The Algebra of Financial Provenance: Lineage Semirings from Source Record to Reviewed Finding

Daniel Ray Edgar
Finsider Labs

October 2026

Abstract

Every figure in a financial diligence report is computed from source records: ledger rows, bank lines, contracts, confirmations. Practice treats the link between a figure and its sources as documentation, a tick mark or a footnote. I argue that it should be treated as algebra. Using the provenance semirings of Green, Karvounarakis and Tannen, I annotate each source record with a token and show that a reported figure is literally a polynomial in those tokens, whose value on any subset of records is the figure that the subset would have produced. Three exact rules follow. First, if evidence grades form a finite chain, with alternatives combined by maximum and joint use by minimum, then the grade of any derived row is a weakest-link expression that can be computed alongside the query, and I characterise exactly which upgrades of source evidence can raise a finding's grade. Second, for an additive figure, dollar-weighted coverage c gives the sharp bound $\rho(1-c)$ times gross amount on the error contributed by untraced lines with relative tolerance ρ, and hence the coverage a materiality threshold requires. Third, recomputing only the figures whose lineage mentions a retracted record is sound, while the natural converse fails, and a modified record needs a two-sided test that the folklore rule omits. I add a standard hash-tree commitment that makes the source layer tamper evident, a closure property for review-ready findings, and a worked illustrative example. The results concern where a number came from and how strong its support is. They do not establish that any source record is true.

1 Introduction

A Quality of Earnings report is an argument made of numbers. Adjusted EBITDA is a sum of ledger lines and add-backs, and each add-back rests on documents. A buyer relies on the report because it cannot inspect the seller's records itself, the information asymmetry of Akerlof [1970] in concrete form. The reviewer's question of every figure is always the same: where did this come from, and how good is that source?

In practice the answer is recorded as documentation: a workpaper reference, a tick mark, a link to a file. Documentation of this kind does not compose. It cannot tell me the grade of a figure when two supported lines are joined with one unsupported assumption, how much of a total is actually traced, or which figures must be recomputed when a source row is

corrected. Those are questions about how annotations propagate through a computation, and that question has a known mathematical structure.

The structure comes from database theory. Cui et al. [2000] defined the lineage of a warehouse view tuple as the set of source tuples that produced it. Buneman et al. [2001] separated why-provenance, the source tuples that witness an output, from where-provenance, the locations from which an output value was copied. Green et al. [2007] showed that such schemes are instances of one construction: annotate source tuples with elements of a commutative semiring, read union and projection as addition and join as multiplication, and the most general annotation is a polynomial with natural-number coefficients in one variable per source tuple. Accounting has itself long been read algebraically, as in the group-theoretic ledger of Ellerman [1985]. This paper joins the two: the ledger is the database and the diligence figure is the query.

Section 2 recalls the semiring framework with proofs and adds an explicit modelling extension for additive figures, which the cited universality result does not cover. Sections 3 to 7 prove the results, and Section 8 works one illustrative example exactly. The paper contains no empirical results.

2 Setup: Annotated Ledgers and Figures as Polynomials

A *commutative semiring* $(K, +, \cdot, 0, 1)$ consists of commutative monoids $(K, +, 0)$ and $(K, \cdot, 1)$ such that multiplication distributes over addition and $0 \cdot a = 0$ for all a. A *homomorphism* of semirings preserves 0, 1, sums and products. Fix a domain $\mathbb{D}$ of values. A *tuple* over a finite attribute set U is a function $t : U \to \mathbb{D}$, and for $V \subseteq U$ I write $t|_V$ for its restriction.

Definition 2.1 (K-relation)**.** A *K-relation* over U is a function $R : \mathbb{D}^U \to K$ whose *support* $\operatorname{supp} R = \{t : R(t) \neq 0\}$ is finite. A *K-database* is a finite family of K-relations indexed by relation names.

Definition 2.2 (Positive relational algebra, following Green et al., 2007)**.** The queries of RA^+ are the expressions built from relation names by the following operations on K-relations: the *empty relation* $\emptyset(t) = 0$; *union* of R_1, R_2 over the same U, $(R_1 \cup R_2)(t) = R_1(t) + R_2(t)$; *projection* onto $V \subseteq U$, $(\pi_V R)(t) = \sum_{t' \in \operatorname{supp} R,\ t'|_V = t} R(t')$; *selection* by a predicate $P : \mathbb{D}^U \to \{0, 1\}$, $(\sigma_P R)(t) = R(t) \cdot P(t)$; *natural join* of R_i over U_i, $(R_1 \bowtie R_2)(t) = R_1(t|_{U_1}) \cdot R_2(t|_{U_2})$ for t over $U_1 \cup U_2$; and *renaming* of attributes along a bijection. For a query q and K-database D, $q(D)$ denotes the resulting K-relation.

The fragment is positive: there is no difference and no negation.

Lemma 2.3 (Homomorphisms commute with positive queries; Green et al., 2007)**.** *Let $h : K \to K'$ be a homomorphism of commutative semirings, and for a K-database D let $h(D)$ be the K'-database obtained by applying h to every annotation. Then for every $q \in \mathrm{RA}^+$ and every tuple t, $q(h(D))(t) = h\big(q(D)(t)\big)$.*

Proof. By induction on the structure of q. For a relation name the claim is the definition of $h(D)$, and $h(0) = 0$ handles the empty relation. Union and join follow from $h(a+b) = h(a) + h(b)$ and $h(ab) = h(a)h(b)$. Selection follows from $h(a \cdot 1) = h(a) \cdot 1$ and $h(a \cdot 0) = 0 = h(a) \cdot 0$.

