Under Transformation

A Mathematical Search for the Self

You are not the person you were ten years ago. The cells are different, the memories have shifted, the beliefs have updated, and that haircut was a mistake you have since corrected. Yet something persists, or seems to. The ancient Greeks had a name for this puzzle. If you replace every plank of the Ship of Theseus, is it the same ship? Two thousand years of philosophy have not settled it. But mathematics has a tool that philosophy does not: the **invariant**.

An invariant is a property that survives transformation. The Euler characteristic of a surface, the Jones polynomial of a knot [1]: each certifies that something remains "the same" through change. This essay asks whether a person has one. It will try four times, with four mathematical frameworks. Each will fail. But each failure will be more precise, and more unsettling, than the last.

Under Deformation

Topology studies properties preserved under continuous deformation: stretching, bending, compressing, but never cutting or gluing. The central tool is the **Euler characteristic** [2]:

$$\chi = V - E + F$$

where V, E, F count vertices, edges, and faces of a polyhedral decomposition. For a cube: 8 vertices, 12 edges, 6 faces.

$$\chi(\text{cube}) = 8 - 12 + 6 = 2$$

Deform that cube into a sphere, a pyramid, any convex polyhedron: χ stays 2. But for a torus, $\chi = 0$. Since the invariants differ, no continuous deformation connects them. One computation, total certainty. This is the power of an invariant: it converts a question about *all possible transformations* into a single number.

Challenge 1.

1.1. A surface has 20 vertices, 30 edges, and 12 faces. What is χ? Could it be a sphere?

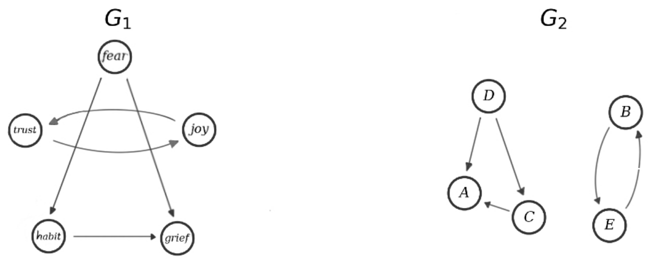
1.2. Can you find a polyhedron where $\chi = -2$? (Hint: think genus 2.)

Invariant found? No. Topology classifies surfaces beautifully, but it cannot tell us what the "surface" of a person is. To test identity, the self needs a more concrete mathematical form.

Under Rewiring

Let $G = (V, E)$ be a directed graph. Model a person's psychology with vertices for memories, beliefs, habits, fears, and desires, and edges for the causal, emotional, and associative connections between them. Losing a memory is a vertex deletion. Therapy restructuring a trauma-response pathway is an edge reweighting. Grief after losing a parent may strengthen some edges (memory to sadness) while severing others entirely. The self, on this model, is a graph under ongoing local transformation. This is either a profound account of personhood or a very sophisticated way of drawing a social network.

Now consider two specific graphs. Let G_1 have five vertices {*fear, trust, habit, grief, joy*} with edges *fear*→*habit*, *trust*→*joy*, *grief*→*fear*, *joy*→*trust*, *habit*→*grief*. Let G_2 have vertices {*A,B,C,D,E*} with different-looking edges. Define the mapping φ(*fear*)=*D*, φ(*trust*)=*B*, φ(*habit*)=*C*, φ(*grief*)=*A*, φ(*joy*)=*E*.



φ: fear ↦ D, trust ↦ B, habit ↦ C, grief ↦ A, joy ↦ E

Figure 1. *G_1 and G_2: visually different, structurally identical under φ. The reader can verify each edge.*

Check: *fear*→*habit* maps to *D*→*C*. Every edge checks out. Formally, two graphs are isomorphic when there exists a bijection

$$\varphi : V_1 \to V_2 \ \text{ such that } \ (u, v) \in E_1 \iff (\varphi(u), \varphi(v)) \in E_2$$

If identity is relational structure, then an isomorphic copy is indistinguishable from the original. This is excellent news for graph theory. It is deeply unfortunate news for everyone else. In *Serial Experiments Lain* [3], copies of a girl propagate across a network, each carrying her full relational structure. Graph theory makes the horror precise: anything isomorphic to you *is* you.

