UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 35

The Social Compression Hypothesis

A Pre-Registered Test of Coordination Demand, Institutional Capacity, and Instability in Historical and Modern Societies, with a Formal Framework and Agent-Based Experiments

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

November 2026

Version 1.0

Abstract

Institutions, standards, records and administrative hierarchies can be read as devices that lower the communication cost of coordinating many people: a form of compression. This paper states that reading formally, derives what it does and does not imply, and tests its empirical corollary—that a mismatch between a society's coordination demand (population, territory, settlement structure) and its coordination capacity (administration, standardized rules and money, communication infrastructure, record-keeping) predicts subsequent instability—under a frozen pre-registration. We build demand, capacity and mismatch indices from the Seshat Polaris-2026 databank (626 polities; 127 CrisisDB crisis records with polity-level predictors, dated facts preceding the crisis where they exist), the Ethnographic Atlas (1,291 societies) and a V-Dem–Maddison country-year panel (14,637 country-years, 901 distinct event-years, 1820–2025), with a transparent block index and a latent-variable alternative, uncertainty propagation and out-of-sample validation. Capacity and scale are positively associated in all three datasets ([EMPIRICAL]). The mismatch index is positively associated with crisis severity in Seshat (0.20 log-odds per SD, stable across regions, eras and cultural spheres) and with instability onsets in modern states (0.39 per SD), but by the pre-registered criterion it fails: it does not beat the scale-only baseline out of sample in Seshat, nor the capacity-only baseline in the modern panel; the restriction it imposes on scale and capacity is rejected in the modern panel; a placebo index is not distinguishable from it in Seshat; the modern association is nearly as strong with past as with future events; and it is not robust to an Oster-type adjustment. No threshold is found in Seshat and only a small nonlinearity in the modern panel. Among Seshat polities with a crisis record, higher capacity is also associated with a *shorter* polity duration, against the hypothesis. Five theorems make the compression reading precise (coding gain, break-even scale, stale-code penalty, hierarchical routing, backlog dynamics), and factorial agent-based experiments (6,048 runs, five imposed institutional regimes, equal administrative budgets) instantiate them in silico on the cost side—bits, messages, queues—while locating the regimes where institutions hurt: unpaid set-up costs, saturated hierarchies, and weakly structured task environments ([SIMULATED]). We conclude that the compression framework is coherent and simulable, that scale and capacity each matter, but that their mismatch, as measurable in existing historical data, carries no identifiable predictive content of its own. The universal theory is not supported; the narrower results are stated with their scope.

Keywords: institutions, coordination, state capacity, social complexity, Seshat, pre-registration, agent-based simulation, information theory

JEL Classification: D02, D83, H11, N40, P48, C63

---

1. Introduction

Large societies coordinate. Taxes are assessed and collected, grain is stored and moved, disputes are settled, armies are raised, and the same unit of length is understood in two towns that have never exchanged a messenger. Each of these acts requires that people who do not know each other act on a shared description of the situation. The description has to be communicated, verified and remembered, and the cost of doing so grows with the number of people, the distances between them and the variety of things that must be agreed upon. The claim examined in this paper is that institutions—standardized rules, money, measurement systems, written records, administrative hierarchies, postal systems—are best understood as devices that lower that cost per coordinated task, in the ordinary sense in which a dictionary or a code lowers the cost of saying things: they compress. We call this reading the *Social Compression Hypothesis* (SCH). Its empirical corollary, which gives it teeth, is that a society whose coordination *demand* outruns its coordination *capacity* should be more likely to experience institutional breakdown.

The idea has a long ancestry, which Section 2 documents. Scalar stress (Johnson, 1982), administrative hierarchies as information-processing structures (WRIGHT and JOHNSON, 1975), the declining marginal returns to complexity (Tainter, 1988), organizational codes (Arrow, 1974; Cremer et al., 2007) and the informational foundations of state capacity (Lee and Zhang, 2017; Brambor et al., 2019) all contain pieces of it. The closest precursor is Shin et al. (2020), who found in the Seshat databank a second axis of variation opposing scale to information-processing characteristics and wrote, explicitly as a conjecture left for future research, that a "mismatch among aspects of social development should be considered as a generic cause of instability". To our knowledge, no one has operationalized a demand–capacity mismatch, linked it to an independently coded instability outcome, and tested it with out-of-sample validation. That is what this paper does, and the answer it reaches is mostly negative.

We proceed in the order a skeptical reader would want. First, we make the compression reading precise. Five theorems (Section 3) state when a shared code lowers expected communication per coordinated task (Shannon's source-coding bound), when the fixed cost of an institution is repaid (a break-even condition in scale and task intensity), what a code optimized for yesterday's tasks costs today (the Kullback–Leibler divergence), what hierarchical routing buys and what it concentrates at the apex, and how a backlog behaves when demand exceeds capacity (Lindley's recursion). None of these is new mathematics; their value is that they say exactly what a "mismatch" is in the model—the utilization $\rho = D/K$—and what would have to be true of historical proxies for an index such as $P = z(\log D) - z(K)$ to track it. We state that requirement as a conjecture, not a result.

Second, we pre-registered the empirical tests before computing any association between predictors and outcomes (the frozen document and its hash are in the replication package). The

pre-registration fixes the datasets, the unit of analysis, disjoint indicator sets for demand D and capacity K, the outcomes, the horizon, the models, the baselines that P must beat out of sample, the multiple-testing correction, a power calculation, the robustness set and the decision rule for abandoning the universal framing. We then built the indices from three real sources with full provenance: the Seshat Polaris-2026 release (626 polities; 56,274 facts, of which only 2,503 are dated; a CrisisDB table of crisis consequences for 150 polities) (Turchin et al., 2015; Hoyer et al., 2023), the D-PLACE Ethnographic Atlas v3.3 (Kirby et al., 2016; Murdock, 1967), and a V-Dem v16 country-year panel (Coppedge et al., 2020) merged with Maddison 2023 population (Bolt and van Zanden, 2024). Measurement is done twice, with a transparent block index and a two-parameter latent-variable model, and uncertainty in ranges and dates is carried through.

Third, we report what the pre-registered tests found (Sections 6 and 7). Capacity and scale are positively associated in every dataset, as the framework and much prior work predict, although the direction is not identified. The mismatch index P is positively associated with how badly a recorded crisis unfolds in Seshat (about 0.20 log-odds per standard deviation, stable when any macro-region or era is held out) and with the probability of a civil war, coup or irregular regime end in the following five years in modern states (about 0.39 per standard deviation). But the pre-registration asked a harder question than whether β_P is positive: it asked whether P beats the obvious baselines out of sample. It does not. In Seshat, a scale-only model does as well, a placebo index that replaces capacity with production infrastructure is not distinguishable from P, and P loses to scale alone when regions are held out. In the modern panel, a capacity-only model does as well or better in most forward-chaining and leave-one-region-out comparisons, the equal-and-opposite restriction that P imposes on scale and capacity is rejected, past instability is associated with P nearly as strongly as future instability, and the coefficient is not robust to an Oster-type adjustment for unobservables. We find no threshold in Seshat and a small nonlinearity in the modern panel. Polity duration and discontinuous succession are unrelated to P, and higher capacity is associated with shorter, not longer, polity durations. Under the decision rule we wrote down in advance, the universal framing is not supported: scale and capacity each carry signal, their difference carries none that we can identify.

Fourth, we report the constructive part. Factorial agent-based experiments with five *imposed* institutional regimes (Section 8; 6,048 main runs, 584 ablation runs, all labelled [SIMULATED]) measure messages, bits, success rates, latency, overhead, welfare and inequality directly. Shared codes and records lower the operating cost per successful task in essentially every cell (their effect on success is a parameter of the model, not a finding); the stale-code penalty after a distribution shock equals the divergence to within the rounding of Theorem 3; at equal administrative budgets, centralized hierarchies fail abruptly when relayed load exceeds the budget, and federated routing keeps most of the benefit at a budget on which the centralized design fails. The same experiments

locate where institutions hurt: a learning cost of 200 bits per agent is repaid only with a large and concentrated task space and a long horizon, saturated centralized hierarchies do worse than no institution at all, and a code can be longer than raw description when the task distribution is nearly uniform. These are the cases the framework predicts, and they are the cases historians describe.

The contribution, then, is threefold. We give the compression reading of institutions a formal statement that distinguishes what is proved from what is conjectured. We provide the first pre-registered, out-of-sample test of the demand–capacity mismatch conjecture on real historical and modern data, with a negative result that we think is more useful than another positive one would have been, because it shows that the mismatch, as it can be measured today, is a residual of two collinear quantities. And we provide a simulation testbed in which the mechanisms that the historical record cannot reveal—message counts, code lengths, queue saturation—can be measured. Section 9 states the scope conditions that survive, Section 10 the limitations, and Section 11 what a credible next test would require.

2. Related work and contribution

Every reference in this section was checked against Crossref by DOI or bibliographic query on 1 November 2026 (`research/prior_art.csv`, `research/bib_overrides.csv`); the abstracts and key results of the eighteen works closest to our hypothesis were read from primary or repository sources in a separate verification pass (`docs/novelty_search_report.md`). Works that we know only through reviews or citing articles are flagged in footnotes and are cited for their general thesis only.

2.1. Scalar stress and information-processing hierarchies

The oldest form of our idea is Johnson's *scalar stress*:[1] as the number of units that must be coordinated grows, the communication and decision load grows faster than linearly, and groups respond either by fission or by adding a hierarchical level, with a modal span of about six units per level (Johnson, 1982). Alberti (2014) estimate a related fission threshold on Hutterite colonies. WRIGHT and JOHNSON (1975) define the state as a specialized administrative hierarchy with at least three decision-making levels and reject population growth and exchange as single sufficient causes of its emergence in Susiana; Johnson (1978) extends the information-processing reading of decision-making organizations. Carneiro (1967) documents the log-log regularity between population and the number of organizational traits, NAROLL (1956) proposes an early index of social development, Feinman (2010) reviews the size–complexity relation, and Flannery (1972)

[1] Known to us through library records and through Alberti (2014); the chapter itself was not read.

frames institutional evolution as a change in information-processing architecture. None of this strand builds a demand/capacity *ratio*, and none tests it against instability.

2.2. Declining returns and collapse

Tainter (1988) argues that complexity is a problem-solving investment with declining marginal returns, so that collapse (simplification) eventually becomes rational and the society loses its capacity to absorb shocks.[2] Middleton (2017), Butzer (2012) and Cline (2014) review the collapse literature and warn against mono-causal accounts. Structural-demographic theory (Turchin, 2004; Turchin and Nefedov, 2009; Turchin, 2023) models secular cycles driven by population pressure, elite overproduction and fiscal distress; Goldstone (1991) is its historical origin. Sng (2014) is the closest economic precedent: in late-imperial China the cost of transmitting information over distance creates a principal–agent problem whose severity rises with territorial size, so that fiscal capacity decays during the prosperous eighteenth century—a demand (territory) versus capacity (monitoring) argument, tested within a single polity with prefecture-level data. Using the Seshat CrisisDB, Hoyer et al. (2024) show that crisis consequences are heterogeneous, largely uncorrelated with one another, and not well predicted by territory, religion or administrative complexity taken one at a time ("no typical societal crisis").

2.3. Seshat-based quantitative history

The Seshat Global History Databank (Turchin et al., 2015; François et al., 2016) underlies a growing quantitative literature. Turchin et al. (2017) found that nine complexity characteristics (population, territory, capital population, hierarchical levels, government, infrastructure, writing, texts, money) load on a single principal component explaining about three quarters of the variance across 414 polity-centuries. Shin et al. (2020) decomposed the remaining variance and identified a second axis opposing *scale* characteristics to *information-processing* characteristics, defined a "scale threshold" and an "information threshold" along the first component, and wrote—explicitly as a conjecture left for future research—that "mismatch among aspects of social development should be considered as a generic cause of instability". Miranda and Freeman (2020) clustered Seshat polities into two super-types with an unstable transitional cluster in which polities dwell only a few centuries. Turchin et al. (2022) compared more than 100,000 predictor combinations for the *growth* of complexity and found agricultural productivity interacting with military technology best supported; they did not model instability. Related Seshat-based work includes Currie et al. (2010) on the rise and fall of political complexity in Austronesia, Sheehan et al. (2018) on the coevolution of intensive agriculture and hierarchy, Kohler et al. (2017); Kohler and Smith, eds (2018) on long-run inequality,

[2] Content known from reviews and secondary sources; the book was not re-read for this paper.

Peregrine (2017, 2018) on resilience to disasters, and Hoyer et al. (2023) on the long-run factors that shape responses to changing climate. Whitehouse et al. (2019) was retracted (Whitehouse et al., 2021) after Beheim et al. (2021) showed that recoding missing moralizing-god data as "absent" drove the result; the Seshat team's "retake" (Whitehouse et al., 2022) reports qualitatively similar conclusions with corrected coding, and Slingerland et al. (2020) record the historians' objections. The methodological lesson—never treat unknown as absent, never forward-fill, report results on extant data—is written into our pre-registration.

2.4. State capacity and its informational foundations

Mann (1984, 1986) distinguish despotic from *infrastructural* power; Soifer (2008) and Hanson and Sigman (2021) discuss its measurement, the latter with a Bayesian latent model over 21 indicators for 1960–2015. Scott (1998) describes *legibility*—standardized names, cadastres, censuses, weights and measures—as the state's way of making society readable, and SCOTT (2017) traces it to the earliest states. Lee and Zhang (2017) operationalize legibility through age-heaping in censuses and show that it predicts tax contributions; Brambor et al. (2019) build an "information capacity" index (census, statistical agency, yearbooks, civil and population registers) for 85 polities since 1789; D'Arcy and Nistotskaya (2016) use cadastral records. Besley and Persson (2009, 2011), Dincecco (2017), Johnson and Koyama (2017), Mayshar et al. (2022), Borcan et al. (2017) and Acemoglu et al. (2001) study the origins and consequences of fiscal and legal capacity; Gennaioli and Voth (2015) link capacity to military competition, as do Tilly (1992), McNeill (1982) and HOFFMAN (2015). STASAVAGE (2010) shows that geographic scale limited representative assemblies in Europe—a direct instance of distance-driven coordination cost—and Stasavage (2020); AHMED and STASAVAGE (2020) and BLAYDES and CHANEY (2013) trace the long-run divergence of political institutions; Ko et al. (2018), Fernández-Villaverde et al. (2023), SCHEIDEL (2019) and Dincecco and Wang (2018) compare China and Europe. Classic comparative treatments of empires and bureaucracy are Weber (1978), Eisenstadt (1963), Trigger (2003), Yoffee (2005), Blanton and Fargher (2008), Fukuyama (2011) and Graeber and Wengrow (2021). None of these operationalize a *ratio* of coordination demand to capacity, nor test it against instability across pre-modern polities.