Renaming only relabels tuples. For projection, write R for the value of the subquery on D, so that by the induction hypothesis its value on $h(D)$ is $h \circ R$. Then

$$h\Big(\sum_{t' \in \operatorname{supp} R,\ t'|_V = t} R(t')\Big) = \sum_{t' \in \operatorname{supp} R,\ t'|_V = t} h(R(t')) = \sum_{t' \in \operatorname{supp}(h \circ R),\ t'|_V = t} h(R(t')),$$

where the last step discards the terms with $R(t') \neq 0$ but $h(R(t')) = 0$, which contribute nothing, and uses $\operatorname{supp}(h \circ R) \subseteq \operatorname{supp} R$. □

Provenance polynomials. Let X be a finite set of *tokens*. A *monomial* is a function $m : X \to \mathbb{N}$, written $\prod_x x^{m(x)}$, and its token set is $\operatorname{tok}(m) = \{x : m(x) > 0\}$. The semiring $\mathbb{N}[X]$ consists of the finite formal sums $p = \sum_m c_m\, m$ with $c_m \in \mathbb{N}$, under the usual addition and multiplication of polynomials. I write $\operatorname{Mon}(p) = \{m : c_m > 0\}$ and $\operatorname{tok}(p) = \bigcup_{m \in \operatorname{Mon}(p)} \operatorname{tok}(m)$. For $c \in \mathbb{N}$ and k in a semiring K, $c \cdot k$ denotes the c-fold sum $k + \cdots + k$.

Lemma 2.4 (Universality of $\mathbb{N}[X]$)**.** *Let K be a commutative semiring and $v : X \to K$ any function. There is exactly one semiring homomorphism $\operatorname{Eval}_v : \mathbb{N}[X] \to K$ with $\operatorname{Eval}_v(x) = v(x)$ for all $x \in X$, namely*

$$\operatorname{Eval}_v\Big(\sum_m c_m\, m\Big) = \sum_m c_m \cdot \prod_{x \in X} v(x)^{m(x)}.$$

Proof. Write $v(m) = \prod_x v(x)^{m(x)}$. The displayed map sends 0 to 0 and the constant polynomial 1 (the monomial with all exponents zero) to 1, and it is additive because $(c + c') \cdot k = c \cdot k + c' \cdot k$ in any semiring. Commutativity of K gives $v(mm') = v(m)v(m')$, and distributivity gives $(c \cdot k)(c' \cdot k') = (cc') \cdot (kk')$, so for $p = \sum c_m m$ and $p' = \sum c'_{m'} m'$ one has $\operatorname{Eval}_v(pp') = \sum_{m,m'} (c_m c'_{m'}) \cdot v(m)v(m') = \operatorname{Eval}_v(p) \operatorname{Eval}_v(p')$. Uniqueness: a homomorphism that agrees with v on X must send m to $v(m)$ and $c_m\, m$, a c_m-fold sum, to $c_m \cdot v(m)$. □

A *ledger database* D is a finite family of finite sets of tuples, the *source records*. I assume each source record carries a record identifier, so that records are distinct tuples, and I assign to each a distinct token; X is the set of all tokens. The *tagged database* $\overline{D}$ is the $\mathbb{N}[X]$-database in which each source record is annotated with its token and every other tuple with 0.

Theorem 2.5 (Factorisation through $\mathbb{N}[X]$; Green et al., 2007)**.** *Let K be a commutative semiring, $v : X \to K$, and $v(\overline{D})$ the K-database in which each source record is annotated with the value of its token. For every $q \in \mathrm{RA}^+$ and every tuple t,*

$$q(v(\overline{D}))(t) = \operatorname{Eval}_v(q(\overline{D})(t)).$$

Proof. $v(\overline{D}) = \operatorname{Eval}_v(\overline{D})$, and Eval_v is a homomorphism by Lemma 2.4. Apply Lemma 2.3. □

I am careful about scope. Theorem 2.5 is the result of Green et al. [2007] for the positive relational algebra, and I use nothing beyond that fragment from the cited work. It says nothing, by itself, about the numerical value of a sum of amounts. That step is mine and is stated next as a modelling choice.

Definition 2.6 (Figure, rows, lineage)**.** A *figure* is a pair $f = (q, a)$ of a query $q \in \mathrm{RA}^+$ with output attributes U and an *amount function* $a : \mathbb{D}^U \to \mathbb{R}$ (typically the signed value of an amount attribute). On a ledger database D with tagged database $\overline{D}$, define the *row polynomials* $p_t = q(\overline{D})(t)$, the *rows* $\mathrm{rows}_f(D) = \{t : p_t \neq 0\}$, the *lineage* $\Lambda_f(D) = \bigcup_t \mathrm{tok}(p_t)$, and the *value*

$$f(D) = \sum_{t \in \mathrm{rows}_f(D)} a(t).$$

So a figure adds the amount of each distinct output row once. This is the modelling extension: the row annotations come from RA^+, and the sum is taken outside the semiring. For $S \subseteq X$ let D_S be the ledger database that keeps only the source records with tokens in S, and let $\nu_S : X \to \{0, 1\}$ be the indicator of S.

Lemma 2.7 (Rows of a sub-ledger)**.** *For every $S \subseteq X$ and tuple t: $t \in \mathrm{rows}_f(D_S)$ if and only if $p_t(\nu_S) \neq 0$, if and only if some $m \in \mathrm{Mon}(p_t)$ has $\mathrm{tok}(m) \subseteq S$.*

Proof. Let $h_S : \mathbb{N}[X] \to \mathbb{N}[X]$ be the homomorphism of Lemma 2.4 that fixes the tokens in S and sends the others to 0. Then $h_S(\overline{D})$ is the tagged database of D_S, so by Lemma 2.3 the row polynomial of t on D_S is $h_S(p_t)$, which is the sum of the terms of p_t whose monomials have $\mathrm{tok}(m) \subseteq S$. It is non-zero exactly when such a monomial exists. Finally $p_t(\nu_S) = \sum_m c_m\, m(\nu_S)$ with natural coefficients and $m(\nu_S) \in \{0, 1\}$, equal to 1 exactly when $\mathrm{tok}(m) \subseteq S$; a sum of natural numbers is non-zero exactly when some term is. □

Theorem 2.8 (A figure is a polynomial in its source records)**.** *For a figure f on D define the real polynomial*

$$F_f = \sum_{t \in \mathrm{rows}_f(D)} a(t)\Big(1 - \prod_{m \in \mathrm{Mon}(p_t)} \big(1 - \textstyle\prod_{x \in \mathrm{tok}(m)} x\big)\Big) \;\in\; \mathbb{R}[X].$$

Then $F_f(\nu_S) = f(D_S)$ for every $S \subseteq X$; in particular $F_f(1, \ldots, 1) = f(D)$. If every p_t is a single square-free monomial with coefficient one, then $F_f = \sum_t a(t)\, p_t$.

