Counting in Token Space: Finite Witnesses, Linear Tests and Regular Heaps

Wuming Pan

8 November 2026 (V13)

Abstract

Token Space is a categorical framework motivated by the structural understanding of large language models (LLMs) and the development of related computational mechanisms. Its classes have carriers, fixed core alphabets and word heaps; maps substitute carrier letters while preserving heaps. We study restricted homomorphism counts and witness resources. For core-free r-ary targets on n letters, a decoder gives the linear-test bound $n + (r-1)\max(t, t') \le rn$, where t, t' count twin classes; linear tests forbid repeated letters within records. Qualitative completeness follows from existing weighted-graph and simple-#CSP theorems; our contribution is the source-variable bound for 0/1 relations. Standard directed cycles give $\lfloor n/2 \rfloor \le b_{\mathfrak{L}}(n, 2) \le 2n$ for $n \ge 4$, hence linear growth. An exact search over 113 230 six-letter types finds two complement pairs agreeing on all five-letter linear tests but separated by one six-letter oriented path, proving $b_{\mathfrak{L}}(6, 2) \ge 6$, not its exact value. Complementation extends the classification to a specified 130 164-type family. Exact integer GPU computations are independently checked on CPU. For arbitrary heaps, finite witnesses use at most n letters and $n!$ records; automata with m, m' states give record length below mm'. For induced counting on regular heaps, the exact test-state threshold is m. We also give a one-pattern reconstruction guarantee and organize kernel and shape information by classical Möbius inversion. These results provide a precise method of structural observation with potential uses in understanding relational information expressed by LLMs; no LLM application effects are established here.

Keywords: homomorphism counts; relational identification; Token Space; linear tests; regular languages; Möbius inversion.

1 The problem and the results

The value of Token Space lies in its choice of basic objects and operations. It brings variable repetition, fixed symbols, different word shapes, substitution, quotients and regular representations into a common mathematical framework, making the resources required to observe a structure a central question throughout this paper. By organizing mathematical concepts and methods for Token structures, it provides a unified and practical language for structural description, counting analysis and the study of computational mechanisms.

The framework is designed to support the structural understanding of LLMs and the development of related computational mechanisms. Encoding shapes as relation symbols connects it to relational and constraint-counting theory. Applications to LLM-expressed relations remain potential uses requiring a suitable observation interface and empirical validation.

Token Space [1] studies classes $T = \langle B, E, \Upsilon \rangle$, with $\Upsilon \subseteq F_E(B) = (B \sqcup E)^*$, and maps $f : B_1 \to B_2$ satisfying $F_E(f)(\Upsilon_1) \subseteq \Upsilon_2$. Its category $\mathbf{TS}_E$ fixes the objects, substitutions and operations within which we ask counting questions. We study

$$\hom(F, G) = |\operatorname{Hom}_{\mathbf{TS}_E}(F, G)|$$

as observations of a target G, with restrictions on the test F and separate budgets for its carrier, records and presentation.

Problem 1.1 (Restricted identification)**.** Let $\mathcal{K}$ be a class of *targets* and $\mathfrak{F}$ a class of *tests*, both with finite carriers. Write $G \equiv_{\mathfrak{F}} G'$ if $\hom(F, G) = \hom(F, G')$ for all $F \in \mathfrak{F}$. The fragment $\mathfrak{F}$ is *complete for* $\mathcal{K}$ if $\equiv_{\mathfrak{F}}$ restricted to $\mathcal{K}$ is isomorphism. When it is, the *witness size* $b_{\mathfrak{F}}(n)$ is the least k, or $+\infty$ if no finite such k exists, such that any two non-isomorphic $G, G' \in \mathcal{K}$ with $|B_G| = |B_{G'}| = n$ are separated by some $F \in \mathfrak{F}$ with $|B_F| \le k$; one also asks for the number and length of the records of F and for the cost of finding F.

For finite relational structures, the unrestricted profile determines isomorphism by Lovász's theorem [2]. The present contribution is quantitative: how large must a test be when coordinates within its records cannot repeat, and how does a finite presentation of an infinite heap affect the cost?

Bounded linear witnesses. Theorem 4.12 reconstructs the twin quotient and its integer multiplicities using at most $n + (r-1)\max(t, t') \le rn$ source letters. The bound extends to fixed finite signatures and named ports. In the binary case, a directed-cycle lower bound gives $b_{\mathfrak{L}}(n, 2) = \Theta(n)$ (Proposition 4.5); exact values through $n = 5$ complement the bounds.

Exact computational classification. Section 4.3 certifies two six-letter complement collisions and proves $6 \le b_{\mathfrak{L}}(6, 2) \le 12$. The search exhausts specified structural families, with independent orbit counts and CPU verification of GPU pattern laws; the exact six-letter value remains open.

Heap presentation and cost. Theorem 3.1 packages a finite-obstruction argument into budgets of n letters, $n!$ records and, for regular heaps, length $< mm'$. Theorem 6.4 gives the exact state threshold $\psi(n, m) = m$ for induced counting under symmetric target/test budgets. Neither result asserts polynomial-time reconstruction.

Observation and decomposition. Proposition 4.17 reconstructs from one complete random pattern of $O(n^2 \log n + rn)$ positions. Section 5 organizes kernel and shape information by two classical inversion operations. These are supporting results, not claims of a new counting identity or sampling paradigm.

Precise relation to prior work. Cai–Govorov [9, Theorem 1] identify twin-free weighted digraphs by simple tests. For a binary class, collapse twins and assign each quotient letter its class size as weight: weighted counts equal the original counts, and a weight-preserving quotient isomorphism expands to an isomorphism. Thus binary linear completeness already follows. More generally, replace each relation by its 0/1 indicator. Young's simple-#CSP instances [10, Definition 4, Corollary 36] form a subfamily of our linear tests, so his theorem supplies qualitative completeness in arbitrary arity. Singleton unary relations cover named ports; nullary relations are checked separately. The contribution developed below is the stated variable budget for Boolean relations and integer multiplicities.

Prior result	What is already available	Focus of this paper
Lovász [8, Theorem 5.33]	Simple witnesses on at most $2(t + t' + 3)^8$ vertices for twin-free weighted undirected graphs.	Linear source-size bound for 0/1 relations with integer multiplicities; no general weighted improvement claimed.
Cai–Govorov [9]; Young [10]	Weighted directed and arbitrary-arity identification; explicit finite witness constructions.	The particular bound $n + (r-1)\max(t, t')$ and its quotient/multiplicity decoder.
ten Cate et al. [11]	Directed-cycle count formula; query-number separations.	Its specialization to a source-size lower bound and combination with our upper bound.

1id1:core:20261109-0002@v1 1id1.net

The restricted-profile results of Dell–Grohe–Rattan [4], Grohe [5] and Böker [6] relate tree, treewidth, treedepth and Berge-acyclic tests to refinement and logic. Tuple-linearity is a different restriction: it permits directed cycles, but not repeated coordinates in one record. Graph-parameter theory [7] supplies the broader algebraic context. The quantitative claims above are compared with these specific results; we do not claim the first finite witness bound or the first constructive identification theorem.

Boundary with the companion audit. The companion preprint [12] studies the empirical value and limits of Token-Space consistency certificates for language models. Here we study exact structural identification and witness resources. The works share Token-map semantics and classical counting identities; no LLM experimental results or reliability claims are reused here. The CPU/GPU experiments below enumerate exact Token structures and certify witness bounds; they are not language-model evaluations or continuations of the companion experiments.

2 Preliminaries

We use the definitions of [1, Section 5]: $F_E(B) = (B \sqcup E)^*$, $F_E(f) = (f \sqcup \mathrm{id}_E)^*$; product $G \otimes H$, coproduct $G \uplus H$, exponential G^H with heap $\Upsilon_G \lhd \Upsilon_H$ [1, Theorem 5.2]. $\mathbf{TS}^{\mathrm{fin}}_E$ is the full subcategory of classes with finite carrier; a class is *finite* if carrier and heap are finite. $D = \langle \{*\}, E, \varnothing \rangle$ is the discrete point and $1_E = \langle \{*\}, E, F_E(\{*\}) \rangle$ the terminal object. $\mathrm{inj}(F, G)$ counts injective Token maps.

