. Moreover, Dowker duality gives:

Lemma 92. *Let $G = (V, \mathfrak{A})$ be a graph as discussed in Section 13, with $V \neq \emptyset$.*

Then the inclusion $\iota : \mathfrak{F}(\Psi_A) \to \mathfrak{F}(\Psi_B)$ is a homotopy equivalence.

Comment: The assumption $V \neq \emptyset$ means Δ_G and $\overline{\Delta}_G$ are not void, so relation B is not void. If $V \neq \emptyset$ but $\mathfrak{A} = \emptyset$, then technically relation A is void, but is is convenient to think of it as an instance of the empty relation instead, with associated empty Dowker complexes.

Proof. Consider the following diagram:

$$\begin{array}{ccc}
\mathfrak{F}(\Psi_A) & \xhookrightarrow{\ \iota\ } & \mathfrak{F}(\Psi_B) \\
{\scriptstyle \psi_A}\big\uparrow & & {\scriptstyle \psi_B}\big\uparrow \\
\mathfrak{F}(\Phi_A) & & \mathfrak{F}(\Phi_B) \\
\| & & \| \\
\mathfrak{F}(\Delta_G) & \xrightarrow{\ \mathrm{src}\ } & \mathfrak{F}(\overline{\Delta}_G).
\end{array}$$

Recall that ψ_A, ψ_B, and src are homotopy equivalences.

Let $\mathfrak{M}$ denote the maximal simplices of Δ_G. Observe the following, for each $\sigma \in \mathfrak{F}(\Delta_G)$:

$$(\iota \circ \psi_A)(\sigma) = \{\sigma' \in \mathfrak{M} \mid \sigma \subseteq \sigma'\}.$$

$$(\psi_B \circ \mathrm{src})(\sigma) = \{\sigma' \in \mathfrak{M} \mid \mathrm{src}(\sigma) \subseteq \mathrm{src}(\sigma')\}.$$

If $\sigma \subseteq \sigma'$, then $\mathrm{src}(\sigma) \subseteq \mathrm{src}(\sigma')$.

Consequently, $(\iota \circ \psi_A)(\sigma) \leq (\psi_B \circ \mathrm{src})(\sigma)$ for every $\sigma \in \mathfrak{F}(\Delta_G)$, where "$\leq$" refers to the partial order on $\mathfrak{F}(\Psi_B)$.

We conclude that the two order-reversing poset maps $\iota \circ \psi_A$ and $\psi_B \circ \mathrm{src}$ are homotopic (see [1], Theorem 10.11) and therefore that ι is a homotopy equivalence. □

Lemma 93. *Let $G = (V, \mathfrak{A})$ be a graph as discussed in Section 13, with $V \neq \emptyset$. Then* src *induces a homotopy equivalence of posets $P_A \to P_B$ with explicit formula*

$$(\tau, \sigma) \mapsto \big((\psi_B \circ \mathrm{src})(\sigma), (\phi_B \circ \psi_B \circ \mathrm{src})(\sigma)\big).$$

Proof. Let cl_A denote the image of the closure operator $\phi_A \circ \psi_A : \mathfrak{F}(\Phi_A) \to \mathfrak{F}(\Phi_A)$ and let cl_B denote the image of the closure operator $\phi_B \circ \psi_B : \mathfrak{F}(\Phi_B) \to \mathfrak{F}(\Phi_B)$. We then have the following diagram of homotopy equivalences:

$$P_A \xrightarrow{\pi_2} \mathrm{cl}_A \xhookrightarrow{\iota} \mathfrak{F}(\Phi_A) = \mathfrak{F}(\Delta_G) \xrightarrow{\mathrm{src}} \mathfrak{F}(\overline{\Delta}_G) = \mathfrak{F}(\Phi_B) \xrightarrow{\phi_B \circ \psi_B} \mathrm{cl}_B \xhookrightarrow{\iota} P_B.$$

(Here π_2 is projection onto the second coordinate, i.e., $\pi_2(\tau, \sigma) = \sigma$, and each of the occurrences of ι is an inclusion.)

The composition of all these maps is an order-preserving poset map with the specified formula. The overall map is a homotopy equivalence because each of its constituent maps is a homotopy equivalence. □

Corollary 94. *If G is fully controllable in Lemma 93, then the formula for the poset map becomes $(\tau, \sigma) \mapsto \big((\psi_B \circ \mathrm{src})(\sigma), \mathrm{src}(\sigma)\big)$.*

Proof. Since G is fully controllable, $\Phi_B = \overline{\Delta}_G = \partial(V) \simeq \mathbb{S}^{n-2}$, with $n = |V|$. So Φ_B has no free faces, implying that $\phi_B \circ \psi_B$ is the identity, by Lemma 62 on page 98. □

Two Observations: Suppose that G is a fully controllable graph $(V, \mathfrak{A})$, with both V and $\mathfrak{A}$ nonempty. (i) No action can appear in all maximal simplices of Δ_G, as that would mean Δ_G would be a cone, so not homotopic to a sphere. Consequently, $\hat{1}_A = (\mathfrak{M}, \gamma)$ has $\gamma = \emptyset$ (recall that $\mathfrak{M}$ is the collection of all maximal simplices of Δ_G). (ii) Even if all actions of $\mathfrak{A}$ appear individually as vertices of Δ_G, $\hat{0}_A = (\tau, \mathfrak{A})$ has $\tau = \emptyset$, since $\mathrm{src}(\mathfrak{A}) = V$ and $V \notin \partial(V)$.

These observations mean that P_A does *not* contain either the top element $\hat{1}_A$ or the bottom element $\hat{0}_A$ of P_A^+, when G is fully controllable.

H.2 Delaying Strategy Identification

We now turn to the proof of the main theorem, the statement of which is replicated here:

Theorem 32 (Delaying Strategy Identification)**.** *Let $G = (V, \mathfrak{A})$ be a fully controllable graph, with $n = |V| > 1$. Let A be the relation constructed as in Lemma 30 on page 65 and let P_A be its associated doubly-labeled poset. Then:*

For each $v \in V$, there exists a maximal strategy $\sigma_v \in \Delta_G$ for attaining singleton goal state v such that P_A contains at least $(n-1)!$ distinct maximal chains for identifying σ_v, with each chain consisting of at least $n-1$ elements.

Proof. Let P_A^{op} be P_A but with the opposite partial order. Then P_A^{op} is almost a join-based lattice, with join operation for elements of P_A^{op} given by

$$(\tau_1, \sigma_1) \vee (\tau_2, \sigma_2) = \begin{cases} \big(\tau_1 \cap \tau_2,\ (\phi_A \circ \psi_A)(\sigma_1 \cup \sigma_2)\big), & \text{when } \tau_1 \cap \tau_2 \neq \emptyset; \\ \hat{1}, & \text{otherwise.} \end{cases}$$

The maximal elements of $P_A^{\rm op}$ are of the form $(\{\sigma\}, \sigma)$, with σ varying over the maximal simplices of Δ_G. Each minimal element of $P_A^{\rm op}$ is of the form $\big(\psi_A(\{\mathtt{a}\}), (\phi_A \circ \psi_A)(\{\mathtt{a}\})\big)$, with action $\mathtt{a}$ some vertex of Δ_G. (Aside: not every element of that form is necessarily minimal.)

Since G is fully controllable, $\Delta(P_A^{\rm op}) \simeq \mathbb{S}^{n-2}$. So $\Delta(P_A^{\rm op})$ has reduced homology in dimension $k = n-2 \geq 0$. By the proof of Theorem 26, on page 126, there exists a reduced homology generator z for $\Delta(P_A^{\rm op})$, with support $S = \|z\|$, such that $P_A^{\rm op}$ contains, for each $p \in S_{\rm max}$, a collection of maximal chains passing through p with the following property: Even if one merely considers the portions of the chains at and below p, the collection contains at least $(n-1)!$ distinct such subchains and each subchain has length at least $n-2$. Each full chain, being maximal, must be a path in $P_A^{\rm op}$ between some maximal element $(\{\sigma\}, \sigma)$ and some minimal element $\big(\psi_A(\{\mathtt{a}\}), (\phi_A \circ \psi_A)(\{\mathtt{a}\})\big)$. Working upward from the bottom in $(P_A^{+})^{\rm op}$ (which is equivalent to working downward from the top in P_A^{+}), each such chain therefore gives rise to an informative action release sequence for identifying σ, consisting of at least $n-1$ actions. Moreover, there are at least $(n-1)!$ different such sequences for that same strategy σ; we can hold fixed the portion of any chain at and above p in $P_A^{\rm op}$, while varying the portion below p in at least $(n-1)!$ different ways.

