The Infrastructure–Isolation Principle

J. R. Landers

Reliance, semantic neighborhoods, and proof complexity in Mathlib

A formal library grows by reuse. Definitions support lemmas, lemmas support later theorems, and automation hides long chains of accumulated infrastructure behind short tactic scripts. This suggests two competing pictures of maturity. Reuse could make deep results resemble one another: shared machinery might concentrate the library into dense families. Or reuse could enable specialization: common infrastructure might support increasingly particular endpoints whose proofs have few close analogues.

Measurements on a seed-0 uniform sample of 10,000 theorem–proof pairs from the 52,187 nonempty tactic traces in LeanDojo Benchmark 4 support the second picture [4]. Every target uses Mathlib commit `3c307701` [5]. Theorems resting on more kernel-visible infrastructure have more involved recorded proofs, but smaller semantic neighborhoods. The contraction is stronger for proofs than for statements. We call this joint pattern the *Infrastructure–Isolation Principle*:

> *In a mature formal library, declarations that depend on more mathematical infrastructure tend to have more involved proofs but fewer close semantic peers; procedural specialization outpaces semantic specialization.*

Intuitive interpretation

The dependency graph measures how much mathematical scaffolding a theorem stands on. The embeddings measure how many close companions it has, and the complexity score measures how involved its recorded construction is. The surprise is that more scaffolding accompanies fewer companions—especially for proofs. This is an empirical pattern, not a theorem of Lean.

Reliance in the kernel graph. Let $G = (V, E)$ be the directed graph of declarations in the pinned Mathlib environment. An edge $u \to v$ means that declaration v uses constant u in its kernel-checked type or value, including the elaborated proof term. The export contains 888,992 declarations and 13,264,773 direct edges. Unlike identifiers written visibly in tactic syntax, these edges include constants introduced by automation such as `simp`.

This graph-first view is also entering learned proving. Graph2Tac builds hierarchical representations of new formal concepts from the dependency structure of a Coq development and updates them online as the library grows [8]. Here the graph becomes an object of measurement in its own right: not only a route to the next tactic, but a record of how accumulated infrastructure positions a theorem.

For target declaration i, let $A(i)$ contain every upstream declaration that can reach i, and let $d(u, i)$ be the length of the shortest directed path from u to i. With discount $0 < \alpha \leq 1$, define dependency mass and its reported reliance score by

$$D_\alpha(i) = \sum_{u \in A(i)} \alpha^{d(u,i)}, \qquad R_i = \log(1 + D_{0.5}(i)) \,.$$

Here log is the natural logarithm. At $\alpha = 1$, dependency mass counts distinct ancestors; $\alpha = 0.5$ emphasizes nearby infrastructure without discarding the transitive closure. All 10,000 searches terminate within 52 levels. The median target has 39 direct dependencies, shortest-path depth 18, and $R_i = 4.668$.

Connectedness in two semantic spaces. For each target, the statement view contains its theorem declaration and the proof view contains its newline-joined traced tactics, each preceded by a fixed descriptive label. Cohere Embed v4 (Amazon Bedrock model `cohere.embed-v4:0`, input type `clustering`) maps the two texts separately to $e_i^{(S)}, e_i^{(P)} \in \mathbb{R}^{1024}$ [6]. For view $v \in \{S, P\}$, normalize

$$x_i^{(v)} = \frac{e_i^{(v)}}{\|e_i^{(v)}\|_2}, \qquad \|e\|_2 = \left(\sum_k e_k^2\right)^{1/2}.$$

The dot product $x_i^{(v)} \cdot x_j^{(v)}$ is then cosine similarity. Let $\mathcal{N}_{100}^{(v)}(i)$ be the exact top-100 neighbors of i in view v. A view-specific threshold τ_v is the 99th percentile of 500,000 fixed random-pair cosines: $\tau_S = 0.690$ and $\tau_P = 0.755$. For $j \in \mathcal{N}_{100}^{(v)}(i)$, set

$$w_{ij}^{(v)} = \max\left\{0, x_i^{(v)} \cdot x_j^{(v)} - \tau_v\right\}, \qquad p_{ij}^{(v)} = \frac{w_{ij}^{(v)}}{\sum_{\ell \in \mathcal{N}_{100}^{(v)}(i)} w_{i\ell}^{(v)}}.$$

The effective semantic neighborhood is

$$S_i^{(v)} = \exp\left(-\sum_{j \in \mathcal{N}_{100}^{(v)}(i)} p_{ij}^{(v)} \log p_{ij}^{(v)}\right),$$

with $S_i^{(v)} = 0$ if every weight is zero. This is the entropy-equivalent number of neighbors sharing the retained similarity weight: one dominant neighbor gives a value near one, and many balanced neighbors give a larger value, capped here at 100 [7]. Median connectedness is 31.3 in statement space but only 8.6 in proof space.