Proposition 2.1 (Identities)**.** *For $F, F' \in \mathbf{TS}^{\mathrm{fin}}_E$ and finite-carrier G, H:*

$$\hom(F, G \otimes H) = \hom(F, G)\ \hom(F, H), \qquad \hom(F \uplus F', G) = \hom(F, G)\ \hom(F', G), \tag{1}$$

$$\hom(F, G^H) = \hom(F \otimes H, G), \qquad \hom(D, G) = |B_G|, \tag{2}$$

$$\hom(1_E, G) = |\{b \in B_G : F_E(\{b\}) \subseteq \Upsilon_G\}|, \qquad \hom(0_E, G) = 1 = \hom(F, 1_E). \tag{3}$$

Proof. Universal properties give bijections of hom-sets; (2) is the currying bijection of [1, Theorem 5.2]; the rest is direct. □

These identities hold in any cartesian closed category of relational structures (directed graphs with loops are one: $u \to v$ in G^H iff $u(x) \to v(y)$ for every edge $x \to y$ of H). We use (2) as a device for trading an enlargement of the test $(F \otimes H)$ for a transformation of the target (G^H), not as a new law. We do *not* claim that $\hom(-, G)$ is a rig homomorphism on isomorphism classes: two tests with equal counts into G can have different counts after a product with a third class, so the multiplicative structure on test classes is not determined by the counts.

Let $\Pi(B)$ be the partition lattice of a finite set B. For $\pi \in \Pi(B_F)$ let F/π be the quotient class with carrier B_F/π and heap $F_E(q_\pi)(\Upsilon_F)$. Every Token map $F \to G$ factors uniquely as q_π followed by an injective Token map $F/\pi \to G$, where π is its kernel; this uses only that F_E preserves composition and surjections, not finiteness of heaps. Hence

$$\hom(F, G) = \sum_{\pi \in \Pi(B_F)} \mathrm{inj}(F/\pi, G), \qquad \mathrm{inj}(F, G) = \sum_{\pi \in \Pi(B_F)} \mu(\hat{0}, \pi)\ \hom(F/\pi, G). \tag{4}$$

Lemma 2.2 (Permutation-order argument)**.** *If $G, G' \in \mathbf{TS}^{\mathrm{fin}}_E$ admit bijective Token maps $f : G \to G'$ and $g : G' \to G$, then f is an isomorphism.*

Proof. $h = g \circ f$ is a Token map and a permutation of the finite set B_G, so $h^k = \mathrm{id}$ for some $k \geq 1$ and $f^{-1} = h^{k-1} \circ g$ is a Token map. □

3 Finite witnesses

Lovász's argument compares heap sizes at its last step; for infinite heaps that step is replaced by Lemma 2.2. The following strengthening shows that infinite heaps do not change the identification problem for finite-heap tests at all: the same carrier and record-number bounds hold, while record length and effective access to the heap remain representation-dependent.

Theorem 3.1 (Finite witnesses)**.** *Let $G, G' \in \mathbf{TS}_E^{\mathrm{fin}}$ be non-isomorphic.*

(i) *If $|B_G| \neq |B_{G'}|$ then $\hom(D, G) \neq \hom(D, G')$.*

(ii) *If $|B_G| = |B_{G'}| = n$, there is a finite class F with $|B_F| \leq n$ and $|\Upsilon_F| \leq n!$ such that $\hom(F, G) \neq \hom(F, G')$. The records of F can be taken from $\Upsilon_G \cup \Upsilon_{G'}$, relabelled.*

(iii) *If moreover $\Upsilon_G, \Upsilon_{G'}$ are regular languages recognised by deterministic automata with m, m' states, the records in* (ii) *can be taken of length $< mm'$.*

In particular $b(n) \leq n$ for $\mathcal{K} = \mathbf{TS}_E^{\mathrm{fin}}$ and $\mathfrak{F}$ the finite classes.

Proof. (i) is (2). (ii) By Lemma 2.2, bijective Token maps cannot exist in both directions; say there is none $G \to G'$. For each bijection $\sigma : B_G \to B_{G'}$ choose $w_\sigma \in \Upsilon_G$ with $F_E(\sigma)(w_\sigma) \notin \Upsilon_{G'}$, and let $F = \langle B_G, E, \{w_\sigma\}_\sigma \rangle$, a finite class with $\leq n!$ records and $\Upsilon_F \subseteq \Upsilon_G$. Then id_{B_G} is an injective Token map $F \to G$, so $\mathrm{inj}(F, G) \geq 1$; an injective map $F \to G'$ would be a bijection σ preserving w_σ, so $\mathrm{inj}(F, G') = 0$. By (4), $\hom(F/\pi, G) \neq \hom(F/\pi, G')$ for some π, and F/π has $\leq n$ letters and $\leq n!$ records. (iii) For each σ the set $\{w \in \Upsilon_G : F_E(\sigma)(w) \notin \Upsilon_{G'}\}$ is recognised by the product of the automaton for Υ_G with the complement automaton for $\Upsilon_{G'}$ (relabelled by σ), which has mm' states; it is non-empty, so it contains a word of length $< mm'$. □

Proposition 3.2 (n letters are necessary)**.** *Let $G_n = \langle \{y_1, \ldots, y_n\}, E, \{y_1 \cdots y_n\} \rangle$ and $G_\varnothing = \langle \{y_1, \ldots, y_n\}, E, \varnothing \rangle$. Then $G_n \ncong G_\varnothing$ and $\hom(F, G_n) = \hom(F, G_\varnothing)$ for every finite F with $|B_F| < n$. Hence $b(n) = n$, and also $b_{\mathfrak{L}}(n) = n$ for linear targets (Proposition 4.1), since $G_n, G_\varnothing$ are linear and the separating test $\langle \{x_1, \ldots, x_n\}, E, \{x_1 \cdots x_n\} \rangle$ is linear.*

Proof. A record of F maps into Υ_{G_n} only if its image is $y_1 \cdots y_n$, which needs n distinct carrier letters and no core letter; with $|B_F| < n$ no record qualifies. So a function $B_F \to B_{G_n}$ is a Token map iff $\Upsilon_F = \varnothing$, the same condition as for $G_\varnothing$. □

Corollary 3.3 (Identification algorithm)**.** *Let G, G' have carriers of size n and heaps presented by complete deterministic automata with m, m' states over the finite alphabet $B \sqcup E$.*

(i) *$G \cong G'$ is decided, and in the negative case a finite test F with $\mathrm{inj}(F, G) \neq \mathrm{inj}(F, G')$ is output, using $O(n!)$ inclusion tests $F_E(\sigma)(\Upsilon_G) \subseteq \Upsilon_{G'}$, each a product-automaton emptiness check with mm' states over an alphabet of size $n + |E|$.*

(ii) *A test F/π with $\hom(F/\pi, G) \neq \hom(F/\pi, G')$ is then found by evaluating both sides of (4); the search ranges over at most B_n (Bell number) quotients, and each evaluation of $\hom(F/\pi, -)$ costs at most $n^n \cdot n!$ membership tests of words of length $< mm'$.*

Proof. (i) Test all bijections in both directions. If some bijection preserves heaps in each direction, Lemma 2.2 gives an isomorphism (and exhibits it). Otherwise the proof of Theorem 3.1 produces F from the failing words, and $\mathrm{inj}(F, G) \geq 1 = 1 + \mathrm{inj}(F, G')$. (ii) By (4) the two sums $\sum_\pi \mu(\hat{0}, \pi) \hom(F/\pi, -)$ differ, so some term differs; $\hom(F/\pi, -)$ is evaluated by testing every function $B_F/\pi \to B$ on every record. □

The injective witness in (i) is already a certificate for the *injective* profile; step (ii) is needed because inj and hom can disagree on which tests separate. Example: $G = \langle \{0, 1\}, E, \{01\} \rangle$, $G' = \langle \{0, 1\}, E, \{00\} \rangle$, $F = G$: $\mathrm{inj}(F, G) = 1 \neq 0 = \mathrm{inj}(F, G')$ but $\hom(F, G) = 1 = \hom(F, G')$, and only the one-letter quotient $\{xx\}$ gives $0 \neq 1$. Whether an ordinary-count witness can be extracted without the search over $\Pi(B_F)$ is open; the bound in (ii) is crude.

1id1:core:20261109-0002@v1 1id1.net

The witness of Theorem 3.1 uses the quotients F/π, which repeat letters. The next section asks what happens when those are forbidden.

4 Linear tests

A class is *linear* if no record in its heap repeats a carrier letter; $\mathfrak{L}$ denotes the finite linear classes. In relational terms, every record has pairwise distinct coordinates; this is not acyclicity, nor the hypergraph convention that two edges meet in at most one vertex. The witnesses of Theorem 3.1 are quotients F/π and are typically not linear.

Proposition 4.1. *$\mathfrak{L}$ is complete for linear targets, with $b_{\mathfrak{L}}(n) = n$.*

Proof. A non-linear F has a record repeating a letter, whose image repeats a letter and lies in no linear heap; so $\hom(F, G) = 0 = \hom(F, G')$ and the hypothesis extends to all finite tests. Theorem 3.1 applies, and its witness F/π is linear here because non-linear quotients contribute 0 on both sides. The lower bound is Proposition 3.2. □

4.1 Lower bounds

The following examples separate the effects of arity, diagonal information and increasing test order.

