Identifying Agentic Substructures via Causal Emergence and Graph Learning

Working Draft

February 27, 2026

Abstract

We present a methodology for identifying agent-like substructures within complex dynamical systems. The approach combines two perspectives: (1) causal emergence analysis of transition probability matrices to identify scales with maximal effective information, and (2) graph structure learning to recover the message-passing architecture that generates the observed dynamics. We argue that for systems under selection pressure, these two analyses converge—the scale of maximal causal emergence corresponds to the scale at which Markov blanket structure naturally forms. This correspondence arises because selection optimises for causal power, and causal power is maximised when boundaries form at information bottlenecks. The framework provides a principled method for discovering where "agency" exists within arbitrary dynamical systems.

1 Selection, Free Energy, and the Emergence of Boundaries

Consider a population of dynamical systems subject to selection for persistence. Systems that continue to exist have, by definition, resisted dissolution into their environment. The question is: what structural features enable this persistence, and where do the boundaries of persistent systems naturally form?

The free energy principle [4] provides a universal answer. Any system that maintains its organisation over time must, in some sense, be minimising variational free energy—the divergence between its model of the world and the actual world. This is not a metaphysical claim but a mathematical tautology: systems that fail to predict and respond to their environment are systems that dissipate.

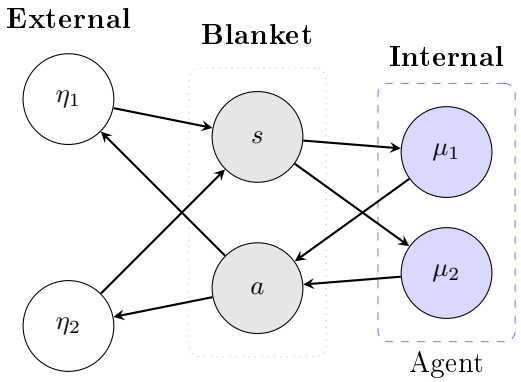


Figure 1: Markov blanket structure. Sensory states s mediate external-to-internal influence; active states a mediate internal-to-external influence. Internal states μ are conditionally independent of external states η given blanket states.

The structure that emerges from this minimisation is the Markov blanket (Figure 1). A Markov blanket is a statistical boundary: a set of states that renders internal states conditionally

independent of external states. Formally, if we partition states into internal (μ), blanket (b), and external (η), then b is a Markov blanket if $p(\mu|b,\eta) = p(\mu|b)$. The blanket screens off the inside from the outside.

Why do persistent systems develop blanket structure? The answer connects to causal power. Define the **causal power** of a subsystem A as the fraction of the future that A determines:

$$\mathrm{CP}(A) = \frac{\mathrm{I}(\mathrm{Future}; A)}{\mathrm{H}(\mathrm{Future})} \tag{1}$$

A system with high causal power controls its future; a system with low causal power is buffeted by external forces. Selection favours high causal power because systems that control their futures persist.

The key insight, developed in [7], is that maximising causal power is equivalent to minimising environmental free energy. If we write the environmental variational free energy as $\mathrm{EVFE} = \mathrm{H}(\eta|\text{observations})$—the uncertainty about the environment given what we can sense—then:

$$\mathrm{CP}(A) = 1 - \frac{\mathrm{H}(\mathrm{Future}|A)}{\mathrm{H}(\mathrm{Future})} \propto 1 - \mathrm{EVFE} \tag{2}$$

High causal power means low environmental uncertainty. A system achieves this by forming a clean boundary—a Markov blanket—that compresses environmental complexity into a tractable interface.