2.5. Transaction costs, codes and communication in organizations

Coase (1937), Williamson (1985) and North (1990, 1991) treat institutions as devices that economize on transaction costs; Hayek (1945) on dispersed knowledge, Simon (1955) on bounded rationality and Simon (1962) on near-decomposable hierarchies are the classic information-processing views. Arrow (1974) introduced organizational *codes*;[3] Cremer et al. (2007) model the firm's language as a

[3]Verified only through secondary citations.

code of bounded size in which precise words go to frequent events—literally lossy compression that trades scope against precision. Marschak and Radner (1972), Radner (1993), Bolton and Dewatripont (1994), Van Zandt (1999) and Garicano (2000) model hierarchies as information-processing networks with explicit communication costs; Dessein (2002) and Dewatripont and Tirole (2005) study authority and modes of communication; Crawford and Sobel (1982), HAGENBACH and KOESSLER (2009), Galeotti et al. (2013) and Calvó-Armengol et al. (2015) analyze strategic communication on networks. Mechanism-design results bound the *message-space size* needed to implement allocations (Mount and Reiter, 1974; Segal, 2007; Nisan and Segal, 2006), and communication complexity (Yao, 1979; Kushilevitz, 1997) gives the computer-science counterpart. The business-history literature documents the same economizing in practice: Chandler (1977), Yates (1989), Beniger (1986) and, earlier, Innis (1950). These results are the formal basis for Section 3, but they concern firms, markets or abstract protocols, not polities.

2.6. Coordination, common knowledge and conventions

Lewis (1969) and Schelling (1960) define conventions and focal points; Young (1993), Kandori et al. (1993) and Ellison (1993) show how conventions are selected and how local interaction speeds convergence; Morris (2000) on contagion and Skyrms (2010) on signalling complete the picture. Chwe (2000, 2001) argues that coordination requires *common knowledge* and that rituals and public communication structures exist to manufacture it cheaply; Rubinstein (1989) shows that almost-common knowledge is not enough. Morris and Shin (2002) and Angeletos and Pavan (2007) quantify the social value of public signals, and DEWAN and MYATT (2008) model leaders as coordination devices. Ostrom (1990, 2010) and Olson (1965) are the collective-action anchors; Boyd and Richerson (1985, 1988), Henrich (2004), Richerson et al. (2014) and Turchin (2016) supply the cultural-evolution mechanisms for norm maintenance in large groups, Gavrilets et al. (2008) and Turchin and Gavrilets (2009) the evolutionary dynamics of hierarchy, and Stanish (2017) the role of ritual in stateless societies; Greif (1993, 2006) and Milgrom et al. (1990) are the historical cases of reputation- and record-based enforcement.

2.7. Scaling and information theory

Bettencourt et al. (2007); Bettencourt (2013), Ortman et al. (2014) and Lobo et al. (2019) show superlinear scaling of socioeconomic outputs with settlement population, ancient and modern; Dunbar (1993) and Hamilton et al. (2007) document scaling constraints in small-scale societies. Scaling theory implies that interaction demand grows faster than population, but it assumes that infrastructure and institutions co-scale; it predicts benefits, not instability. Dedeo (2017) is the closest explicit statement that transitions in political order are driven by information-processing

innovations and that coarse-graining (lossy compression) is the central mechanism; Wolpert et al. (2017) surveys state-space compression, and Marzen and DeDeo (2017) give a rate-distortion model with a high-fidelity regime (cost linear in environmental complexity) and a low-fidelity regime. Flack (2012, 2017), Daniels et al. (2017) and Krakauer et al. (2020) develop coarse-graining and collective computation in animal societies; Hidalgo (2015) popularizes the information-growth view of economies. Shannon (1948) and Cover and Thomas (2005) are cited only for the definitions used in the simulations, where messages and distributions are actually defined. To our knowledge no published work uses the name "Social Compression Hypothesis" (search of 1 November 2026), but the idea has the clear ancestors listed above, which this paper credits.

2.8. Instability forecasting and methods

Goldstone et al. (2009) forecast instability onsets for 1955–2003 with regime type as the dominant predictor; FEARON and LAITIN (2003) and Hegre (2001) are the civil-war baselines, and CEDERMAN (2003) models the size distribution of wars. For binary time-series–cross-section data we follow Beck et al. (1998) and Carter and Signorino (2010); for event history, Cox (1972) and Box-Steffensmeier and Jones (2004); for logistic calibration, Hosmer et al. (2013); for clustered and spatial inference, Colin Cameron and Miller (2015) and Conley (1999); for sensitivity to unobservables, Oster (2017) and Cinelli and Hazlett (2019); for Galton's problem, Mace et al. (1994) and Dow (2007); for latent measurement, Bollen (1989), Samejima (1969), Treier and Jackman (2008) and Pemstein et al. (2018); for agent-based modelling, Epstein and Axtell (1996), Axtell et al. (2002), Macy and Willer (2002), Watts and Strogatz (1998) and Jackson (2008); and for pre-registration and the statistical crisis, Nosek et al. (2018) and Gelman and Loken (2014). Data sources are documented by Seshat: Global History Databank (2026), Kirby et al. (2016) and Murdock (1967), Bolt and van Zanden (2024), Coppedge et al. (2020) and Coppedge et al. (2026), and Hammarström et al. (2026).

2.9. Novelty matrix

Table 1 summarizes which elements of the Social Compression Hypothesis already exist in the closest works. The ranking by overlap is: Shin et al. (2020) (a verbal mismatch–instability conjecture, declared untested); Johnson (1982) (scalar stress); Sng (2014) (distance-driven information cost, one polity); Tainter (1988); the codes-as-compression micro-foundation of Arrow (1974), Cremer et al. (2007) and Dedeo (2017); Miranda and Freeman (2020); the CrisisDB results of Hoyer et al. (2024); and the settlement-scaling literature as a foil.

Table 1. Novelty matrix: elements of the Social Compression Hypothesis in the closest prior work

Work	Demand/scale	Capacity	Mismatch	Instability
Johnson (1982)	group size	hierarchy levels	implicit	fission (ethnogr.)
WRIGHT and JOHNSON (1975)	settlement hier.	admin. levels	–	–
Tainter (1988)	–	complexity	–	collapse (cases)
Sng (2014)	territory, distance	monitoring	implicit	fiscal decay
Turchin et al. (2017)	pop., territory	govt., writing	–	–
Shin et al. (2020)	PC2 scale side	PC2 info. side	verbal	conjectured
Miranda and Freeman (2020)	cluster features	cluster features	–	dwell time
Turchin et al. (2022)	complexity (outcome)	–	–	–
Hoyer et al. (2023, 2024)	territory (covariate)	admin. (covariate)	–	CrisisDB
Lee and Zhang (2017); Brambor et al. (2019)	–	legibility, info. capacity	–	–
Hanson and Sigman (2021)	–	latent capacity	–	–
Goldstone et al. (2009)	pop. threshold	regime type	–	onsets
Cremer et al. (2007); Arrow (1974)	org. scope	code size	trade-off	–
Chwe (2000, 2001)	network	common knowledge	–	–
Bettencourt (2013); Ortman et al. (2014)	settlement pop.	infrastructure	–	–
Dedeo (2017); Marzen and DeDeo (2017)	–	coarse-graining	–	–
This paper	log pop., territory, largest settlement, hierarchy	4 disjoint blocks (mean, IRT, PC)	$P = z(\log D) - z(K)$	CrisisDB severity; haz

2.10. Precise contributions and non-claims

1. *Empirical.* The first pre-registered test of the scale/information mismatch conjecture of Shin et al. (2020) against an independently coded instability outcome (the Seshat CrisisDB), with a transparent and a latent measurement model, propagated uncertainty, and out-of-sample validation, replicated in form on a modern country-year panel. The result is negative in the pre-registered sense: the mismatch index P is associated with crisis severity and with modern instability onsets, but it does not predict either outcome beyond its own components (scale in the historical panel, capacity in the modern panel), fails the out-of-sample rule against those baselines, and performs no better than a placebo index in the historical panel. What survives is that scale predicts crisis severity among crisis cases and that capacity predicts fewer onsets in modern states—neither of which is new, and neither of which is the Social Compression Hypothesis.
2. *Theoretical.* Formal propositions stating when a shared code or record system lowers the expected communication cost per coordinated task as a function of scale, when fixed institutional overhead makes it worse, by how much a stale code raises cost after a shock, and when hierarchies saturate; each is stated as a theorem with proof or as a conjecture, never as an empirical fact.
3. *Simulated.* A factorial agent-based comparison of five imposed institutional regimes with directly measured messages, bits, success rates and overhead, including an explicit search for regimes in which institutions reduce efficiency or add fragility (set-up costs that are never repaid, saturated hierarchies, weakly concentrated plan spaces).
4. *Not claimed.* No causal identification in either panel; no literal Shannon compression of historical institutions; no universal theory. The name "Social Compression Hypothesis" has no prior academic use that we could find, but the idea has clear ancestors (Table 1), which this paper credits, and the data do not support the universal framing under which the project was mandated.

3. A formal framework: institutions as codes

This section states the theory behind the Social Compression Hypothesis in a form that separates what is *proved* from what is *conjectured*. Definitions are conventions; theorems are statements proved below or standard results whose proofs are reproduced; conjectures are statements we believe worth testing, each with its falsifier. Nothing in this section is an empirical finding. The results concern an idealized communication system; Section 3.8 states what is not claimed and which pre-registered test, if any, can speak to each statement.

3.1. Primitives and definitions

Definition 1 (agents, tasks, plans)**.** A society has N agents. In each period, coordination tasks arrive at rate λ per agent. A task requires that $k \geq 2$ participants settle on a common element of a finite *plan set* S, $|S| = s$, the required plan being drawn independently across tasks from a distribution p on S with Shannon entropy $H(p) = -\sum_{x \in S} p(x) \log_2 p(x) \leq \log_2 s$ (Shannon, 1948; Cover and Thomas, 2005). Communication is counted in *messages* (one transmission between two agents) and in *bits* (the length of what is transmitted). Verification and record-keeping use a second channel whose cost is counted separately as *administrative overhead* (Definition 6).

Definition 2 (coordination demand)**.** The coordination demand of a period is the task volume that must be settled: $D = N\lambda k$ participant-slots per period. In the worst case in which every pair of agents must be reconciled, D is bounded by the number of pairs, $N(N-1)/2$, i.e. $D = \Theta(N^2)$; under bounded λ and k, $D = \Theta(N)$. Either way D is increasing in N.

Definition 3 (coordination capacity)**.** Fix a tolerated error rate $\varepsilon \in (0,1)$. The coordination capacity K of an institutional system is the maximal number of tasks it can settle per period such that the fraction of tasks settled on a wrong or inconsistent plan is at most ε. K depends on the available communication budget (bits per period), the code in use (Section 3.2), the routing structure (Section 3.5) and the overhead (Section 3.3).

Definition 4 (utilization)**.** $\rho = D/K$. When D and K are measured in the same units (tasks per period), $\log D - \log K = \log \rho$ exactly.

Limitation (not a theorem). The pre-registered empirical index $P = z(\log D) - z(K)$ uses standardized proxies (population, territory, settlement hierarchy for D; shares of administrative, legal/monetary, communication and record-keeping indicators for K). It is a monotone proxy for $\log \rho$ only under Conjecture 1; it is not a measurement of ρ, and the Seshat and V-Dem indicators are not Shannon quantities.

3.2. Coding gain of a shared code

Definition 5 (code)**.** A *shared code* is an injective map $c : S \to \{0,1\}^*$ known to all participants, with lengths $\ell(x) = |c(x)|$. It is *prefix-free* if no codeword is a prefix of another. Institutions in the sense of this paper (standardized names, units, legal forms, accounting categories, calendars) are shared codes in exactly this sense (Arrow, 1974; Cremer et al., 2007).

Assumption 1. Settling a task means that the initiating participant identifies the required plan $x \in S$ to the other $k-1$ participants without ambiguity; messages to different participants are counted separately in messages but identically in bits per recipient, so the result is stated per recipient.

Theorem 1 (coding gain)**.** *(i) Without a shared code, identifying an arbitrary element of S without error requires at least $\lceil \log_2 s \rceil$ bits per recipient in the worst case.*

(ii) With a shared prefix-free code that is optimal for p, the expected length per recipient $L_p^* = \min_{c \text{ prefix-free}} \sum_x p(x)\ell(x)$ satisfies

$$H(p) \le L_p^* < H(p) + 1. \tag{1}$$

(iii) Hence the expected saving per recipient, $G(p) = \lceil \log_2 s \rceil - L_p^*$, obeys $G(p) > \lceil \log_2 s \rceil - H(p) - 1 \ge 0$ whenever $\lceil \log_2 s \rceil \ge H(p) + 1$, and $G(p)$ is strictly positive as soon as p is sufficiently concentrated ($H(p) < \lceil \log_2 s \rceil - 1$).

Proof. (i) Without a shared code there is no prior agreement on which binary string denotes which plan, so the only error-free strategy is a fixed-length enumeration of S agreed by the trivial convention of listing; any scheme that distinguishes s alternatives with binary strings of a fixed length m needs $2^m \ge s$, i.e. $m \ge \lceil \log_2 s \rceil$. (The point is not that $\log_2 s$ is a lower bound for *all* conceivable codes; it is the cost of the *codeless* baseline, in which the message must be self-describing.)

(ii) *Lower bound.* For any prefix-free code the Kraft inequality holds: $\sum_x 2^{-\ell(x)} \le 1$ (Cover and Thomas, 2005, Thm. 5.2.1). Put $q(x) = 2^{-\ell(x)}/Z$ with $Z = \sum_x 2^{-\ell(x)} \le 1$. Then

$$\sum_x p(x)\ell(x) - H(p) = \sum_x p(x) \log_2 \frac{p(x)}{2^{-\ell(x)}} = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} - \log_2 Z = D_{KL}(p\|q) - \log_2 Z \ge 0,$$

because $D_{KL} \ge 0$ (Gibbs' inequality) and $-\log_2 Z \ge 0$. *Upper bound.* Choose $\ell(x) = \lceil -\log_2 p(x) \rceil$. Then $\sum_x 2^{-\ell(x)} \le \sum_x p(x) = 1$, so by the converse Kraft theorem (Cover and Thomas, 2005, Thm. 5.2.1) a prefix-free code with these lengths exists, and $\sum_x p(x)\ell(x) < \sum_x p(x)(-\log_2 p(x) + 1) = H(p) + 1$. The optimal code does at least as well.

(iii) Subtract (ii) from (i). □

Remark. Because the fixed-length code of $\lceil \log_2 s \rceil$ bits is itself prefix-free, the optimal code satisfies $L_p^* \le \lceil \log_2 s \rceil$ and the saving $G(p)$ is never negative; (iii) is only a lower bound. A suboptimal prefix code with Shannon lengths $\lceil -\log_2 p(x) \rceil$ also satisfies $H(p) \le L < H(p) + 1$ but can exceed $\lceil \log_2 s \rceil$ when p is nearly uniform; the simulations of Section 8 use Shannon lengths so that Theorem 3 applies verbatim, and this is why one cell there shows a code longer than raw description.

Remark. Theorem 1 is the source-coding theorem of Shannon (1948) applied to the plan set. Its content for the paper is only this: *a shared code converts a worst-case identification cost into an average-case cost that is lower exactly to the extent that tasks are predictable.* The "compression" in the name of the hypothesis refers to this step and to nothing else. The same economy appears in the mechanism-design bounds on message-space size (Mount and Reiter, 1974; Segal, 2007; Nisan and

Segal, 2006) and in communication complexity (Yao, 1979; Kushilevitz, 1997); the rate–distortion view of Marzen and DeDeo (2017) and the coarse-graining view of Dedeo (2017) are the closest conceptual neighbours.