Proof. Fix S and a row t. Each factor $1 - \prod_{x \in \mathrm{tok}(m)} \nu_S(x)$ is 0 if $\mathrm{tok}(m) \subseteq S$ and 1 otherwise, so the bracket evaluates to 1 if some $m \in \mathrm{Mon}(p_t)$ has $\mathrm{tok}(m) \subseteq S$ and to 0 otherwise. By Lemma 2.7 that is the indicator of $t \in \mathrm{rows}_f(D_S)$. Tuples outside $\mathrm{rows}_f(D)$ are never rows of D_S, since $h_S(0) = 0$. Summing $a(t)$ over t gives $f(D_S)$. The last sentence is the case of a single factor. □

The polynomial F_f answers every deletion question about f at once: set to zero the variables of the records one wishes to disregard. If a figure is instead defined as the bag sum $\sum_t a(t)\, q(D)(t)$ over multiplicities, Theorem 2.5 with $v = \nu_S$ gives its value on D_S directly as $\big(\sum_t a(t)\, p_t\big)(\nu_S)$; I use the distinct-row convention because under bag semantics a line corroborated by two independent derivations would be counted twice.

3 The Evidence-Grade Semiring

Definition 3.1 (Evidence grades)**.** Let $\mathcal{G} = \{g_0 < g_1 < \cdots < g_n\}$ be a finite totally ordered set with $n \geq 1$, $\bot = g_0$ and $\top = g_n$. The running example is

unsupported $<$ management-asserted $<$ document-supported $<$ third-party-confirmed.

A *grade assignment* is a function $\gamma : X \to \mathcal{G}$.

Proposition 3.2. $(\mathcal{G}, \max, \min, \bot, \top)$ *is a commutative semiring in which both operations are idempotent. It has no zero divisors and no additive inverses of non-zero elements:* $\min(a, b) = \bot$ *implies* $a = \bot$ *or* $b = \bot$*, and* $\max(a, b) = \bot$ *implies* $a = b = \bot$.

Proof. On a chain, max and min are associative, commutative and idempotent, with $\max(a, \bot) = a$ and $\min(a, \top) = a$, so both are commutative monoids with the stated units, and $\min(\bot, a) = \bot$. For distributivity, take a, b, c with $b \leq c$ without loss of generality. Then $\min(a, \max(b, c)) = \min(a, c)$, and since $\min(a, b) \leq \min(a, c)$ also $\max(\min(a, b), \min(a, c)) = \min(a, c)$. The last two claims hold because in a total order $\min(a, b)$ and $\max(a, b)$ are each one of a, b. □

Addition is max: among alternative derivations of a row the best sets its grade. Multiplication is min: among records used jointly the weakest does. The zero $\bot$ means that a record carries no evidential weight and is treated as absent, the reading of zero in every K-relation.

Theorem 3.3 (Weakest link)**.** *Let* γ *be a grade assignment and write* $\mathrm{Eval}_\gamma : \mathbb{N}[X] \to \mathcal{G}$ *for the homomorphism of Lemma 2.4.*

(a) For every $p \in \mathbb{N}[X]$*,*

$$\mathrm{Eval}_\gamma(p) = \max_{m \in \mathrm{Mon}(p)} \min_{x \in \mathrm{tok}(m)} \gamma(x),$$

with the conventions $\max \emptyset = \bot$ *and* $\min \emptyset = \top$*. Coefficients and exponents do not affect the value.*

(b) For every $q \in \mathrm{RA}^+$ *and tuple* t*, the grade obtained by running* q *directly on the graded database* $\gamma(\overline{D})$ *equals the evaluated polynomial:* $q(\gamma(\overline{D}))(t) = \mathrm{Eval}_\gamma\big(q(\overline{D})(t)\big)$.

(c) $\mathrm{Eval}_\gamma(p) = \bot$ *if and only if every monomial of* p *contains a token of grade* $\bot$*. In particular, if* $\gamma(x) > \bot$ *for all* x*, then* $\mathrm{Eval}_\gamma(p) = \bot$ *only for* $p = 0$.

Proof. (a) By Lemma 2.4 with the operations of $\mathcal{G}$, $\mathrm{Eval}_\gamma(\sum c_m m) = \max_{m \in \mathrm{Mon}(p)} c_m \cdot \gamma(m)$, where $c_m \cdot g$ is a c_m-fold maximum of g, equal to g by idempotence, and $\gamma(m)$ is the minimum of $\gamma(x)$ taken $m(x)$ times for each x, equal to $\min_{x \in \mathrm{tok}(m)} \gamma(x)$ by idempotence. Terms with $c_m = 0$ contribute $\bot$, the unit of max. (b) This is Theorem 2.5 with $K = \mathcal{G}$, which applies because $\mathcal{G}$ is a commutative semiring (Proposition 3.2) and q is positive. (c) By (a) and Proposition 3.2, the maximum is $\bot$ exactly when every inner minimum is $\bot$, and a minimum over a chain is $\bot$ exactly when one argument is. The empty monomial has minimum $\top \neq \bot$. □

Definition 3.4 (Grade of a figure)**.** The *row grade* of $t \in \mathrm{rows}_f(D)$ is $\mathrm{gr}_\gamma(t) = \mathrm{Eval}_\gamma(p_t)$. The *grade of the figure* is

$$\mathrm{gr}_\gamma(f) = \min_{t \in \mathrm{rows}_f(D)} \mathrm{gr}_\gamma(t) = \mathrm{Eval}_\gamma(J_f), \qquad J_f = \prod_{t \in \mathrm{rows}_f(D)} p_t.$$

A sum uses all of its rows jointly, so its lineage polynomial J_f is the product of the row polynomials, not their sum: annotating an aggregate with the formal sum of its rows' annotations would, under $(\max, \min)$, give a total the grade of its best line. A figure with no rows has $J_f = 1$ and grade $\top$ vacuously; Definition 7.1 excludes that case.