Complexity of the recorded proof. Let s_i be tactic-step count, r_i direct kernel-reference count, h_i maximum dependency depth, and q_{it} the fraction of proof steps whose tactic head is t. Tactic-head entropy is $H_i = -\sum_t q_{it} \log q_{it}$. If $z(y) = (y - \bar{y})/\operatorname{sd}(y)$ standardizes a component over the 10,000 targets, define

$$C_i = \frac{1}{4}\left[z(\log(1 + s_i)) + z(\log(1 + r_i)) + z(h_i) + z(H_i)\right].$$

Thus C_i is an equal-weight index of observed structural involvement. It is not proof length alone, and it is not a claim about intrinsic mathematical difficulty.

Across targets, write R, C, S_S, S_P for the four score variables obtained from $R_i, C_i, S_i^{(S)}, S_i^{(P)}$, respectively.

The empirical sign pattern. Write $\rho(X, Y)$ for Spearman rank correlation, the ordinary Pearson correlation of the ranks of X and Y. The central observation is

$$\rho(R, C) = 0.734 > 0, \qquad \rho(R, S_P) < \rho(R, S_S) < 0, \qquad \rho(C, S_P) < \rho(C, S_S) < 0,$$

with measured values

$$\rho(R, S_S) = -0.314, \quad \rho(R, S_P) = -0.431, \qquad \rho(C, S_S) = -0.398, \quad \rho(C, S_P) = -0.584.$$

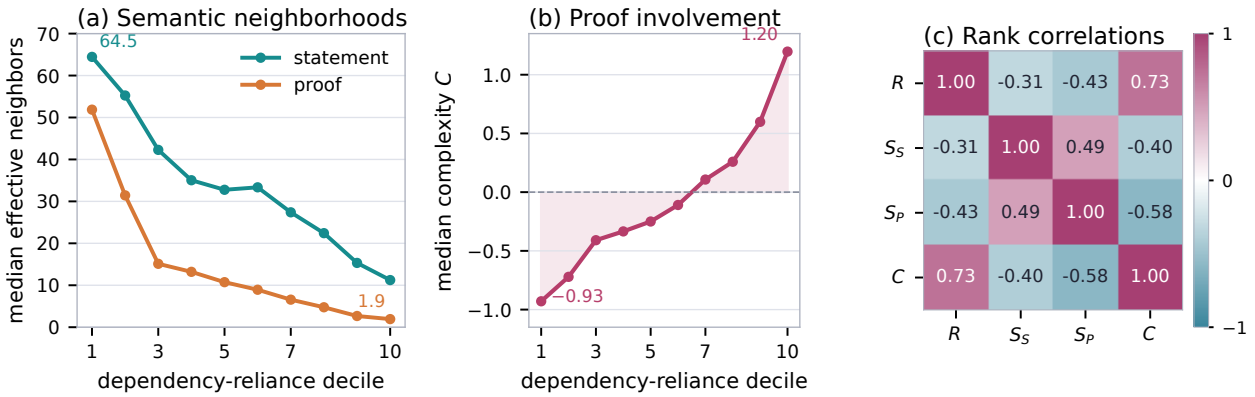

Figure 1: The Infrastructure–Isolation Principle in the frozen 10,000-target sample. Equal-count reliance deciles show (a) shrinking statement and proof neighborhoods and (b) increasing proof complexity. Panel (c) gives Spearman rank correlations among reliance R, statement connectedness S_S, proof connectedness S_P, and complexity C. Effective neighborhood medians, rather than embedding projections, are shown.

Statement and proof connectedness remain positively but incompletely aligned: $\rho(S_S, S_P) = 0.490$. The pattern is therefore not "everything difficult is far from everything else." It says that structural accumulation and semantic density point in opposing directions, while statement and proof geometries preserve a partial bridge.

The positive R–C association is partly built into the definitions: C contains direct reference count and dependency depth. The negative connectedness associations are not. Neither S_S nor S_P uses the dependency graph or tactic counts. They supply the nontrivial half of the principle.

A graded pattern. Figure 1 orders targets by R_i and divides them into ten equal-count groups. From lowest to highest decile, median statement neighborhood falls from 64.5 to 11.2, proof neighborhood from 51.9 to 1.9, and complexity rises from -0.93 to 1.20. Proof connectedness decreases at every step; statement connectedness has one small reversal. Alternative dependency discounts, semantic thresholds, and leave-one-component-out complexity scores preserve rankings with correlations at least 0.904, 0.926, and 0.941, respectively.