3.3. Break-even scale of an institution

Definition 6 (institutional cost structure)**.** Adopting a code costs a fixed set-up $F \geq 0$ (designing and promulgating the standard), a learning cost $\ell \geq 0$ per agent (each of the N agents must learn the code), and an administrative overhead $a \geq 0$ bits per task (verification, record-keeping, enforcement). It saves $G > 0$ bits per task (Theorem 1, per recipient; multiply by $k-1$ if desired; the argument is unchanged).

Theorem 2 (break-even)**.** *Over a horizon with T tasks, the institution lowers total communication cost if and only if*

$$T\,(G-a) > F + N\ell. \tag{2}$$

With $T = N\lambda\tau$ over τ periods, this is equivalent to

$$\lambda\tau > \frac{F/N+\ell}{G-a} \quad \textit{when } G > a, \tag{3}$$

and the inequality can never hold when $a \geq G$.

Proof. Total cost without the institution is $C_0 = T\lceil \log_2 s \rceil$. With it, $C_1 = F + N\ell + T(L_p^* + a)$. Then $C_1 < C_0$ if and only if $F + N\ell < T(\lceil \log_2 s \rceil - L_p^* - a) = T(G-a)$. If $a \geq G$ the right-hand side is $\leq 0 < F + N\ell$ unless $F = \ell = 0$ and $a = G$, in which case the costs are equal, so strict improvement is impossible. If $G > a$, substitute $T = N\lambda\tau$ and divide both sides by $N(G-a) > 0$. □

Corollary 1 (scale advantage)**.** *The threshold $(F/N+\ell)/(G-a)$ is decreasing in N: holding λ, τ fixed, a larger society crosses the break-even point sooner, because the set-up cost is spread over more agents. It is* not *decreasing in N through ℓ: per-agent learning costs do not benefit from scale.*

This is the formal version of the observation that institutions "pay off at scale" (Johnson, 1982; Garicano, 2000; Bolton and Dewatripont, 1994), and it also makes precise when they do not: when overhead a is close to the coding gain G, or when horizons τ are short.

3.4. Mismatched code penalty: institutional rigidity

Theorem 3 (cost of a stale code)**.** *Let the code be optimal for p in the sense of using lengths $\ell_p(x) = \lceil -\log_2 p(x) \rceil$, and let the true task distribution be q with $q \ll p$ (every plan with positive*

q-probability has positive p-probability). Then the expected length under q satisfies

$$H(q)+D_{KL}(q\|p) \le \mathbb{E}_q[\ell_p(X)] < H(q)+D_{KL}(q\|p)+1. \tag{4}$$

The excess cost of using the p-code when the world is q, relative to the q-optimal code, lies in $(D_{KL}(q\|p)-1,\ D_{KL}(q\|p)+1)$ *bits per task.*

Proof. Since $-\log_2 p(x) \le \ell_p(x) < -\log_2 p(x)+1$,

$$\mathbb{E}_q[\ell_p(X)] \ge \sum_x q(x)\big(-\log_2 p(x)\big) = \sum_x q(x)\log_2\frac{q(x)}{p(x)} + \sum_x q(x)\big(-\log_2 q(x)\big) = D_{KL}(q\|p)+H(q),$$

and the same computation with the $+1$ gives the strict upper bound. The q-optimal code has expected length in $[H(q),H(q)+1)$ by Theorem 1(ii), which yields the stated interval for the excess. This is the "wrong code" result of Cover and Thomas (2005, Thm. 5.4.3). □

Corollary 2 (fragility)**.** *A shock that moves the task distribution from p to q raises the per-task communication cost of an existing institution by about* $D_{KL}(q\|p)$ *bits. Combined with Theorem 2, the institution remains worthwhile after the shock only if* $G(p)-D_{KL}(q\|p) > a$ *(up to the* ± 1 *rounding).*

An institution can therefore *lower* efficiency after a shock, and the more specialized the code (the smaller $H(p)$ relative to $\log_2 s$), the larger the divergence a given shock can produce. This is the precise sense in which standardization trades efficiency for rigidity; it echoes the declining marginal returns of Tainter (1988) but gives the mechanism a unit (bits per task) and a magnitude (a divergence).

3.5. Hierarchical routing bound and root load

Definition 7 (hierarchy)**.** A hierarchy is a rooted tree whose N leaves are the agents and whose internal nodes are administrative units; every internal node has at most $b \ge 2$ children; the depth (leaf-to-root distance) is $d = \lceil \log_b N \rceil$ for a complete b-ary tree. A task between two leaves is *settled by relay* when each participant sends a message up to their lowest common ancestor (LCA) and the LCA sends the settlement down.

Theorem 4 (routing and root load)**.** *(i)* Message bound. *Any pairwise task can be settled with at most* $2d$ *messages (up-path plus down-path through the LCA), versus up to* $N-1$ *messages for a flat broadcast in which the initiator must reach every other agent. For M pairwise tasks the total is at most* $2M\lceil \log_b N \rceil$.

(ii) Root load. *In a complete b-ary tree with $N = b^d$ leaves, if the two participants of a task are drawn uniformly at random among leaves (independently, with replacement), the probability that their LCA is the root is $1 - 1/b$. Hence the expected number of tasks that pass through the root is $M(1 - 1/b)$.*

(iii) Federated or modular coordination. *If a fraction $\theta \in [0,1]$ of tasks is* local *(both participants belong to the same first-level module, i.e. child subtree of the root, by construction of the task process) and the remaining $1 - \theta$ are drawn uniformly as in (ii), the expected root load is $M(1 - \theta)(1 - 1/b)$.*

Proof. (i) In a tree there is a unique simple path between two leaves; it passes through their LCA and has length at most $2d$ since each leaf is at depth $\leq d$ and the LCA is at depth ≥ 0. Sending one message per edge along the two leaf-to-LCA paths settles the task in $\leq 2d$ messages. A flat broadcast from one initiator to the other $N - 1$ agents uses $N - 1$ messages. Summing over M tasks gives (i).

(ii) The root has b children, each the root of a subtree with b^{d-1} leaves. The first participant lies in some child subtree j; the second participant, drawn uniformly and independently, lies in the same subtree j with probability $b^{d-1}/b^d = 1/b$. The LCA is the root exactly when the two participants lie in different child subtrees, which has probability $1 - 1/b$. Expectation of a sum of M indicators gives $M(1 - 1/b)$.

(iii) Local tasks never reach the root (their LCA lies inside a module); the non-local fraction behaves as in (ii). By linearity of expectation the load is $M(1 - \theta)(1 - 1/b)$. □

Remark. Theorem 4 formalizes the trade-off between centralized administration and federated coordination (Radner, 1993; Bolton and Dewatripont, 1994): a hierarchy cuts the *message* cost of coordination from linear to logarithmic in N, but concentrates a constant fraction $1 - 1/b$ of all non-local tasks on the apex, whose processing rate then bounds K (Definition 3). Modularity reduces apex load by the locality fraction θ at the price of *not* reconciling across modules. These are the five regimes compared in the pre-registered simulations (no institution, shared norms, standardized records, centralized administration, federated), and the simulations measure messages and bits directly rather than assuming the bounds.

3.6. Backlog dynamics under mismatch

Definition 8 (backlog)**.** Let D_t be the number of tasks arriving in period t and K_t the number the system can settle in period t. The backlog of unsettled tasks follows the Lindley recursion

$$B_{t+1} = \max\{0,\ B_t + D_t - K_t\}, \qquad B_0 \geq 0. \tag{5}$$

Theorem 5 (backlog dynamics)**.** *(i)* Deterministic case. *If $D_t = D$ and $K_t = K$ for all t, then: if $D \le K$, $B_t \le \max\{B_0 - t(K-D), 0\}$, so the backlog reaches 0 after at most $\lceil B_0/(K-D) \rceil$ periods when $D < K$ and stays constant at B_0 when $D = K$; if $D > K$, $B_t = B_0 + t(D-K) \to \infty$ linearly.*

(ii) Stochastic case. *If $(D_t - K_t)_{t \ge 0}$ is i.i.d. with finite mean, then when $\mathbb{E}[D_t] < \mathbb{E}[K_t]$ the backlog has a proper stationary distribution to which B_t converges from any B_0, and when $\mathbb{E}[D_t] > \mathbb{E}[K_t]$, $B_t \to \infty$ almost surely.*

(iii) Delay under a Markovian approximation. *If arrivals are Poisson with rate λ_D and settlements are exponential with rate μ (an M/M/1 queue), with $\rho = \lambda_D/\mu < 1$, the expected sojourn time is $W = 1/(\mu - \lambda_D) = (1/\mu)/(1-\rho)$, which is increasing and convex in ρ on $[0,1)$ and diverges as $\rho \uparrow 1$.*

Proof. (i) If $D \le K$, then $B_{t+1} = \max\{0, B_t - (K-D)\}$; by induction $B_t \le \max\{0, B_0 - t(K-D)\}$, with equality as long as the argument is nonnegative, which gives the hitting time when $D < K$ and constancy when $D = K$. If $D > K$, then $B_t + D - K > 0$ for all t (since $B_t \ge 0$), so the maximum is never binding and $B_{t+1} = B_t + (D-K)$, hence $B_t = B_0 + t(D-K)$.

(ii) The recursion is the single-server waiting-time recursion of Lindley (1952), and the stability dichotomy is the theorem of Loynes (1962); the proof is not reproduced here and the result is cited.

(iii) The M/M/1 stationary number in system is geometric with parameter ρ, so the mean is $\rho/(1-\rho)$ and Little's law gives $W = \rho/((1-\rho)\lambda_D) = 1/(\mu(1-\rho)) = 1/(\mu - \lambda_D)$. The function $f(\rho) = (1-\rho)^{-1}$ has $f' = (1-\rho)^{-2} > 0$ and $f'' = 2(1-\rho)^{-3} > 0$ on $[0,1)$, and $f(\rho) \to \infty$ as $\rho \uparrow 1$. □

Scope of Theorem 5. This is a *mechanism*, not evidence. Polities are not single-server queues: capacity is endogenous (institutions are built in response to backlog), arrivals are correlated across periods and agents, and "unsettled tasks" are not observed in the historical record. The theorem is used only to motivate two qualitative predictions, that costs rise smoothly and convexly in utilization and that sustained $\rho > 1$ is qualitatively different from $\rho < 1$, and both are stated as conjectures below.

3.7. Conjectures

The following statements are unproved. Each is given with its reason and its falsifier.

Conjecture 1 (proxy monotonicity)**.** *The pre-registered Seshat and V-Dem proxies are monotone in the model's primitives: the scale variables (population, territory, largest settlement, settlement hierarchy) increase with D of Definition 2, and the four indicator blocks (administration, standardized*

rules and money, communication infrastructure, record-keeping) increase with K of Definition 3, so that $P = z(\log D) - z(K)$ is monotone in $\log \rho$ up to measurement error.

Reason: population and territory raise the number and the dispersion of participants; written records, legal codes, money and measurement systems are shared codes (Theorem 1); roads, couriers and postal systems raise bits per period; administrative levels implement the routing of Theorem 4. *Falsifier:* the proxies for D and K are so collinear (as Seshat's first principal component suggests; Shin et al., 2020) that P retains almost no variance, or alternative defensible K weightings reverse the sign of P for a substantial share of polities; both are reported in the measurement section.

Conjecture 2 (sustained overload raises breakdown probability)**.** *Periods with $\rho > 1$ sustained longer than the time needed to build capacity raise the probability of institutional breakdown.*

Mechanism family (not proved): unsettled tasks accumulate (Theorem 5), so participants cannot form the common knowledge required for coordinated action (Chwe, 2000) and local defection spreads by contagion through the interaction network (Morris, 2000); institutions whose codes are stale after a shock lose their cost advantage (Theorem 3) and are abandoned. *Falsifier:* the pre-registered H2 tests find β_P not positive, or P does not beat the population-only and capacity-only baselines out of sample.

Conjecture 3 (smooth, not step)**.** *The relation between P and instability is convex and smooth away from $\rho \approx 1$ and steepens near it, rather than a discrete threshold.*

Reason: Theorem 5(iii) gives a smooth, convex delay in ρ; discontinuities would require an additional tipping mechanism (e.g. a critical mass in a contagion model) that the framework does not supply. *Falsifier:* the pre-registered H4 comparison finds that a hinge or step model beats the smooth model out of sample in at least 80% of folds, which would favour a tipping mechanism; conversely, if neither nonlinear form beats the linear model, the "steepening" clause of the conjecture is unsupported.

3.8. What is not claimed, and which test speaks to each statement

- **No Shannon measurement of history.** Theorems 1 and 3 are about codes for a plan set with a known distribution. Seshat indicators (presence of writing, coins, courts, roads) are not message lengths, and no entropy or divergence is computed from historical data anywhere in this paper. Shannon quantities appear only in the simulations, where the plan distribution p and the messages are defined by the experimenter and can be measured.
- **No causal claim.** The theorems describe an idealized communication system; the empirical tests are associational and were pre-registered as such; the identification discussion is in Section 5.5.

- **Endogeneity of capacity.** The framework takes K as given within a period; historically, K is built in response to D (Johnson, 1982; Shin et al., 2020). This is why H1 is pre-registered as an association without a direction.

Table 2. Which evidence can speak to each statement of the framework.

Statement	What can speak to it	Pre-registered test
Theorem 1 (coding gain)	Simulated messages and bits with vs. without a shared code	H3 (regimes b, c vs. a)
Theorem 2 (break-even)	Simulated total cost across N, λ, horizon, overhead a	H3 factorial cells; search for cells where institutions lose
Theorem 3 (stale code)	Simulated shocks to the task distribution	H3 adverse regimes (shock treatment)
Theorem 4 (routing, root load)	Simulated messages and apex load under centralized vs. federated regimes	H3 (regimes d, e)
Theorem 5 (backlog)	Not testable with the historical data (backlog unobserved); illustrated in simulation only	none (stated as mechanism)
Conjecture 1	Measurement diagnostics: collinearity of D and K proxies, sensitivity of P to weights	Measurement section
Conjecture 2	Lagged P and subsequent instability vs. baselines, out of sample	H2 (Seshat Y1–Y3; V-Dem Y4–Y6), H5
Conjecture 3	Linear vs. hinge vs. spline in P, out of sample	H4

4. Data and historical measurement

Every source was downloaded by a script that records the release, URL, retrieval time, licence, size and SHA-256 hash of the raw file (Table 3); raw files are preserved unmodified and never redistributed. All counts in this section are produced by the parsers and written to machine-readable diagnostics from which the tables are generated. Missing values are never recoded as absent, no fact is forward-filled across polities or periods, and no outcome is imputed—the lesson of the moralizing-gods episode (Beheim et al., 2021; Whitehouse et al., 2022).

Table 3. Sources, releases and provenance. Hashes are the first twelve hexadecimal digits of the SHA-256 of the raw file as downloaded on 2026-11-01; the full registry is in the replication package.