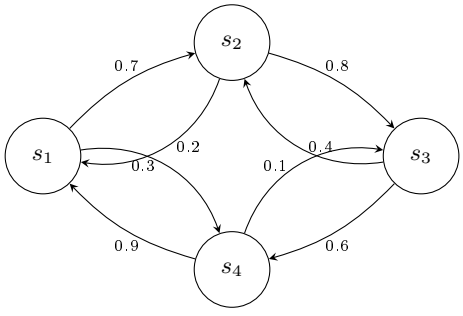
This gives us the first piece of the puzzle: **selection drives systems toward Markov blanket structure because blankets maximise causal power**. But this doesn't tell us *where* the blankets form. Not every partition of a system into inside/outside is equally good. The question becomes: which boundaries are optimal?

2 Causal Emergence and the Scale of Agency

To answer where boundaries form, we need to examine the dynamics more carefully. Consider a system specified by its transition probability matrix (TPM):

$$T_{ij} = P(X_{t+1} = s_j | X_t = s_i) \tag{3}$$

The TPM encodes everything about the one-step dynamics. It is agnostic about agent structure—there is no privileged partition into self and environment at this level. The TPM just describes what happens.



$$T = \begin{pmatrix} 0 & 0.7 & 0 & 0.3 \\ 0.2 & 0 & 0.8 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0.9 & 0 & 0.1 & 0 \end{pmatrix}$$

Figure 2: A Markov chain and its transition probability matrix. The TPM specifies the complete one-step dynamics without assuming any agent structure.

Hoel's framework of causal emergence [1, 2] asks: at what scale of description does the system have maximal causal power? The answer is measured by **effective information**:

$$\mathrm{EI}(T) = \mathrm{I}(X_t^{\text{max ent}}; X_{t+1}) \tag{4}$$

This quantifies how much the cause tells you about the effect when you have maximal uncertainty about which cause occurred. High EI means tight cause-effect relationships; low EI means diffuse, noisy dynamics.

The remarkable finding is that EI can *increase* under coarse-graining. If we group micro-states into macro-states via a partition $\pi : S \to S'$, the coarse-grained TPM T' sometimes has higher effective information than the original T. When this happens, the system exhibits **causal emergence**—the macro-level description has more causal power than the micro-level.

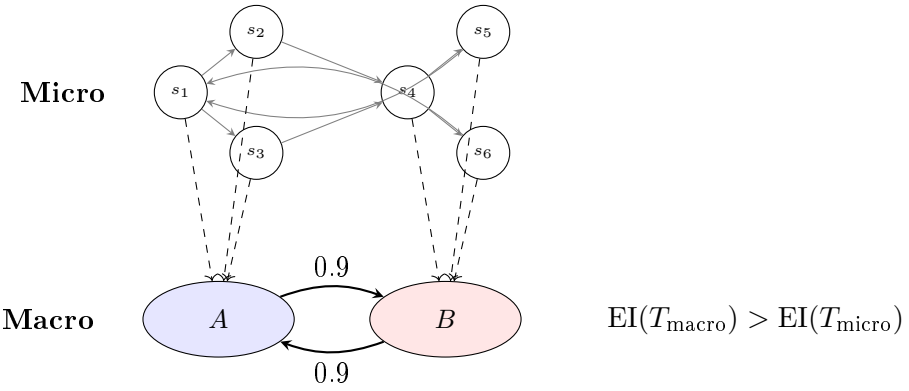


Figure 3: Causal emergence through coarse-graining. The micro-level has diffuse transitions; the macro-level has sharp, deterministic transitions. The macro description has greater causal power.

Why does this happen? High EI requires two things: **determinism** (causes tightly constrain effects) and **non-degeneracy** (different causes produce distinguishable effects). At the micro-level, thermal noise and redundant pathways can obscure causal relationships. Coarse-graining can eliminate this noise by grouping states that behave similarly, revealing the clean causal structure underneath.

The scale at which EI is maximised is, in a precise sense, the "natural" scale of description—the level at which the system's causal structure is most apparent. This is where we should expect agents to live.

3 From TPM to Message-Passing Structure

The TPM gives us a "view from outside"—a complete specification of dynamics as state-to-state transitions. But this isn't how the system sees itself. From the inside, a complex system is a collection of subsystems exchanging messages through channels. The question is: given only the TPM, can we recover this internal structure?