Let $p \in S_{\rm max}$ and suppose c is some maximal chain of $P_A^{\rm op}$ that passes through p and touches maximal element $(\{\sigma\}, \sigma)$. Pick $q \in S$, with $q \leq p$ (here, "$\leq$" is the partial order on $P_A^{\rm op}$). Write $q = (\tau_q, \sigma_q)$ and $p = (\tau_p, \sigma_p)$. Even though q may not be part of chain c, we can still conclude that $\sigma_q \subseteq \sigma_p \subseteq \sigma$. If additionally ${\rm src}(\sigma_q) = V \setminus \{v\}$, then σ at the top of c must be a maximal strategy for attaining singleton goal state v. In order to prove the theorem, it is therefore enough to show that, for any $v \in V$, some such $q \in S$ (and thus $p \in S_{\rm max}$) exists.

Recall the source relation B from Lemma 91. Let $P_B^{\rm op}$ be P_B but with the opposite partial order. Referring back to the notation in the proof of Lemma 93, and using the fact that G is fully controllable, one sees that $\Delta(P_B^{\rm op}) \cong \Delta({\rm cl}_B) = \Delta(\mathfrak{F}(\Phi_B)) = {\rm sd}(\partial(V))$, with "$\cong$" meaning "isomorphic" and "sd" meaning "first barycentric subdivision". The isomorphism holds by definition of P_B. The first equality holds because $\phi_B \circ \psi_B$ is the identity when G is fully controllable, as we saw in the proof of Corollary 94. The second equality amounts to the definition of first barycentric subdivision, bearing in mind that $\Phi_B = \overline{\Delta_G} = \partial(V)$.

The homotopy equivalence of Lemma 93 carries over to this setting as $\theta : \Delta(P_A^{\rm op}) \to {\rm sd}(\partial(V))$. Corollary 94 (or inspection of the diagram in the proof of Lemma 93) provides an explicit formula. Specifically, for vertices (τ, σ) of $\Delta(P_A^{\rm op})$, one has $\theta(\tau, \sigma) = {\rm src}(\sigma)$.

Since θ is a homotopy equivalence, the induced homomorphism θ_* on reduced homology must map the reduced homology generator z to a reduced homology generator for the triangulated $(n-2)$-sphere ${\rm sd}(\partial(V))$. Consequently, $\|\theta_*(z)\|$ must consist of all vertices in ${\rm sd}(\partial(V))$, meaning all nonempty proper subsets of V. In particular, for each $v \in V$, there is some $q = (\tau_q, \sigma_q) \in \|z\|$ such that ${\rm src}(\sigma_q) = \theta(q) = V \setminus \{v\}$, as desired. $\square$

H.3 Delaying Goal Recognition

The next lemma establishes the "yes" assertion in the bullet that starts near the top of page 66.

Definition 95 (Complete Strategy). *Let $G = (V, \mathfrak{A})$ be a graph as discussed in Section 13. A* complete strategy for attaining state v *is a strategy σ that has at least one action at every state other than v. In other words, $\sigma \in \Delta_G$ and $\mathrm{src}(\sigma) = V \setminus \{v\}$.*

Lemma 96 (Delaying Goal Recognition). *Let $G = (V, \mathfrak{A})$ be a fully controllable graph. Let $n = |V|$. Suppose $n > 1$. Let $s \in V$ be some desired goal state.*

There exists a sequence of actions $a_1, a_2, \ldots, a_{n-1}$ in $\mathfrak{A}$ satisfying the following conditions:

(i) $\{a_1, \ldots, a_{n-1}\}$ is a complete strategy for attaining s.

(ii) For each $i = 1, \ldots, n-1$, let $\tau_i = \{a_1, \ldots, a_i\}$ and $W_i = \mathrm{src}(\tau_i)$. Then, for each such i and each $v \in V \setminus W_i$, there exists a complete strategy σ for attaining v, such that $\tau_i \subseteq \sigma \in \Delta_G$.

Comments:

(a) Condition (i) implies that no two of the actions $a_1, \ldots, a_{n-1}$ have the same source state.

(b) Condition (ii) implies that an observer cannot predict the final goal after seeing only a proper prefix of the sequence $a_1, a_2, \ldots, a_{n-1}$.

(c) Condition (ii) further implies that the sequence $a_1, \ldots, a_{n-1}$ forms an informative attribute release sequence for the relation A defined in Lemma 30 on page 65. Again, the reason is that an observer cannot even predict any specific source state for the remaining actions to be released after seeing only a proper prefix of the sequence $a_1, a_2, \ldots, a_{n-1}$.

Proof. For the proof, we assume that $\mathfrak{A}$ contains only deterministic and nondeterministic actions, not stochastic ones. The proof generalizes to graphs that include stochastic actions (possibly in addition to deterministic and nondeterministic actions), by an argument in [7]. The essence of that argument is that the source complex of a graph does not change if one replaces stochastic transitions by deterministic ones.

We sketch the rest of the proof, assuming all actions are deterministic or nondeterministic.

Since G is fully controllable, for each state in V there must be a deterministic transition *to* that state (from some other state). Backchaining such transitions gives rise to a directed cycle of deterministic actions, since the graph is finite. If that cycle is Hamiltonian, then we may choose $a_1, \ldots, a_{n-1}$ to be any ordering of those n deterministic actions except that we omit the action whose source state is s.

Suppose instead that the directed cycle of deterministic actions covers only a proper subset W of the state space V. Form a quotient graph with state space $V' = \{\diamond\} \cup V \setminus W$, where $\diamond$ represents all of W collapsed to a point. Inductively, the lemma's assertions hold for the quotient graph. One then needs to show how to combine the actions determined by the quotient graph with the cycle on W in order to satisfy the lemma's assertions for the original graph G. That argument is straightforward if a bit tedious, so we omit it. □

H.4 Hamiltonian Flexibility for Strategy Obfuscation

The next lemma establishes the Hamiltonian "yes" in the bullet that starts near the bottom of page 65.

Definition 97 (Hamiltonian Action Cycle)**.** *Let $G = (V, \mathfrak{A})$ be a graph as in Section 13, possibly with a mix of deterministic, nondeterministic, and stochastic actions. Let $n = |V|$ and assume $n > 1$. A sequence of actions $a_1, \ldots, a_n$ in $\mathfrak{A}$ is a* Hamiltonian cycle of actions *whenever all three of the following conditions hold:*

(i) No two of the actions $a_1, \ldots, a_n$ have the same source state.

(ii) Each action a_i is either deterministic or stochastic.

(iii) The source of action a_{i+1} is a target of action a_i, for all $i = 1, \ldots, n-1$, and the source of a_1 is a target of a_n.

Observe: Any proper subset of a Hamiltonian cycle of actions is a simplex in Δ_G.

(That observation requires understanding the definition of Δ_G when stochastic actions are involved: stochastic cycles are fine, so long as they are not recurrent. See [7] for details.)

Lemma 98 (Delaying Identification of a Given Strategy)**.** *Let $G = (V, \mathfrak{A})$ be a fully controllable graph. Assume $\mathfrak{A}$ contains a Hamiltonian cycle of actions $a_1, \ldots, a_n$, with $n = |V| > 1$.*

Let $v \in V$ and suppose σ_v is a maximal and complete strategy in Δ_G for attaining v. Then σ_v contains actions $b_{n-1}, \ldots, b_1$ that constitute a complete strategy for attaining v and that form an informative attribute release sequence for relation A.

(Recall: Relation A was defined in Lemma 30 on page 65; it models the maximal simplices of Δ_G in terms of their constituent actions.)

Proof. Let v and σ_v be as specified.

We can assume without loss of generality that $V = \{1, \ldots, n\}$, that the source of a_i is i for all $i \in V$, and that $v = n$.

Now let $b_1, \ldots, b_{n-1}$ be any actions in σ_v chosen so that the source of b_i is i, for $i = 1, \ldots, n-1$. (If $b_i = a_i$ for some or all i, that is fine.)

Then $\{b_1, \ldots, b_{n-1}\}$ is itself a complete strategy for attaining v.

We claim that the release order $b_{n-1}, \ldots, b_1$ constitutes an informative attribute release sequence for relation A. In fact, we will prove the stronger assertion:

> **Claim:** Pick some $i \in \{1, \ldots, n\}$. Then:
> For each $s \in \{n\} \cup \{1, \ldots, i-1\}$, there exists a complete strategy $\sigma_s \in \Delta_G$ for attaining s, with $\{b_i, b_{i+1}, \ldots, b_{n-1}\} \subseteq \sigma_s$.
> (Notation: $\{b_i, b_{i+1}, \ldots, b_{n-1}\} = \emptyset$ when $i = n$. Similarly for other sets.)

Consequently, an observer cannot predict a specific source state for the remaining actions to be released after seeing a proper prefix of $b_{n-1}, \ldots, b_1$, so the action sequence is informative.