Proposition 3.5 (Figure grades computed in the database)**.** *If $\gamma(x) > \bot$ for every $x \in X$, then* $\operatorname{supp} q(\gamma(\overline{D})) = \operatorname{rows}_f(D)$ *and* $\operatorname{gr}_\gamma(f) = \min_{t \in \operatorname{supp} q(\gamma(\overline{D}))} q(\gamma(\overline{D}))(t)$*. Without the hypothesis the two sides can differ.*

Proof. By Theorem 3.3(b), $q(\gamma(\overline{D}))(t) = \operatorname{Eval}_\gamma(p_t)$, and by (c) this is $\bot$ exactly when $p_t = 0$. So the supports agree and the minima are over the same values. For the failure, let q return a single relation whose only record has token x with $\gamma(x) = \bot$. Then $\operatorname{rows}_f(D)$ has one row and $\operatorname{gr}_\gamma(f) = \bot$, while $q(\gamma(\overline{D}))$ has empty support and the minimum over it is $\top$. □

Corollary 3.6 (Monotonicity)**.** *If $\gamma \le \gamma'$ pointwise, then* $\operatorname{Eval}_\gamma(p) \le \operatorname{Eval}_{\gamma'}(p)$ *for all $p \in \mathbb{N}[X]$. Hence upgrading source records never lowers the grade of any row or figure.*

Proof. max and min are monotone; apply Theorem 3.3(a) and Definition 3.4. □

Corollary 3.7 (Which upgrades raise a grade)**.** *Let $p \neq 0$ with $g = \operatorname{Eval}_\gamma(p) < \top$, and let $\gamma' \ge \gamma$ pointwise. Then* $\operatorname{Eval}_{\gamma'}(p) > g$ *if and only if there is a monomial $m \in \operatorname{Mon}(p)$ such that $\gamma'(x) > g$ for every $x \in \operatorname{tok}(m)$ with $\gamma(x) \le g$. Consequently:*

(i) upgrading a single token x raises the grade of p if and only if $\gamma'(x) > g$ and some $m \in \operatorname{Mon}(p)$ has x as its only token of grade at most g; in particular a token with $\gamma(x) > g$ can never help;

(ii) for a figure with $\operatorname{gr}_\gamma(f) = g < \top$, $\operatorname{gr}_{\gamma'}(f) > g$ if and only if the condition above holds for p_t for every *row t with $\operatorname{gr}_\gamma(t) = g$.*

Proof. By Theorem 3.3(a), $\operatorname{Eval}_{\gamma'}(p) > g$ exactly when some $m \in \operatorname{Mon}(p)$ has $\gamma'(x) > g$ for all $x \in \operatorname{tok}(m)$. Tokens with $\gamma(x) > g$ satisfy this automatically since $\gamma' \ge \gamma$, which leaves the stated condition on the tokens with $\gamma(x) \le g$. For (i), let γ' differ from γ only at x. Every $m \in \operatorname{Mon}(p)$ has at least one token of grade at most g, because $\min_{y \in \operatorname{tok}(m)} \gamma(y) \le g$ and $\operatorname{tok}(m) \neq \emptyset$ (an empty monomial would give $g = \top$). The condition then requires a monomial whose tokens of grade at most g all have $\gamma' > g$; only x changed, so x must be the only such token and $\gamma'(x) > g$. For (ii), $\operatorname{gr}_{\gamma'}(f)$ is the minimum of the row grades, rows with $\operatorname{gr}_\gamma(t) > g$ stay above g by Corollary 3.6, and the minimum exceeds g exactly when each row at grade g rises. □

The two quantifiers are the practical content: across alternative derivations of one row it suffices to clear the bottleneck of *one* monomial; across the rows of a sum, *every* row at the minimum must be raised. The simpler conjecture, that the upgraded token must attain the minimum in every maximal monomial, is false: for $p = x + y$ with both tokens at grade g, upgrading x alone raises the grade.

4 Dollar-Weighted Coverage

The grade of a figure is deliberately harsh: one weak line of trivial size pulls the whole figure down. Coverage is the quantitative companion. Let the rows of an additive figure have reported amounts $a_1, \dots, a_n$ with $f = \sum_i a_i$ and gross amount $A = \sum_i |a_i| > 0$, and let the true amounts be a_i^* with $f^* = \sum_i a_i^*$.

Theorem 4.1 (Tolerance bound)**.** *If $|a_i^* - a_i| \le \rho_i |a_i|$ with $\rho_i \ge 0$ for each i, then*

$$|f^* - f| \;\le\; \sum_{i=1}^{n} \rho_i |a_i|,$$

and for every choice of the a_i and ρ_i there are admissible a_i^* attaining equality.

Proof. $|f^* - f| = |\sum_i (a_i^* - a_i)| \le \sum_i |a_i^* - a_i| \le \sum_i \rho_i |a_i|$ by the triangle inequality. Taking $a_i^* = a_i + \rho_i |a_i|$ for all i satisfies the hypothesis and gives $f^* - f = \sum_i \rho_i |a_i|$. □

Definition 4.2 (Coverage)**.** For a set $T \subseteq \{1, \dots, n\}$ of *traced* rows, the *dollar-weighted coverage* is $c = \sum_{i \in T} |a_i| / A \in [0, 1]$. A natural choice is $T = \{i : \mathrm{gr}_\gamma(t_i) \ge \theta\}$ for a grade threshold θ.

Corollary 4.3 (Coverage bound)**.** *If traced rows are exact ($a_i^* = a_i$ for $i \in T$) and each untraced row has relative error at most ρ ($|a_i^* - a_i| \le \rho |a_i|$ for $i \notin T$), then*

$$|f^* - f| \;\le\; \rho\,(1 - c)\,A,$$

with equality attained when every untraced row is misstated by $+\rho |a_i|$.

Proof. Theorem 4.1 with $\rho_i = 0$ on T and $\rho_i = \rho$ off T, using $\sum_{i \notin T} |a_i| = (1 - c)A$. □

For instance, with $a_1 = 90$ traced, $a_2 = 10$ untraced and $\rho = 0.2$, the admissible truth $a_2^* = 12$ gives $|f^* - f| = 2 = 0.2 \times 0.1 \times 100$.

The bound is in terms of the gross amount A, not the net figure. Relative to the figure it is $\rho(1 - c)\kappa$ with $\kappa = A/|f| \ge 1$ the gross-to-net ratio (for $f \ne 0$): a margin built from large offsetting lines needs more coverage than its net size suggests.

Corollary 4.4 (Coverage required by materiality)**.** *Let $M > 0$ be a materiality threshold and $\rho > 0$. Under the hypotheses of Corollary 4.3, the guarantee $|f^* - f| \le M$ holds for all admissible truths if and only if*

$$c \;\ge\; c^* = 1 - \frac{M}{\rho A}.$$

If $M = \mu |f|$ with $f \ne 0$, this reads $c \ge 1 - \mu/(\rho\kappa)$, and no tracing is needed when $\mu \ge \rho\kappa$.

