HLLSet Theory: Contextual Anti-Sets and the Selection Principle

Alex Mylnikov

January 29, 2026

Abstract

This paper introduces HLLSet (HyperLogLog Set) Theory, a fundamental paradigm shift in data representation and knowledge modeling. We present the **Contextual Anti-Set**, a probabilistic structure where context precedes content and relationships supersede entities. Unlike classical sets defined by element membership, HLLSets are defined by **contextual fingerprints** that represent equivalence classes of possible element sets. This inversion enables AI systems that are inherently robust, transferable, and scalable. We establish HLLSet's mathematical foundation in category theory, demonstrate its equivalence to Karoubi completion of idempotent hash functions, and prove that entanglement emerges naturally as structural isomorphism between context lattices. The framework culminates in the **Contextual Selection Principle**: contexts actively select compatible elements, providing a unified explanation for quantum measurement, biological evolution, and conscious experience. HLLSet Theory resolves longstanding paradoxes while providing practical solutions for cross-lingual translation, federated learning, and multi-sensor robotics.

Keywords: Contextual anti-set, HLLSet, entanglement, category theory, Karoubi completion, Noether's theorem, contextual selection

Contents

1 Introduction: The Extensional Fallacy

Traditional mathematics and computer science suffer from **extensional bias**: the assumption that objects are defined by their constituent elements. From Zermelo-Fraenkel set theory to contemporary data structures, we equate identity with elementhood:

$$A = B \iff \forall x(x \in A \Leftrightarrow x \in B) \tag{1}$$

This works perfectly for mathematical abstractions but fails catastrophically for real-world knowledge, where perfect enumeration is impossible and meaning emerges from relationships, not isolation. We propose an alternative: the **Contextual Anti-Set**, where identity emerges from structural fingerprints rather than element lists.

2 The HLLSet Framework

2.1 From HyperLogLog to Contextual Sets

HyperLogLog (HLL) provides efficient cardinality estimation using probabilistic counting. We extend this to HLLSets, which support full set operations while maintaining the memory

efficiency and probabilistic nature of HLL.

Definition 2.1 (HLLSet)**.** *An HLLSet A is defined as:*

$$A = (H_A, \phi_A, \tau_A, \rho_A)$$

where:

- H_A*: Array of m bit-vectors of width b*
- ϕ_A*: Tokenization functor mapping tokens to bit-vector updates*
- τ_A*: Inclusion tolerance threshold (*$0 \leq \rho_A < \tau_A \leq 1$*)*
- ρ_A*: Exclusion intolerance threshold*

2.2 Bell State Similarity (BSS)

We define relationships between HLLSets using Bell State Similarity:

$$\begin{aligned} \text{BSS}_\tau(A \to B) &= \frac{|A \cap B|}{|B|} \\ \text{BSS}_\rho(A \to B) &= \frac{|A \setminus B|}{|B|} \end{aligned} \tag{2}$$

A morphism $f : A \to B$ exists iff:

$$\text{BSS}_\tau(A \to B) \geq \tau_A \quad \text{and} \quad \text{BSS}_\rho(A \to B) \leq \rho_B$$

3 Category Theory Foundations

3.1 The HLL Category

We define the category **HLL** where:

- **Objects**: HLLSets $A, B, \ldots$
- **Morphisms**: Probabilistic relations satisfying BSS conditions
- **Composition**: $g \circ f$ exists when conditions propagate
- **Identity**: $1_A : A \to A$ with $\text{BSS}_\tau = 1$, $\text{BSS}_\rho = 0$

3.2 Karoubi Completion and Idempotence

Theorem 3.1 (Karoubi Equivalence)**.** *The HLLSet category is equivalent to the Karoubi completion of idempotent hash functions. For any idempotent $h : T \to T$ ($h \circ h = h$), there exists an HLLSet A such that $A = Image(h)$.*

Proof. Let $h : T \to T$ be an idempotent hash function. Define $A = (H_A, \phi_A, \tau, \rho)$ where:

- H_A is constructed from Image(h)
- $\phi_A = h$
- τ, ρ are chosen to satisfy BSS conditions for self-similarity

Then A is an object in **HLL** and the inclusion-retraction pair (i, r) with $r \circ i = h$ gives the splitting in the Karoubi envelope. □