Dataset	Release	Licence	Bytes	SHA-256
Seshat Global History Databank — Polaris-2026 snapshot	Polaris-2026 (repository commit 55ca5fbc2b, 2026	MIT (repository LICENSE, Copyright (c) 2	2,899,093	`cf60c9f76eed`
Seshat Global History Databank — Equinox-2020 snapshot (spre	Equinox-2020, spreadsheet dated 05.2023 (reposit	MIT (repository LICENSE); original Equin	1,720,038	`7964d74c3ae3`
Seshat Databank: Equinox Release with Culturally and Institu	Zenodo 8120128 (2023-07-06)	CC-BY-4.0	70,197	`4eb6c1a72de9`
Maddison Project Database 2023	MPD 2023 (Bolt & van Zanden 2024)	Free for academic use with citation (GGD	4,903,804	`ecc5916ca127`
Maddison Project Database 2023 (Stata file)	MPD 2023	Free for academic use with citation	10,892,389	`2bc62ccbd795`
D-PLACE dataset derived from Murdock et al. 1999 'Ethnograph	v3.3 (2025-12-19)	CC BY-NC 4.0	7,039,773	`9ec031e01101`
V-Dem Country-Year Dataset v16 (vdemdata R package copy)	v16 (package commit f4dd26922e, 2026-03-17 'v16	V-Dem data: free for non-commercial use	33,878,255	`39b412d39a06`
V-Dem codebook (R object)	v16	GPL-3 (package)	148,633	`2840d879cb90`
Glottolog CLDF languages table (language family for Galton c	glottolog-cldf v5.3 (2026-03-02)	CC BY 4.0	2,477,017	`1a50a393bc81`

4.1. Three panels, kept separate

Historical polities (Seshat Polaris-2026). The Seshat Global History Databank codes historical polities on hundreds of variables from the published historiography (Turchin et al., 2015; François et al., 2016). We use the Polaris-2026 snapshot, the current official release, which contains 626 polities in 42 regions (35 natural geographic areas) spanning −13,600 to 2020 CE, and 56,274 coded facts in the social-complexity, warfare, religion and general sheets. Each fact carries a value or a value range (`value_from`, `value_to`), an optional date range, and flags for disputed and uncertain codes; presence variables are coded present (P), absent (A), inferred present (IP), inferred absent (IA) or unknown (SU). Two structural features of the release shape the design. First, only 2,503 facts carry a date; the other 93.7% describe the whole polity period. A polity-year panel would therefore be an interpolation that the data do not support, and the historical unit of analysis is the polity, with region and era effects. Second, the release contains the CrisisDB "crisis consequences" table

(Hoyer et al., 2023, 2024): for 166 crisis episodes in 150 polities, historians coded whether each of 21 consequences occurred, among them decline, collapse, civil war, uprising, fragmentation, loss of the capital, conquest, assassination and deposition. This is the independently coded instability outcome. It is also a crisis-selected sample: a polity without a CrisisDB record is one for which no crisis was coded, not one that had no crisis, so the outcome measures how badly a recorded crisis unfolded rather than whether one occurred. Table 4 summarizes coverage; Figure 1 shows the temporal distribution of polities and crises and the pattern of coded and missing indicators.

Ethnographic societies (D-PLACE Ethnographic Atlas v3.3). The Ethnographic Atlas (Murdock, 1967), as curated in D-PLACE (Kirby et al., 2016), describes 1,291 societies at an ethnographic focal year (between −2000 and 1965, most in the nineteenth and twentieth centuries) on a fixed set of categorical variables. We use mean size of local communities (EA031, eight ordinal categories converted to geometric midpoints of the category ranges printed in the codebook), population size (EA202, continuous), and jurisdictional hierarchy beyond the local community (EA033, from acephalous to four levels), with subsistence, agricultural intensity and class differentiation as covariates. Galton's problem—societies are related by descent and diffusion—is handled with Glottolog 5.3 language families (173 families among the 1,291 societies), regions, and spatial heteroskedasticity-and-autocorrelation-consistent standard errors (Conley, 1999; Mace et al., 1994; Dow, 2007). The Atlas has no instability outcome and is used only for the capacity–scale association.

Modern states (V-Dem v16 and Maddison 2023). The Varieties of Democracy country-year dataset (Coppedge et al., 2020; Pemstein et al., 2018) provides expert-coded institutional variables for 202 countries from 1789 to 2025; the Maddison Project Database 2023 (Bolt and van Zanden, 2024) provides population and GDP per capita, which we merge by ISO3 code (163 countries matched; the 39 unmatched V-Dem entities are mostly historical polities and micro-states and drop out of the population-based index). Outcomes are the first year of a civil war (Correlates-of-War coding carried in V-Dem, 1816–2006), a coup (1789–2008), and the last year of a V-Dem regime whose coded end type is a military or non-military coup, a self-coup, an assassination, a loss in civil or inter-state war, a foreign intervention or a popular uprising (1789–2025); natural deaths and guided or negotiated transitions are not events. We never map modern countries onto historical polities: the two panels answer the same hypotheses separately.

Table 4. Coverage of the three panels after parsing (counts produced by the build script).

Quantity	Value	Detail
Seshat Polaris-2026 polities	626	42 regions / 10 macro-regions
Seshat facts (dated)	56,274 (2,503)	unparsed 132, inferred 8,094
Polities with population / territory / any D	299 / 355 / 561	
Polities with $\geq 50\%$ of K indicators	376	
CrisisDB records (polities)	166 (150)	Y1 coded 166; Y1b coded 158, positive 122
Y3 succession coded (discontinuous)	207 (37)	polities censored at 1900: 30
D-PLACE EA societies	1291	EA031 611, EA202 953, EA033 1155; 173 language families
Modern country-years (countries)	28,092 (202)	1789–2025; Maddison population 15,821
Modern events	civil-war onsets 176 / 13,608	coups 599 / 13,813; irregular regime ends 1041 / 27,846

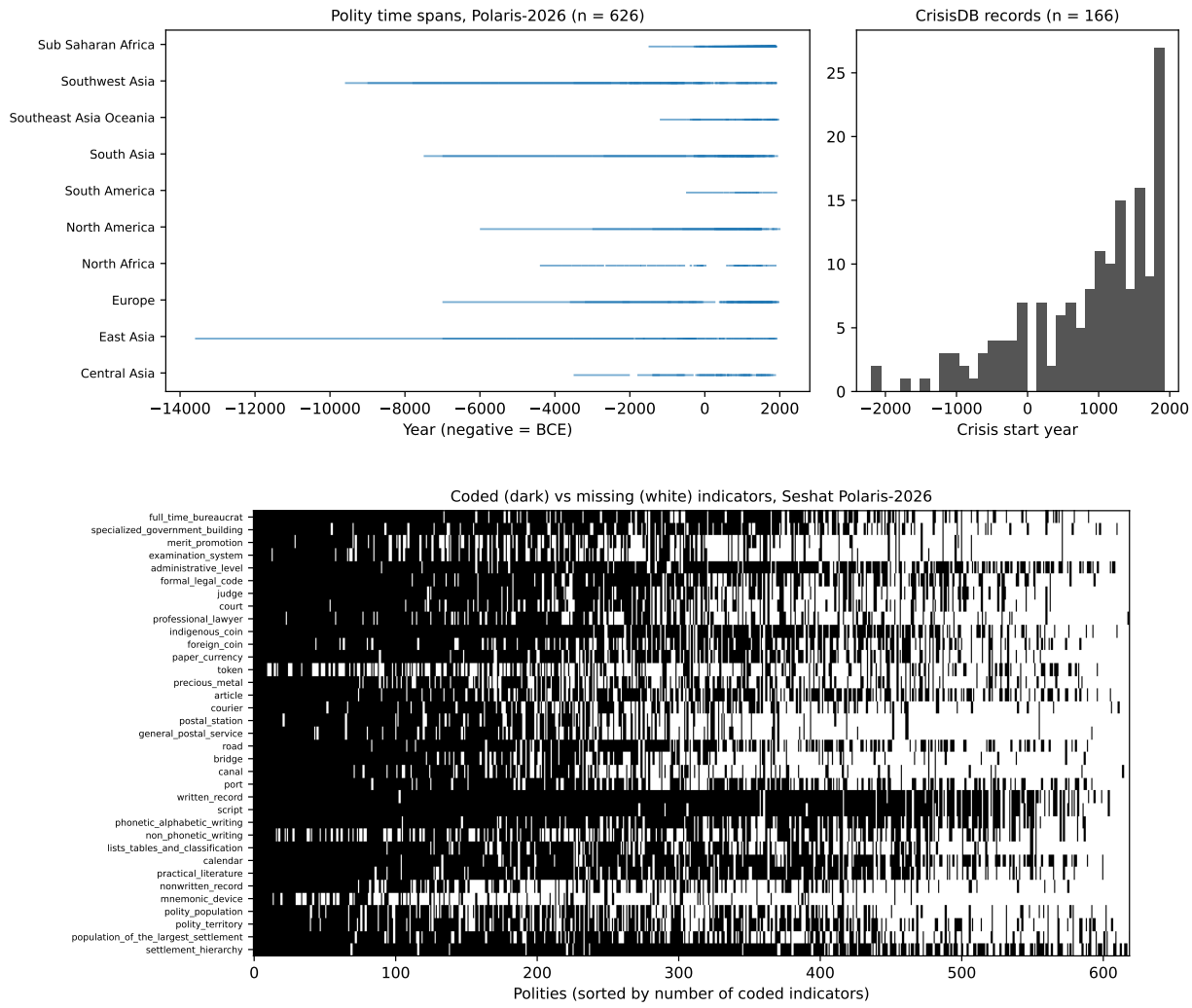


Figure 1. Seshat Polaris-2026. Top: polity time spans by macro-region (left) and CrisisDB crisis start years (right). Bottom: coded (dark) versus missing (white) pre-registered indicators, polities sorted by number of coded indicators. Missing is never treated as absent.

4.2. Coding rules and uncertainty

Presence codes are mapped to 1 (P, IP), 0 (A, IA) and 0.5 (disputed codes of the form A∼P), with inferred codes flagged so that a sensitivity run can drop them; unknown (SU), blank and unparseable values are missing (132 of 56,274 facts were unparseable and are logged). Numeric ranges are summarized by their geometric midpoint for population, territory and largest-settlement population and by the arithmetic midpoint for level counts, and the bounds are kept: the paper re-estimates its main model at the lower and upper bounds. Dates are kept as intervals; a crisis is dated by the start of its coded interval; nothing is interpolated. When several facts exist for a polity and a variable, the polity value is the mean over facts (the mean log for logged variables); for each crisis

record, "pre-crisis" aggregates use dated facts ending before the crisis start together with the undated whole-polity facts; only 3.6% of the facts entering these aggregates are dated, so "pre-crisis" means, in practice, "whole-polity", and temporal precedence is weak. We keep the label for consistency with the pre-registration and flag it wherever it is used.

4.3. Demand, capacity, and the mismatch index

The pre-registration fixed disjoint indicator sets before any outcome was examined. Coordination demand D is scale: the first principal component of the pairwise-complete correlation matrix of log population, log territory, log largest-settlement population and settlement-hierarchy levels, scored as the loading-weighted mean of whatever is available for a polity (at least two of the four), with log population alone as the secondary measure. The component explains 82% of the variance with nearly equal loadings and is defined for 442 polities. Coordination capacity K is the unweighted mean of four block scores, each the share of coded indicators present and each requiring at least half of its indicators coded, with at least three blocks present: administration (full-time bureaucrats, specialized government buildings, merit promotion, examination system, rank-scaled administrative levels), standardized rules and money (legal code, judges, courts, professional lawyers, indigenous and foreign coins, paper currency, tokens, precious metals, articles), communication infrastructure (couriers, postal stations, general postal service, roads, bridges, canals, ports) and record-keeping (written records, script, phonetic and non-phonetic writing, lists and tables, calendar, practical literature, non-written records, mnemonic devices). K is defined for 344 polities. Military, religious and production-infrastructure indicators are excluded from K by design and serve as rival-explanation covariates. The mismatch index is $P = z(\log D) - z(K)$, standardized over the polity sample; because K is a bounded share that can be zero, it is standardized rather than logged, a pre-registered deviation from the candidate formula in the mandate. In the modern panel D is log population and K the mean of six standardized V-Dem items—rigorous and impartial public administration, meritocratic appointment, transparent and predictably enforced laws, fiscal capacity, territorial control and election-management capacity—none of which is a democracy index; in the Atlas the capacity-side measure is the number of jurisdictional levels.

Table 5. Measurement model: construction and diagnostics. Values from the measurement diagnostics written by the index-building script; no outcome variable was read by that script.

Index	Construction and diagnostics
D: PC1 of log population, log territory, log largest settlement, settlement hierarchy	variance share 0.821; loadings log_pop 0.26, log_terr 0.24, log_cap 0.26, settl_hier 0.24; $n = 442$
K: mean of four blocks (admin, rules/money, communication, records)	inter-block corr. 0.75–0.79; PC1 share 0.824; $n = 344$
K (2PL IRT, 30 items, MAP)	corr. with block mean 0.963; posterior SD median 0.34; $n = 579$
$\mathrm{corr}(z(D), z(K))$	0.838 (IRT: 0.800)
P = $z(D) - z(K)$	variance share of $z(D) + z(K)$ retained 0.166; corr. with D 0.39, with K -0.18, with complexity PC1 -0.04; $n = 317$ polities, 127 crisis records
IRT invariance (Old World vs Americas/Oceania fits)	difficulty correlation 0.88 over 30 items
Modern K (6 V-Dem items)	PC1 share 0.605; $\mathrm{corr}(D, K)$ 0.062; $\mathrm{corr}(K, \text{polyarchy})$ 0.754; $n = 15,697$
Crisis sample (127 records)	K mean 0.71 (all polities 0.58); share of records with K > 0.8: 0.40; $\mathrm{corr}(P, D)$ 0.63, $\mathrm{corr}(P, K)$ 0.02; dated share of pre-crisis facts 0.036

Two alternatives to the block mean were pre-registered and are reported side by side: the first principal component of the four blocks, which correlates 0.999 with the mean, and a two-parameter logistic item-response model estimated by marginal maximum likelihood with mild priors on the thirty binary items, which handles item missingness through the likelihood rather than by imputation, is defined for 579 polities, and correlates 0.96 with the block mean. The latent score carries a median posterior standard deviation of 0.34 on a unit-variance scale, which is the honest measure of how uncertain an individual polity's capacity is. Item parameters refitted separately on the Old World and on the Americas and Oceania have difficulties correlated 0.88, so the ordering of institutions by rarity is similar across the two worlds, although the New-World sample (94 polities) is too small for a strict invariance test and the leave-region-out analyses are the operational safeguard. The block structure is stable: inter-block correlations are 0.75–0.79 overall, Cronbach's α across the four blocks is between 0.83 and 0.97 in every macro-region, and dropping any block leaves K correlated at least 0.98 with the full index.

The central measurement fact of the paper is collinearity (Figure 2, Table 5). Scale and capacity

correlate 0.84 across polities, as the single-dimension result of Turchin et al. (2017) implies, so the mismatch index retains only 17% of the variance of $z(D)+z(K)$; it correlates 0.39 with D, -0.18 with K and -0.04 with a Turchin-style first principal component of 57 social-complexity indicators. P is a residual, nearly orthogonal to overall complexity. That is what would make it informative if it carried signal, and what makes it fragile: measurement error in either index leaks into it, and any association with P can be a disguised association with one of its components. The pre-registered baselines (scale only, capacity only, complexity only, scale and capacity entered separately) exist to detect exactly this. The problem is sharper in the crisis sample: polities with a CrisisDB record are better documented and more complex than the average polity, their mean capacity share is 0.71 against 0.58 overall, and 40% of records have K above 0.8, so within that sample P correlates 0.63 with D and 0.02 with K—the mismatch index is, among crisis cases, close to a rescaled measure of scale (last row of Table 5).