The claim certainly holds for $s = n$, using the original σ_v. Now consider an $s \in \{1, \ldots, i-1\}$ and let $\sigma_s = \{a_1, \ldots, a_{s-1}\} \cup \{b_{s+1}, \ldots, b_{n-1}\} \cup \{a_n\}$. By arguments from [7], $\sigma_s \in \Delta_G$. Finally, observe that $\mathrm{src}(\sigma_s) = V \setminus \{s\}$ and that σ_s contains $\{b_i, b_{i+1}, \ldots, b_{n-1}\}$. □

Caution: As mentioned on page 66, just because $b_{n-1}, \ldots, b_1$ as produced by Lemma 98 is an informative attribute release sequence for A, that does not mean one should always release actions in that fashion. If the release protocol were so rigid, an adversary familiar with the protocol would be able to infer much about the goal. In particular, the target set of a_{n-1} includes the goal state, so if that action is deterministic and if always $b_{n-1} = a_{n-1}$, then the adversary would be able to infer at least the goal from the first action released.

H.5 Example: A Rapidly Inferable Strategy

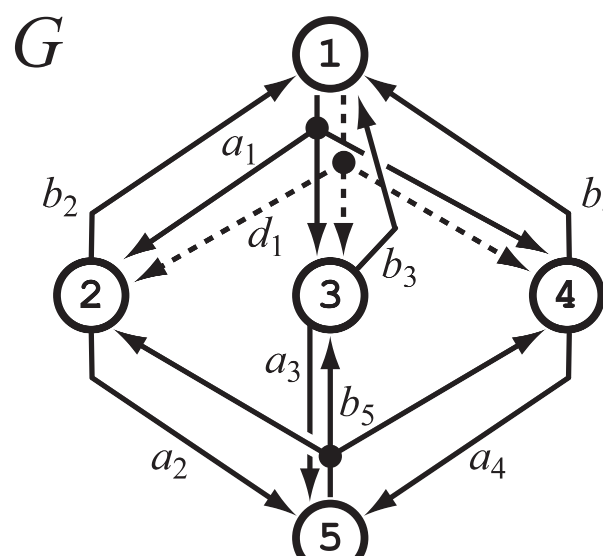


Figure 67: A graph with five states $\{1, 2, 3, 4, 5\}$, six deterministic actions $\{a_2, a_3, a_4, b_2, b_3, b_4\}$, two nondeterministic actions $\{a_1, b_5\}$, and one stochastic action $\{d_1\}$.

Figure 67 shows a fully controllable graph on fives states. The graph contains six deterministic actions, two nondeterministic actions, and one stochastic action. (See [6, 7] to learn more about graphs that contain deterministic, nondeterministic, and stochastic actions.) The action relation A for the graph appears in Figure 68.

Strategy σ_5 is a maximal and complete strategy for attaining state $\{5\}$, consisting of actions $\{a_1, a_2, a_3, a_4, d_1\}$. Observe that a longest informative action release sequence for this strategy has length 3. For instance, d_1, a_2, a_1 is such a sequence. Why is there no informative action release sequence longer than 3? Answer: As soon as one releases any one of the three actions $\{a_2, a_3, a_4\}$, an observer can infer that the strategy cannot contain the action b_5 and therefore, being maximal, must contain the other two actions in the set $\{a_2, a_3, a_4\}$ as well.

On page 65 we asked whether one can always find an informative action release sequence of length $n-1$ for a maximal complete strategy. (Here n is the number of states in the graph.) We see now that the answer is "no", not in general. Key to the current counterexample are two characteristics:

1. The graph contains three equivalent actions, namely $\{a_2, a_3, a_4\}$: The actions have completely identical columns in relation A. One could prune the graph to remove

A	a_1	a_2	a_3	a_4	d_1	b_2	b_3	b_4	b_5	Goal
σ_1						•	•	•	•	1
σ_2					•		•	•	•	2
σ_3					•	•		•	•	3
σ_4					•	•	•		•	4
σ_{234}	•				•				•	$\{2, 3, 4\}$
σ_5	•	•	•	•	•					5
σ_{54}		•	•	•	•	•	•			5
σ_{53}		•	•	•	•	•		•		5
σ_{52}		•	•	•	•		•	•		5
σ_{15}		•	•	•		•	•	•		$\{1, 5\}$

Figure 68: Relation A describes the strategy complex for the graph of Figure 67 in terms of its maximal strategies and their constituent actions. The rightmost column further shows each maximal strategy's goal.

redundant such actions while preserving full controllability, but this would not ensure a "yes" answer in general. The reason is that action equivalence is merely one of an infinite family of inferences one can construct with graphs. For instance, we could build a subgraph with an "any-2-actions-imply-4" character and construct a new counterexample.

2. The actions a_1 and d_1 both have source state 1 and target states $\{2, 3, 4\}$, yet are not equivalent. The difference is that a_1 is nondeterministic, meaning that potentially an adversary could determine the target state attained any time the action is executed. In contrast, d_1 is merely stochastic, meaning each target state has some fixed nonzero probability of occurring any time the action is executed, independent over repeated execution instances.

   If we disallowed the stochastic action d_1, full controllability would necessitate more deterministic actions in the graph, thereby changing both the collection of maximal simplices and consequent inferences. For instance, we might replace d_1 by three deterministic transitions. Doing so would similarly change the problematic strategy σ_5, providing more actions and increasing the length of an informative action release sequence for identifying σ_5. We then would indeed obtain an informative action release sequence of length 4. Alternatively, if we disallowed the nondeterministic action a_1, then the problematic strategy σ_5 would simply disappear. We would not even need to add any other actions, as the graph would remain fully controllable. Whether and when such alternatives are realistic is application dependent. We leave for future work exploration of more detailed conditions under which each maximal complete strategy in a graph has an informative action release sequence of length at least $n - 1$. Such characterizations might be useful in designing systems that either are or are not obfuscation-friendly. Appendix H.6 initiates this exploration with a discussion of pure graphs.

H.6 Pure Nondeterministic Graphs and Pure Stochastic Graphs

Let $G = (V, \mathfrak{A})$ be a graph whose actions may have uncertain outcomes. We say that G is *pure nondeterministic* if each action in $\mathfrak{A}$ is nondeterministic (for the purposes of this definition, we regard a deterministic action as a special instance of a nondeterministic action). We say that G is *pure stochastic* if each action in $\mathfrak{A}$ is stochastic (now viewing a deterministic action as a special instance of a stochastic action). We may simply say that G is *pure* if it is either pure nondeterministic or pure stochastic.

One can prove that every maximal strategy in the strategy complex of a fully controllable pure graph contains an informative action release sequence of length at least $n-1$, with $n = |V|$. This result means that one can release at least $n-1$ actions informatively before an observer can identify the strategy.

We will omit proof details, merely sketch the approach. First, one can show that if the Hamiltonian cycle in Lemma 98 is formed from deterministic actions, then every maximal strategy with a multi-state goal actually contains an informative action release sequence (iars) of length n. More specifically, suppose σ is such a strategy. The iars for σ consists of all the Hamiltonian cycle edges that are in σ, along with one action in σ for each cycle edge that is missing from σ. At least two such cycle edges are missing, since σ has at least two goal states. Consequently, the cycle breaks nicely into at least two intervals. For each missing cycle edge e, let I_e denote the interval of states between e's target and the next missing edge's source (based on the circular ordering induced by the Hamiltonian cycle). Include the endpoints (these could be identical, which is fine). One knows that every minimal nonface of Δ_G containing e must include some other action whose source lies in I_e and whose targets do *not* all lie inside I_e. One can therefore release one such action from σ informatively, in place of e.

Next, observe that every fully controllable pure nondeterministic graph contains a hierarchical decomposition of nested directed cycles, whose union covers V, such that each cycle is deterministic Hamiltonian in the quotient graph formed when one regards each of that cycle's subcycles as a singleton state. There are some details to verify, but, given a maximal strategy σ in the graph's strategy complex, this hierarchical decomposition and the previous Hamiltonian result yield an informative action release sequence of length at least $n-1$ for σ.

The approach is different for pure stochastic graphs. A nice property of minimal nonfaces in Δ_G, when G is pure stochastic, is that they form irreducible recurrent Markov chains and thus define fully controllable subgraphs of G. Moreover, each nonempty proper subset of a minimal nonface defines an isotropic simplex of actions with respect to G's action relation. When G itself is fully controllable, one can patch such subgraphs together expansively. In particular, given a maximal strategy σ in Δ_G, one can choose as building blocks fully controllable subgraphs that each consist solely of actions in σ plus one action not in σ. One starts by considering an action that moves off a goal state of σ. That action cannot be in σ, so gives rise to a minimal nonface all of whose other actions do lie in σ. One then expands outward repeatedly. Once no additional expansion is possible, one passes to a quotient graph by identifying all states covered thus far. Inductively, one can then repeat the expansion in the quotient graph. Again, there are some details to verify, but ultimately this process covers the graph's state space and produces an informative action release sequence of length at least $n-1$ for σ.