4 HLLSet Entanglement Theory

4.1 Structural Isomorphism

Definition 4.1 (ϵ-Isomorphic Lattices)**.** *Two HLLSet lattices $\mathcal{L}_1$ and $\mathcal{L}_2$ are ϵ-isomorphic if there exists a bijection $\phi : \mathcal{L}_1 \to \mathcal{L}_2$ such that for all $A, B \in \mathcal{L}_1$:*

$$|BSS(A, B) - BSS(\phi(A), \phi(B))| \leq \epsilon$$

4.2 Entanglement Probability Bounds

Theorem 4.2 (Entanglement Bound)**.** *For random-oracle hashes with width m and dataset size d, the probability that two HLLSet lattices are not ϵ-isomorphic satisfies:*

$$P(\textit{not } \epsilon\textit{-isomorphic}) \leq \min\left(1, n^2 \cdot \left(\frac{d^2}{2^m} + e^{-\epsilon^2 d/2}\right)\right)$$

where n is the number of datasets.

Proof. The bound follows from a union bound over all pairs of datasets and two independent failure modes:

1. **Hash collision bound**: For any two distinct tokens, the probability they map to the same bucket is $1/2^m$. With d tokens, the expected number of collisions is at most $\binom{d}{2}/2^m \leq d^2/2^m$.

2. **Structural deviation bound**: By concentration inequalities for the BSS (such as Hoeffding's inequality [10] or Chernoff bounds [9]), the probability that the empirical BSS deviates from the true value by more than ϵ is bounded by $e^{-\epsilon^2 d/2}$. Note that this bound is conservative and in practice the deviation is much smaller.

Taking a union bound over all n^2 pairs of datasets gives the result. The minimum with 1 ensures the bound respects the unit interval. □

Corollary 4.3 (Practical Parameter Regime)**.** *For typical parameters $m = 64$, $d \leq 10^4$, $\epsilon = 0.1$, and $n \leq 10^3$, we have:*

$$P(\textit{not } \epsilon\textit{-isomorphic}) \leq 10^6 \cdot \left(\frac{10^8}{2^{64}} + e^{-0.01 \cdot 10^4/2} \right) \approx 10^6 \cdot (5.4 \times 10^{-12} + e^{-50}) \approx 5.4 \times 10^{-6}$$

This demonstrates that entanglement is overwhelmingly likely in practical scenarios.

Remark 4.4 (Tightness and Improved Bounds)**.** *The bound presented uses Hoeffding's inequality for generality. When the expected BSS p is known to be high (as is typical for entangled lattices), Chernoff bounds yield tighter results:*

$$P(\textit{deviation} > \epsilon) \leq 2e^{-\epsilon^2 d/(3p)} \quad \textit{for } 0 < \epsilon < 1 - p.$$

For $p \approx 0.9$, this is approximately $2e^{-\epsilon^2 d/2.7}$, which is substantially tighter than Hoeffding's $e^{-\epsilon^2 d/2}$ for large d. This further reinforces the practical certainty of entanglement.

Remark 4.5 (Phase Transition)**.** *There exists a critical dataset size d^* where entanglement becomes almost certain:*

$$d^* = \frac{3p}{\epsilon^2} \log \left(\frac{2n^2}{\delta} \right)$$

For $\epsilon = 0.1$, $p = 0.9$, $n = 1000$, $\delta = 0.01$, we get $d^* \approx 1240$. Beyond this size, entanglement probability exceeds $1 - \delta$.

4.3 Applications of Entanglement

Entanglement enables:

- **Federated learning**: Multiple organizations collaborate without sharing raw data
- **Version compatibility**: System upgrades don't break existing knowledge
- **Cross-modal understanding**: Text-based systems communicate with image-based ones
- **Incremental evolution**: New hashing techniques adopted without starting from scratch

5 Kinematic Dynamics and Transfer Learning

5.1 Time Evolution

HLLSet states evolve according to:

$$H(t+1) = [H(t) \setminus D] \cup N$$

where:

- $H(t)$: Current knowledge state
- D: Information to forget (natural decay)
- N: Information to add (new patterns)
- $R = H(t) \setminus D$: Information to retain