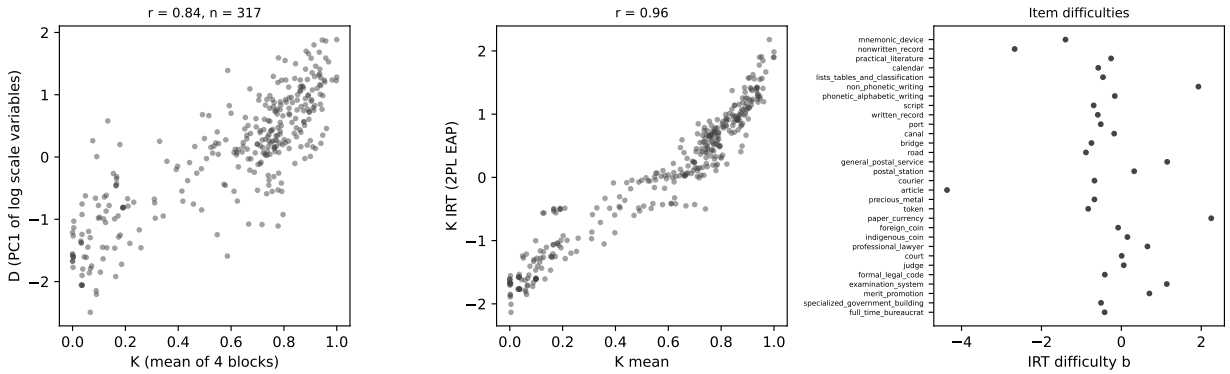


Figure 2. Measurement. Left: capacity index against the scale component across Seshat polities. Middle: block-mean against latent (IRT) capacity. Right: item difficulties from the two-parameter model (items ordered by block).

4.4. A-priori power

Before estimation, we drew random outcomes with the actual distribution of pre-crisis P among the 127 crisis records with a defined index and computed power at the 5% level with region-clustered standard errors (Table 6). For the binary "severe outcome" indicator (base rate 0.77) the minimum detectable effect at 80% power is about 0.6 log-odds per standard deviation, above the pre-registered threshold of 0.5, so that outcome was declared underpowered before any result was seen and is reported as secondary. For the fractional severity share (base rate 0.29 over fourteen items) the minimum detectable effect is about 0.15, and that outcome is primary.

Table 6. A-priori power for the Seshat crisis-severity test (random outcomes, actual P distribution, $n = 127$, 400 replications per effect size).

Outcome	β per SD of P	Power	n
Y1b binary	0.20	0.19	127
Y1b binary	0.30	0.31	127
Y1b binary	0.40	0.47	127
Y1b binary	0.50	0.62	127
Y1b binary	0.60	0.79	127
Y1b binary	0.80	0.94	127
Y1b binary	1.00	0.99	127
Y1 fractional	0.10	0.49	127
Y1 fractional	0.15	0.81	127
Y1 fractional	0.20	0.96	127
Y1 fractional	0.30	1.00	127
Y1 fractional	0.40	1.00	127
Y1 fractional	0.50	1.00	127

5. Econometric methods and identification

5.1. Pre-registration

The analysis plan was frozen before any predictor–outcome association was computed; its SHA-256 hash is recorded in the decision log and every deviation is a numbered amendment. The plan fixes five hypotheses with falsifiers (H1: capacity and scale are positively associated conditional on era, region and rival covariates, with no direction claimed; H2: lagged mismatch predicts subsequent instability *beyond* scale-only, capacity-only, complexity-only and persistence baselines, judged out of sample; H3: in simulation, shared protocols and institutions lower communication cost per successfully coordinated task; H4: a threshold form outperforms a smooth form out of sample; H5: estimates keep their sign in every held-out macro-region and era, otherwise the domain is narrowed), the samples and exclusions, the disjoint indicator sets of Section 4.3, three primary tests with a Holm correction (H1 in Seshat, H2 on Seshat crisis severity, H2 on the modern composite outcome), a robustness set, and the rule for abandoning the universal framing: if H2 fails in both panels, the paper reports the hypothesis as not supported and keeps only what survived.

5.2. Historical polities

For H1 we regress the scale component on the capacity index with macro-region and era fixed effects (ten macro-regions grouping the 42 Seshat regions; six era blocks on the polity midpoint), with and without the rival covariates (warfare-technology index, production-infrastructure index, military professionalization), with standard errors clustered by Seshat region, and we repeat the exercise with a region random intercept, with the latent capacity score, with the four blocks entered separately, and without quasi-polities.

For H2 the primary outcome is crisis severity Y_1, the share of the fourteen instability consequences coded present among those coded at all for a CrisisDB record (unknown codes are missing, never zero). The model is a fractional logit,

$$\mathbb{E}[Y_{1,c}] = \Lambda\big(\alpha + \beta\, P_{i(c)} + \gamma' X_{i(c)} + \mu_{r(c)} + \tau_{e(c)}\big)\,, \tag{6}$$

where $P_{i(c)}$ is the pre-crisis mismatch of the polity of crisis c, X contains the warfare-technology and production-infrastructure indices (an epidemic indicator was pre-registered as a covariate but dropped after review because it is itself a CrisisDB consequence item, amendment A-006), and μ and τ are macro-region and era effects; standard errors are clustered by Seshat region (alternatively by natural geographic area and by polity). The binary secondary outcome Y_{1b}—any of collapse, fragmentation, century-long conflict, conquest or civil war coded present, defined only when all five are coded—is fitted by logit; several region–era cells are all-severe, so era effects are dropped for that outcome and heteroskedasticity-robust errors are used. Baselines are fitted on a common sample on which every variable is available: fixed effects only, scale only, capacity only, a Turchin-style complexity component, a persistence indicator (an earlier CrisisDB crisis in the same natural geographic area), and scale and capacity entered separately, with a Wald test of the restriction $\beta_D + \beta_K = 0$ that the mismatch index imposes. A placebo index replaces K with the production-infrastructure share that was excluded from K by design, fitted without that share as a control (with it, the placebo would be an exact reparametrization of the scale-only baseline). Out-of-sample comparison uses leave-one-macro-region-out folds (region effects removed) and ten repetitions of five-fold cross-validation with folds blocked by polity, test rows whose fixed-effect level is absent from the training fold being assigned the training mode; the pre-registered rule is that P "beats" a baseline if its mean out-of-sample log-likelihood is higher in at least 80% of folds. For H4, a hinge model whose knot is chosen by profile likelihood on the training fold only and a natural cubic spline with three degrees of freedom are compared with the linear form on the same folds. For H5, β is re-estimated with each macro-region and each era held out, and, where the held-out block has at least twelve crisis records, within the block.

Two secondary historical outcomes use all polities. Polity duration (end year minus start year)

is modelled by a Cox proportional-hazards regression stratified by macro-region, with the polity's end as the event and right-censoring for polities that reach 1900, robust errors clustered by region and a Schoenfeld test (Cox, 1972; Box-Steffensmeier and Jones, 2004); discontinuous succession—the successor polity's relationship to its predecessor coded as anything other than continuity—is modelled by logit. Both are reported with the caveat that Seshat polity boundaries are periodization decisions.

5.3. Ethnographic societies

Log community size is regressed on jurisdictional levels with subsistence, agricultural intensity, class differentiation and region controls; standard errors are Conley spatial HAC with 500 and 1,000 km cut-offs (Conley, 1999) and, alternatively, clustered by Glottolog language family. The reverse specification, an ordered logit of jurisdictional levels on log community size, is reported with family-clustered errors.

5.4. Modern states

The modern outcome Y_{456} equals one if a civil-war onset, a coup or an irregular regime end occurs in $(t, t+5]$ and is missing if no year in the window is coded. The primary model is a pooled logit with region and year fixed effects and country-clustered standard errors,

$$\Pr(Y_{456,i,t+1..t+5} = 1) = \Lambda\big(\alpha + \beta P_{it} + \gamma' X_{it} + \mu_{r(i)} + \tau_t\big), \tag{7}$$

with X containing log GDP per capita, the V-Dem electoral-democracy index (because the capacity items correlate 0.75 with it) and a cubic in years since the last event (Carter and Signorino, 2010; Beck et al., 1998); a conditional logit with country and decade effects and a linear probability model with country and year effects are secondary (Colin Cameron and Miller, 2015). Baselines are population only, capacity only, development only, persistence (an event in the previous five years), and population and capacity entered separately; the placebo index replaces K with the democracy index. Two negative controls are pre-registered: events in $[t-5, t-1]$ regressed on P_t with the same controls should not load more strongly than future events, and the placebo should not perform as well as P. Out-of-sample validation is by forward chaining (train to 1945, test 1946–1975; train to 1975, test 1976–2005; train to 2005, test 2006 onward; year effects replaced by decade effects so that test years have a fitted level) and by leave-one-region-out on the six V-Dem politico-geographic regions, with log-likelihood, area under the ROC curve, Brier score and calibration slope (Goldstone et al., 2009). H4 and H5 proceed as in the historical panel with country-blocked folds and with the six regions and four periods (pre-1900, 1900–1945, 1946–1989, 1990 onward) as blocks. The robustness set varies the horizon (1, 3, 10 years), the outcome components, the

capacity aggregation, the standardization window, the inclusion of land area, the sample period and the control set, and reports Oster's bias-adjusted coefficient and the Cinelli–Hazlett robustness value on the linear-probability version of the primary model (Oster, 2017; Cinelli and Hazlett, 2019).

5.5. What the designs identify

Nothing in this paper is causal, and the text does not use causal language for H1 or H2. The estimands are conditional associations with pre-specified direction, baselines, out-of-sample validation and falsification tests. The reasons are stated in advance in the identification note of the replication package and summarized here. Scale creates the demand for administration and administration enables scale, so H1 is symmetric by construction. The historical outcome is crisis-selected, most Seshat facts are undated so that "pre-crisis" capacity is mostly whole-polity capacity, the institutional codes and the crisis codes are produced by the same project from overlapping historiography, and the mismatch index is a residual of two indices that correlate 0.84. In the modern panel the capacity items are expert-coded with knowledge of later events and compete with regime type as an explanation, which the pre-registered placebo, negative control and sensitivity analysis are designed to expose rather than to remove. A credible design would move demand or capacity for reasons unrelated to local instability; no such variation is coded at the polity level in Seshat, and the candidate instruments in the modern panel (printing, railways, telegraphy) do not satisfy an exclusion restriction for instability. The paper therefore claims association, prediction and robustness, and declines a causal claim.

6. Results: historical polities and ethnographic societies

All estimates in this section are [EMPIRICAL] and associational. Coefficients are per standard deviation of the standardized regressor unless noted; stars denote $p < 0.10$ (*), 0.05 (**), 0.01 (***) with the clustering stated in each table.

6.1. H1: capacity and scale

Capacity and scale move together in every dataset (Tables 7 and 8). In Seshat, a one-unit increase in the capacity share (from no indicator present to all present) is associated with 2.36 units of the scale component in the primary specification with region and era effects (SE 0.17, $n = 317$), which is 0.75 standard deviations of D per standard deviation of K; with the warfare, production-infrastructure and military covariates the coefficient falls to 1.21 (SE 0.43, $n = 191$), still clearly positive, and it is 2.18 (SE 0.13) with a region random intercept instead of fixed effects. With the latent capacity score the association is 0.37 scale units per IRT standard deviation (SE 0.09). Entering the four

blocks separately, communication infrastructure carries most of the association (1.09, SE 0.35) while administration, rules and money, and records are individually indistinguishable from zero given the others—which, with inter-block correlations near 0.8, is a statement about collinearity rather than about mechanisms. In the Ethnographic Atlas, each additional level of jurisdictional hierarchy beyond the local community is associated with 0.70 log units of mean community size ($n = 579$; Conley standard errors 0.09 and 0.10 at 500 and 1,000 km; 0.12 clustered by language family) and with 0.59 log units of population size ($n = 799$); the ordered-logit specification in the other direction gives 0.49 log-odds of one more level per log unit of community size (SE 0.12). In the modern panel the within-country association of log population with capacity is 0.15 (SE 0.06) and the fifty-year-difference association 0.13 (SE 0.05), while the pooled cross-sectional coefficient is positive but imprecise (0.19, SE 0.24). The sign is stable across datasets, specifications and the two worlds; the first pre-registered primary test is passed. Nothing here identifies a direction.

Table 7. H1 in Seshat: scale regressed on capacity. OLS, standard errors clustered by Seshat region (42 clusters); region random intercept in the last row. K is a share in $[0, 1]$; the last column rescales the coefficient per standard deviation of K.

Specification	Term	Coef.	SE	n	R^2
H1-S1 primary: D_pc1 ∼ K_mean + FE	K_mean	2.356***	0.173	317	0.73
H1-S2: D_pc1 ∼ K_mean + X + FE	K_mean	1.214***	0.432	191	0.83
H1-S3: log_pop ∼ K_mean + X + FE	K_mean	3.482*	1.934	146	0.80
H1-S4: log_terr ∼ K_mean + X + FE	K_mean	2.103	2.376	158	0.60
H1-S5: D_pc1 ∼ K_irt + X + FE	K_irt	0.366***	0.094	208	0.81
H1-S6: D_pc1 ∼ K blocks + X + FE	K_admin	0.148	0.335	164	0.86
H1-S6: D_pc1 ∼ K blocks + X + FE	K_rules_money	-0.290	0.294	164	0.86
H1-S6: D_pc1 ∼ K blocks + X + FE	K_comm_infra	1.085***	0.350	164	0.86
H1-S6: D_pc1 ∼ K blocks + X + FE	K_records	0.220	0.253	164	0.86
H1-S7: D_pc1 ∼ K_mean + X + FE, no quasi-polities	K_mean	1.360**	0.530	155	0.79
H1-S8: D_pc1 ∼ K_mean + era, region random intercept	K_mean	2.180***	0.126	317	–

Table 8. H1 in the Ethnographic Atlas (top) and in the modern panel (bottom).