H.7 Strategy Obfuscation Summary

We summarize the key points of this appendix with the following theorem:

Theorem 99 (Obfuscation). *Let $G = (V, \mathfrak{A})$ be a fully controllable graph, with $n = |V| > 1$. Let σ be a maximal strategy in Δ_G.*

(a) If G is pure, then σ contains at least $n-1$ actions $a_1, \ldots, a_{n-1}$ that form an informative attribute release sequence (iars) with respect to G's action relation.

(This means that releasing the actions in the order $a_1, \ldots, a_{n-1}$ reduces the possible maximal strategies consistent with the actions released each time an action is released, but prevents identification of σ until at least all $n-1$ actions have been released.)

(b) This result can fail if G contains a mix of deterministic, nondeterministic, and stochastic actions.

(c) Even if G contains such a mix, the following is true:

Let $v \in V$. Then there exist maximal strategies σ_v and τ_v in Δ_G such that each strategy is a complete strategy for attaining state v and:

(i) σ_v contains at least $(n-1)!$ distinct iars of length at least $n-1$ each. (These iars may or may not be permutations of the same underlying $n-1$ actions.)

(ii) τ_v contains an iars of length at least $n-1$ that narrows the possible goal states consistent with the actions released by at most one state with each action released.

Caution: The release order of the actions in an iars need not correspond to the order in which actions might be executed at runtime.

I Morphisms and Lattice Generators

The aim of this appendix is to prove the claims of Section 14, ending with Theorem 41. That theorem shows how a surjective morphism of relations can use lattice operations to fully cover its codomain's poset even when the poset maps induced by the morphism are not themselves surjective.

I.1 Morphisms

Notation reminder: We frequently will be working with two relations: R is a relation on $X^R \times Y^R$ and Q is a relation on $X^Q \times Y^Q$. In order to distinguish rows and columns between the two relations, we also use notation of the form X^R_y, Y^R_x, X^Q_y, and Y^Q_x.

Also, recall that a *set map* is a function between two sets.

Now recall the definition of *morphism* from page 70:

Definition 33 (Morphism)**.** *Let R be a relation on $X^R \times Y^R$ and let Q be a relation on $X^Q \times Y^Q$. A* <u>morphism of relations</u> *$f : R \to Q$ is a pair of set maps:*

$$\begin{aligned} f_X &: X^R \to X^Q \\ f_Y &: Y^R \to Y^Q \end{aligned}$$

such that $\big(f_X(x),\, f_Y(y)\big) \in Q$ whenever $(x, y) \in R$.

Throughout this appendix, `'morphism'` refers to Definition 33. When the time comes, we will refer to `'G-morphism'` explicitly (see again Definitions 36 and 39 on pages 75 and 76).

Morphism Equality: Before proving properties about morphisms, we should give a notion of morphism equality. Suppose $g, h : R \to Q$ are two morphisms of relations. We say that $g = h$ if and only if $\big(g_X(x),\, g_Y(y)\big) = \big(h_X(x),\, h_Y(y)\big)$ for all $(x, y) \in R$. In particular, we do not care what the constituent set maps do on elements that are not relevant to the relations viewed as sets of ordered pairs. (Note: The condition stated is equivalent to requiring $g_X(x) = h_X(x)$ and $g_Y(y) = h_Y(y)$ for all $(x, y) \in R$.)

The following lemma shows that the component maps of a morphism between relations may be viewed as simplicial maps:

Lemma 34 (Induced Simplicial Maps)**.** *A morphism $f : R \to Q$ between nonvoid relations induces simplicial maps between the Dowker complexes:*

$$\begin{aligned} f_X &: \Psi_R \to \Psi_Q \\ f_Y &: \Phi_R \to \Phi_Q \end{aligned}$$

Proof. We need to show that $f_X(\sigma) \in \Psi_Q$ for all $\sigma \in \Psi_R$.

If $\sigma = \emptyset$, then $f_X(\sigma) = \emptyset \in \Psi_Q$, since Q is nonvoid.

If $\sigma = \{x_1, \ldots, x_k\}$, then $f_X(\sigma) = \{f_X(x_1), \ldots, f_X(x_k)\}$.

Since $\emptyset \neq \sigma \in \Psi_R$, there exists $y \in Y^R$ such that $(x, y) \in R$ for all $x \in \sigma$. Thus

$(f_X(x),\, f_Y(y)) \in Q$ for all $x \in \sigma$, by the definition of morphism. So $f_Y(y)$ is a witness for $f_X(\sigma)$ in Q, telling us $f_X(\sigma) \in \Psi_Q$.

The argument for the map $f_Y : \Phi_R \to \Phi_Q$ is similar. □

Comment: The nonvoid requirement is an artifact, arising because we sometimes regard void relations as having empty rather than void Dowker complexes, in the context of links (see Definitions 7 and 8 on page 24, Definition 19 on page 40, the comments about void relations on page 88, and the hypotheses of Lemma 55 on page 94). The nonvoid requirement of Lemma 34 avoids having to worry about mapping from an artificially empty complex into a void one.

Lemma 35 (Morphism Properties). *Assume the notation from before and that all relevant relations are nonvoid. Let $f : R \to Q$ be a morphism of relations. Then:*

(i) f_X and f_Y *are injective set maps* $\Longrightarrow$ f *is injective* $\Longleftrightarrow$ f *is a monomorphism.*

(ii) f *surjective* $\Longrightarrow$ f *epimorphism* $\Longleftrightarrow$ f_X and f_Y *are surjective set maps.*

(Additional conditions for that last $\Longleftrightarrow$ *: The* $\Longrightarrow$ *direction assumes that Q has no blank rows or columns, while the* $\Longleftarrow$ *direction assumes that R has no blank rows or columns.)*

The two uni-directional implications $\Longrightarrow$ *above are strict.*

(iii) If $f_X : \Psi_R \to \Psi_Q$ is surjective and Q has no blank rows, then $f_X : X^R \to X^Q$ is surjective.

Similarly for f_Y, now assuming that Q has no blank columns.

The converses need not hold. Indeed, f itself can be surjective but the maps of simplicial complexes need not be (as we saw with the maps of page 73 and as one can see with simpler examples as well).

(iv) If $f_X : X^R \to X^Q$ is injective, then $f_X : \Psi_R \to \Psi_Q$ is injective. The converse holds if R has no blank rows.

Similarly for f_Y, now assuming that R has no blank columns for the converse.

Proof. We will prove the various implications. Strictness, i.e., failure of converses, where mentioned above, can be seen readily with simple examples.

Part (i):

(a) Let f_X and f_Y be injective set maps.

Suppose $(f_X(x'), f_Y(y')) = (f_X(x), f_Y(y))$. Then $f_X(x') = f_X(x)$, so $x' = x$.
And $f_Y(y') = f_Y(y)$, so $y' = y$. So f is injective as a set map of ordered pairs.

(b) Let f be injective as a set map of ordered pairs.

Suppose $g, h : S \to R$ are morphisms such that $f \circ g = f \circ h$.
Suppose $(x, y) \in S$. By assumption, $\big(f_X(g_X(x)),\, f_Y(g_Y(y))\big) = \big(f_X(h_X(x)),\, f_Y(h_Y(y))\big)$.
Since f is injective, $(g_X(x),\, g_Y(y)) = (h_X(x),\, h_Y(y))$.
So $g = h$, by our notion of equality. Consequently, f is a monomorphism.

(c) Let f be a monomorphism.

Suppose $f(x,y) = f(x',y')$ but $(x,y) \neq (x',y')$. Let S be the relation consisting of the single element $\{(I,\alpha)\}$, with I and α new symbols:

S	α
I	$\bullet$

Define two morphisms $g, h : S \to R$ by:

$$\begin{aligned} g_X &: I \mapsto x, & h_X &: I \mapsto x', \\ g_Y &: \alpha \mapsto y, & h_Y &: \alpha \mapsto y'. \end{aligned}$$

Then $g \neq h$, but $f \circ g = f \circ h$, a contradiction. So f is injective.

<u>Part (ii):</u>

(a) Let f be surjective as a set map of ordered pairs.

Suppose $g, h : Q \to S$ are morphisms such that $g \circ f = h \circ f$.
Suppose $(x',y') \in Q$.
By surjectivity, there exists $(x,y) \in R$ such that $(f_X(x), f_Y(y)) = (x',y')$. So:
$\big(g_X(x'),\, g_Y(y')\big) = \big(g_X(f_X(x)),\, g_Y(f_Y(y))\big) = \big(h_X(f_X(x)),\, h_Y(f_Y(y))\big) = \big(h_X(x'),\, h_Y(y')\big)$.
Thus $g = h$ and we see that f is an epimorphism.

(b) Assume Q has no blank rows or columns and let f be an epimorphism.