Proof. If $c \ge c^*$ then $\rho(1 - c)A \le M$ and Corollary 4.3 applies. If $c < c^*$ then $\rho(1 - c)A > M$ and the extremal truth of Corollary 4.3 violates the guarantee. Substituting $A = \kappa |f|$ gives the second form. □

When rows differ in tracing cost, choosing what to trace is a knapsack problem [Dantzig, 1957]; when untraced rows are sampled, the deterministic bound gives way to the probabilistic ones of audit sampling [International Auditing and Assurance Standards Board, 2009d] and concentration inequalities [Hoeffding, 1963]. I pursue neither. Figure 1 plots the bound.

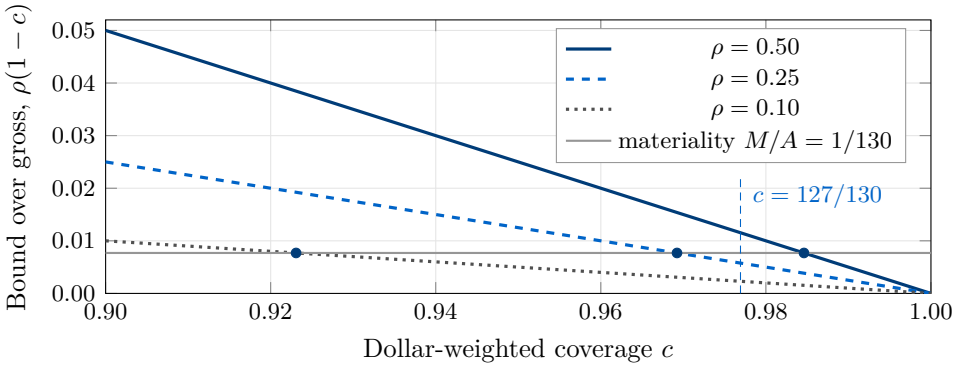


Figure 1: The coverage bound $\rho(1-c)$ of Corollary 4.3 for three tolerances, with the materiality line of the example in Section 8 ($M/A = 1/130$). Each curve meets the line at the required coverage c^* of Corollary 4.4: 64/65, 63/65 and 12/13 for $\rho = 0.50$, 0.25 and 0.10. The example's coverage 127/130 suffices at $\rho = 0.25$ but not at $\rho = 0.50$.

5 Sound Incremental Invalidation

Source records get corrected during an engagement. The question is which figures must be recomputed. I model immutably: a record is never edited in place; it is *retracted* (removed), or *replaced* by a new record that carries a fresh token.

Theorem 5.1 (Sound invalidation)**.** *Let $f = (q, a)$ be a figure on a ledger database D with token set X, and let $x \in X$.*

(a) Retraction. *If $x \notin \Lambda_f(D)$, then retracting the record of x changes no row polynomial of f; hence its rows, its value, and its grade under every grade assignment are unchanged.*

(b) Replacement. *Let D' be D with the record of x replaced by a record with fresh token x'. If $x \notin \Lambda_f(D)$ and $x' \notin \Lambda_f(D')$, then f has the same row polynomials on D' as on D, with the same consequences.*

(c) Regrading. *If γ and γ' agree on $\Lambda_f(D)$, then $\mathrm{gr}_\gamma(f) = \mathrm{gr}_{\gamma'}(f)$ and all row grades agree.*

(d) The condition $x \notin \Lambda_f(D)$ alone does not suffice in (b).

Hence it is sound to recompute exactly the figures whose lineage mentions a retracted or regraded token, and for a replacement those whose old lineage mentions x or whose new lineage mentions x'.

Proof. (a) Let $h_x : \mathbb{N}[X] \to \mathbb{N}[X]$ fix every token except x and send x to 0. Then $h_x(\overline{D})$ is the tagged database of D without the record of x, and by Lemma 2.3 the row polynomial of t after retraction is $h_x(p_t)$. If $x \notin \mathrm{tok}(p_t)$ then $h_x(p_t) = p_t$. Rows, value and grades are functions of the row polynomials (Definitions 2.6 and 3.4). (b) Let D_0 be D without the record of x, which is also D' without the record of x'. By (a) applied to D and x, the row polynomials on D equal those on D_0. By (a) applied to D' and x', the row polynomials on

D' equal those on D_0. (c) By Theorem 3.3(a), $\mathrm{Eval}_\gamma(p_t)$ depends only on the values of γ on $\mathrm{tok}(p_t) \subseteq \Lambda_f(D)$. (d) Let $q = \sigma_P(R)$ select the rows dated within a period, and let the record of x be dated outside it, so $x \notin \Lambda_f(D)$. Replace it by a record with the same amount $a_0 \neq 0$ dated inside the period. The new record is a row of f on D' and the value changes by a_0. □

Part (d) corrects the rule as usually stated, "recompute what mentions the changed record": a replacement also inserts, and an insertion can create derivations that no index built on the old state knows about.

The converse fails. Lineage-based invalidation is sound but not necessary: a figure may mention x and not change when x is retracted. Three ordinary mechanisms produce this. *Zero amounts*: a row t with $x \in \mathrm{tok}(p_t)$ and $a(t) = 0$ contributes nothing whether present or not. *Offsetting rows*: let an invoice of $+500$ (token r_1) and its reversal of -500 (token r_2) both be joined to a vendor master record (token x). The row polynomials are xr_1 and xr_2, so $F_f = 500\,xr_1 - 500\,xr_2$; retracting x removes both rows and the value is 0 before and after. *Redundant derivations*: if $p_t = x + y$, the row survives the retraction of x because y still derives it; here $F_f = a(t)(x + y - xy)$, which at $y = 1$ equals $a(t)$ for both values of x.

6 Tamper Evidence for the Source Layer

Lineage presumes that the record a token resolves to today is the record that was there when the figure was computed. The standard device is a hash tree [Merkle, 1988], stated here as a reduction so that the cryptographic assumption is explicit.