Two conditional arguments operate at different scales. Orthogonal extension models how one proof can leave a local semantic community; the later rank-sign law explains the population correlation directions under a latent infrastructure coordinate.

Proposition (orthogonal-extension attrition). Fix unit vectors $y_1, \dots, y_m$ representing one proof's local semantic community and a nonzero semantic core μ. Suppose proof growth adds components $\delta_1, \dots, \delta_k$ of common norm σ that are mutually orthogonal and orthogonal to μ and every y_j. For

$$e_k = \mu + \sum_{\ell=1}^{k} \delta_\ell, \qquad x_k = \frac{e_k}{\|e_k\|_2},$$

at any fixed threshold $\tau > 0$, the thresholded strength, surviving-neighbor count, and effective neighborhood computed on this fixed community are all nonincreasing in k.

Proof. Writing $b_j = \mu \cdot y_j$ and $a_k = (\|\mu\|_2^2 + k\sigma^2)^{-1/2}$ gives $x_k \cdot y_j = a_k b_j$. Since a_k decreases, each $w_j(k) = \max\{0, a_k b_j - \tau\}$ decreases, proving the first two claims. Scaling weights by $1/a_k$ does not change their normalized entropy. Put $\lambda = \tau / a_k$ and, between threshold crossings, let m_λ be the

active-set size and B_λ the sum of its b_j. Then

$$\pi_j(\lambda) = \frac{b_j - \lambda}{B_\lambda - m_\lambda \lambda}, \qquad \pi_j'(\lambda) = \frac{m_\lambda b_j - B_\lambda}{(B_\lambda - m_\lambda \lambda)^2}.$$

Consequently the derivative of Shannon entropy is nonpositive: after the common normalization term cancels, it is a negative multiple of the covariance between b_j and $\log(b_j - \lambda)$, two similarly ordered sequences. Entropy is continuous when a vanishing weight drops out, and falls to zero when none survives. □

This additive path models embedding motion, not Embed v4's architecture, and connects to C only when greater structural involvement contributes such directions.

Intuitive interpretation

Orthogonal novelty diverts normalized embedding energy from the old community, attenuating its cosines; the positive cutoff removes weak neighbors first. Tactic count alone is insufficient, and global isolation additionally requires incomplete replacement by new neighbors.

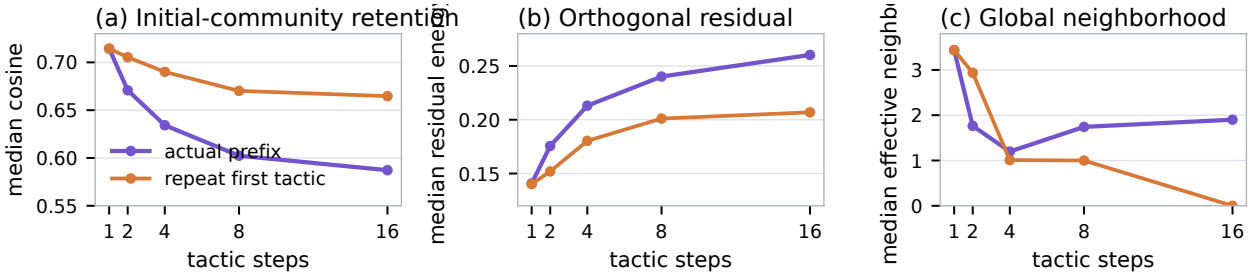

Figure 2: Proof-prefix trajectories ($n = 300$). Actual prefixes leave the initial community faster than repetition (a,b) but find new peers; repetition finds none and is more globally isolated at 16 steps (c).

A within-proof mechanism check. At fixed seed we select 300 of 396 targets with at least 16 tactics, embedding five prefixes $(1, 2, 4, 8, 16$ tactics) against the frozen 10,000 full-proof embeddings; matched controls repeat the first tactic. If L_i spans the initial top-100 neighbors and P_{L_i} is orthogonal projection, let $\eta_i(k) = 1 - \|P_{L_i} x_i(k)\|_2^2$. From 1 to 16 tactics, median cosine to the initial neighbors fell from 0.714 to 0.587 for actual proofs, versus 0.665 for repetition. Median η rose from 0.141 to 0.260 (in 81.3% of proofs), versus 0.207 under repetition: actual prefixes depart more than repeated length alone.