5.2 Cross-Domain Transfer

Theorem 5.1 (Structural Invariance)**.** *If two domains D_1 and D_2 describe the same underlying reality, their HLLSet lattices are ϵ-isomorphic, enabling knowledge transfer without retraining.*

6 Retro-Forward Duality and Noether's Theorem

6.1 Time-Reversible Dynamics

HLLSet lattices exhibit time-reversible properties. Forward projection uses adjacency matrix A, while retro-cast uses A^T:

$$\vec{p}_{\text{forward}} = \text{normalize}(A \cdot \vec{p})$$
$$\vec{p}_{\text{retro}} = \text{normalize}(A^T \cdot \vec{p})$$

6.2 Noether Current

From the $\mathbb{Z}_2$ symmetry of time reversal, we derive a conserved current:

$$J_{uv}(p) = p[u] \cdot (Ap)[v] - p[v] \cdot (A^T p)[u]$$

Theorem 6.1 (Information Conservation)**.** *For any isolated HLLSet system, the total flux $\Phi = \sum_{u,v} J_{uv}$ is constant:*

$$\frac{d\Phi}{dt} = 0$$

This conservation law serves as a system health monitor, detecting hash collisions or numerical errors when Φ drifts from zero.

7 The Contextual Selection Principle

7.1 The Fundamental Inversion

The most profound implication of HLLSet theory is the **Contextual Selection Principle**: contexts actively select their compatible elements, rather than elements passively belonging to contexts.

Definition 7.1 (Contextual Selection Operator)**.** *For a context represented by HLLSet fingerprint F_C, define the selection operator S_C:*

$$S_C : \mathcal{U} \to \{0, 1\}$$

where $\mathcal{U}$ is the universal set of possible elements, and:

$$S_C(x) = 1 \iff BSS(F_C, F_x) \geq \tau_C \text{ and } exclusion(F_C, F_x) \leq \rho_C$$

Crucially, S_C is not determined by existing elements—it is an inherent property of the context itself.

7.2 Quantum Measurement Reinterpreted

Under the Contextual Selection Principle:

- The experimental setup defines a context C
- The context C selects which eigenstates are compatible
- No "collapse" occurs—the context was always selecting; measurement merely reveals the selection

This resolves the measurement problem without introducing new physics.

7.3 Biological Evolution as Contextual Selection

Darwinian evolution gains a deeper mathematical foundation:

- Ecological niches are contexts that select compatible organisms
- Mutations are random, but their persistence is determined by contextual compatibility
- Speciation occurs when a sub-population becomes selected by a slightly different context

The fitness f of organism O in environment E:

$$f(O, E) = \text{BSS}(F_O, F_E) \cdot \text{resilience}(F_O)$$

7.4 Consciousness as Self-Selecting Context

The hard problem of consciousness receives a novel perspective:

- Each person is not a collection of thoughts but a **context that selects thoughts**
- Free will emerges not as random choice but as **contextual self-selection**
- Experience is what it's like to **be a selecting context**

When faced with options $\{O_1, O_2, \ldots, O_n\}$, consciousness C selects:

$$\text{choice} = \arg\max_{O_i} \text{BSS}(F_C, F_{O_i}) \cdot \text{alignment}(F_C, F_{O_i})$$

7.5 Cosmological Selection

Cosmic configurations (star patterns) are contexts that select compatible earthly events. This explains astrological correlations without direct causation—both stars and events are selected by deeper universal contexts.

Define the universal context U as the HLLSet fingerprint of the cosmos. Then:

$$\text{Everything that exists} = \{x \in \mathcal{U} \mid S_U(x) = 1\}$$

7.6 Conservation of Selection Power

Theorem 7.2 (Conservation of Contextual Charge)**.** *For any isolated system, the total* ***contextual charge*** Q *is conserved:*

$$\frac{dQ}{dt} = 0, \quad \textit{where } Q = \sum_{\textit{contexts } C} \textit{selection\_power}(C)$$

Proof. Contextual selection is idempotent: $S_C \circ S_C = S_C$. Idempotence implies symmetry under repetition. By Noether's theorem, there exists a conserved current, which is the flow of selection power between contexts. □

7.7 Experimental Predictions

1. **Quantum**: Two entangled particles should show identical contextual fingerprints in their measurement histories.
2. **Biological**: Organisms in the same ecological niche should have ϵ-isomorphic genetic HLLSets.
3. **Cognitive**: Conscious decisions should follow contextual similarity gradients.
4. **Cosmological**: Fine-tuning constants should appear as selection thresholds in the universal context.