Let $H : \{0,1\}^* \to \{0,1\}^\lambda$ be a hash function. A *collision* is a pair $u \neq u'$ with $H(u) = H(u')$. Let $r = (r_1, \ldots, r_n)$ be the source records as bit strings, with $n = 2^d$ (pad with a fixed empty record). Define leaf values $H(0 \,\|\, r_i)$, internal values $H(1 \,\|\, v_L \,\|\, v_R)$ from the values of the two children, and $\mathrm{root}(r)$ as the value at the root. The commitment is the pair $(n, \mathrm{root}(r))$. A *membership proof* for position i is the list $s_1, \ldots, s_d$ of sibling values along the path from leaf i to the root. The verifier, given (n, w), i, a candidate record r' and the proof, computes $v_0 = H(0 \,\|\, r')$ and $v_k = H(1 \,\|\, v_{k-1} \,\|\, s_k)$ or $H(1 \,\|\, s_k \,\|\, v_{k-1})$ according to the k-th bit of i, and accepts if and only if $v_d = w$.

Proposition 6.1 (Binding by reduction)**.** *(a) If $r \neq r'$ are sequences of the same length $n = 2^d$ with $\mathrm{root}(r) = \mathrm{root}(r')$, then a collision of H occurs among the inputs hashed in building the two trees, and it is found with at most $2(2n - 1)$ evaluations of H.*

(b) If the verifier accepts, against $(n, \mathrm{root}(r))$ and position i, a record $r' \neq r_i$ with some proof $s'_1, \ldots, s'_d$, then a collision of H occurs among the $2(d + 1)$ inputs hashed along the honest and the claimed paths.

(c) A membership proof consists of $d = \log_2 n$ hash values, that is $\lambda \log_2 n$ bits.

Proof. (a) By induction on d. For $d = 0$, $H(0 \,\|\, r_1) = H(0 \,\|\, r'_1)$ with $r_1 \neq r'_1$ is a collision. For $d > 0$, the roots are $H(1 \,\|\, v_L \,\|\, v_R)$ and $H(1 \,\|\, v'_L \,\|\, v'_R)$. If $(v_L, v_R) \neq (v'_L, v'_R)$ the two inputs differ, since all values have the fixed length λ, and they collide. Otherwise $v_L = v'_L$ and

$v_R = v'_R$, and since $r \neq r'$ the sequences differ in their left halves or in their right halves; the induction hypothesis applied to that half gives a collision. Each tree has $2n - 1$ nodes. (b) Let $v_0, \ldots, v_d$ be the honest path values for r_i and $v'_0, \ldots, v'_d$ the values the verifier computes, so $v'_d = v_d$. If $v'_0 = v_0$ then $0 \,\|\, r'$ and $0 \,\|\, r_i$ collide. Otherwise let $k \geq 1$ be least with $v'_k = v_k$. Then $v'_{k-1} \neq v_{k-1}$, and the two inputs hashed at level k place these values in the same slot (determined by i), so the inputs differ and collide. (c) The path from a leaf to the root of a complete binary tree with 2^d leaves has d edges and one sibling per edge. □

Assumption (collision resistance). No computationally feasible procedure outputs a collision of H except with negligible probability in λ. This is an assumption about the hash function, not a theorem. Under it, any feasible procedure that alters a committed source record without changing the root, or that convinces a verifier that a different record sits at a committed position, would produce a collision by Proposition 6.1, so it succeeds only with negligible probability. Two caveats: the commitment says nothing about a record that was wrong before it was committed, and the root must be held out of reach of whoever could alter the records.

7 Review-Ready Findings

A *resolver* is a partial function from tokens to retrievable source rows; write $X_{\text{res}} \subseteq X$ for its domain. Fix a grade assignment γ and a threshold $\theta > \bot$.

Definition 7.1 (Review-ready)**.** A polynomial $p \in \mathbb{N}[X]$ is *review-ready* if (i) $p \neq 0$, (ii) $\text{tok}(p) \subseteq X_{\text{res}}$, and (iii) $\text{Eval}_\gamma(p) \geq \theta$. An $\mathbb{N}[X]$-relation is review-ready if the annotation of every tuple in its support is. A *finding*, that is a figure f presented for review, is review-ready if $\text{rows}_f(D) \neq \emptyset$ and J_f is review-ready; equivalently, if it has at least one row and every row polynomial is review-ready.

The equivalence holds because $\mathbb{N}[X]$ has no zero divisors, $\text{tok}(J_f) = \Lambda_f(D)$, and $\text{Eval}_\gamma(J_f) = \min_t \text{Eval}_\gamma(p_t)$.

Proposition 7.2 (Closure and checking)**.** *(a) A source record with token x is review-ready if and only if $x \in X_{\text{res}}$ and $\gamma(x) \geq \theta$.*

(b) If every relation of an $\mathbb{N}[X]$-database is review-ready, then so is q of it, for every $q \in \text{RA}^+$. Hence a figure with at least one row, computed by a positive query over review-ready relations, is a review-ready finding.

(c) Given the row polynomials as lists of monomials, with constant-time lookup of γ and of membership in X_{res}, review-readiness of a finding can be decided in time $O(\sum_t |p_t|)$, where $|p|$ is the number of monomials of p plus the total number of token occurrences in them.

Proof. (a) Immediate from the definition with $p = x$. (b) By induction on q; it suffices that each operation preserves the property on supports. The empty relation has empty support. For union and projection, an output annotation in the support is a finite sum of input annotations, of which the non-zero ones are review-ready and at least one is non-zero; a

sum of polynomials with natural coefficients, one of them non-zero, is non-zero, its tokens are the union of the summands' tokens, and its grade is the maximum of the summands' grades, at least θ. For selection the annotation is unchanged or zero. For join, a non-zero output annotation is a product $p_1 p_2$ of non-zero review-ready annotations; it is non-zero, $\mathrm{tok}(p_1 p_2) = \mathrm{tok}(p_1) \cup \mathrm{tok}(p_2)$, and its grade is $\min(\mathrm{Eval}_\gamma(p_1), \mathrm{Eval}_\gamma(p_2)) \geq \theta$. Renaming relabels tuples. The statement about figures follows from the remark after Definition 7.1. (c) For each row, one pass over the monomials tests non-emptiness of $\mathrm{Mon}(p_t)$, tests each token occurrence for membership in X_{res}, and computes $\max_m \min_x \gamma(x)$ by Theorem 3.3(a); then compare with θ. Each token occurrence and each monomial is visited once. □

Size. I should be modest about the size of these polynomials. For a fixed query a row polynomial has polynomially many monomials in the size of the database, with a degree that grows with the number of joins, and the expanded J_f can have exponentially many, which is why the check works row by row. The standard remedies are to store a shared circuit of sums and products, or only what a purpose needs: grades can be computed inside the query (Theorem 3.3(b)), and for invalidation the token set Λ_f, the lineage of Cui et al. [2000], suffices.