Suppose f_Y is not surjective, so there exists $y^* \in Y^Q \setminus (f_Y(Y^R))$.
Let S be the relation consisting of two elements $\{(I,\alpha),(I,\beta)\}$, with I, α, β new symbols:

S	α	β
I	$\bullet$	$\bullet$

Define two morphisms $g, h : Q \to S$ by:

$$\begin{aligned} g_X(x) &= I & &\text{and} & h_X(x) &= I, & &\text{for every } x \in X^Q; \\ g_Y(y) &= \alpha & &\text{and} & h_Y(y) &= \alpha, & &\text{for every } y \in Y^Q \setminus \{y^*\}; \\ g_Y(y^*) &= \alpha & &\text{and} & h_Y(y^*) &= \beta. \end{aligned}$$

Since $y^* \in Y^Q$ and Q has no blank columns there is at least one $x^* \in X^Q$ such that $(x^*, y^*) \in Q$. So $g \neq h$.

Observe that $g \circ f = h \circ f$ since y^* does not appear in the image of f_Y, contradicting f being an epimorphism.

The argument showing that f_X is surjective is similar.

(c) Assume R has no blank rows or columns and let f_X and f_Y be surjective.

Suppose $g, h : Q \to S$ are morphisms such that $g \circ f = h \circ f$.

Suppose $(x,y) \in Q$. We need to show that $g_X(x) = h_X(x)$ and $g_Y(y) = h_Y(y)$, as that means $g = h$, given our definition of equality. We will make the argument for the X coordinate; the Y argument is similar.

Since f_X is surjective, there exists $\overline{x} \in X^R$ such that $f_X(\overline{x}) = x$. Since R has no blank rows, there exists $\overline{y} \in Y^R$ such that $(\overline{x}, \overline{y}) \in R$.

Since $(g \circ f)(\overline{x}, \overline{y}) = (h \circ f)(\overline{x}, \overline{y})$, one obtains $g_X(x) = g_X(f_X(\overline{x})) = h_X(f_X(\overline{x})) = h_X(x)$.

Part (iii):

Suppose Q has no blank rows and suppose $f_X : \Psi_R \to \Psi_Q$ is surjective as a simplicial map.

Suppose $x \in X^Q$. Since Q has no blank rows, $\{x\}$ is a vertex of Ψ_Q, so there is some simplex $\sigma \in \Psi_R$ such that $f_X(\sigma) = \{x\}$, with f_X viewed as a simplicial map. Necessarily, $\sigma \neq \emptyset$, so pick some $\overline{x} \in \sigma$. Then $f_X(\overline{x}) = x$, with f_X now viewed as a set map. Thus the set map $f_X : X^R \to X^Q$ is surjective.

The argument for f_Y assuming Q has no blank columns is similar.

Part (iv):

(a) Let f_X be injective as a set map $X^R \to X^Q$. Consider f_X as a simplicial map $\Psi_R \to \Psi_Q$.

Suppose $f_X(\sigma) = \kappa = f_X(\tau)$, with $\sigma, \tau \in \Psi_R$ and $\kappa \in \Psi_Q$.

If $\kappa = \emptyset$, then necessarily $\sigma = \tau = \emptyset$. Otherwise, $\sigma \neq \emptyset$ and $\tau \neq \emptyset$, so let $x \in \sigma$. Then $f_X(x) \in \kappa$. So there exists $x' \in \tau$ such that $f_X(x') = f_X(x)$. Since f_X is injective as a set map, that says $x' = x$. Thus $\sigma \subseteq \tau$. A similar argument shows the reverse inclusion, so $\sigma = \tau$. Thus f_X is injective as a simplicial map.

(b) Assume R has no blank rows and let f_X be injective as a simplicial map $\Psi_R \to \Psi_Q$.

Consider f_X as a set map $X^R \to X^Q$ and suppose $f_X(x) = f_X(x')$. Since R has no blank rows, both $\{x\}$ and $\{x'\}$ are vertices in Ψ_R. That means $f_X(\{x\}) = f_X(\{x'\})$ when we view f_X as a simplicial map, so $\{x\} = \{x'\}$ by injectivity, i.e., $x = x'$. So we see that f_X is injective as a set map.

A similar argument holds for the assertions regarding f_Y. □

I.2 G-Morphisms

Recall the material of Section 14.4, starting on page 75.

Lemma 37 (Witness Containment)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. Then:*

(a) $(f_Y \circ \phi_R)(\sigma) \subseteq (\phi_Q \circ f_X)(\sigma)$, *for every* $\sigma \in \Psi_R$,

(b) $(f_X \circ \psi_R)(\gamma) \subseteq (\psi_Q \circ f_Y)(\gamma)$, *for every* $\gamma \in \Phi_R$.

Proof. Observe that $(f_Y \circ \phi_R)(\emptyset) = f_Y(Y^R) \subseteq Y^Q = \phi_Q(\emptyset) = (\phi_Q \circ f_X)(\emptyset)$.

Now let $\emptyset \neq \sigma \in \Psi_R$. Let $y \in \phi_R(\sigma) \neq \emptyset$. Then $(x, y) \in R$ for every $x \in \sigma$. Thus $(f_X(x), f_Y(y)) \in Q$ for every $x \in \sigma$. So $f_Y(y) \in \phi_Q(f_X(\sigma))$. This is true for all $y \in \phi_R(\sigma)$, telling us $f_Y(\phi_R(\sigma)) \subseteq \phi_Q(f_X(\sigma))$.

The argument for assertion (b) is similar. □

Corollary 38 (Homotopic Face Maps)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. Then:*

(a) f_X *and* $\psi_Q \circ f_Y \circ \phi_R$ *are homotopic poset maps* $\mathfrak{F}(\Psi_R) \to \mathfrak{F}(\Psi_Q)$,

(b) f_Y *and* $\phi_Q \circ f_X \circ \psi_R$ *are homotopic poset maps* $\mathfrak{F}(\Phi_R) \to \mathfrak{F}(\Phi_Q)$.

Proof. Let $\sigma \in \mathfrak{F}(\Psi_R)$.

By Lemma 37, $(f_Y \circ \phi_R)(\sigma) \subseteq (\phi_Q \circ f_X)(\sigma)$.

Therefore $(\psi_Q \circ f_Y \circ \phi_R)(\sigma) \supseteq (\psi_Q \circ \phi_Q \circ f_X)(\sigma)$.

So $(\psi_Q \circ f_Y \circ \phi_R)$ and $(\psi_Q \circ \phi_Q \circ f_X)$ are homotopic maps (see [1], Theorem 10.11).

Since $\psi_Q \circ \phi_Q$ is homotopic to the identity on $\mathfrak{F}(\Psi_Q)$, part (a) follows.

The proof of (b) is similar. □

Corollary 40 (Homotopic G-Morphisms)**.** *Let $f : R \to Q$ be a morphism of nonvoid relations. The induced G-morphisms, as given by the poset maps $f^g_X, f^g_Y : P_R \to P_Q$ of Definition 39 on page 76, are homotopic.*

Proof. See Figure 53 on page 75 for the underlying maps comprising the G-morphisms. The G-morphisms are defined as follows:

For all $(\sigma, \gamma) \in P_R$:

$$f^g_X(\sigma,\gamma) \;=\; (\sigma',\gamma'), \quad \text{with} \quad \sigma' \;=\; (\psi_Q \circ f_Y \circ \phi_R)(\sigma) \quad \text{and} \quad \gamma' \;=\; \phi_Q(\sigma').$$

$$f^g_Y(\sigma,\gamma) \;=\; (\sigma'',\gamma''), \quad \text{with} \quad \gamma'' \;=\; (\phi_Q \circ f_X \circ \psi_R)(\gamma) \quad \text{and} \quad \sigma'' \;=\; \psi_Q(\gamma'').$$

These definitions make sense because f_X and f_Y map nonempty simplices to nonempty simplices and because the images of ψ_Q and ϕ_Q may be viewed as lying in P_Q, by Corollary 46 on page 91. (Similarly, the images of ψ_R and ϕ_R may be viewed as lying in P_R — In fact, as used above, these maps are simply switching between the σ and γ components ("labels") of the given element (σ, γ) in P_R.) Observe that f^g_X and f^g_Y are order-preserving poset maps.

Applying Lemma 37 and since $(\sigma, \gamma) \in P_R$:

$$(f_Y \circ \phi_R)(\sigma) \;\subseteq\; (\phi_Q \circ f_X)(\sigma) \;=\; (\phi_Q \circ f_X \circ \psi_R)(\gamma) \;=\; \gamma''.$$

Consequently:

$$\sigma' \;=\; (\psi_Q \circ f_Y \circ \phi_R)(\sigma) \;\supseteq\; \psi_Q(\gamma'') \;=\; \sigma''.$$

So the maps are homotopic (see [1], Theorem 10.11). □

I.3 Lattice Generators

We turn now to the main result.