Proposition 4.2 (Arity lower bound)**.** *Let $G_r = \langle \{a\}, E, \{a^r\} \rangle$ and $G_\varnothing = \langle \{a\}, E, \varnothing \rangle$. Then $G_r \not\cong G_\varnothing$, $\hom(F, G_r) = \hom(F, G_\varnothing)$ for every linear F with $|B_F| < r$, and $F_r = \langle \{x_1, \dots, x_r\}, E, \{x_1 \cdots x_r\} \rangle$ separates them. Hence $b_{\mathfrak{L}}(1) \geq r$ whenever targets with r-ary records are admitted, while the non-linear test $\langle \{x\}, E, \{x^r\} \rangle$ separates them with one letter.*

Proof. A linear record with fewer than r letters, or with a core letter, cannot map onto a^r. So the unique map $F \to G_r$ is a Token map iff $\Upsilon_F = \varnothing$, the same condition as for $G_\varnothing$. $\hom(F_r, G_r) = 1$, $\hom(F_r, G_\varnothing) = 0$. □

Witness size for linear tests is therefore a function $b_{\mathfrak{L}}(n, r)$ of carrier size and maximal arity, and the identification problem has to be posed per signature. A class is *core-free r-ary* if $\Upsilon \subseteq B^r$; for $r = 2$ these are directed graphs with loops, and linear tests are loopless digraphs.

Proposition 4.3 (Linear witnesses may need more letters than the target)**.** *Let G, G' be the binary classes on $\{0, 1, 2, 3\}$ with heaps*

$$\Upsilon_G = \{00, 01, 02, 10, 13, 21, 22, 30\}, \qquad \Upsilon_{G'} = \{01, 02, 10, 11, 13, 21, 22, 30\}.$$

Then $G \not\cong G'$; $\hom(F, G) = \hom(F, G')$ for all 4165 loopless digraphs F with at most four vertices; and the oriented tree $T = \{01, 03, 04, 12\}$ gives $\hom(T, G) = 98 \neq 96 = \hom(T, G')$. Consequently $b_{\mathfrak{L}}(4, 2) = 5$ (computer-assisted), and $b_{\mathfrak{L}}(n, 2) \geq 5$ for all $n \geq 4$.

Proof. The heaps differ in the position of one loop. In G' the looped letter of out-degree 3 (namely 1) has an out-neighbour of out-degree 1 ($1 \to 3$); in G the looped letter of out-degree 3 is 0 with $N^+(0) = \{0, 1, 2\}$, none of out-degree 1; so $G \not\cong G'$. Agreement on the 4165 tests is check C12 of Appendix A, exhaustive over all 2^{16} heaps on four letters: exactly two non-isomorphic pairs collide, this one and $\Upsilon = \{00, 01, 02, 11, 13, 21, 23, 30\}$ versus $\{00, 02, 10, 11, 13, 21, 23, 30\}$, and both are separated by five-letter loopless tests, which gives $b_{\mathfrak{L}}(4, 2) = 5$. For T, $\hom(T, G) = \sum_x d(x)^2 \sum_{y \in N^+(x)} d(y)$ with d the out-degree: $63 + 16 + 16 + 3 = 98$ for G and $20 + 54 + 20 + 2 = 96$ for G'. For $n > 4$ add $n - 4$ isolated letters to both: a loopless test H maps a weak component with an edge into the original letters, and an isolated vertex anywhere, so $\hom(H, G \uplus kD) = \prod_K \big(\hom(K, G) + k[\,|K| = 1\,]\big)$ over the weak components K of H, which agrees for G and G' whenever all components have ≤ 4 vertices. □

Proposition 4.4 (Five-letter targets)**.** $b_{\mathfrak{L}}(5,2) = 5$*, by computer (check C20): no two of the* 291 968 *isomorphism types of binary classes on five letters agree on all loopless tests with at most five letters.*

Proof. By Lemma 4.7 below, counts on five-letter loopless tests determine, and are determined by, the labelled sample-pattern law. Smaller tests are recovered by adding isolated vertices. Check C20 enumerates all 291 968 target types and finds distinct pattern-orbit signatures. Its four-letter calibration reproduces the two pairs of Proposition 4.3. Together with that proposition's lower bound, this gives equality. □

Proposition 4.5 (A cycle lower bound)**.** *Let $n \geq 4$, $a = \lfloor n/2 \rfloor$ and $b = n - a$. The targets C_n and $C_a \uplus C_b$, where C_ℓ is the directed ℓ-cycle, have equal counts on every loopless test with fewer than a letters. The test C_a separates them. Consequently*

$$\lfloor n/2 \rfloor \leq b_{\mathfrak{L}}(n,2) \leq 2n,$$

where the upper bound is Theorem 4.12; in particular $b_{\mathfrak{L}}(n,2) = \Theta(n)$.

Proof. We specialize the standard directed-cycle count criterion [11, Proposition 2.4, Lemma 4.2]. Let F be weakly connected with k vertices. Choose a spanning tree of its underlying graph, one directed edge for each tree edge, and integer heights $h(v)$ increasing by 1 along each chosen edge. A map to C_ℓ exists exactly when every edge $u \to v$ satisfies $h(v) - h(u) \equiv 1 \pmod{\ell}$; if it exists, there are exactly ℓ maps, one per image of a root. For a non-tree edge the defect $h(v) - h(u) - 1$ has absolute value at most k. Thus when $\ell > k$ the congruences hold iff every defect is zero, independently of ℓ. This includes opposite edges, whose two-edge defect has magnitude 2.

If $k < a$, the counts into C_n, C_a, C_b are therefore either n, a, b, respectively, or all zero. Since a connected test maps into one component of a disjoint union, its counts into the two targets agree. Products over weak components handle arbitrary F. Finally $\hom(C_a, C_n) = 0$, whereas $\hom(C_a, C_a \uplus C_b) = a$ if $a < b$, and $2a$ if $a = b$. The least separating order of this pair is exactly a. □

The lower bound is an application of known cycle theory, not a new cycle-counting theorem. It replaces an elementary profile-counting lower bound by an explicit family and fixes the binary growth order. The leading constants and the excess over target size at small n remain to be understood.

Proposition 4.6 (Trees and cycles are not enough)**.** *Let R be the rook graph $K_4 \,\square\, K_4$ and S the Shrikhande graph, read as symmetric binary classes on* 16 *letters. Then* $\hom(T,R) = \hom(T,S) = 16 \cdot 6^{|\Upsilon_T|}$ *for every oriented tree T,* $\hom(C_k,R) = \hom(C_k,S)$ *for every directed cycle, and* $\hom(K_4,R) = 192 \neq 0 = \hom(K_4,S)$.

Proof. Both are 6-regular on 16 vertices with $A^2 = 4I + 2J$, hence cospectral. For a d-regular target on n vertices, $\hom(T,G) = n\, d^{|\Upsilon_T|}$ by choosing images leaf by leaf; $\hom(C_k,G) = \operatorname{tr} A^k$. R has 8 cliques K_4 (rows and columns), giving $8 \cdot 4! = 192$ maps; in S the two common neighbours of an edge are non-adjacent (each neighbourhood is a 6-cycle), so S is K_4-free. □

Tree counts are the information of colour refinement [4], cycle counts that of the spectrum; the proposition rules out this particular subfamily, not other small complete subfamilies.

1id1:core:20261109-0002@v1 1id1.net

4.2 A bounded-variable reconstruction

Fix $r \geq 1$ and a core-free r-ary target G with nonempty carrier B, $|B| = n \geq 1$. The empty-carrier case is immediate. Call $u, v \in B$ *twins*, $u \sim v$, if for every position $p \in \{1, \dots, r\}$ and every $(w_1, \dots, w_{r-1}) \in B^{r-1}$, the record with u inserted at position p lies in Υ_G iff the one with v inserted does. This is an equivalence relation, and membership of a record in Υ_G depends only on the sequence of twin classes of its letters (replace the letters by twins one position at a time). Hence G is determined up to isomorphism by the quotient set $B/\!\sim$, the multiplicities $|C|$ of the classes, and the set $P_G \subseteq (B/\!\sim)^r$ of class sequences of records; P_G includes sequences with repeated classes, which encode loops and all other records with repeated letters.

Let $N \geq 1$ and let $X_1, \dots, X_N$ be independent uniform letters of B. The *sample pattern* is the linear r-ary class M on carrier $[N] = \{1, \dots, N\}$ with

$$\Upsilon_M = \{(i_1, \dots, i_r) \in [N]^r \text{ pairwise distinct} : X_{i_1} \cdots X_{i_r} \in \Upsilon_G\}.$$

Lemma 4.7 (Law of the sample pattern)**.** *For every linear r-ary class H on carrier $[N]$,*

$$\Pr(M = H) = n^{-N} \sum_{S \subseteq I_N \setminus \Upsilon_H} (-1)^{|S|} \hom(H + S, G),$$

where I_N is the set of injective r-tuples of $[N]$ and $H + S$ is H with the records S added. In particular the law of M is determined by the counts $\hom(F, G)$ over linear r-ary tests F with N letters.