8 An Illustrative Worked Example

The following example is illustrative: the company, records and amounts are invented to exercise the definitions, and every number is exact. An adjusted EBITDA figure is built from six source records through four ledger lines (Table 1 and Figure 2). Grades are abbreviated U, M, D, C for the chain of Definition 3.1. Three records come from the company's own system and are graded D: the general-ledger revenue detail x_1, the general-ledger expense detail x_3 and the payroll register x_5. Two come from outside parties and are graded C: a bank confirmation of deposits x_2 and a settlement agreement supplied by the counterparty's counsel x_4. One is a management assertion of a market-rate salary, x_6, graded M.

Table 1: Illustrative ledger lines of the adjusted EBITDA figure (USD); share is $|a_i|/A$.

Line	Description	Amount a_i	Polynomial p_t	Grade	Share
ℓ_1	Revenue	4,200,000	$x_1 + x_2$	C	7/13
ℓ_2	Operating expenses	−3,300,000	x_3	D	11/26
ℓ_3	Add-back: legal settlement	120,000	$x_3 x_4$	D	1/65
ℓ_4	Add-back: owner compensation	180,000	$x_5 x_6$	M	3/130
	Adjusted EBITDA f	1,200,000	J_f	M	

Polynomials. The revenue line is derived independently from the ledger and from the bank confirmation, so $p_{\ell_1} = x_1 + x_2$. Each add-back joins the ledger entry it adjusts with the record that justifies the adjustment. By Theorem 2.8,

$$F_f = 4{,}200{,}000\,(x_1 + x_2 - x_1 x_2) - 3{,}300{,}000\,x_3 + 120{,}000\,x_3 x_4 + 180{,}000\,x_5 x_6,$$

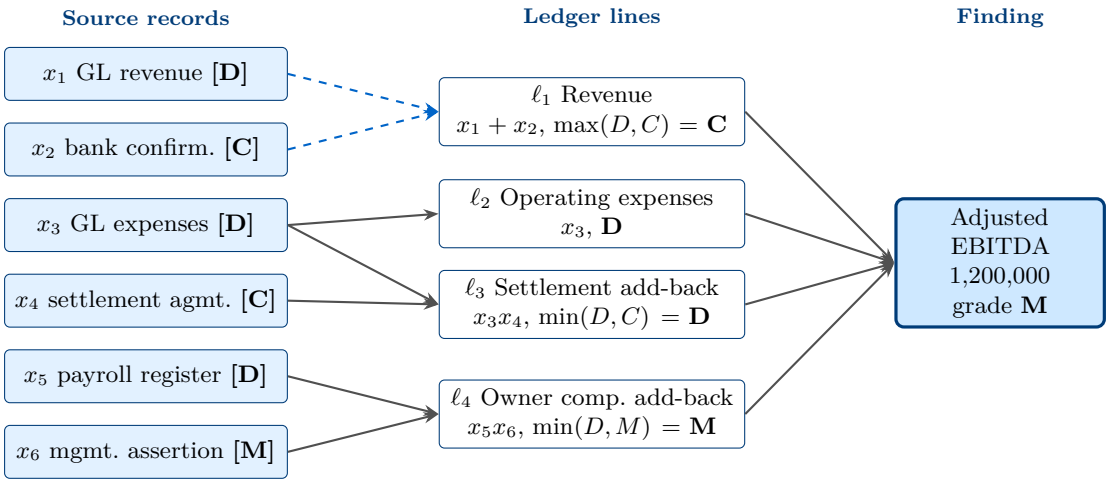


Figure 2: Lineage graph of the illustrative example. Dashed arrows are alternative derivations of one line (sum, graded by maximum); solid arrows into a line are joint use (product, graded by minimum); the finding uses all four lines jointly. The single management-asserted record x_6 sets the grade of the finding.

and $F_f(1, \ldots, 1) = 4{,}200{,}000 - 3{,}300{,}000 + 120{,}000 + 180{,}000 = 1{,}200{,}000$. Retracting x_2 leaves the value unchanged (a redundant derivation). Retracting x_3 removes both ℓ_2 and ℓ_3 and gives $4{,}200{,}000 + 180{,}000 = 4{,}380{,}000$. The joint lineage polynomial is

$$J_f = (x_1 + x_2) \cdot x_3 \cdot x_3 x_4 \cdot x_5 x_6 = x_1 x_3^2 x_4 x_5 x_6 + x_2 x_3^2 x_4 x_5 x_6.$$

Grade. By Theorem 3.3(a), $\mathrm{gr}_\gamma(f) = \max\big(\min(D, D, C, D, M),\ \min(C, D, C, D, M)\big) = M$, the minimum of the line grades C, D, D, M. By Corollary 3.7, x_6 is the only token of grade at most M in either monomial, so upgrading x_6 alone to D raises the finding to D, and no other single upgrade raises it at all. To reach C, every line at grade D must then rise, which requires x_3, x_5 and x_6 at C. Upgrading x_1 never matters while x_2 is at C.

Coverage and error bound. With threshold $\theta = \mathrm{D}$ the traced set is $T = \{\ell_1, \ell_2, \ell_3\}$ and

$$c = \frac{4{,}200{,}000 + 3{,}300{,}000 + 120{,}000}{7{,}800{,}000} = \frac{7{,}620{,}000}{7{,}800{,}000} = \frac{127}{130} \approx 0.9769.$$

The only untraced line is ℓ_4. With tolerance $\rho = 0.25$ on it, Corollary 4.3 bounds the error by $0.25 \times \frac{3}{130} \times 7{,}800{,}000 = 45{,}000$. Take materiality at five per cent of the figure, $M = 60{,}000$. The gross-to-net ratio is $\kappa = 7{,}800{,}000/1{,}200{,}000 = 6.5$, and Corollary 4.4 requires $c \geq 1 - 60{,}000/(0.25 \times 7{,}800{,}000) = 1 - 2/65 = 63/65 \approx 0.9692$. Since $127/130 > 126/130$, the coverage suffices. At $\rho = 0.50$ the bound is $90{,}000 > 60{,}000$ and the required coverage $64/65 = 128/130$ is not met: the conclusion depends on the tolerance one is prepared to defend for a management assertion.