Replacement is incomplete. Componentwise medians show retained initial neighbors falling from five to zero, one new neighbor, and two globally; effective size falls from 3.44 to 1.90, while its change correlates −0.637 with residual change. Repetition retains more old geometry but gains no threshold neighbor and ends at effective size zero: this encoder also displaces repeated text. The trajectory supports local attrition plus incomplete entry into a smaller new community, not a claim that novelty must isolate more than any length increase.

Proposition (the infrastructure rank-sign law). Let X' be an independent copy of X, and let $U_X = \Pr(X' < X) + \frac{1}{2}\Pr(X' = X)$ be its population midrank for $X \in \{R, C, S_S, S_P\}$. Suppose a latent infrastructure coordinate Z makes $m_X(z) = \mathbb{E}[U_X \mid Z = z]$ nondecreasing for $X = R, C$ and nonincreasing for $X = S_S, S_P$. If the rank residuals are pairwise conditionally uncorrelated given Z, then

$$\rho(R, C),\ \rho(S_S, S_P) \geq 0, \qquad \rho(X, Y) \leq 0 \quad \text{for } X \in \{R, C\},\ Y \in \{S_S, S_P\}.$$

Proof. Total covariance reduces the left side below to the right; for an independent copy Z' of Z,

$$2\,\mathrm{Cov}(U_X, U_Y) = \mathbb{E}[(f(Z) - f(Z'))(g(Z) - g(Z'))].$$

Here $f = m_X$ and $g = m_Y$. The product has the sign claimed when the functions move in the same or opposite directions; rank standardization preserves it. Strict sign on a positive-probability set gives strict inequality. □

Intuitive interpretation

One hidden infrastructure dial that raises reliance and complexity while lowering both neighborhoods forces the observed signs, but not their strengths; faster proof isolation remains empirical.

Corollary (the one-factor tetrad test). Specialize further to standardized rank variables

$$\widetilde{U}_R = a_R Z + \varepsilon_R, \quad \widetilde{U}_C = a_C Z + \varepsilon_C, \quad \widetilde{U}_{S_S} = -a_S Z + \varepsilon_S, \quad \widetilde{U}_{S_P} = -a_P Z + \varepsilon_P,$$

where $a_R, a_C, a_S, a_P > 0$, $\mathrm{Var}(Z) = 1$, and the residuals are mutually uncorrelated and uncorrelated with Z. Pairwise correlations are then signed products of loadings, so

$$\rho(R, C)\rho(S_S, S_P) = \rho(R, S_S)\rho(C, S_P) = \rho(R, S_P)\rho(C, S_S).$$

Intuitive interpretation

A linear one-factor model forces the three products to agree. Their measured values—0.360, 0.183, and 0.172—require nonlinearity, correlated residuals, or another procedural axis.

Interpretation and retrieval. Statements recur in mathematical families, while tactic scripts reflect available lemmas, automation, and intermediate goals; companion studies likewise separate content from procedure while retaining a local bridge [1, 2]. Greater dependency depth accompanies smaller neighborhoods, more strongly in proof space. The high-reliance, low-connectedness tail therefore needs broad library coverage but offers few procedural precedents. Retrieval should combine semantic proximity, which can aid generation [3], with dependency coverage; LeanSearch v2 frames a related challenge as *global premise retrieval* [9].

Scope of the principle. These scores describe one snapshot, one recorded proof per theorem, and one encoder; dependency mass includes formalization machinery, scripts are not canonical, and neighborhoods are model-dependent. Prefixes and repetition controls are encoder probes, not completed Lean proofs, and the observational local mechanism does not causally explain statement-space isolation. Replication across snapshots, libraries, encoders, and multiple proofs should test the principle. Within those limits, common foundations accompany specialization without homogenization.

References

[1] J. R. Landers. “Textures of Modern Formal Mathematics.” 2026.
[2] J. R. Landers. “What Statements Know, What Proofs Do.” 2026.
[3] J. R. Landers. “Nearby Examples, New Proofs.” 2026.
[4] K. Yang et al. “LeanDojo: Theorem Proving with Retrieval-Augmented Language Models.” NeurIPS, 2023.
[5] The mathlib Community. “The Lean Mathematical Library.” CPP, 2020.
[6] Amazon Web Services. “Cohere Embed v4 Model Details.” Amazon Bedrock User Guide, 2025.
[7] M. O. Hill. “Diversity and Evenness: A Unifying Notation and Its Consequences.” Ecology 54(2), 1973.
[8] L. Blaauwbroek et al. “Graph2Tac: Online Representation Learning of Formal Math Concepts.” 2024.
[9] G. Gao et al. “LeanSearch v2: Global Premise Retrieval for Lean 4 Theorem Proving.” 2026.