Figure 2. *The same five-node self-graph, replicated three times across a network, each in a different spatial arrangement. All three are isomorphic. Graph theory cannot identify the original.*

And verifying isomorphism is not easy. For two graphs on n vertices, there are $n!$ candidate bijections. For $n = 100$, that exceeds 10^{157}. In 2015, the mathematician László Babai announced he had placed graph isomorphism in quasipolynomial time [4], faster than anything known before. Within days, a flaw was found. Within weeks, he fixed it. The result held. The problem *still* resists full tractability. The question "are these two selves structurally the same?" is not just philosophically hard. It may be computationally hard in a precise complexity-theoretic sense.

Challenge 2.

2.1. Verify the remaining edges: does φ preserve trust → joy, grief → fear, joy → trust, habit → grief?

2.2. For n = 5, there are 5! = 120 candidate bijections. How many would you need to check for your own self-graph, if it had 1000 nodes?

Invariant found? No. Graph structure gives the self a concrete form, but it cannot distinguish original from copy, and it captures only a single moment. A graph is a snapshot. We need dynamics.

Under Evolution

A **dynamical system** [5] models how a state evolves. Model a person as a point in state space, each axis a psychological variable. Identity is no longer about what you are, but where your trajectory goes. Consider:

$$\frac{dx}{dt} = x - x^3 = x(1-x)(1+x)$$

Three equilibria: $x^* = -1, 0, +1$. To classify stability, linearise via $f'(x) = 1 - 3x^2$:

$$f'(\pm 1) = -2 < 0 \ \ (\text{stable}), \qquad f'(0) = 1 > 0 \ \ (\text{unstable})$$

A negative derivative means perturbations decay; a positive one means they grow. So ±1 are **stable attractors** and 0 is the **basin boundary**. Let us trace a trajectory. At $x = 0.8$, the derivative is $0.8 - 0.512 = 0.288 > 0$, so the trajectory moves rightward toward +1. At $x = 0.95$, the derivative is $0.95 - 0.857 = 0.093$: still positive, still approaching, but slowing down. The system converges. This is **Lyapunov stability**: small perturbations decay. In the language of selfhood, it is resilience.

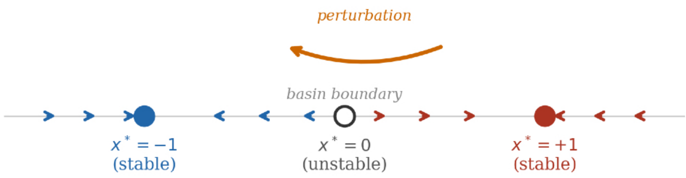


Figure 3. *Phase line for $dx/dt = x - x^3$. Trajectories converge toward ±1. A perturbation crossing $x = 0$ switches the destination entirely.*

But suppose grief or revelation pushes the trajectory past $x = 0$. Now at $x = -0.1$, the derivative is $-0.1 + 0.001 = -0.099$: the system moves *leftward*, toward −1. A different attractor. A different stable self. In Satoshi Kon's *Perfect Blue*, a woman's identity shifts continuously through performance and dissociation until the original is irretrievable. The mathematics makes this precise: the transition is smooth, the path continuous, and there is no identifiable point of rupture. Continuity of trajectory does not guarantee continuity of identity.

Challenge 3.

3.1. Compute dx/dt at $x = 0.01$ and $x = -0.01$. Which attractor does each approach?

3.2. At what value of x is the "pull" toward +1 strongest? (Hint: maximise $x - x^3$ on (0,1).)

One comfort is that most of us are at least *locally* stable. The discomfort is the word "locally."

Invariant found? No. Dynamics models how identity evolves, but it cannot prevent basin crossing. And it assumes the information sustaining the system stays intact. Over time, it does not.