Proof. $n^N \Pr(M = H)$ counts functions $f : [N] \to B$ that send every record of H into Υ_G and no injective non-record of H into Υ_G. Inclusion–exclusion over the second condition gives the sum, since the functions sending $\Upsilon_H \cup S$ into Υ_G are the Token maps $H + S \to G$. □

We now define the deterministic objects that replace the random sample. For a function $s : Q \to \mathbb{N}$ on a finite set Q and a set $P \subseteq Q^r$, the *linear blow-up* $\mathrm{Bl}_{\mathrm{lin}}(Q, P; s)$ is the linear r-ary class on $N = \sum_q s(q)$ positions, each position labelled by an element of Q so that exactly $s(q)$ positions carry q, whose heap consists of the injective position tuples whose label sequence lies in P. (Different labellings give isomorphic classes; the counts below do not depend on the choice.) For a linear r-ary class K on $[N]$ write $i \approx j$ if for every position p and every injective $(r-1)$-tuple τ of positions in $[N] \setminus \{i, j\}$, inserting i or j into τ at p gives records that are both in or both out of Υ_K. For the target G let $Q = B/\!\sim$, $t = |Q|$, $x(q) = |q|$ the multiplicity, and P_G as above. Write $\mathrm{Bl}_{\mathrm{full}}(Q, P_G; x)$ for the *full blow-up*, in which every position tuple is retained when its class sequence belongs to P_G, including tuples with repeated positions. Then $G \cong \mathrm{Bl}_{\mathrm{full}}(Q, P_G; x)$. The two blow-ups must be distinguished: sample patterns are linear, whereas the reconstructed target need not be.

Lemma 4.8 (Patterns with large classes)**.** *Let $v \in B^N$ and let $M(v)$ be its sample pattern. For a twin class C put $s_C = |\{i : v_i \in C\}|$.*

(i) *If $v_i \sim v_j$ then $i \approx j$ in $M(v)$.*

(ii) *If $s_C \geq r$ for* every twin class of the target*, then $i \approx j$ iff $v_i \sim v_j$. In particular $\approx$ is an equivalence relation, its classes are the sets $\{i : v_i \in C\}$, every sequence of P_G is realised by an injective position tuple, and $M(v) \cong \mathrm{Bl}_{\mathrm{lin}}(Q, P_G; s)$.*

(iii) *If $\approx$ is an equivalence relation on $M(v)$, its classes are unions of the sets $\{i : v_i \in C\}$, so $M(v)$ has at most t of them, with equality iff every twin class is met and no two positions with non-twin letters are $\approx$-related.*

Proof. (i) is the fact that membership depends only on twin classes. For (ii) let $v_i = u \not\sim v_j = w$ and choose a position p and a context $(c_1, \dots, c_{r-1}) \in B^{r-1}$ on which u, w differ. Each twin class D occurs at most $r - 1$ times in the context; it has $s_D \geq r$ positions, of which at most

1id1:core:20261109-0002@v1 1id1.net

one is i or j (if $D \ni u$ then $D \not\ni w$), so the context can be realised, class by class, by distinct positions outside $\{i, j\}$, and since membership depends only on classes, $i \not\approx j$. The remaining assertions follow: the class partition is $\{i : v_i \in C\}$, a sequence of P_G needs at most r positions per class and $s_C \geq r$, and the heap of $M(v)$ is by definition the set of injective tuples whose class sequence lies in P_G. (iii) Under the stated equivalence hypothesis, (i) places each nonempty set $\{i : v_i \in C\}$ inside one equivalence class, so the partition coarsens the partition by the target classes that were met. Equality in the bound t holds exactly when all target classes are met and none are merged. □

Remark 4.9 (Why the coverage and equivalence hypotheses are needed). If the target has only the edge $u \to z$, then u and an isolated vertex v are not twins. The sample (u, u, v, v) meets both sampled classes twice, but its pattern is empty; the missing class of z contains the distinguishing context. Also, on the three-position pattern with edges $01, 02, 12$, one has $0 \approx 1$ and $1 \approx 2$ but not $0 \approx 2$. Hence arbitrary sample patterns do not have an $\approx$-partition, and the decoder below explicitly filters for equivalence before counting classes.

Lemma 4.10 (Counting blow-ups). *Let $s : Q \to \mathbb{N}$ with $s(q) \geq r$ for all q, $N = \sum_q s(q)$, and let K be a labelled copy of $\mathrm{Bl}_{\mathrm{lin}}(Q, P_G; s)$ on $[N]$. Then*

$$c_N(K) := |\{v \in B^N : M(v) = K\}| = \sum_{\alpha \in \mathrm{Aut}(Q, P_G)} \prod_{q \in Q} x(\alpha q)^{s(q)}.$$

Proof. K has exactly t $\approx$-classes (the label classes, by Lemma 4.8(ii) applied to any v realising K, e.g. one letter from each class, repeated). If $M(v) = K$ then by Lemma 4.8(iii) v meets all t twin classes without spurious identifications, so v is constant in twin class on each label class, inducing a bijection β from the label classes to Q that carries P_G to itself (every sequence is realised, by (ii)); conversely every such β and every choice of $v_i \in \beta(\text{label of } i)$ gives $M(v) = K$. Fixing one such β_0, the others are $\alpha \circ \beta_0$ with $\alpha \in \mathrm{Aut}(Q, P_G)$, and the number of choices of v is $\prod_q x(\alpha q)^{s(q)}$ after relabelling. □

Write $F_s(x)$ for the right-hand side; it is the $\mathrm{Aut}(Q, P_G)$-symmetrised monomial of x with exponent vector s.

Lemma 4.11 (Reading the multiplicities). *Let A be a group of permutations of Q, $x : Q \to \mathbb{N}_{\geq 1}$, $z = x - 1$, $k = \sum_q z(q)$. For $e : Q \to \mathbb{N}$ with $\sum_q e(q) = k$,*

$$\Big(\prod_q x(q)^r\Big) \cdot |\{\alpha \in A : z \circ \alpha = e\}| \;=\; \sum_{\alpha \in A} \prod_q x(\alpha q)^r \binom{x(\alpha q) - 1}{e(q)},$$

and the right-hand side is a fixed rational linear combination (depending on e, r only) of the symmetrised monomials $F_s(x)$ with $r \leq s(q) \leq r + e(q)$ for all q, hence with $\sum_q s(q) \leq rt + k$. After multiplication by $\prod_q e(q)!$ the combination has integer coefficients. Also $\prod_q x(q)^r = F_{r\mathbf{1}}(x)/|A|$.

Proof. $\prod_q \binom{z(\alpha q)}{e(q)}$ is nonzero iff $e \leq z \circ \alpha$ coordinatewise, which, as both sum to k, means $e = z \circ \alpha$, and then the product equals 1. Multiplying by the A-invariant $\prod_q x(q)^r = \prod_q x(\alpha q)^r$ gives the identity. Each factor $y^r \binom{y-1}{e(q)}$ is a polynomial in y with terms of degree $r, \ldots, r + e(q)$ and rational coefficients with denominator $e(q)!$; expanding the product and summing over α gives the stated combination. □

Theorem 4.12 (Linear source-variable bound). *Let G, G' be core-free r-ary classes on n letters with t, t' twin classes. If $\hom(F, G) = \hom(F, G')$ for all linear r-ary tests F with at most $n + (r-1)\max(t, t')$ letters, then $G \cong G'$. Hence $\mathfrak{L}$ is complete for core-free r-ary classes and*

$$b_{\mathfrak{L}}(n, r) \leq rn, \qquad \textit{in particular} \qquad b_{\mathfrak{L}}(n, 2) \leq 2n.$$

1id1:core:20261109-0002@v1 1id1.net

Proof. Put $N^* = n + (r-1)\max(t, t')$. By Lemma 4.7 the hypothesis gives $c_N(K)$ the same value under G and G' for every $N \le N^*$ and every K on $[N]$ (the counts for N letters involve only tests with N letters). We describe a decoder that outputs the isomorphism type of G from the family $(c_N)_{N \le N^*}$ and n; applied to the identical family of G' it outputs G', so $G \cong G'$.

Step 1: the quotient. Among all K with $c_N(K) > 0$, $N \le N^*$, for which $\approx$ is an equivalence relation with all classes of size $\ge r$, take one with the largest number of classes. By Lemma 4.8(ii),(iii) and since the blow-up $\mathrm{Bl}_{\mathrm{lin}}(Q, P_G; r\mathbf{1})$ has $rt \le N^*$ positions and $c_{rt} > 0$ on it, this maximum is t, and any maximiser K_0 has as $\approx$-classes the sets $\{i : v_i \in C\}$ for any v with $M(v) = K_0$, with every twin class met. Hence the class set of K_0 with the set of class sequences of its records is a copy of (Q, P_G); this yields t, (Q, P_G), $\mathrm{Aut}(Q, P_G)$ and $k = n - t$.