Specification	Term	Coef.	SE	n	R^2
H1-E1 OLS log community size ∼ hierarchy + X + region, Conley 500 km	hier	0.702***	0.089	579	0.75
H1-E1 OLS log community size ∼ hierarchy + X + region, Conley 1000 km	hier	0.702***	0.103	579	0.75
H1-E2 same, SE clustered by language family	hier	0.702***	0.123	579	0.75
H1-E3 log population (EA202) ∼ hierarchy + X + region, family-clustered	hier	0.587***	0.107	799	0.67
H1-E4 ordered logit hierarchy ∼ log community size + X + region (family-clustered)	log_community_size	0.489***	0.116	579	–
H1-M1 within-country: log pop ∼ K + log gdppc + polyarchy + year FE	K_modern_dm	0.148**	0.062	14647	0.65
H1-M2 pooled: log pop ∼ K + log gdppc + polyarchy + year FE + region FE	K_modern	0.188	0.240	14647	0.23
H1-M3 50-year differences: Δlog pop ∼ ΔK + Δlog gdppc + Δpolyarchy + year FE	dK_modern	0.135**	0.053	6836	0.17

6.2. H2: crisis severity

Table 9 and Figure 3 report the pre-registered primary test. In the primary specification, a one-standard-deviation higher mismatch is associated with 0.20 higher log-odds that a given adverse consequence is recorded for the crisis (SE 0.06, $n = 127$ crisis records in 117 polities and 31 Seshat regions with a defined index); on the common sample used for every specification with covariates and every baseline ($n = 93$), and with the warfare and production-infrastructure covariates, the estimate is 0.24 (SE 0.09), unchanged when the overall complexity component is added. The coefficient is positive, survives the Holm correction over the three primary tests (adjusted $p < 0.001$), and is stable under every pre-registered perturbation of the measurement model (Table 10): latent capacity (0.18, SE 0.08), population only (0.28, SE 0.11), capacity as a principal component, lower and upper range bounds (0.24 either way), inferred codes dropped (0.13, SE 0.06), within-era and within-region standardization, ranks (0.19, SE 0.08), quasi-polities dropped (0.21, SE 0.10), crises starting at least 25 years after the polity's start (0.31, SE 0.10), records with at least twelve coded items, no era effects, no region effects, no effects at all, and clustering by natural geographic area or by polity. Removing settlement hierarchy from D, administrative levels from K, or both—the two indicators that measure hierarchy depth on either side of the index—leaves the estimate at 0.25. Held-out macro-regions, eras and—in a scheme added after review, grouping Seshat regions by cultural sphere rather than geography—held-out spheres (Table 12, Figure 5) give pooled coefficients between 0.24 and 0.37, every one with a confidence interval excluding zero except the block that drops the

first millennium BCE (0.26, SE 0.15), and the within-block estimates are positive wherever they can be computed, if imprecise. By the standards of a conventional regression table, the mismatch hypothesis would be reported as confirmed.

Table 9. H2 in Seshat: crisis severity Y_1 (fractional logit; share of fourteen instability consequences coded present) with macro-region and era effects; standard errors clustered by Seshat region. A1 uses every record with a defined index; all other rows use the common sample ($n = 93$) on which every covariate and baseline is available. The binary outcome Y_{1b} uses region effects only (era cells are all-severe) and robust errors.

Specification	Term	Coef.	SE	n
A1 primary: Y1 ∼ P + FE	P_main_z	0.196***	0.056	127
A2: Y1 ∼ P + X + FE	P_main_z	0.243***	0.086	93
A3: Y1 ∼ P + complexity PC1 + X + FE	P_main_z	0.243***	0.087	93
B1 baseline D only	D_pc1_z	0.416**	0.173	93
B2 baseline K only	K_mean_z	0.041	0.248	93
B3 baseline complexity PC1	complexity_pc1_z	0.078	0.313	93
B4 baseline persistence	prior_crisis_nga	0.373*	0.202	93
B5 D and K separately	D_pc1_z	0.440***	0.170	93
B5 D and K separately	K_mean_z	-0.126	0.188	93
B5 Wald test of beta_D + beta_K = 0 (restriction implied by P)	D+K	0.314	0.257	93
Placebo: P built with production infrastructure instead of K (X = warfare only)	P_placebo	0.192	0.166	93
Placebo comparison: P with the same controls (X = warfare only)	P_main_z	0.290***	0.079	93
A1b: Y1b ∼ P + region FE (logit, HC1; era FE dropped: separation)	P_main_z	-0.178	0.203	123
A2b: Y1b ∼ P + X + region FE (logit, HC1)	P_main_z	0.254	0.316	91
A3b: Y1b ∼ P (logit, HC1, no FE)	P_main_z	-0.068	0.180	123
Exploratory: reform consequences ∼ P + X + FE	P_main_z	-0.013	0.164	91

The pre-registration asked a different question, and the rest of Table 9 answers it. On the same 93 records, scale alone is associated with severity at 0.42 per standard deviation (SE 0.17); capacity alone at 0.04 (SE 0.25); the Turchin-style complexity component at 0.08 (SE 0.31); the persistence indicator at 0.37 (SE 0.20). When scale and capacity are entered separately, scale carries 0.44 (SE 0.17) and capacity -0.13 (SE 0.19); the restriction that the mismatch index imposes, $\beta_D + \beta_K = 0$, is not rejected (Wald $p = 0.22$), but neither is $\beta_K = 0$: capacity has the sign the hypothesis predicts and is not distinguishable from zero. The placebo index, which replaces capacity by the share of production-infrastructure indicators that were excluded from K precisely because they are not coordination capacity, is associated with severity at 0.19 (SE 0.17) against 0.29 (SE 0.08) for P with the same controls—weaker, and not distinguishable from zero, but not distinguishable from P either.

Out of sample (Table 11; fifty polity-blocked folds), the model with P has a higher log-likelihood than the fixed-effects-only, capacity-only and complexity-only models in 84–86% of folds, but beats the scale-only model in 42%, the persistence model in 78% and the scale-plus-capacity model in 44%, below the 80% rule; with macro-regions held out it loses to scale alone and to scale plus capacity in eight of ten folds. For the binary outcome, which was declared underpowered in advance, the coefficient is negative and imprecise (-0.18, SE 0.20 with region effects; -0.07, SE 0.18 without), and no model discriminates severe from non-severe crises out of sample much better than the fixed-effects baseline (areas under the curve 0.69–0.72 for every model). Reform consequences are unrelated to P. The conclusion we are obliged to draw is that what the mismatch index predicts about crisis severity is what polity scale predicts, and that nothing in the data distinguishes a capacity shortfall from a large population.

The severity share also hides heterogeneity that Hoyer et al. (2024) emphasize. The fourteen consequences are almost uncorrelated with each other (mean inter-item correlation 0.08; the first eigenvalue of their correlation matrix carries 19% of the variance), so the share is a count of weakly related events rather than a measurement of one latent severity. Item by item (Table 13), P and D are associated with *decline*, *assassination*, *deposition*, *successful revolution* and *civil war*—the elite-political consequences that larger polities record more often—and negatively with *century-long conflict*; *collapse* is present in only two of 126 coded records and its estimate is meaningless. Capacity K is, if anything, positively associated with the same elite-political items. Larger and more documented polities have more recorded elite turmoil; whether that is severity or visibility, the data cannot say.

Table 10. Pre-registered robustness set for the Seshat primary coefficient (fractional logit of Y_1 on the alternative index, warfare and production-infrastructure covariates, macro-region and era effects; region-clustered standard errors unless stated).

Variant	Term	Coef.	SE	n
R: K = IRT latent	_P	0.175**	0.082	100
R: D = log population only	_P	0.279**	0.109	76
R: K = PC1 of blocks	_P	0.247***	0.084	93
R: D at lower bounds	_P	0.237***	0.087	93
R: D at upper bounds	_P	0.244***	0.085	93
R: inferred codes dropped	_P	0.131**	0.060	77
R: standardized within era	_P	0.265***	0.091	93
R: standardized within macro-region	_P	0.224***	0.085	93
R: rank-based P	_P	0.185**	0.076	93
R: D without settlement hierarchy	_P	0.249***	0.074	84
R: K without administrative levels	_P	0.250***	0.088	94
R: both hierarchy indicators removed	_P	0.247***	0.068	85
R: quasi-polities dropped	P_main_z	0.205**	0.101	87
R: crises starting $\geq$25 years after polity start	P_main_z	0.305***	0.095	79
R: $\geq$12 of 14 items coded	P_main_z	0.250***	0.091	88
R: no era FE	P_main_z	0.296***	0.111	93
R: no region FE	P_main_z	0.277***	0.079	93
R: no FE	P_main_z	0.288***	0.095	93
R: SE clustered by NGA	P_main_z	0.243***	0.074	93
R: SE clustered by polity	P_main_z	0.243***	0.088	93

Table 11. Out-of-sample comparison in Seshat. Top: share of folds in which the model with P (rival covariates included) has a higher out-of-sample log-likelihood than the baseline, and the mean difference per observation; the pre-registered rule requires 0.80. Bottom: out-of-sample metrics for the binary outcome Y_{1b} (ten repetitions of five-fold cross-validation, folds blocked by polity, unseen fixed-effect levels mapped to the training mode).

Baseline	Scheme	Folds	Win share	Mean Δ log-lik.
B0 FE only	loro	10	1.00	+0.0161
B1 D only	loro	10	0.20	-0.0038
B2 K only	loro	10	1.00	+0.0167
B3 complexity	loro	10	1.00	+0.0166
B4 persistence	loro	10	0.70	+0.0110
B5 D+K	loro	10	0.20	-0.0039
B0 FE only	kfold	50	0.86	+0.0081
B1 D only	kfold	50	0.42	-0.0005
B2 K only	kfold	50	0.84	+0.0106
B3 complexity	kfold	50	0.84	+0.0105
B4 persistence	kfold	50	0.78	+0.0051
B5 D+K	kfold	50	0.44	+0.0008

Model (Y_{1b})	OOS log-lik.	AUC	Brier
A2 P	-1.192	0.702	0.219
B0 FE only	-1.172	0.716	0.214
B1 D only	-1.196	0.703	0.218
B2 K only	-1.191	0.708	0.218
B3 complexity	-1.194	0.703	0.220
B4 persistence	-1.206	0.702	0.224
B5 D+K	-1.216	0.693	0.223

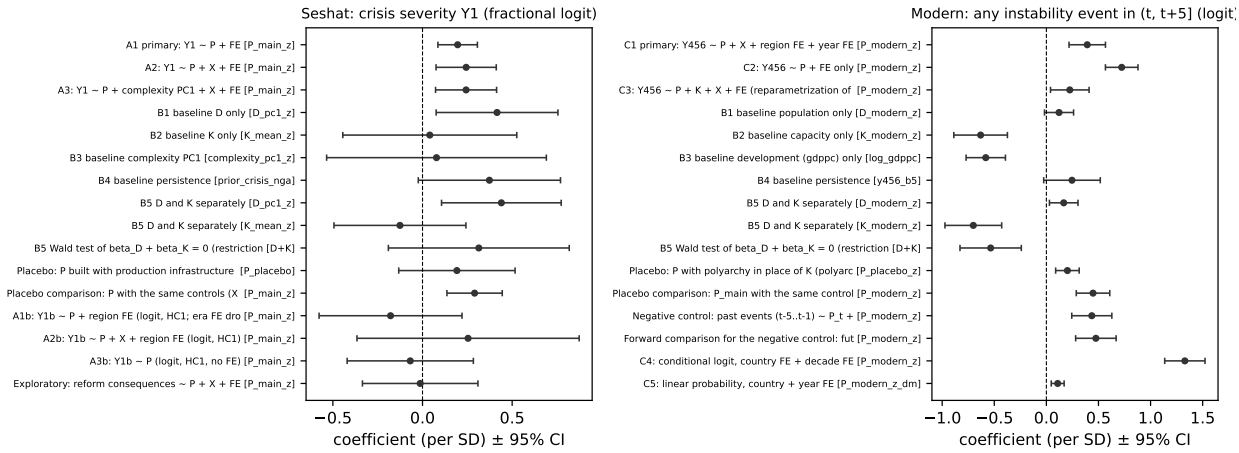


Figure 3. Coefficients (per standard deviation) with 95% confidence intervals for the pre-registered specifications, baselines, placebos and negative controls. Left: Seshat crisis severity. Right: modern instability events.

Table 12. H5 in Seshat: β_P with one macro-region or era held out (pooled column), the number of crisis records in the held-out block, and the coefficient estimated within the held-out block when it has at least twelve records.

Held-out block	Pooled β_P	SE	n	Held-out n	Within-block (SE)
leave-region-out: Central_Asia	0.275***	0.080	89	13	–
leave-region-out: East_Asia	0.269***	0.103	70	30	0.362 (0.083)
leave-region-out: Europe	0.366***	0.096	78	39	0.187 (0.164)
leave-region-out: North_Africa	0.281***	0.080	83	12	–
leave-region-out: North_America	0.266***	0.088	89	13	–
leave-region-out: South_America	0.285***	0.085	90	6	–
leave-region-out: South_Asia	0.269***	0.078	88	5	–
leave-region-out: Southeast_Asia_Oceania	0.255***	0.090	86	13	–
leave-region-out: Southwest_Asia	0.264***	0.087	75	29	0.437 (0.234)
leave-region-out: Sub_Saharan_Africa	0.236***	0.065	89	6	–
leave-era-out: 1-1000	0.336**	0.130	71	40	0.193 (0.182)
leave-era-out: 1000BCE-0	0.255*	0.146	73	28	0.390 (0.145)
leave-era-out: 1001-1500	0.310***	0.112	70	43	0.316 (0.152)
leave-era-out: 1501-1800	0.267***	0.084	79	31	0.325 (0.483)
leave-era-out: post-1800	0.287**	0.121	86	16	0.208 (0.264)
leave-era-out: pre-1000BCE	0.293**	0.114	86	8	–
leave-sphere-out: Andes_Amazonia	0.285***	0.085	90	6	–
leave-sphere-out: East_Asia	0.269***	0.103	70	30	0.362 (0.083)
leave-sphere-out: Europe	0.366***	0.096	78	38	0.187 (0.164)
leave-sphere-out: Mesoamerica_North_America	0.266***	0.088	89	13	–
leave-sphere-out: Oceania	0.278***	0.091	91	6	–
leave-sphere-out: South_Asia	0.269***	0.078	88	5	–
leave-sphere-out: Southeast_Asia	0.272***	0.077	88	7	–
leave-sphere-out: Steppe_Central_Asia	0.279***	0.079	91	12	–
leave-sphere-out: Sub_Saharan_Africa	0.236***	0.065	89	6	–
leave-sphere-out: West_Asia_North_Africa	0.274***	0.089	63	43	0.401 (0.393)

Table 13. Item-level associations in CrisisDB: logit of each consequence (present vs absent) on P, on D and on K separately, each with era effects and robust errors; coefficients per standard deviation (SE).

Consequence	n	Share present	β_P (SE)	β_D (SE)	β_K (SE)
decline	93	0.41	1.11 (0.36)	0.59 (0.29)	0.10 (0.26)
collapse	126	0.02	5.70 (2.60)	-0.49 (0.48)	-0.99 (0.39)
downward_mobility	123	0.45	0.28 (0.18)	0.20 (0.19)	0.05 (0.20)
extermination	122	0.14	0.32 (0.34)	0.54 (0.28)	0.46 (0.24)
uprising	123	0.46	0.15 (0.19)	0.47 (0.23)	0.52 (0.24)
revolution	117	0.17	0.50 (0.25)	0.66 (0.41)	0.57 (0.38)
successful_revolution	121	0.10	0.59 (0.27)	1.10 (0.58)	1.44 (0.94)
civil_war	124	0.54	0.31 (0.21)	0.45 (0.19)	0.35 (0.20)
century_plus	126	0.06	-0.73 (0.36)	-0.87 (0.38)	-0.52 (0.32)
fragmentation	127	0.35	0.21 (0.19)	0.12 (0.20)	-0.00 (0.19)
capital	116	0.18	0.06 (0.23)	0.12 (0.23)	0.13 (0.23)
conquest	124	0.40	-0.12 (0.19)	-0.05 (0.19)	0.03 (0.19)
assassination	115	0.40	0.52 (0.21)	0.66 (0.24)	0.39 (0.25)
depose	113	0.50	0.22 (0.20)	0.54 (0.21)	0.56 (0.23)
Mean inter-item correlation 0.08 (median 0.07); first eigenvalue share 0.19 of 14 items					

6.3. H4: no threshold

A hinge in P with its knot chosen on the training fold, and a three-degree-of-freedom spline, improve on the linear model out of sample in only 10% and 8% of the fifty polity-blocked folds (Table 20); the in-sample hinge places its knot at 0.92 standard deviations with a negative post-knot slope, the opposite of a tipping point. The smooth form is retained for Seshat.