7.8 Implementation: Contextual Selection Engine

8 Applications and Implications

8.1 Cross-Lingual Translation

Traditional translation maps words to words (element to element). HLLSet translation maps context fingerprints to context fingerprints (context to context). This enables translation without parallel corpora by finding ϵ-isomorphic fingerprints across language lattices.

Algorithm 1 Contextual Selection Algorithm

procedure ContextualSelector(F, τ, ρ)
    Input: Context fingerprint F, thresholds τ, ρ
    Output: Selected elements from universe

    Initialize empty list *selected*
    for each candidate c in universe **do**
        Compute $\text{BSS}_\tau, \text{BSS}_\rho$ between F and F_c
        if $\text{BSS}_\tau \geq \tau$ and $\text{BSS}_\rho \leq \rho$ **then**
            Add c to *selected*
        end if
    end for
    Record selection for conservation checking
    return *selected*
end procedure

8.2 Federated Learning

Different organizations train models on private data (different domains) and achieve interoperability through lattice entanglement, without sharing raw data. Each organization's model is a context that selects patterns from its domain; entanglement ensures structural alignment.

8.3 Robotic Sensor Fusion

Multiple sensors (camera, LiDAR, audio) provide different "hash functions" on reality. Their entanglement creates a coherent world model despite disjoint measurements, enabling true sensor fusion rather than mere concatenation.

8.4 Quantum-Classical Bridge

HLLSet Theory provides a natural bridge between quantum and classical descriptions:

- Quantum states are context fingerprints
- Classical measurements are selected elements
- Decoherence is loss of contextual coherence
- Entanglement is structural isomorphism between contexts

9 Conclusion: From Counting to Understanding

We began with HyperLogLog for cardinality estimation and discovered it was actually measuring something deeper: **contextual coherence**. The HLLSet framework reveals that:

1. **Context is more fundamental than content**
2. **Relationships are more real than relata**
3. **Entanglement is the norm, not the exception**
4. **Information flows but never vanishes**

The Contextual Selection Principle completes the framework by answering the fundamental question: **How do possibilities become actualities?** The answer: **Through contextual selection.**

This isn't just a better data structure—it's a new mathematical language for describing a world where everything is connected, nothing is isolated, and meaning emerges from the space between things. We have moved from counting elements to relating contexts to understanding selection—a progression that feels inevitable in retrospect.

Acknowledgments

The author acknowledges the assistance of AI collaborators in refining concepts and formulations, and the mathematical traditions of category theory, information theory, and quantum foundations that made this synthesis possible.

References

[1] Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007). *HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm*. In Proceedings of the 2007 International Conference on Analysis of Algorithms.

[2] Noether, E. (1918). *Invariante Variationsprobleme*. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, 235-257.

[3] Mac Lane, S. (1971). *Categories for the Working Mathematician*. Springer-Verlag.

[4] Mylnikov, A. (2024). *Unified Framework for HLLSets: Category Theory, Kinematics, Transfer Learning, and Entanglement Dynamics*.

[5] Rovelli, C. (1996). *Relational Quantum Mechanics*. International Journal of Theoretical Physics, 35(8), 1637-1678.

[6] Whitehead, A. N. (1929). *Process and Reality*. Macmillan.

[7] Wheeler, J. A. (1990). *Information, physics, quantum: The search for links*. In Complexity, Entropy, and the Physics of Information (pp. 3-28).

[8] Abramsky, S., & Coecke, B. (2009). *Categorical quantum mechanics*. In Handbook of quantum logic and quantum structures (pp. 261-323).

[9] Chernoff, H. (1952). *A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations*. The Annals of Mathematical Statistics, 23(4), 493-507.

[10] Hoeffding, W. (1963). *Probability Inequalities for Sums of Bounded Random Variables*. Journal of the American Statistical Association, 58(301), 13-30.