(Recall that a relation is *tight* when it has no blank rows or columns.)

Lemma 100 (Generators in Image)**.** *Let $f : R \to Q$ be a surjective morphism between nonvoid tight relations.*

Suppose $q \in P_Q$ is of the form $\left(X^Q_y,\, (\phi_Q \circ \psi_Q)(\{y\})\right)$, for some $y \in Y^Q$.

Then there exist $q_1, \ldots, q_k$ in the image of $f^g_X : P_R \to P_Q$, with $k \geq 1$, such that $q = \bigvee_{i=1}^{k} q_i$. (Here, $\bigvee$ is the join operation of P^+_Q.)

Proof. By Lemma 35(ii), the component functions $f_X : X^R \to X^Q$ and $f_Y : Y^R \to Y^Q$ are surjective. Since f_Y is surjective, $f_Y^{-1}(\{y\}) = \{y_1, \ldots, y_k\} \subseteq Y^R$, for some $k \geq 1$.

For each $i = 1, \ldots, k$, observe and define the following:

- Since R has no blank columns, $X^R_{y_i} \neq \emptyset$, so $\bigl(X^R_{y_i}, (\phi_R \circ \psi_R)(\{y_i\})\bigr) \in P_R$.
- Define σ_i as the "σ'-component" of $f^g_X\bigl(X^R_{y_i}, (\phi_R \circ \psi_R)(\{y_i\})\bigr)$, meaning:

$$\sigma_i \;=\; (\psi_Q \circ f_Y \circ \phi_R)\bigl(X^R_{y_i}\bigr) \;=\; \psi_Q(\gamma) \;=\; \bigcap_{\overline{y} \in \gamma} X^Q_{\overline{y}}, \quad \text{with } \gamma = f_Y\bigl((\phi_R \circ \psi_R)(\{y_i\})\bigr).$$

- Observe that $y \in \gamma$, since $y = f_Y(y_i)$ and $y_i \in (\phi_R \circ \psi_R)(\{y_i\})$. Therefore $\sigma_i \subseteq X^Q_y$.
- Define $q_i = (\sigma_i, \gamma_i) \in P_Q$, with $\gamma_i = \phi_Q(\sigma_i)$. So q_i is in the image of $f^g_X : P_R \to P_Q$.

We need to show that $q = \bigvee_{i=1}^k q_i$. Expanding, we see:

$$\bigvee_{i=1}^{k} q_i \;=\; \Bigl((\psi_Q \circ \phi_Q)\Bigl(\bigcup_{i=1}^{k} \sigma_i\Bigr),\; \bigcap_{i=1}^{k} \gamma_i\Bigr).$$

By the third bullet above, $\bigcup_{i=1}^k \sigma_i \subseteq X^Q_y$, so:

$$\bigcup_{i=1}^{k} \sigma_i \;\subseteq\; (\psi_Q \circ \phi_Q)\Bigl(\bigcup_{i=1}^{k} \sigma_i\Bigr) \;\subseteq\; (\psi_Q \circ \phi_Q)\bigl(X^Q_y\bigr) \;=\; X^Q_y.$$

We will establish $X^Q_y \subseteq \bigcup_{i=1}^k \sigma_i$, thereby completing the proof.

Let $\overline{x} \in X^Q_y$. So $(\overline{x}, y) \in Q$.
By surjectivity of f, there exists $(\hat{x}, \hat{y}) \in R$ such that $f_X(\hat{x}) = \overline{x}$ and $f_Y(\hat{y}) = y$.
Now $\hat{y} = y_j$, for some $j \in \{1, \ldots, k\}$, as defined earlier. Thus $\hat{x} \in X^R_{y_j}$.
Consequently, for every $z \in (\phi_R \circ \psi_R)(\{y_j\})$, $(\hat{x}, z) \in R$ and so $(f_X(\hat{x}), f_Y(z)) \in Q$.
That means $(\overline{x}, \overline{y}) \in Q$ for every $\overline{y} \in f_Y\bigl((\phi_R \circ \psi_R)(\{y_j\})\bigr)$.
Therefore, $\overline{x} \in \sigma_j \subseteq \bigcup_{i=1}^k \sigma_i$ and we conclude that $X^Q_y \subseteq \bigcup_{i=1}^k \sigma_i$.
(Note: X^Q_y need not lie in a single σ_j, since j depends on $\overline{x}$.) □

Corollary 101. *Assume the hypotheses of Lemma 100.*
Suppose further that for some $y_i \in f_Y^{-1}(\{y\})$, $(\phi_R \circ \psi_R)(\{y_i\}) = \{y_i\}$.
Then q *is itself in the image of* $f^g_X : P_R \to P_Q$.

Proof. In the proof of Lemma 100, we see that now $f_Y\bigl((\phi_R \circ \psi_R)(\{y_i\})\bigr) = \{y\}$, so $\sigma_i = X^Q_y$. □

Comment: Corollary 101 helps to explain the example of pages 73 and 77, in which a surjective morphism generated the entire poset of its codomain even though the induced maps on the Dowker complexes were not surjective. Namely:

In the Möbius relation M of page 72, singletons are unmoved by the closure operators. In the tetrahedral relation T of page 42, maximal simplices are dual to singletons. Intersections of maximal simplices in the tetrahedral relation generate all of P_T. These maximal simplices come from dualizing images of singletons of the Möbius relation. Consequently:

- Even though one merely has $f_X(\{1,2,5\}) = \{1,4\}$, one further sees that $f^g_X(\{1,2,5\},\{\mathtt{a}\}) = (\{1,3,4\},\{\mathtt{a}\})$. The G-morphism f^g_X therefore supplies the apparently uncovered simplex $\{1,3,4\}$ of Ψ_T. Similarly, f^g_Y supplies $\{\mathtt{a},\mathtt{b},\mathtt{d}\}$ in Φ_T.
- Previously, in Table 9 on page 78, we saw that the element $(13,\mathtt{ac})$ of P_T did not itself appear in the images of the maps f^g_Y and f^g_X. However, $(13,\mathtt{ac})$ *does* appear as the join or meet of elements in the images:

$$\begin{aligned}(13,\mathtt{ac}) &= (134,\mathtt{a}) \wedge (123,\mathtt{c}), \quad \text{with both arguments to } \wedge \text{ in the image of } f^g_X;\\ (13,\mathtt{ac}) &= (1,\mathtt{abc}) \vee (3,\mathtt{acd}), \quad \text{with both arguments to } \vee \text{ in the image of } f^g_Y.\end{aligned}$$

More generally, the following theorem describes the process:

Theorem 41 (Lattice Surjectivity)**.** *Let R and Q be nonvoid tight relations. Suppose $f: R \to Q$ is a surjective morphism (in the sense of Definition 33). For any $q \in P_Q$:*

$$q = \bigwedge_j \bigvee_i q_{ji}, \quad \textit{with each } q_{ji} \textit{ in the image of } f^g_X : P_R \to P_Q,$$

$$q = \bigvee_k \bigwedge_\ell q'_{k\ell}, \quad \textit{with each } q'_{k\ell} \textit{ in the image of } f^g_Y : P_R \to P_Q.$$

(Here, $\bigvee$ and $\bigwedge$ are the lattice operations of P^+_Q.)

Proof. Write $q = (\sigma, \gamma) \in P_Q$. Then $\sigma = \psi_Q(\gamma) = \bigcap_{y\in\gamma} X^Q_y$.

So $q = \bigwedge_{y\in\gamma} q_y$, with each $q_y \in P_Q$ of the form $\big(X^Q_y,\, (\phi_Q \circ \psi_Q)(\{y\})\big)$.

By Lemma 100, for each $y \in \gamma$, we have that $q_y = \bigvee_i q_{y,i}$ with each $q_{y,i}$ in the image of $f^g_X : P_R \to P_Q$ and with i in some finite index set $\mathcal{I}(y)$. Thus:

$$q = \bigwedge_{y\in\gamma} \; \bigvee_{i\in\mathcal{I}(y)} q_{y,i}.$$

The other form follows by dualizing the previous arguments. □

J A Few More Examples

J.1 Local Spheres versus Global Contractibility

The reader may wonder whether preservation of attribute privacy always requires a relation to exhibit homology in its Dowker complexes. The answer is that links of individuals must have homology, by Theorems 10 and 11 on page 26, but the overall relation need not.