This is a graph learning problem. We observe the aggregate dynamics T and want to infer the underlying message-passing architecture $G = (V, E, w)$ that generates these dynamics. The nodes V are subsystems; the edges E are communication channels; the weights w encode the information content of messages.

The learning problem has a natural loss function derived from the TPM. We want a graph G whose implied dynamics match the observed dynamics:

$$\mathcal{L}(G) = D_{\mathrm{KL}}\left(T \,\|\, T_G\right) + \lambda \cdot \mathrm{Complexity}(G) \tag{5}$$

where T_G is the TPM implied by message-passing on graph G, and the complexity term favours sparse, hierarchical structures. This is related to variational inference: we're finding the simplest graph that explains the observed dynamics.

The key insight is that this learning problem should be solved *at the scale of maximal causal emergence*. Why? Because causal emergence identifies where the dynamics have clean

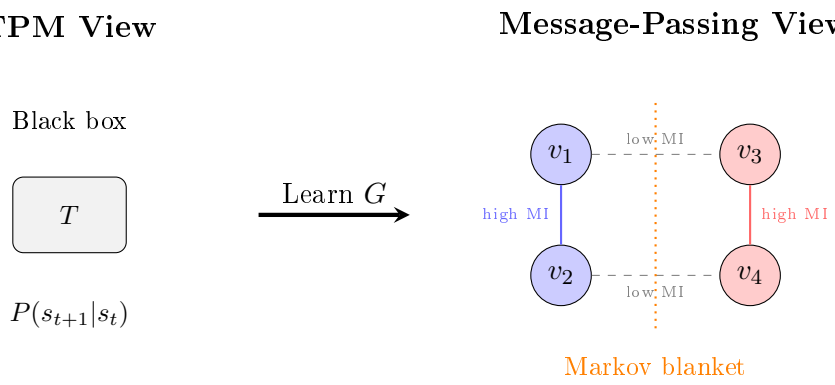


Figure 4: The graph learning problem. Given observed dynamics T, infer the message-passing structure G with subsystems as nodes and information channels as edges. Markov blankets appear as low-MI boundaries between tightly-coupled clusters.

structure. At scales with low EI, the dynamics are noisy and the graph structure is obscured. At the EI-optimal scale, the causal relationships are sharpest, making the underlying graph most identifiable.

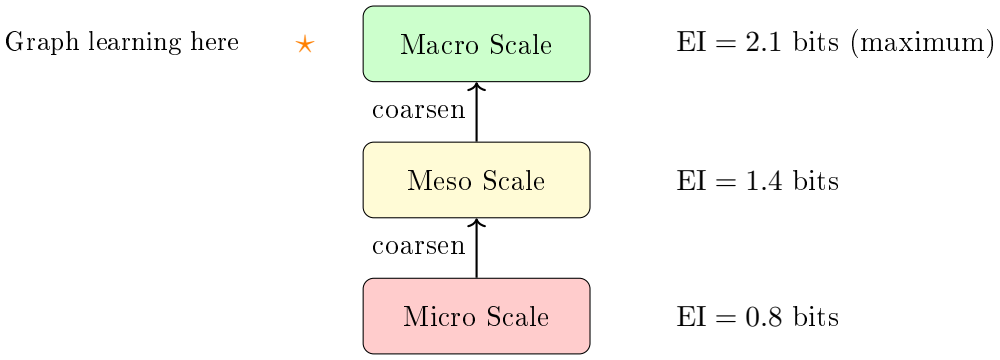


Figure 5: Hierarchical scale selection. The graph structure is most identifiable at the scale of maximum effective information, where causal relationships are sharpest.