Under Compression

Memory is **lossy compression**. Every act of recall reconstructs from incomplete data, introducing distortion. Over decades, vertices in the self-graph fade and edges disconnect. The question becomes quantitative: how much of the original self survives? **Shannon entropy** [6] gives the measuring tool:

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

Consider a simplified self with four equally likely memory-states (call them *a*, *b*, *c*, *d*). The entropy is:

$$H = -4 \times \tfrac{1}{4} \log_2 \tfrac{1}{4} = \log_2 4 = 2 \text{ bits}$$

Two bits: that is the information needed to specify which state the self is in. Now apply a lossy transformation: memory degrades and states c and d become indistinguishable. Before compression, observing the output told you everything: $H(X|Y) = 0$. After merging, observing the combined state leaves ½ bit of uncertainty about which original state produced it. The **mutual information** between past and present self drops:

$$I(X; Y) = H(X) - H(X \mid Y) = 2 - \tfrac{1}{2} = \tfrac{3}{2} \text{ bits}$$

Half a bit of selfhood, gone. The transformation is many-to-one: distinct pasts collapse to the same present, and the mapping cannot be inverted. Scale this across decades of degraded, reconstructed memory and the mutual information between "you at twenty" and "you at seventy" strictly decreases with every passing year. At this point, identity is beginning to look less like a soul and more like version control.

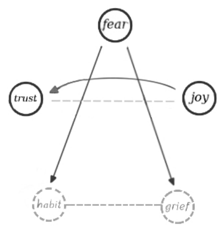


Figure 4. *The graph from Figure 1, after decades of information loss. Dashed nodes and edges mark degraded connections. Enough structure to suggest the original; not enough to verify isomorphism with it.*

Challenge 4.

4.1. If three of the four states merge into one, what is I(X; Y)? At what point is the original self entirely unrecoverable?

4.2. Is there a transformation on G_1 that deletes vertices and edges but preserves isomorphism with G_2? What is the minimum structure that must survive?

Invariant found? No. And this is the deepest failure. The self's information does not merely change: it degrades irreversibly. The past self does not exist inside the present one. It has been compressed beyond reconstruction. This is reassuring right up until someone asks for a backup.

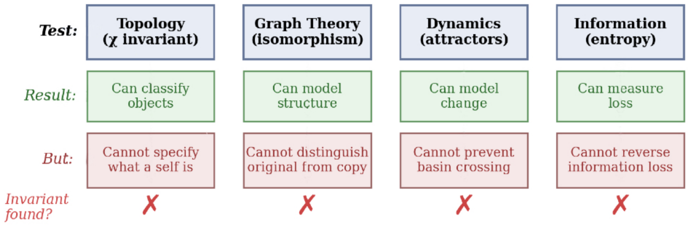


Figure 5. *Four frameworks, four failures. Each test sharpened the question; none produced an invariant.*

Perhaps the self is not an invariant at all. It may be something more fragile: a pattern that persists only while its distortions remain within tolerable bounds, held together by the fact that the transformations acting on it have, so far, been gentle enough to preserve something recognisable.

This leaves a precise open question. Define a person-graph G_t evolving over discrete time steps, subject to random vertex deletions and edge degradation. Let $I(G_0\,;\,G_t)$ be the mutual information between the original and the current structure. Is there a critical threshold I^* below which isomorphism with the original becomes computationally unverifiable? If such a threshold exists, then the boundary of identity is not a philosophical question. It is a **phase transition**, and in principle, it can be calculated.

The mathematics does not yet tell us where that boundary lies. But it tells us that "the same person" is a claim with a tolerance threshold. You cannot always tell when it has been crossed.

References

[1] V. F. R. Jones, "A polynomial invariant for knots via von Neumann algebras," *Bull. Amer. Math. Soc.*, 12(1), 1985.

[2] L. Euler, "Elementa doctrinae solidorum," *Novi Commentarii Acad. Sci. Petrop.*, 4, 1758.

[3] Y. Ueda (producer), *Serial Experiments Lain*, TV Tokyo, 1998.

[4] L. Babai, "Graph isomorphism in quasipolynomial time," *Proc. 48th ACM STOC*, 2016.

[5] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, 2nd ed., Westview Press, 2015.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27, 1948.