Step 2: the multiplicities. For every $s : Q \to \mathbb{N}$ with $s \ge r$ and $\sum_q s(q) \le rt + k \le N^*$, build a labelled copy of $\mathrm{Bl}_{\mathrm{lin}}(Q, P_G; s)$ and read $c(K) = F_s(x)$ from the data (Lemma 4.10). For every e with $\sum e = k$, Lemma 4.11 with $A = \mathrm{Aut}(Q, P_G)$ gives $|\{\alpha : z \circ \alpha = e\}|$ from these values. The set of e with nonzero count is the A-orbit of $z = x - 1$; pick any member e and put $x^* = e + 1$.

Step 3: output $\mathrm{Bl}_{\mathrm{full}}(Q, P_G; x^*)$ with all records (repeated letters allowed) whose class sequence lies in P_G. Since $x^* = x \circ \alpha$ for some automorphism α of (Q, P_G), this is isomorphic to G.

The letters used are $\max(rt, rt + k) = n + (r-1)t \le N^*$ for G, and symmetrically for G'; $t \le n$ gives $N^* \le rn$. □

Remark 4.13. (1) The decoder is explicit: Step 1 is a search over patterns with at most rt positions, Step 2 evaluates $\binom{k+t-1}{t-1}$ linear combinations of counts $c(K)$ of blow-ups, Step 3 is immediate. Its cost is not claimed to be polynomial in n; what the theorem shows is that a separating test with $\le rn$ letters *exists*, and that the quotient structure and the multiplicities are recovered by separate mechanisms. (2) For $n = 1$ the bound r is attained (Proposition 4.2), so $b_{\mathfrak{L}}(1, r) = r$. (3) Step 2 is a reconstruction problem for the permutation group $\mathrm{Aut}(Q, P_G)$: the multiset z of size k is recovered from its symmetrised monomials of degree $\le k$ through Lemma 4.11. Lower degrees do not suffice in general (two different orbits can share all symmetrised monomials of degree $< k$); whether this step, rather than Step 1, governs the true value of $b_{\mathfrak{L}}$ is open. (4) The argument proves slightly more than stated: the law of the sample pattern with N^* positions determines G, and the law is a function of the counts on tests with exactly N^* letters. (5) In tensor notation, a linear test evaluates $\sum_f \prod_{(i_1,\dots,i_r)\in\Upsilon_F} A_{f(i_1),\dots,f(i_r)}$, with distinct source positions within each factor. These permutation-invariant polynomials determine a 0/1 tensor using tests with at most rn source variables. This is a bound on source variables, not on the number of tensor factors or the running time, and it is not a statement about arbitrary real tensors. Qualitative separation is covered by [9, 10]; our assertion concerns the displayed source-variable budget, using Boolean entries and positive integer multiplicities. It is not a claim that constructive witnesses were previously unavailable.

Exact values, by exhaustive enumeration of all isomorphism types of n-letter binary classes against loopless tests of increasing size (checks C16, C12, C20 of Appendix A):

$$b_{\mathfrak{L}}(1,2) = 2, \qquad b_{\mathfrak{L}}(2,2) = 3, \qquad b_{\mathfrak{L}}(3,2) = 3, \qquad b_{\mathfrak{L}}(4,2) = 5, \qquad b_{\mathfrak{L}}(5,2) = 5,$$

These are exact global values. For $n = 6$, Proposition 4.15 proves only the interval $[6, 12]$; it does not append an exact value to this sequence. Among the five known values, the upper bound $2n$ is attained only at $n = 1$. Proposition 4.5 supplies linear growth independently of this short table; the table does not determine a formula or a limiting constant.

Question 4.14 (Binary witness constants)**.** What are the optimal constants between $\lfloor n/2 \rfloor$ and $2n$? Can the decoder's source budget be reduced, and which targets require more source letters than target letters? The cycle family explains growing witness order without loops; the

1id1:core:20261109-0002@v1 1id1.net

four-letter collisions show a different, small-scale obstruction involving diagonal information. These mechanisms should not be conflated.

4.3 Exact computation and complement collisions

Scope and access. For a binary target G with adjacency matrix A, encode each ordered k-tuple of target letters by the $k(k-1)$ bits $A_{x_i x_j}$, $i \neq j$. Sorting all n^k codes retains the complete labelled pattern law, including multiplicities. By Lemma 4.7, equality of these integer vectors is equivalent to agreement on all loopless tests with at most k letters. This is exact computation on supplied targets, not statistical sampling or model inference.

Let $\mathcal{C}_6$ contain the six-letter binary targets whose adjacency matrix is symmetric or has constant row sum. Loops are allowed in both families. The frozen search directly enumerates the constant-out-degree families $d = 0, 1, 2, 3$ and all symmetric relations:

Family	Labelled targets	Isomorphism types
Constant out-degree 0	1	1
Constant out-degree 1	46 656	130
Constant out-degree 2	11 390 625	16 816
Constant out-degree 3	64 000 000	91 212
Symmetric, with loops	2 097 152	5 096

The families overlap: their union has 113 230 types. Independent Burnside counts reproduce every row. Relation complementation includes diagonal entries and transforms every off-diagonal sample bit, so it preserves equality of pattern laws. It covers degrees $4, 5, 6$ without further enumeration; the full class $\mathcal{C}_6$ has 130 164 types (check C28). This reduction does not cover arbitrary nonsymmetric targets with nonconstant out-degree.

Proposition 4.15 (Six-letter collisions; computer-assisted)**.** *Let $\overline{G}$ have the same carrier as G and binary heap $B^2 \setminus \Upsilon_G$. There are two pairs $G_i, \overline{G}_i$ whose least linear separating order is exactly 6. In particular*

$$\boxed{6 \leq b_{\mathfrak{L}}(6,2) \leq 12.}$$

Within $\mathcal{C}_6$, these are the only two unordered pairs of non-isomorphic types agreeing through order 5, and order 6 is necessary and sufficient. The exact global value $b_{\mathfrak{L}}(6,2)$ is not determined.

Proof. Take the adjacency matrices

$$A_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

For each, enumerating the $6^5 = 7776$ ordered assignments gives a five-position pattern law invariant under complement: every 20-bit pattern and its bitwise complement have the same multiplicity (check C27). Thus all linear tests on at most five letters agree with $J - A_i$. The common separating test is the oriented path $T : 0 \to 1 \to 2 \to 3 \leftarrow 5 \leftarrow 4$. If $u_j = (A^{\mathsf{T}})^j \mathbf{1}$, then $\hom(T, G) = u_2^{\mathsf{T}} u_3$, giving

$$(\hom(T, G_1), \hom(T, \overline{G}_1)) = (1648, 1640), \qquad (\hom(T, G_2), \hom(T, \overline{G}_2)) = (1616, 1624).$$

These unequal counts also prove non-isomorphism. The upper bound 12 is Theorem 4.12, not an inference from the search.

1id1:core:20261109-0002@v1 1id1.net

Complete-law grouping of the 113 230 enumerated types leaves 66, 25 385, 329 and 2 nonsingleton groups at orders $2, 3, 4, 5$, respectively; the last two are exactly the displayed pairs (check C26). For a comparison involving a constant-out-degree target with $d > 3$, complement both targets if the other is symmetric or also has degree > 3. This reduces to the enumerated union. Otherwise the two constant degrees differ, and a single edge separates them. Thus the two pairs are exhaustive in $\mathcal{C}_6$, and T separates both. □

Why complement hides information. This mechanism has a standard moment description. Set $S = 2A - J$ and $Z_G(F) = \sum_f \prod_{uv \in \Upsilon_F} S_{f(u)f(v)}$. Products of sign bits form a basis for functions on binary patterns. Hence a k-position law is invariant under complement iff $Z_G(F) = 0$ for every loopless F on at most k letters with an odd number of edges. This is a parity identity, not a new general identification theorem. In the two examples, every proper edge-subset of T has components on at most five vertices. Its odd centered moment vanishes by component factorization. Expanding $A = (J + S)/2$ therefore gives

$$\hom(T, G) - \hom(T, \overline{G}) = Z_G(T)/16, \qquad Z_{G_1}(T) = 128, \quad Z_{G_2}(T) = -128.$$

Thus the collision hides all five-position information, while the product of incoming-walk counts of lengths 2 and 3 exposes it at six positions.