D	a	b	c	d	e	f	g	h
1				•		•		•
2				•	•	•		
3						•	•	•
4			•			•	•	
5	•				•	•		
6	•		•		•			
7			•	•				•
8		•		•	•			
9	•	•					•	
10		•	•				•	
11	•						•	•
12	•	•						•
13		•	•					•
14	•		•	•				
15	•	•		•				
16		•	•		•			
17	•		•			•		

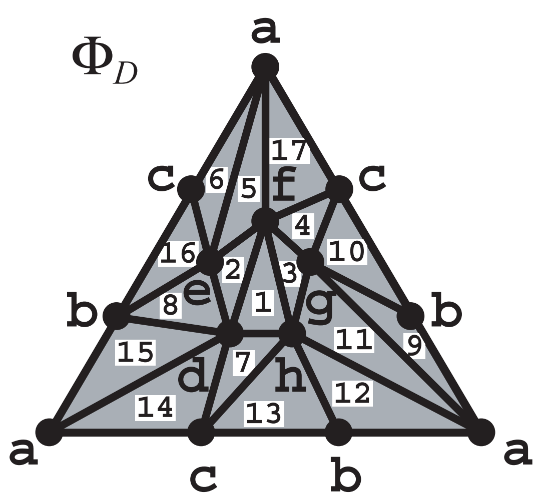


Figure 69: Relation D and its Dowker complex Φ_D. The complex is a triangulation of the Dunce Hat, a contractible space (the seemingly bounding edges actually touch, as suggested by the vertex labels). The Dunce Hat has no free faces, indicating that D preserves attribute privacy. (Vertices of Φ_D are attributes. Triangles are labeled with their generating individuals.)

Consider for example the relation D of Figure 69. There are 17 individuals, each with three attributes. The figure also shows Φ_D. We can see that there are no free faces, so the relation preserves attribute privacy by Lemma 62 on page 98. Moreover, each link $\text{Lk}(\Psi_D, x)$ is homotopic to a circle $\mathbb{S}^1$. Indeed, viewed from attribute space, that link is exactly the boundary of a triangle for each individual. Figure 70 shows such a link for individual #10. The link relation has a large number of individuals but only three attributes. So Theorem 10 holds and there is homology in the link. There is however no homology in the attribute complex of the relation D itself; the simplicial complex Φ_D is a triangulation of the Dunce Hat, a nontrivially contractible space.

Although R preserves attribute privacy, it does not preserve association privacy. For example: Individuals #1 and #12 share exactly one attribute (namely `h`), but do so with four additional individuals (namely #3, #7, #11, and #13). If attributes represent shared dinners, then in some cases one can infer all the guests at a dinner after having seen as few

Q	b	c	g
3			•
4		•	•
6		•	
7		•	
8	•		
9	•		•
11			•
12	•		
13	•	•	
14		•	
15	•		
16	•	•	
17		•	

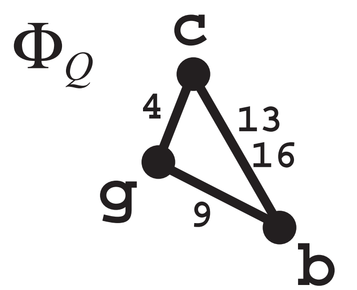


Figure 70: Relation Q represents $\mathrm{Lk}(\Psi_D, 10)$ from Figure 69. Also shown is the attribute Dowker complex Φ_Q. It is the boundary of a triangle, so homotopic to $\mathbb{S}^1 = \mathbb{S}^{k-2}$. Since individual #10 has three attributes and $1 = 3 - 2$, that means relation R preserves attribute privacy for individual #10. (Vertices of Φ_Q are attributes. Edges are labeled with their generating individuals. Notice that the edge $\{\mathtt{b}, \mathtt{c}\}$ is generated by two individuals. Whereas most edges in Φ_D are shared by only two triangles, edge $\{\mathtt{b}, \mathtt{c}\}$ is shared by three triangles; it is one of those edges that are glued to two others in the Dunce Hat representation. — Individuals who generate just vertices are not shown in the drawing of Φ_Q here.)

as two guests. (Attribute privacy means that one cannot definitively infer additional dinners attended by a guest simply from having observed that guest at a particular dinner or two.)

J.2 Disinformation

Privacy loss is possible when there is a free face in the relevant Dowker complex. One way to preserve privacy is to eliminate such free faces. Earlier in the report, we studied morphisms between relations as a possible way to transform data so as to reduce privacy loss. Ideally, for attribute privacy, the goal of such a transformation is to map onto a relation whose attribute complex has no free faces. We saw that such transformations need not always exist, for topological reasons, unless one is willing to introduce discontinuities, that is, discard knowledge of some relationships in the underlying spaces.

Alternatively, one could imagine embedding a relation within another that does preserve privacy. Of course, at the extreme, one simply embeds the given relation in a huge relation that looks like a perfect sphere. Now there is privacy but the same mechanism that provides privacy reduces utility. Nonetheless, one has not discarded relationships, merely surrounded them with disinformation. We saw an example of that early on, when we added a single attribute to relation H in the example of Section 3.1, in order to remove the inference that someone had cancer. If one has a separate mechanism for discerning fake entries from legitimate entries, then one can see past the disinformation — in the earlier example that would entail having a (presumably safely encrypted) memory of which single entry in the relation is false.

M	a	b	c	d	e
1	•	•			•
2	•	•	•		
3		•	•	•	
4			•	•	•
5	•			•	•

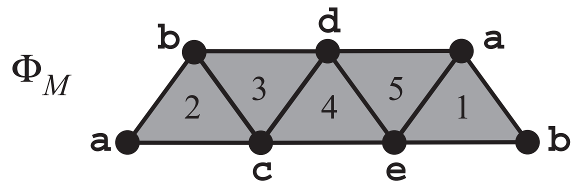


Figure 71: Relation M revisited, along with its attribute complex Φ_M.

Figure 71 revisits our earlier Möbius strip relation, showing the relation M and its attribute complex Φ_M. Loss of attribute privacy occurs when someone observes two attributes that form a free edge on the boundary of the Möbius strip, such as the edge $\{\mathtt{b}, \mathtt{d}\}$. Given the relation, the observer can then infer a third attribute and identify the underlying individual, in this case infer attribute `c` and identify individual #3.

In order to preserve attribute privacy, one might consider adding decoy individuals whose so-called attributes include those edges, making the edges nonfree, thus removing that inference mechanism. Relation MM in Figure 72 does so by doubling the number of individuals.

MM	a	b	c	d	e
1	•	•			•
2	•	•	•		
3		•	•	•	
4			•	•	•
5	•			•	•
6	•	•		•	
7		•	•		•
8	•		•	•	
9		•		•	•
10	•		•		•

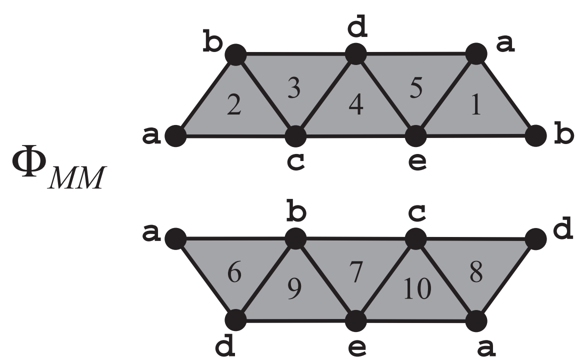


Figure 72: Relation MM adds five decoy individuals. The attribute complex Φ_{MM} entails gluing two Möbius strips together.

The additional five individuals form their own Möbius strip. The figure therefore describes the overall attribute complex Φ_{MM} as two Möbius strips, with edges shared between the two strips, as suggested by the vertex labels. The overall attribute complex amounts to gluing the two Möbius strips together, boundary to zigzag spine. The resulting attribute complex is the 2-skeleton of the full complex on the attribute set $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$. It therefore is homotopic to a wedge sum of four 2-spheres: $\mathbb{S}^2 \vee \mathbb{S}^2 \vee \mathbb{S}^2 \vee \mathbb{S}^2$.

Each of Φ_{MM}'s edges is now shared by three triangles. There are no free faces.

No attribute inference is possible. (Association inference is possible.)

Moreover, the complex is sufficiently isotropic that one cannot say *a priori* which individuals are real and which are decoys, even if one knows that there might be decoys. Of course, the curator of the relation likely would want some secure mechanism to separate truth from fiction, that is, to peel apart the gluing. Regardless, real individuals may be identified via MM upon seeing all their attributes (and only then).

J.3 Insufficient Representation

In this subsection we show that if there are fewer than 2^k individuals in a nonvoid relation with $2k$ attributes that model k bits for each individual, then the relation cannot preserve attribute privacy for everyone. The reason is that fewer than 2^k individuals amounts to removing some generating simplices from the potential attribute complex $\mathbb{S}^0 * \mathbb{S}^0 * \cdots * \mathbb{S}^0$, thereby creating free faces in Φ_R. By similar intuition, it may be possible to preserve attribute privacy even if there are fewer than, say, 3^k individuals in a relation with $3k$ attributes representing k trivalent pieces of information. After all, bits are a special case of tri-values, so one can preserve attribute privacy with certain 2^k individuals. Thinking simplicially, the potential attribute complex for tri-values is $(\mathbb{S}^0 \vee \mathbb{S}^0) * (\mathbb{S}^0 \vee \mathbb{S}^0) * \cdots * (\mathbb{S}^0 \vee \mathbb{S}^0)$. Removing some generating simplices from that space does not necessarily create free faces, as one can see by simple example.