9 Relation to Audit-Evidence Standards

The ordering of grades by source is not my invention. Auditing standards state, as generalisations, that evidence from independent sources outside the entity is generally more reliable than evidence generated internally, that evidence obtained directly is generally more reliable than evidence obtained indirectly or by inference, and that documentary evidence is generally more reliable than oral representation [International Auditing and Assurance Standards Board, 2009b, Public Company Accounting Oversight Board, 2010]. External confirmation has its own standard [International Auditing and Assurance Standards Board, 2009c], and materiality, my threshold M, is a planning concept in International Auditing and Assurance Standards Board [2009a].

I want to be exact about the relationship. The standards describe general tendencies subject to exceptions; they do not define a four-point scale or prescribe combining evidence by maximum and minimum, and they govern audits, which a diligence engagement is not. The chain of Definition 3.1 is my ordinal model, consistent in direction with those generalisations; the theorems hold for any finite chain.

10 Implications for Practice

Finsider keeps immutable original source files and traces reported figures to source rows; the token and the resolver of this paper abstract that design. Finsider also computes a deterministic core from ledger data and keeps any figure extracted by an AI layer separate, labelled as needing verification, so that it never feeds headline metrics until verified. In the language of this paper such a figure behaves like a token below the threshold θ: by the weakest-link theorem it would cap the grade of everything joined to it. Finally, a company's results are held behind a release gate until verification passes, with human reviewers in the loop. Definition 7.1 is a candidate formal predicate for the mechanical part of such a gate; it does not describe the gate itself, and it does not replace the reviewer, whose judgment supplies the grades.

11 Falsifiable Predictions

Whether the model is useful is an empirical matter. I commit to four predictions about what a future study on real engagement data would find. None is a claimed result.

Prediction 1 (Bottleneck concentration)**.** Among findings graded below document-supported, at most ten per cent of the tokens in their lineage are bottlenecks in the sense of Corollary 3.7(i). *Refuted if* in a sample of at least twenty engagements the median share of bottleneck tokens exceeds twenty-five per cent.

Prediction 2 (Selective invalidation)**.** When a single source record is retracted, the figures whose lineage mentions it are a small fraction of all reported figures. *Refuted if* the median fraction invalidated per retraction exceeds one half across engagements with at least one hundred reported figures each.

Prediction 3 (Slack in the coverage bound)**.** When untraced lines are later traced, the realised error of a figure is typically well inside the bound of Corollary 4.3, because errors on different lines do not all share a sign. *Refuted if* the realised error exceeds half of the bound in a majority of figures for which the per-line tolerance ρ held.

Prediction 4 (Grades order revision risk)**.** The rate at which a ledger line is revised between a draft and a final report decreases weakly with its row grade. *Refuted if* lines graded third-party-confirmed are revised at a higher rate than lines graded management-asserted in a sample large enough to separate the two rates at the five per cent significance level.

12 Limitations

Provenance is not truth. Lineage establishes where a number came from, not that the source is truthful. A fabricated invoice has impeccable lineage, and a hash commitment faithfully preserves a false record. "Traced lines are exact" in Corollary 4.3 means exact agreement with the source. A lineage graph that renders well also invites the over-reliance on automation documented by Parasuraman and Riley [1997]: a review-ready flag states that support is traceable and graded, not that the finding is right.

Grades are judgments. The chain $\mathcal{G}$ is ordinal, and assigning a grade to a record is a human or automated classification that can be wrong. Because max and min use only the order, the model cannot express that two independent document-supported derivations together deserve more confidence than one. Corroboration in that stronger sense needs a probabilistic model and independence assumptions that I have not made.

The aggregation step is a modelling choice. The cited universality result covers positive relational algebra. My treatment of sums (distinct rows, value outside the semiring, joint lineage as a product) is one consistent choice, and by Proposition 3.5 it commutes with direct computation only when no token has grade $\bot$. Difference, and aggregates other than sums, are not covered.

Resolution can fail. I assumed that a token resolves to one source row. In practice linking a ledger line to the document that supports it is a record-linkage problem with its own error rates [Fellegi and Sunter, 1969], and a wrong link gives a clean lineage to the wrong evidence. The tolerance ρ is likewise an input that nothing here estimates.

13 Conclusion

Once source records are tokens, a figure is a polynomial, and the questions reviewers ask have exact answers: the grade of a finding is a weakest-link expression; the effective upgrades are the bottlenecks of one derivation of every weakest row; the error left by untraced lines is at most $\rho(1 - c)$ of the gross amount, and no better bound holds; retractions invalidate along lineage, soundly but not necessarily, and replacements need a second test. None of this makes a source record true. It makes precise what is being relied on, and how much.

References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Database Theory: ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pages 316–330. Springer, 2001.

Yingwei Cui, Jennifer Widom, and Janet L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Transactions on Database Systems*, 25(2):179–227, 2000.

George B. Dantzig. Discrete-variable extremum problems. *Operations Research*, 5(2):266–288, 1957.

David P. Ellerman. The mathematics of double entry bookkeeping. *Mathematics Magazine*, 58(4):226–233, 1985.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 31–40, 2007.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

International Auditing and Assurance Standards Board. International standard on auditing 320: Materiality in planning and performing an audit. IFAC, 2009a.

International Auditing and Assurance Standards Board. International standard on auditing 500: Audit evidence. IFAC, 2009b.

International Auditing and Assurance Standards Board. International standard on auditing 505: External confirmations. IFAC, 2009c.

International Auditing and Assurance Standards Board. International standard on auditing 530: Audit sampling. IFAC, 2009d.

Ralph C. Merkle. A digital signature based on a conventional encryption function. In *Advances in Cryptology: CRYPTO '87*, volume 293 of *Lecture Notes in Computer Science*, pages 369–378. Springer, 1988.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Public Company Accounting Oversight Board. AS 1105: Audit evidence. PCAOB Auditing Standards, 2010.