Computation and verification. A C generator uses complete S_6 orbits and exact 36-bit keys for isomorphism reduction. A Tesla V100-SXM2-32GB on a 10-vCPU VM batches integer pattern generation and sorting; hashes only index full keys or full vectors, never certify equality. All 113 230 types were regrouped independently on CPU using adjacency indexing. A separate standard-library Python verifier checks the final pairs by assignment enumeration and all 720 permutations; it also permits the lower-bound certificate to be checked without a GPU. Independent orbit counts, a full VM rerun and resource measurements are supplied with the artifact. Batching accelerates pattern construction; the exhaustive scope, independently reproduced partitions and exact certificates supply the computational evidence.

Persistent collisions are a different mechanism. The earlier targeted extensions of the four-letter pairs cover 9356 parameter cases: 1024 single-vertex extensions, 8192 matched two-vertex extensions and 140 positive blow-ups. All 588 pairs surviving order 4 are separated at order 5 (C23). For either base pair, full twin classes of sizes (a, a, b, b) have identical four-position laws for all positive a, b, whereas the tree $\{01, 03, 04, 12\}$ has count difference $a^2b^2(a + b) > 0$ (C24). This exact coefficient certificate gives arbitrarily large targets of least separating order 5. Such persistence, six-position complement collisions, and the growing cycle obstruction are distinct phenomena.

4.4 Finite signatures and named ports

A finite relational signature $\mathcal{S}$ has relation symbols σ of arity $a_\sigma \le r$. A finite $\mathcal{S}$-structure is encoded in Token Space by using a fixed core label for each relation symbol and its ordered carrier tuple. The word length and the carrier arity are different quantities: the label does not add a carrier position. A test is linear if the carrier entries within each relation tuple are distinct. Nullary relations, if present, are compared separately by tests on the empty carrier.

Corollary 4.16 (Finite signatures and named ports)**.** *Let $\mathcal{S}$ be a fixed finite relational signature of maximum positive arity $r \ge 1$. Two non-isomorphic $\mathcal{S}$-structures on $n \ge 1$ letters, with t, t' twin classes defined using all relations, are separated by a linear $\mathcal{S}$-test with at most $n + (r - 1)\max(t, t') \le rn$ letters. The same bound holds for isomorphism preserving any fixed finite list of named ports, encoded by singleton unary relations in both targets.*

1id1:core:20261109-0002@v1 1id1.net

Proof. For N sampled positions, record membership of every injective a_σ-tuple in every relation σ. Inclusion–exclusion now ranges over the finite set of pairs (σ, position tuple), so the analogue of Lemma 4.7 recovers the full joint pattern law from linear counts. Define twins by substitution in every relation and every coordinate. Relation membership then factors through the common twin quotient, with one quotient relation P_σ per symbol. With at least r positions from every twin class, a context of any relation uses at most $r-1$ positions and can be realised as in Lemma 4.8. The equivalence filter, maximal-class argument and reconstruction of the quotient in Theorem 4.12 therefore apply simultaneously to all relations. The counting formula for a linear blow-up is still

$$c_N(K) = \sum_{\alpha \in \mathrm{Aut}(Q,(P_\sigma)_\sigma)} \prod_{q \in Q} x(\alpha q)^{s(q)},$$

and Lemma 4.11 recovers the multiplicities with $\sum s \le rt + n - t$. Adding a unary relation $P_j = \{a_j\}$ for each named port a_j does not increase the maximum arity; a structure isomorphism preserves these relations exactly when it preserves the ports. □

Named ports matter because an unlabelled profile cannot distinguish an edge $a \to b$ from the edge obtained by exchanging the names a, b, although the fixed query "does a point to b?" changes its answer. A source may contain several vertices carrying the same port label; all are forced to the corresponding target port. The corollary concerns the complete family of such constrained counts; it does not claim completeness for a chosen small bank of patterns.

4.5 Identification from one sample

Theorem 4.12 uses the exact *law* of the sample pattern. A single realisation already identifies the target with constant probability, at the price of a quadratic rather than linear number of positions.

Proposition 4.17 (One-sample identification)**.** *Let G be a core-free r-ary class on $n \ge 1$ letters and let $N \ge \max\bigl(2n^2 \log(16n),\, 2rn\bigr)$. There is a deterministic procedure which, given n and one sample pattern M on N positions, outputs a class isomorphic to G with probability at least $3/4$. Consequently two non-isomorphic such targets on n letters have different laws of the N-position sample pattern, and hence are separated by a linear test with N letters.*

Proof. For each letter b and each twin class C let s_b, s_C be their numbers of occurrences among $X_1, \dots, X_N$. By Hoeffding's inequality, $\Pr\bigl(|s_b/N - 1/n| \ge 1/(2n)\bigr) \le 2e^{-N/(2n^2)}$, and likewise for s_C/N against $|C|/n$. A union bound over the at most n letters and at most n classes gives failure probability at most $4n\, e^{-N/(2n^2)} \le 4n\, e^{-\log(16n)} = 1/4$. On the complementary event every letter occurs $s_b > N/(2n) \ge r$ times, so every twin class has $s_C \ge r$; by Lemma 4.8(ii), $\approx$ is an equivalence relation on M, its classes are the sets $\{i : X_i \in C\}$, and the class sequences of the records of M form a copy of (Q, P_G). The procedure computes $\approx$, reads off (Q, P_G) and the class sizes s_C, and outputs $\mathrm{Bl}_{\mathrm{full}}(Q, P_G; x^*)$ with $x^*(C)$ the integer nearest to $n\, s_C/N$. Since $|n\, s_C/N - |C|| < 1/2$ on the good event, $x^* = x$ and the output is isomorphic to G (and the procedure may output anything when $\approx$ fails to be an equivalence relation with all classes of size $\ge r$). If $G \not\cong G'$ had the same law of M, the output distributions would coincide, but the same distribution cannot give probability $\ge 3/4$ to two disjoint isomorphism types. The last assertion is Lemma 4.7. □

The access model supplies every relation bit on distinct sampled positions: up to $(N)_r$ bits for one r-ary relation, or $\sum_\sigma (N)_{a_\sigma}$ for a fixed signature. Neither the sample pattern nor its exact law is one scalar count query. The guarantee assumes known n and exact relation observations; it is not an LLM-query or noisy-count guarantee.

1id1:core:20261109-0002@v1 1id1.net

The two statements answer different questions. Theorem 4.12 bounds the number of letters of a test whose *exact* count separates two targets; Proposition 4.17 bounds the size of a *single random observation* of the target that identifies it. A third quantity, the number of independent observations needed to decide between two specified targets with a prescribed error, depends on the access model: for the four-letter pair of Proposition 4.3 and its witness T, exhaustive evaluation checks $4^5 = 1024$ carrier functions; a Bernoulli oracle returning whether a uniform random function $B_T \to B_G$ is a Token map sees a density difference of $(98 - 96)/4^5 = 1/512$ and, by Hoeffding, needs about $2 \cdot 512^2 \log 8 \approx 1.1 \cdot 10^6$ queries for error $1/4$; while the law of the 8-position sample pattern already determines the type by Theorem 4.12. Which counts must be estimated, to what precision, at what cost, to certify which structural conclusion, is a question to be posed per access model and is not pursued here.

5 Decomposition of the counts through one test

Identity (2) with $F = 1_E$ gives $\hom(1_E, G^H) = \hom(H, G)$: the map space counts its own maps at the terminal test. With $H = D$ it forgets the heap: $\Upsilon_G \lhd \varnothing = F_E(B_G)$, so G^D is the chaotic class on B_G and $\hom(F, G^D) = |B_G|^{|B_F|}$ for F without all-core words. Between D and 1_E lie the one-point classes $P_K = \langle \{*\}, E, K \rangle$, K a set of shapes.

Lemma 5.1. *Let* $F|_K = \langle B_F, E, \{w \in \Upsilon_F : \sigma(w) \in K\} \rangle$ *be the shape restriction. Then* $F \otimes P_K \cong F|_K$, *hence* $\hom(F, G^{P_K}) = \hom(F|_K, G) =: \hom_K(F, G)$.

Proof. By the product-heap formula of [1, Section 5], a word over $B_F \times \{*\}$ lies in the product heap iff its B_F-projection lies in Υ_F and its shape lies in K. □

Proposition 5.2 (Transfer)**.** *If* $\mathfrak{F}$ *is closed under* $- \otimes H$ *then* $G \equiv_{\mathfrak{F}} G'$ *implies* $G^H \equiv_{\mathfrak{F}} G'^H$. *Fragments closed under shape restriction are those closed under* $- \otimes P_K$ *for all* K.

Proof. $\hom(F, G^H) = \hom(F \otimes H, G)$; the second statement is Lemma 5.1. □

For a finite test F let $\Sigma = \sigma(\Upsilon_F)$ be its set of shapes. A carrier function $f : B_F \to B_G$ has a *violation set* $S_f = \{\sigma \in \Sigma : f \text{ violates some record of shape } \sigma\}$ and a *kernel* $\ker f \in \Pi(B_F)$.