Concretely, the procedure is:

1. Compute the causal emergence hierarchy by evaluating $\mathrm{EI}(T_\pi)$ for candidate coarse-grainings π
2. Identify the scale π^* with maximal effective information
3. At scale π^*, solve the graph learning problem to recover the message-passing structure G^*
4. The clusters in G^* with high internal connectivity and low external connectivity are the natural agent boundaries—the Markov blankets

This connects geometric deep learning to the causal emergence framework. Graph neural networks and related methods provide the computational machinery for step 3, while causal emergence analysis (step 1-2) tells us at what scale to apply these methods.

4 Why the Frameworks Converge

We now have two ways to identify agent structure:

- **Causal emergence**: Find the scale where EI is maximised

- **Graph learning**: Find the message-passing structure that explains the dynamics

The central claim is that these methods converge to the same answer for systems under selection pressure. The Markov blanket boundaries discovered by graph learning at the EI-optimal scale are precisely the boundaries that maximise causal power.

The correspondence can be expressed categorically. Let **Markov** be the category of Markov processes (objects are TPMs, morphisms are lumpable coarse-grainings) and let **InfoGraph** be the category of weighted graphs (objects are message-passing architectures, morphisms are graph contractions). There is a functor $F : \mathbf{Markov} \to \mathbf{InfoGraph}$ that maps a TPM to its implied information structure:

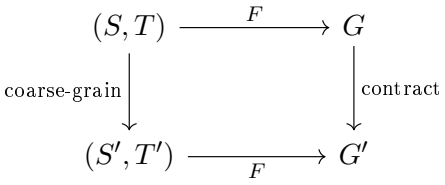


Figure 6: The functor F maps Markov processes to information graphs, preserving the coarse-graining/contraction structure.

The deeper claim is that for evolved systems, EI-optimal coarse-grainings in **Markov** map to Markov blanket structures in **InfoGraph**:

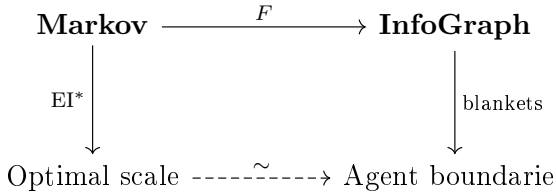


Figure 7: The correspondence: EI-optimal scales correspond to natural Markov blanket boundaries.

Why should this be true? The answer lies in what selection optimises. Systems under selection pressure are implicitly maximising causal power—their ability to determine their own futures. Causal power has two components:

1. **Internal coherence**: The agent's states should tightly predict each other (high internal MI)

2. **External separation**: The agent's boundary should minimise information flow with the environment (low cross-boundary MI)

These are exactly the conditions that produce both high EI (deterministic, non-degenerate dynamics at the agent scale) and clean Markov blankets (statistical separation between inside and outside). Selection finds the partition that optimises both simultaneously because *they are the same optimisation target viewed from different angles.*

The TPM perspective asks: at what scale does the system have maximal causal power?

The graph perspective asks: where are the information bottlenecks that separate subsystems?

For a system that has converged under selection, these questions have the same answer. The boundaries that maximise causal power are the boundaries at information bottlenecks—the natural joints where the system carves itself into agents.

References

References

[1] Hoel, E.P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *PNAS*, 110(49), 19790-19795.

[2] Hoel, E.P. (2017). When the map is better than the territory. *Entropy*, 19(5), 188.

[3] Hoel, E.P. (2024). Causal emergence 2.0. *arXiv preprint*.

[4] Friston, K. (2019). A free energy principle for a particular physics. *arXiv:1906.10184*.

[5] Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life. *J. R. Soc. Interface*, 15(138), 20170792.

[6] Pearl, J. (2014). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.

[7] Hallgren, J. (2023). Power-seeking = minimising free energy. *LessWrong*.

[8] Bronstein, M.M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478*.

[9] Rosas, F.E., Mediano, P.A., et al. (2020). Reconciling emergences: An information-theoretic approach. *PLoS Comp. Biol.*, 16(12), e1008289.