Definition 102 (Binary Attribute Pair)**.** *By a* <u>binary attribute pair</u> *we mean two mutually exclusive attributes, written y and $\overline{y}$. No individual can have both attributes. Moreover, in what follows we will assume that every individual has exactly one attribute from any such pair.*

Lemma 103 (Privacy Requires Many Individuals)**.** *Suppose $Y = \{y_1, \overline{y}_1, y_2, \overline{y}_2, \ldots, y_k, \overline{y}_k\}$, with $\{y_i, \overline{y}_i\}$ a binary attribute pair, for $i = 1, \ldots, k$, and $k \geq 1$.*

Let R be a relation on $X \times Y$, with $X \neq \emptyset$, such that every individual $x \in X$ has as attributes exactly one of $\{y_i, \overline{y}_i\}$, for each $i = 1, \ldots, k$. Let n be the number of distinct rows of R.

Then R preserves attribute privacy if and only if $n = 2^k$.

Proof. Observe that each row of R has exactly k nonblank entries, so each maximal simplex of Φ_R consists of exactly k vertices. Moreover, no row of R is contained in another row unless the two rows are identical. We may therefore assume, without loss of generality, that all rows of R are distinct and incomparable. Consequently, every $x \in X$ is uniquely identifiable. We can think of each individual $x \in X$ as defining a unique and identifying k-bit number, with one bit per binary attribute pair, as determined by that individual's row, Y_x. All possible k-bit numbers are represented by X if and only if $n = 2^k$.

I. <u>Suppose that $n = 2^k$.</u>

Showing that Φ_R contains no free faces would establish that R preserves attribute privacy, by Lemma 62 on page 98. To show that Φ_R contains no free faces, it is enough to show that, for every maximal $\gamma \in \Phi_R$ and every $y \in \gamma$, the simplex $\gamma \setminus \{y\}$ is contained in some maximal simplex of Φ_R other than just γ.

Write $\chi = \gamma \setminus \{y\}$. Since y is part of a binary attribute pair, we can construct a new set γ' from γ by replacing y with its "opposite". Specifically: $\gamma' = \chi \cup \{y_i\}$, if $y = \overline{y}_i$; and $\gamma' = \chi \cup \{\overline{y}_i\}$, if $y = y_i$. Since $n = 2^k$, there is an $x \in X$ for which $Y_x = \gamma'$. So $\gamma' \in \Phi_R$, telling us χ is not free.

II. Suppose that that R preserves attribute privacy.

By Lemma 63 on page 99, Φ_R contains no free faces.

Let γ be a maximal simplex of Φ_R and $y \in \gamma$. Define $\chi = \gamma \setminus \{y\}$. Construct γ' as in part I above. Consider the collection $\Gamma = \{\eta \in \Phi_R \mid \chi \subsetneq \eta\}$. The only possible set that might be in Γ besides γ is γ'. Since Φ_R contains no free faces, $\Gamma = \{\gamma, \gamma'\}$.

Now vary y across γ and then repeat the process for all γ' thus constructed. The transitive closure of this process generates 2^k distinct maximal simplices in Φ_R, each of which corresponds to a unique $x \in X$. So $n = 2^k$. □

J.4 A Structural Inference Example: Passengers on Ferries

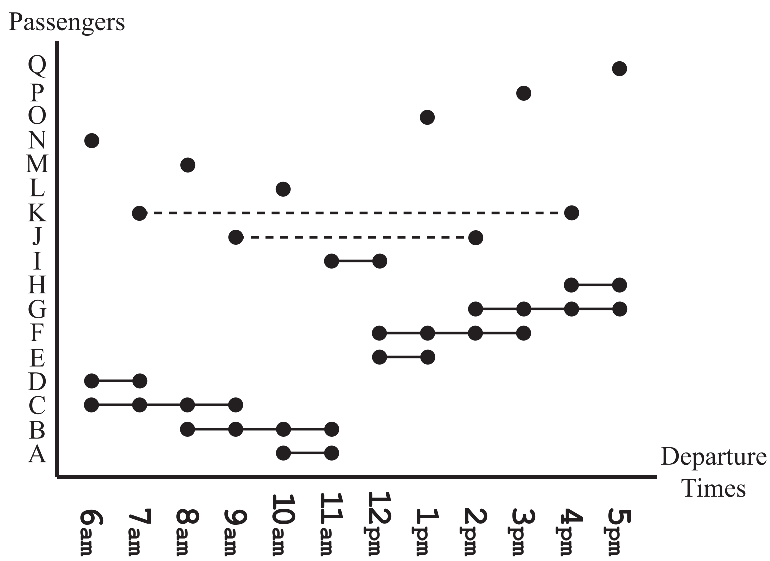


Figure 73: This time series represents 17 different passengers on 12 different ferry crossings. Each dot represents a passenger on a crossing. As a visual aid, solid lines connect multiple crossings by the same passenger at consecutive departure times, while dashed lines connect multiple crossings by the same passenger at non-consecutive departure times.

Imagine a commuter ferry that crosses back and forth between downtown and an island. Passengers pay electronically as they enter the ferry, so there is a record of who is on which crossing. Figure 73 shows a hypothetical time series for 12 crossings during a day in which 17 passengers took the ferry, some of whom crossed several times. Figure 74 shows the corresponding Ψ_R complex: vertices are individuals; each triangle represents a particular crossing. (Each ferry crossing had three passengers in this simplified example.)

The waitress in the ferry's coffee shop observes four individuals ordering coffee and conversing during the day, appearing in pairs on four different crossings. She remembers seeing four distinct pairs, but does not remember the crossing times. Who are the individuals?

It is convenient to also represent the waitress's observations as a simplicial complex. Figure 75 does so. Vertices are now the four unknown individuals; edges are their (unknown) common crossing times. One can interpret who the individuals are by embedding the complex of Figure 75 into the complex of Figure 74, using injective maps in both the passenger and time

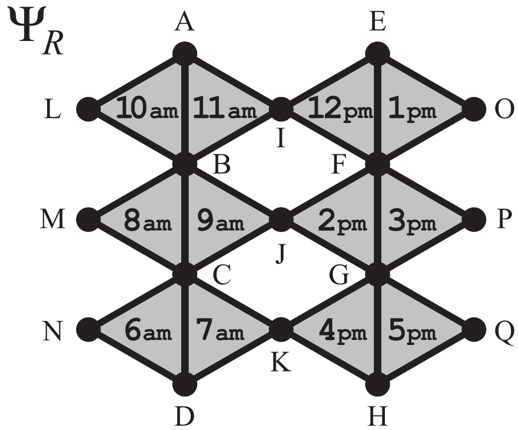


Figure 74: Simplicial complex Ψ_R determined by viewing the time series of Figure 73 as a relation R. Vertices represent passengers, labeled with letters. Triangles represent ferry crossings, labeled with departure times.

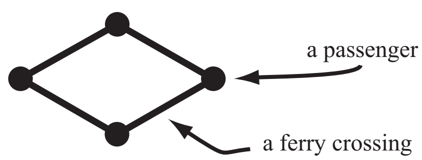


Figure 75: A waitress's observations of passengers drinking coffee together at various times, represented by a simplicial complex. Vertices represent unknown but distinct passengers. Edges represent unknown but distinct crossing times.

domains. There are exactly two such embeddings (modulo index permutations), given by the two ways one can wrap a rectangle around the two holes in the complex of Figure 74. (Those are the only two "diamonds" touching four different crossing times.) Thus the individuals are either $\{C, G, J, K\}$ or $\{B, F, I, J\}$, as indicated by Figure 76. Either way, one knows for sure that individual "J" twice had a conversation over coffee that day.

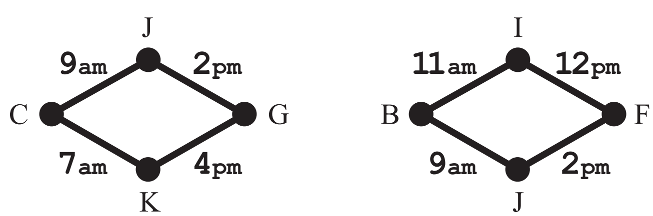


Figure 76: The two possible embeddings of the complex of Fig. 75 into the complex of Fig. 74.

Moreover, each of these embeddings places a time ordering on the embedded edges, from which one can make inferences as to who might have transmitted information to whom. For instance, for the embedding involving individuals $\{B, F, I, J\}$, one sees that individual "J" could have been both the initial source and final recipient of information.