Theorem 5.3 (Shape decomposition)**.** *Let* $V_S(F, G) = \#\{f : S_f = S\}$. *For every* $K \subseteq \Sigma$,

$$\hom_K(F, G) = \sum_{S \subseteq \Sigma \setminus K} V_S(F, G), \qquad V_S(F, G) = \sum_{J \subseteq S} (-1)^{|S \setminus J|} \hom_{\Sigma \setminus J}(F, G).$$

Along a maximal chain $\varnothing = K_0 \subset \cdots \subset K_m = \Sigma$ *with* $K_i \setminus K_{i-1} = \{\sigma_i\}$, *the difference* $\hom_{K_{i-1}} - \hom_{K_i}$ *counts the maps whose first violated shape in the order* $\sigma_1, \ldots, \sigma_m$ *is* σ_i.

Proof. f preserves $F|_K$ iff $S_f \cap K = \varnothing$; the second formula is Möbius inversion on 2^Σ. □

Shape restriction sees record types, not the identity of occurrences within a record (*aa* and *ab* have the same shape). The finer invariant comes from quotients. Put $A_F(\pi, K; G) = \hom((F/\pi)|_K, G) = \#\{f : \ker f \geq \pi,\ S_f \cap K = \varnothing\}$ and $W_F(\rho, S; G) = \#\{f : \ker f = \rho,\ S_f = S\}$.

Theorem 5.4 (Pattern decomposition)**.** *For finite* F *and finite-carrier* G,

$$A_F(\pi, K; G) = \sum_{\rho \geq \pi} \sum_{S \cap K = \varnothing} W_F(\rho, S; G),$$

$$W_F(\pi, S; G) = \sum_{\rho \geq \pi} \mu(\pi, \rho) \sum_{J \subseteq S} (-1)^{|S \setminus J|} A_F(\rho, \Sigma \setminus J; G).$$

1id1:core:20261109-0002@v1 1id1.net

Thus the counts of G on the fragment generated by F under quotients and shape restrictions are linearly equivalent to the joint distribution $(W_F(\rho, S; G))_{\rho,S}$ of (kernel, violated shapes) over the $|B_G|^{|B_F|}$ carrier functions.

Proof. Möbius inversion on $\Pi(B_F) \times 2^{\Sigma}$, whose Möbius function is the product. □

The dependence of A_F on π is where occurrence identity of the *target* is read: for $G_1 = \langle\{a, b\}, E, \{aa\}\rangle$, $G_2 = \langle\{a, b\}, E, \{ab\}\rangle$ and $F = \langle\{x, y\}, E, \{xy\}\rangle$, $A_F(\bot, \Sigma) = 1$ for both while $A_F(\top, \Sigma)$ (constant maps) is 1 and 0. Theorem 4.12 shows that this dependence can nevertheless be recovered without ever using quotients, by enlarging the carrier of the test instead. The decomposition theorems are organisational: they state what information a single test, closed under its own quotients and shape restrictions, carries, and they are the identities behind the inversion steps of Sections 3 and 6. The nullary part $\Upsilon_G \cap E^*$ is invisible to fragments without all-core records, since $\hom(F, G) = [\Upsilon_F \cap E^* \subseteq \Upsilon_G] \cdot \hom(F^\circ, G)$ with F° the class without all-core records.

6 Regular heaps

$\mathbf{TS}_E$ is a quasitopos [1, Section 5], [3]; its regular monomorphisms are the injective maps with $\Upsilon_1 = F_E(f)^{-1}(\Upsilon_2)$, the *heap-reflecting* injections. Let $\mathrm{inj}^{\mathrm{r}}(F, G)$ count them; the *regular profile* $\mathrm{inj}^{\mathrm{r}}(-, G)$ is the Token analogue of induced-substructure counting. For finite targets it is as good as the profile:

Proposition 6.1 (Regular profile, finite heaps)**.** *For F finite and G finite,*

$$\mathrm{inj}(F, G) = \sum_K \mathrm{inj}^{\mathrm{r}}\big(\langle B_F, E, K\rangle, G\big), \tag{5}$$

the sum over finite K with $\Upsilon_F \subseteq K \subseteq F_E(B_F)$, of which only finitely many terms are non-zero. Hence a finite class is determined by its regular profile on finite tests.

Proof. An injective Token map $f : F \to G$ determines $K_f = F_E(f)^{-1}(\Upsilon_G) \supseteq \Upsilon_F$, finite because Υ_G is finite and $F_E(f)$ is injective on words; f is heap-reflecting from $\langle B_F, E, K_f\rangle$, and conversely. There are finitely many injections $B_F \to B_G$, so finitely many K_f occur and the sum has finite support. Thus $\mathrm{inj}(F, G)$, hence $\hom(F, G)$ by (4), is a function of the regular profile; Theorem 3.1 finishes. □

For infinite heaps the regular profile on finite tests is not complete, and the reason is structural rather than a defect of the proof:

Proposition 6.2 (Finite tests do not cover regular heaps)**.** *$H_1 = \langle\{a\}, E, \{a^n : n \geq 1\}\rangle$ and $H_2 = \langle\{a\}, E, \{a^{2n} : n \geq 1\}\rangle$ have regular heaps, are non-isomorphic, and $\mathrm{inj}^{\mathrm{r}}(F, H_1) = \mathrm{inj}^{\mathrm{r}}(F, H_2)$ for every finite F.*

Proof. For $B_F = \varnothing$ the empty map reflects the heap iff $\Upsilon_F = \varnothing$, independently of the target. For $B_F \neq \varnothing$ any injection sends a letter x to a and the preimage of either heap contains infinitely many powers of x, so it is not the finite Υ_F: both counts are 0. $H_1 \ncong H_2$ since $a \in \Upsilon_{H_1} \setminus \Upsilon_{H_2}$. □

(The ordinary count $\hom(\langle\{x\}, E, \{x\}\rangle, -)$ separates them; the failure is specific to induced counting, since a finite induced pattern never occurs in an infinite heap over the same letters.) The natural repair is to let tests have regular heaps and to bound their automaton size. This must be formulated with care:

1id1:core:20261109-0002@v1 1id1.net

Proposition 6.3 (Two boundary cases)**.** *Let targets and tests have finite carriers and regular heaps.*

(i) *If both targets G, G' have heaps recognised by $\leq m$ states and tests with $\leq m$ states are allowed, then $\mathrm{inj}^{\mathrm{r}}(-, G) = \mathrm{inj}^{\mathrm{r}}(-, G')$ on such tests implies $G \cong G'$.*

(ii) *If only G is constrained to $\leq m$ states and G' is arbitrary, no function ϕ makes tests with $\leq \phi(m)$ states complete: for $k = \phi(1)$, $G = \langle \{a\}, E, \varnothing \rangle$ and $G' = \langle \{a, b\}, E, \{b^{k+1}\} \rangle$ have the same regular profile on all tests with $\leq k$ states.*

Proof. (i) Take $F = G$ and $F = G'$: $\mathrm{inj}^{\mathrm{r}}(G, G) \geq 1$ gives $\mathrm{inj}^{\mathrm{r}}(G, G') \geq 1$ and symmetrically, so there are injective Token maps both ways; Lemma 2.2 applies. (ii) A heap-reflecting injection into G' hitting b forces the induced language of the test on that letter to be exactly $\{x^{k+1}\}$, which needs more than k states; so every admissible test sees only the letter a of G', where both targets are empty, and the counts agree (1 or 0 according to Υ_F, times the number of injections into $\{a\}$). □

So "is there a state bound" is the wrong question: with the target's own size allowed it is trivially yes, without any constraint on the competitor it is no. Within the symmetric setting of (i) the threshold can be determined exactly. Fix a finite core E; a heap on carrier B is *m-state* if its minimal complete deterministic automaton over $B \sqcup E$ has m states. Let $\mathcal{K}_{n,m}$ be the classes with n letters and heap with $\leq m$ states, and let $\psi(n, m)$ be the least k such that the regular tests with $\leq n$ letters and $\leq k$ states are complete for $\mathcal{K}_{n,m}$ under $\mathrm{inj}^{\mathrm{r}}$.