6.4. Polity duration and succession

The mismatch index is unrelated to the hazard of a polity's end (Table 14): hazard ratios of 1.03 (SE 0.06) and 1.00 (SE 0.07, Schoenfeld test $p = 0.98$) with 302 and 201 events. Capacity, entered with scale, is associated with a *higher* hazard of a polity's end (ratio 1.75, SE 0.20 on the log scale). Our first reading was periodization: well-documented complex polities are divided into more, shorter Seshat polities than poorly documented simple ones, and the number of polities coded in a natural geographic area does correlate 0.29 with capacity and does raise the hazard (ratio 1.37, SE 0.14). But controlling for it leaves the capacity coefficient almost unchanged (ratio 1.71, SE 0.20), so periodization is at most part of the story. The pre-registered secondary outcome therefore yields a result that runs against the hypothesis: in Seshat, higher-capacity polities end sooner, not later, and the paper reports it as such rather than explaining it away. Candidate explanations—more frequent dynastic turnover being coded as polity change in complex states, or genuine fragility of elaborate

administrations (Tainter, 1988)—are not distinguishable here. Discontinuous succession, coded for 207 polities with 37 discontinuities, is unrelated to P (0.02 and 0.24 log-odds, SE 0.39 and 0.33).

Table 14. Secondary historical outcomes. Cox proportional hazards for polity duration, stratified by macro-region with region-clustered robust errors (coefficients on the log-hazard scale, hazard ratio in the last column), and logit for discontinuous succession with region and era effects.

Specification	Term	Coef.	SE	n	HR	Events
Y2-1 Cox: P + strata(macro-region)	P_z	0.031	0.060	317	1.03	302
Y2-2 Cox: P + X + midpoint + strata	P_z	0.000	0.068	208	1.00	201
Y2-3 Cox: D + K + X + midpoint + strata	D_pc1_z	0.128	0.133	208	1.14	201
Y2-3 Cox: D + K + X + midpoint + strata	K_mean_z	0.559***	0.200	208	1.75	201
Y2-4 Cox: complexity PC1 + X + strata	complexity_pc1_z	0.399	0.291	208	1.49	201
Y2-5 Cox: D + K + X + midpoint + polities per NGA (periodization) + strata	D_pc1_z	0.082	0.135	206	1.09	200
Y2-5 Cox: D + K + X + midpoint + polities per NGA (periodization) + strata	K_mean_z	0.535***	0.203	206	1.71	200
Y2-5 Cox: D + K + X + midpoint + polities per NGA (periodization) + strata	n_polities_in_nga_z	0.317**	0.140	206	1.37	200
Y3-1 logit: discontinuous succession ∼ P + FE	P_z	0.018	0.393	145	–	24
Y3-2 logit: + X	P_z	0.239	0.325	117	–	18

7. Results: modern states, 1820–2025

The modern panel is the form of the test that the mandate's event-risk model has in mind: a country-year panel with lagged predictors, fixed effects, forward-chaining validation and held-out regions. All estimates are [EMPIRICAL] and associational.

7.1. H2: instability events in the next five years

In the primary specification (Table 15), a one-standard-deviation higher mismatch is associated with 0.39 higher log-odds of at least one civil-war onset, coup or irregular regime end in the following five years (SE 0.09; 14,637 country-years, of which 2,692 have an event in their five-year window; the windows overlap, and the sample contains 901 distinct event-years out of 1,676 in the full panel), conditional on development, electoral democracy, time since the last event (with an indicator for countries whose spell is left-censored at panel entry), region and year. The coefficient is 0.72 without covariates, 1.33 in a conditional logit with country and decade effects (which uses only countries that experience at least one event window), and 0.11 in probability units in a linear model with country and year effects. It is positive for every horizon (0.32 at one year, 0.38 at

three, 0.45 at ten), largest for civil-war onsets (1.02, SE 0.19), small and imprecise for coups (0.19, SE 0.14) and intermediate for irregular regime ends (0.24, SE 0.09); it does not depend on the capacity aggregation, on standardizing within years, on adding land area to demand, on dropping the nineteenth century, on removing the democracy control or on adding the persistence control (Table 16). Every held-out region and period gives a pooled coefficient of similar size, and every within-block estimate is positive (Table 18, Figure 5). The second H2 primary test also passes the conventional bar.

Table 15. H2 in the modern panel: any instability event in $(t, t+5]$. Logit with region and year effects and country-clustered standard errors unless stated; X = log GDP per capita, electoral democracy, cubic in years since the last event. "Events" counts country-years whose five-year window contains an event.

Specification	Term	Coef.	SE	n	Events
C1 primary: Y456 ~ P + X + region FE + year FE	P_modern_z	0.392***	0.089	14637	2692.00
C2: Y456 ~ P + FE only	P_modern_z	0.723***	0.080	15687	2981.00
C3: Y456 ~ P + K + X + FE (reparametrization of B5: beta_P = beta_D)	P_modern_z	0.225**	0.094	14637	2692.00
B1 baseline population only	D_modern_z	0.122*	0.071	14754	2710.00
B2 baseline capacity only	K_modern_z	-0.631***	0.131	15675	2957.00
B3 baseline development (gdppc) only	log_gdppc	-0.582***	0.096	15862	2980.00
B4 baseline persistence	y456_b5	0.246*	0.138	15806	2968.00
B5 D and K separately	D_modern_z	0.166**	0.070	14637	2692.00
B5 D and K separately	K_modern_z	-0.701***	0.139	14637	–
B5 Wald test of beta_D + beta_K = 0 (restriction implied by P)	D+K	-0.535***	0.150	14637	–
Placebo: P with polyarchy in place of K (polyarchy not separately controlled)	P_placebo_z	0.202***	0.058	14754	2710.00
Placebo comparison: P_main with the same controls as the placebo	P_modern_z	0.448***	0.082	14676	2706.00
Negative control: past events (t-5..t-1) ~ P_t + gdppc + polyarchy + FE (no duration terms)	P_modern_z	0.436***	0.098	14608	2757.00
Forward comparison for the negative control: future events ~ P_t + gdppc + polyarchy + FE	P_modern_z	0.475***	0.099	14637	2692.00
C4: conditional logit, country FE + decade FE	P_modern_z	1.329***	0.099	14637	2692.00
C5: linear probability, country + year FE	P_modern_z_dm	0.109***	0.032	14637	2692.00

The baselines, placebo and negative control tell the rest (Table 15, Figure 3). Population alone is weakly associated with instability (0.12, SE 0.07); capacity alone is strongly and negatively associated (-0.63, SE 0.13), as is development (-0.58 per log unit of GDP per capita). With population and capacity entered separately, population carries 0.17 (SE 0.07) and capacity -0.70 (SE 0.14), and the restriction $\beta_D + \beta_K = 0$ that the mismatch index imposes is rejected (sum -0.54,

SE 0.15, Wald $p < 0.001$): in the modern panel P is a misspecified summary of two components of unequal weight, and the row that adds K to P (which is algebraically the same model) shows the mismatch adding 0.23 (SE 0.09) once capacity is held fixed—that is, the population component. The signal in P is, here, overwhelmingly the capacity component with the sign reversed. The placebo that replaces capacity with electoral democracy is associated at 0.20 (SE 0.06) against 0.45 (SE 0.08) for P with the same controls, so the capacity items do carry information beyond regime type, which is reassuring for the measurement model but not for the hypothesis. The negative control, read by the letter of the pre-registration, passes: past events in the five years before t are associated with P_t at 0.44 (SE 0.10), against 0.48 (SE 0.10) for future events with the same controls, so the forward association is not weaker than the backward one. But the margin is within one standard error, and a predictor that is nearly as strongly related to what has already happened as to what will happen behaves like a marker of chronically unstable countries rather than a leading indicator; we report both the literal verdict and this reading.

Table 16. Pre-registered robustness set for the modern primary coefficient (same specification as the primary row of Table 15 unless stated).

Variant	Term	Coef.	SE	n	Events
R: horizon h=1	P_modern_z	0.321***	0.091	14635	902.00
R: horizon h=3	P_modern_z	0.380***	0.087	14636	1911.00
R: horizon h=10	P_modern_z	0.448***	0.099	14639	4143.00
R: component y4_f5	P_modern_z	1.016***	0.193	10279	566.00
R: component y5_f5	P_modern_z	0.193	0.137	10853	1460.00
R: component y6_f5	P_modern_z	0.240**	0.094	14636	2068.00
R: K = PC1	P_modern_pc1_z	0.391***	0.090	14637	2692.00
R: standardized within year	P_modern_within_year_z	0.404***	0.091	14637	2692.00
R: D = (log pop + log area)/2	P_modern_area_z	0.346***	0.098	11735	2359.00
R: years $\geq$ 1900	P_modern_z	0.388***	0.091	13083	2475.00
R: polyarchy not controlled	P_modern_z	0.448***	0.082	14676	2706.00
R: + persistence control	P_modern_z	0.389***	0.088	14598	2682.00

Out of sample (Table 17, Figure 4) the verdict is the same as in Seshat with the roles of the components exchanged. Under forward chaining (three temporal splits, the first of which trains on 1820–1945 and tests on 1946–1975), the model with P has the highest mean area under the curve (0.732 against 0.722 for the fixed-effects model, 0.724 for population or capacity alone and 0.726 for both), but it beats the capacity-only and the scale-plus-capacity baselines in only one of the three splits on either metric; with regions held out it beats capacity alone in one of six folds on log-likelihood and three of six on discrimination. All models are reasonably calibrated under forward chaining (slopes 0.77–0.79) and all are badly calibrated when a whole region is held out

(slopes about 0.40), which is itself a finding: the level of instability risk is regional, and no index built from scale and capacity transports it. Under the pre-registered 80% rule, P fails against the capacity-only and the scale-plus-capacity baselines in both schemes (the latter nests the former, so this is the expected outcome whenever the restriction is wrong), and the second H2 primary test is therefore not passed.

Table 17. Out-of-sample validation in the modern panel. Top: share of folds in which the model with *P* beats the baseline on the metric; the rule requires 0.80. Bottom: mean metrics by scheme and model (forward chaining: three temporal splits; LORO: six held-out regions).

Baseline	Scheme	Metric	Folds	Win share	Mean Δ
B0 FE+X	forward	oos_loglik	3	0.67	+0.0031
B0 FE+X	forward	auc	3	1.00	+0.0093
B1 D	forward	oos_loglik	3	0.67	+0.0022
B1 D	forward	auc	3	1.00	+0.0080
B2 K	forward	oos_loglik	3	0.33	+0.0221
B2 K	forward	auc	3	0.33	+0.0073
B4 persistence	forward	oos_loglik	3	0.67	+0.0049
B4 persistence	forward	auc	3	1.00	+0.0136
B5 D+K	forward	oos_loglik	3	0.33	+0.0215
B5 D+K	forward	auc	3	0.33	+0.0057
B0 FE+X	loro	oos_loglik	6	0.67	-0.0028
B0 FE+X	loro	auc	6	0.83	+0.0071
B1 D	loro	oos_loglik	6	0.67	-0.0014
B1 D	loro	auc	6	1.00	+0.0070
B2 K	loro	oos_loglik	6	0.17	-0.0128
B2 K	loro	auc	6	0.50	+0.0014
B4 persistence	loro	oos_loglik	6	0.50	-0.0047
B4 persistence	loro	auc	6	0.83	+0.0073
B5 D+K	loro	oos_loglik	6	0.17	-0.0108
B5 D+K	loro	auc	6	0.50	-0.0011

Scheme	Model	OOS log-lik.	AUC	Brier	Calibration slope
forward	B0 FE+X	-0.425	0.722	0.137	0.79
forward	B1 D	-0.424	0.724	0.137	0.79
forward	B2 K	-0.444	0.724	0.146	0.77
forward	B4 persistence	-0.427	0.718	0.138	0.78
forward	B5 D+K	-0.443	0.726	0.146	0.77
forward	C1 P	-0.422	0.732	0.136	0.79
loro	B0 FE+X	-0.527	0.671	0.142	0.40
loro	B1 D	-0.528	0.671	0.143	0.40
loro	B2 K	-0.517	0.677	0.140	0.39
loro	B4 persistence	-0.525	0.671	0.142	0.40
loro	B5 D+K	-0.519	0.679	0.141	0.39
loro	C1 P	-0.530	0.678	0.143	0.39

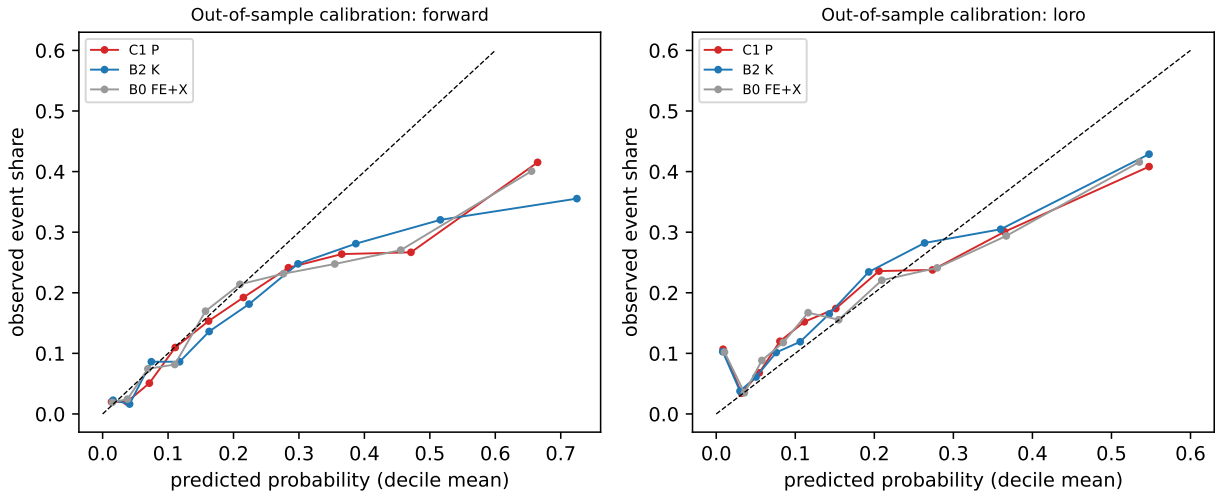


Figure 4. Out-of-sample calibration in the modern panel (decile means of predicted probability against observed event shares) for the model with P, the capacity-only baseline and the fixed-effects baseline, under forward chaining (left) and leave-one-region-out (right).

Table 18. H5 in the modern panel: β_P with one region or period held out (pooled), and estimated within the held-out block.