Theorem 6.4 (State threshold)**.** *$\psi(n, m) = m$ for all $n \geq 1$, $m \geq 2$, and $\psi(n, 1) = 1$.*

Proof. $\psi(n, m) \leq m$ is Proposition 6.3(i). For the lower bound let $\nu_B(w)$ be the number of carrier letters in w and put, on one carrier B with $|B| = n$,

$$L_{\geq} = \{w \in F_E(B) : \nu_B(w) \geq m - 1\}, \qquad L_{\equiv} = \{w \in F_E(B) : \nu_B(w) \equiv m - 1 \pmod{m}\}.$$

Both are m-state: carrier letters advance a saturating, respectively cyclic, counter with m values, core letters fix it, and the m counter values are pairwise Myhill–Nerode inequivalent (the suffix b^{m-1-j} separates value j from any smaller value, respectively from any other residue). Any isomorphism between the two classes is a bijection of B, which preserves ν_B and therefore each language; but $b^m \in L_{\geq} \setminus L_{\equiv}$ for $m \geq 2$, so the classes are not isomorphic. Both lie in $\mathcal{K}_{n,m}$. Let F be a regular test with carrier C. If $C \neq \varnothing$, for any injection $\iota : C \to B$ one has $\nu_B(F_E(\iota)(w)) = \nu_C(w)$, so $F_E(\iota)^{-1}(L_{\geq}) = \{w : \nu_C(w) \geq m - 1\}$ and likewise for $L_{\equiv}$, again m-state languages. A heap-reflecting injection requires Υ_F to equal one of these, so if Υ_F has fewer than m states then $\mathrm{inj}^{\mathrm{r}}(F, -) = 0$ on both targets. If $C = \varnothing$, the empty map reflects the heap iff $\Upsilon_F = L \cap E^* = \varnothing$, the same condition for both (all-core words have $\nu_B = 0$, and $0 < m - 1$, $0 \not\equiv m - 1$ for $m \geq 2$). Hence tests with fewer than m states do not separate the pair, and $\psi(n, m) \geq m$. For $m = 1$ the heaps are $\varnothing$ or $F_E(B)$ and the one-letter empty-heap test separates them (n versus 0). □

The lower-bound pair exploits only the counting power of the heap language, with no interaction between letters. The number of *states* is thus settled; the quantities still open are the number of tests and the total cost.

Question 6.5. Let $\tau(n, m)$ be the least size of a family of regular tests (with $\leq n$ letters and $\leq m$ states) that is complete for $\mathcal{K}_{n,m}$ under $\mathrm{inj}^{\mathrm{r}}$, uniformly in the target. Proposition 6.3(i) uses the targets themselves, so $\tau(n, m) \leq |\mathcal{K}_{n,m}/\!\cong|$, the number of isomorphism types; is τ polynomial in m for fixed n? What is the total query cost (number of tests times the cost of one count $\mathrm{inj}^{\mathrm{r}}(F, G)$, itself $O(n!)$ automaton equivalence checks of size $O(m^2)$) against the ordinary route of Corollary 3.3?

1id1:core:20261109-0002@v1 1id1.net

7 Open questions

The binary growth order is linear, but the exact value at $n = 6$ remains open outside the certified families. Extending the complement obstruction beyond six letters, or explaining why it terminates, is a concrete computational and structural problem. Questions 4.14 and 6.5 concern sharper constants and query cost. Three further directions remain.

Fragments. For graphs, the correspondence between test fragments and equivalences is well developed (trees, bounded treewidth, cycles, bounded tree depth, hypergraphs [4, 5, 6]). The fragments specific to Token Space are the tree Tokens, prefix classes and operator-generated classes of [1]. Relating the prefix fragment to the prefix-naturality statement of [1] requires first defining the equivalence on targets that the naturality square induces; this definition is to be supplied before any conjecture is stated.

Efficient witness extraction. Corollary 3.3 finds an injective witness in $O(n!)$ automaton checks but extracts an ordinary-count witness only by a search over $\Pi(B_F)$; the decoder of Theorem 4.12 is explicit but not polynomial. Whether separating tests can be found in time polynomial in n and the automaton sizes, for either count, is open.

Densities and limits. $t(F, G) = \hom(F, G)/|B_G|^{|B_F|}$ is the probability that a uniform carrier function is a Token map; (1) become multiplicativity statements, and left-convergence can be defined on any fragment. A limit theory compatible with the chosen Token operations and heap presentations, particularly infinite heaps, is left for future work.

A Finite checks

C1–C25 are archived checks; C26–C28 support the six-letter classification and its independent verification. Exhaustive claims apply only to the stated ranges. Scripts and outputs for C1–C25 are in `checks`. Six-letter data and standalone verifiers are in `experiments/round2_20261108`; orbit counts and the full VM rerun are in `experiments/v13_validation`. Finite exhaustive certificates establish the stated finite classifications, whereas tests at finitely many parameters do not replace a general proof.

Check	Scope	Result
C1–C10	Product, exponential, injective and two-inversion identities on the recorded finite instances	All inherited checks pass
C11–C19	Small-target completeness, explicit collisions, witness construction, sampling, state bounds and decoder checks	Includes the exact $n \leq 4$ values and 4552 witness pairs
C20	All five-letter binary target types, complete order-5 pattern signatures	291 968 types; no collision
C21–C22	Finite signatures, named ports and twin-recovery hypotheses	All inherited checks pass
C23–C25	Targeted extensions, symbolic blow-ups and cycle-family checks	9356 cases; coefficient and cycle checks pass
C26	Full six-letter partition; independent CPU regrouping; 64 witness audits	113 230 types; two order-5 collision pairs
C27	Complete five-position laws, all 720 relabellings, path counts and centered moments	Both pairs first separate at order 6
C28	Burnside counts and byte-identical VM regeneration	130 164 types in the complement-closed family

The full check-by-check catalog is supplied as `audit/full_check_catalog.tex`; earlier checks are reported from their archived outputs, not rerun here.

Independent orbit count. For $p \in S_6$, each row cycle of length ℓ permits $[x^d] \prod_{C \in \mathrm{Cyc}(p^\ell)} (1 + x^{|C|})$ invariant rows of weight d. Multiply over row cycles and average over all 720 permutations to count constant-out-degree types. For symmetric matrices, the fixed count is $2^{o(p)}$, where $o(p)$ counts orbits on unordered vertex pairs, including loops. A multivariate orbit polynomial, tracking each row sum, independently gives symmetric/constant-degree intersection sizes $(1, 4, 8, 12, 8, 4, 1)$ for degrees $0, \ldots, 6$, yielding 130 164 types in $\mathcal{C}_6$. These computations use no graph-generation or GPU routines.

Resources and reproducibility. The complete rerun used an Intel Xeon Platinum 8255C VM (10 vCPUs), a Tesla V100-SXM2-32GB, GCC 11.4.0 and PyTorch 2.5.1 with CUDA 12.1. Sequential generation of the five families took 10.56 seconds in total; the degree-3 run peaked at 1 050 140 KiB resident CPU memory. For a fixed batch of 256 targets at order 5, three warm measurements gave median CPU/GPU times 0.1385/0.01236 seconds, including output transfer: about 11.2× for this step, not an end-to-end speedup claim. All regenerated representative files and certificates matched the original run. Scripts, exact outputs, full timing records and standalone verifiers are included in the two experiment directories named above.

References

[1] W. Pan, *Token Space: A Category Theory Framework for AI Computations*, arXiv:2404.11624v4, revised 2 October 2026.

[2] L. Lovász, Operations with structures, *Acta Math. Acad. Sci. Hungar.* 18 (1967) 321–328.

[3] J. Adámek, H. Herrlich, G. Strecker, *Abstract and Concrete Categories*, Wiley, 1990.

[4] H. Dell, M. Grohe, G. Rattan, Lovász meets Weisfeiler and Leman, *ICALP* 2018.

[5] M. Grohe, Counting bounded tree depth homomorphisms, *LICS* 2020.

[6] J. Böker, Color refinement, homomorphisms, and hypergraphs, *WG* 2019.

[7] M. Freedman, L. Lovász, A. Schrijver, Reflection positivity, rank connectivity, and homomorphism of graphs, *J. Amer. Math. Soc.* 20 (2007) 37–51.

[8] L. Lovász, *Large Networks and Graph Limits*, AMS Colloquium Publications 60, 2012.

[9] J.-Y. Cai, A. Govorov, On a theorem of Lovász that $\hom(\cdot, H)$ determines the isomorphism type of H, *ITCS*, LIPIcs 151 (2020), 17:1–17:15. doi:10.4230/LIPIcs.ITCS.2020.17.

[10] B. Young, Equality on all #CSP instances yields constraint function isomorphism via interpolation and intertwiners, *Electron. J. Combin.* 32(4) (2025), P4.42. doi:10.37236/13879.

[11] B. ten Cate, P. G. Kolaitis, A. Á. Kristjánsson, Adaptive query algorithms for relational structures based on homomorphism counts, *MFCS*, LIPIcs 345 (2025), 34:1–34:18. doi:10.4230/LIPIcs.MFCS.2025.34.

[12] W. Pan, *The Limits of Structural Consistency: A Token-Space Audit of LLM Counting*, 1id1 preprint, v1, 8 November 2026. `1id1:core:20261108-0001@v1`. `https://1id1.net/id/core/20261108-0001/v1/`.