Held-out block	Pooled β_P	SE	n	Within n	Within-block (SE)
leave-region-out: 1.0	0.284***	0.094	13252	1385	1.034 (0.436)
leave-region-out: 2.0	0.318***	0.100	12017	2620	0.517 (0.001)
leave-region-out: 3.0	0.311***	0.097	13175	1462	0.441 (0.354)
leave-region-out: 4.0	0.310***	0.117	11363	3274	0.407 (0.163)
leave-region-out: 5.0	0.279***	0.096	10717	3920	0.804 (0.314)
leave-region-out: 6.0	0.462***	0.096	12661	1976	0.246 (0.201)
leave-period-out: 1900-1945	0.407***	0.101	12616	2021	0.256 (0.213)
leave-period-out: 1946-1989	0.367***	0.115	8927	5710	0.436 (0.124)
leave-period-out: 1990+	0.409***	0.108	9285	5352	0.346 (0.161)
leave-period-out: pre-1900	0.388***	0.091	13083	1554	0.621 (0.287)

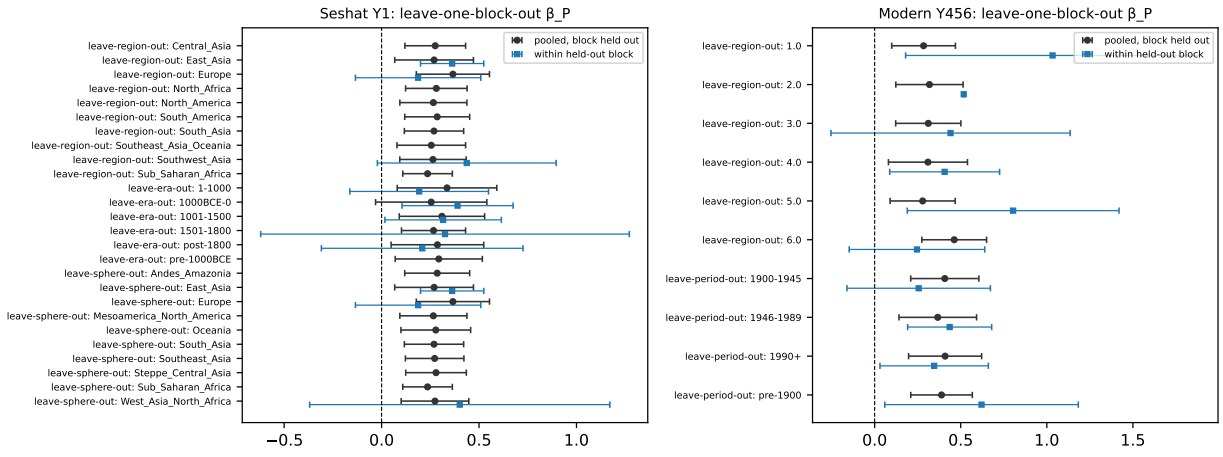


Figure 5. H5: the mismatch coefficient with each macro-region or era (Seshat, left) and each region or period (modern panel, right) held out, and the coefficient estimated inside the held-out block where possible.

7.2. Sensitivity to unobservables

Table 19 applies Oster's and Cinelli–Hazlett's diagnostics to the linear-probability version of the primary model. Adding the controls moves the coefficient from 0.082 to 0.037 while raising R^2 from 0.127 to 0.161; the bias-adjusted coefficient at $R_{\max} = 1.3R^2$ and $\delta = 1$ is -0.030, on the other side of zero. The partial R^2 of P is 0.005 and the robustness value is 0.07: an unobserved confounder that explained seven percent of the residual variance of both the mismatch and the outcome would suffice to remove the association. Two caveats apply. Proportional selection ($\delta = 1$) is a strong assumption when one of the controls, electoral democracy, correlates 0.75 with a component of the regressor, so the Oster adjustment overstates the plausible bias; and the robustness value, while twelve times the partial R^2 of P, has no benchmark covariate against which to judge it. The coefficient is fragile under these conventions, not shown to be spurious.

Table 19. Sensitivity of the modern primary coefficient to unobservables (linear probability model with the primary controls).

Quantity	Value
LPM β_P with controls	0.0367
LPM β_P without controls	0.0819
R^2 with / without controls	0.159 / 0.127
Oster bias-adjusted β_P ($R_{\max} = 1.3R^2$, $\delta = 1$)	-0.0303
Partial R^2 of P	0.0051
Cinelli–Hazlett robustness value $RV_{q=1}$	0.069

7.3. H4: a nonlinearity of negligible size

With country-blocked folds, the hinge improves on the linear form in 11 of 15 folds and the spline in 12 (Table 20); the spline meets the pre-registered 80% rule exactly and the hinge does not. By the letter of the pre-registration, H4 is therefore not rejected in the modern panel. The improvement is 0.003 nats per observation; the in-sample curve (Figure 6) is a mild convexity around the median rather than a step, and the hinge knot is at -0.08 standard deviations, i.e. at the centre of the distribution. We report this as a nonlinearity that survives the rule but is small, consistent with the smooth convexity of Theorem 5 rather than with a tipping point; the pre-registration's falsifier for H4 was a hinge or step model that *outperforms* the smooth one, and the form that passes is the smooth spline.

Table 20. H4: linear against hinge (knot chosen on the training fold) and natural cubic spline, out of sample.

Panel	Folds	Hinge wins	Spline wins	Log-lik. linear	hinge	spline
Seshat Y1 (10×5-fold)	50	0.10	0.08	-0.5998	-0.6046	-0.6015
Modern Y456 (3×5 country-blocked folds)	15	0.73	0.80	-0.4114	-0.4108	-0.4088

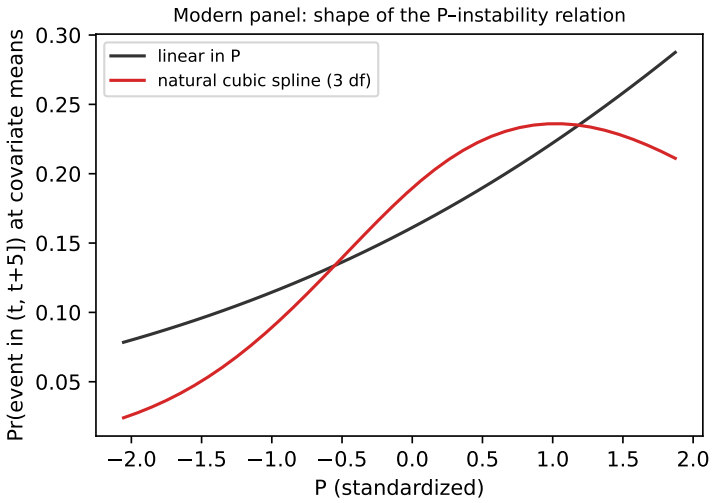


Figure 6. Shape of the relation between P and the five-year event probability in the modern panel, at covariate means, linear and spline fits.

8. Agent-based experiments

Everything in this section is [SIMULATED]. The simulation exists to measure what the historical record cannot: messages, bits, queues, overhead and failures under institutional regimes that are

imposed by the experimenter. Nothing emerges in these experiments, and nothing in them is evidence about history; they instantiate the theorems of Section 3 in a system where the quantities they speak of are defined, and they search for the conditions under which institutions reduce efficiency or add fragility, as the pre-registration requires. Two things should be said before any number. First, the error structure is a parameter, not a finding: agents without a shared code misdecode with a probability proportional to their dialect distance (0.25 per unit), agents with a code with probability 0.02, and agents who can verify against a record not at all. Any statement that institutions "raise success" in this model is therefore a statement about those parameters, and the ablation that varies them is the one to read. What the model measures that is not assumed is the *cost* side: bits, messages, queueing delay, overhead, and how they scale and fail. Second, the regimes are compared at an equal total administrative budget, so that federation can gain only by relaying less.

8.1. Design

N agents are placed at random in the unit square and linked on a lattice-like nearest-neighbour graph, a small-world rewiring of it, or a random graph of the same degree. Each period, tasks arrive at rate $\lambda = 0.5$ per producing agent; a task is initiated by one agent and requires $k = 3$ participants—the initiator and two network neighbours—to hold the same plan, drawn from a set of $s = 64$ plans with a Zipf distribution of exponent 0 (uniform, entropy 6 bits) or 1 (entropy 4.86 bits). The initiator must communicate the plan to its two partners (there is no broadcast). Each agent can process a bounded number of messages per period (2 or 8); excess messages queue first-in-first-out, and a task whose last message is served more than three periods after arrival fails. Five regimes are compared. *None*: raw messages of $\lceil \log_2 s \rceil$ bits per recipient, dialect errors, and a one-bit acknowledgement round that doubles the messages each agent must process. *Norms*: a shared Shannon code optimal for the plan distribution (Shannon rather than Huffman lengths, so that the stale-code penalty of Theorem 3 applies verbatim), lower decoding error, a one-time learning cost of 200 bits per agent, and, after a shift in the plan distribution, a stale code until the institution re-learns it ten periods later. *Records*: norms plus a written record per task (four bits of administrative overhead) against which recipients verify, which removes decoding errors and replaces the acknowledgement round. *Central*: records plus an eight-ary administrative tree; every task is relayed from the initiator up to the root and down to each partner, the codeword is transmitted on every hop, and every administrator on the path processes one message per hop. *Federated*: the same tree, but a task is relayed only up to the lowest common ancestor of its participants, so that local tasks never reach the root. Administrators are withdrawn from production. The total administrative budget is a multiple (0.5 or 2 in the main design; 0.25 to 4 in the ablation) of the expected relayed load of the *central* design, and it is allocated to administrators in proportion to the load each one carries in a five-period warm-up with unconstrained capacity ("capacity follows load"); central and federated designs therefore face the

same budget, and the budget multiple is the inverse utilization of the central design in the sense of Theorem 5. Shocks at period 20 of 40 either shift the plan distribution (a cyclic permutation reaching a Kullback–Leibler divergence of 1.03 bits in the non-uniform cells; in uniform cells the shift is a no-op and those cells are excluded from the shock statistics) or remove half of the links among agents (the administrative tree is not touched). The main design crosses $N \in \{50, 200, 800\}$, three networks, two bandwidths, two entropies, three shock conditions and five regimes with eight seeds (6,048 runs); the ablations (584 runs) vary verification, adaptation speed, horizon, learning cost, heterogeneity, administrative budget, branching and the plan space. Measured outcomes are task success, messages and bits per task, bits per successful task with and without set-up, latency, failure shares by cause, administrative overhead share, welfare (task value minus a cost per bit), the Gini coefficient of benefits among producers, and root load. Shock effects are difference-in-differences against the unshocked run of the same cell and seed. Entropies, code lengths and divergences are computed because the plan distribution and the code are defined by the experimenter.

8.2. H3: the cost side

Table 21 pairs each institutional run with the no-institution run of the same cell and seed, and Table 22 counts the design cells in which an institution does worse than nothing. Shared norms lower the operating bits per successful task in 99.7% of pairs; once the 200-bit learning cost is included they are never cheaper within the forty-period horizon, because the coding gain of a six-bit plan space is about one bit per recipient (Table 25: the Shannon code averages 5.44 bits at Zipf exponent 1 against 6 raw bits), exactly the break-even arithmetic of Theorem 2; the ablation finds the code repaying its learning cost only with a 256-plan space, a Zipf exponent of 1.5 and a 160-period horizon (Figure 8). Records lower operating cost in 71% of pairs and raise welfare in every pair, the acknowledgement round they replace being what saturates agents' queues at low bandwidth in the flat regimes (at bandwidth 2, flat regimes complete about half of their tasks and records 96%; at bandwidth 8 all complete more than 90%). The hierarchies are expensive in bits because the codeword travels on every hop: at the reference budget, central administration costs 83 bits per successful task against 25 for records and 15 for nothing, federated administration 53, and both lower welfare below the flat institutional regimes in half the cells because administrators are withdrawn from production. With a budget at half the central design's load, central administration fails almost every task (success down by 0.49 relative to nothing; welfare loss in all 108 cells), and federation, which relays fewer hops, does only somewhat better (success down by 0.29; welfare loss in 80% of cells). Figure 7 shows the scaling with N: messages and bits per task grow with the depth of the tree for the hierarchies and are flat for the flat regimes; inequality of benefits is slightly higher under the hierarchies.

Table 21. [SIMULATED] Paired comparison of each institutional regime with the no-institution run of the same cell and seed (main design, 864 pairs per row). Shares of pairs in which the regime has lower bits per successful task (operating; including set-up), higher success and higher welfare, and mean differences. The budget multiple is relative to the central design's expected relayed load.

Regime	Budget	Pairs	Lower bits	Lower incl. set-up	Higher success	Higher welfare	Δ bits/success	Δ success	Δ welfare
norms	2.0	864	1.00	0.00	0.99	0.90	-5.5	+0.032	+1.5
records	2.0	864	0.71	0.50	1.00	1.00	-43.6	+0.434	+25.7
central	0.5	864	0.00	0.00	0.00	0.00	+2450.9	-0.493	-32.1
central	2.0	864	0.33	0.25	1.00	0.50	+19.5	+0.409	+9.7
federated	0.5	864	0.09	0.04	0.22	0.18	+692.9	-0.293	-22.4
federated	2.0	864	0.50	0.46	1.00	0.50	-4.6	+0.412	+10.5

Table 22. [SIMULATED] Fragility search: share of the 108 design cells (averaged over seeds) in which the regime has lower welfare, higher bits per success including set-up, or lower success than no institution.

Regime	Budget	Cells	Welfare loss	Bits loss	Success loss
central	0.5	108	1.00	1.00	1.00
central	2.0	108	0.50	0.75	0.00
federated	0.5	108	0.80	0.95	0.78
federated	2.0	108	0.50	0.54	0.00
norms	2.0	108	0.06	1.00	0.00
records	2.0	108	0.00	0.50	0.00

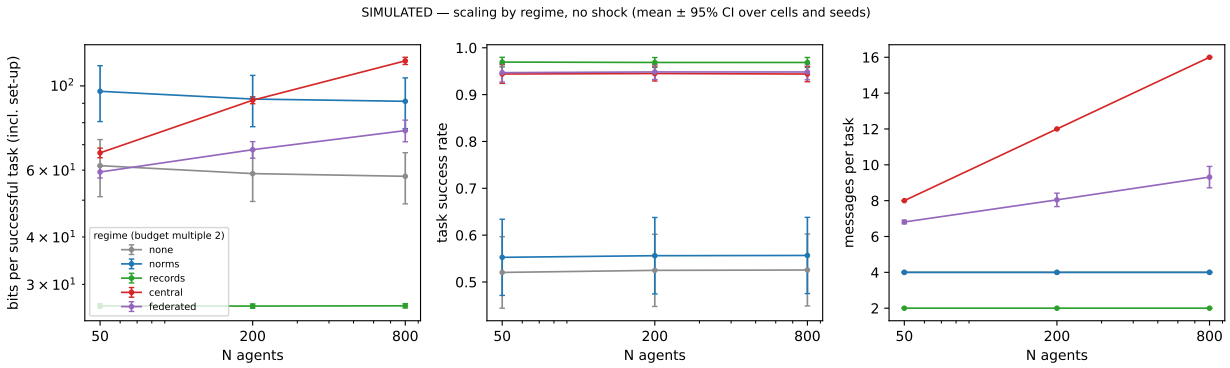


Figure 7. [SIMULATED] Scaling by regime without shocks (budget multiple 2): bits per successful task including set-up (log scale), success rate and messages per task against N; means and 95% confidence intervals over cells and seeds.

8.3. Theorems in silico

The coding-gain sweep reproduces Theorem 1: the expected code length lies in $[H(p), H(p)+1)$ for every plan space and exponent, and it is below the raw length whenever $H(p) \leq \log_2 s - 1$. In

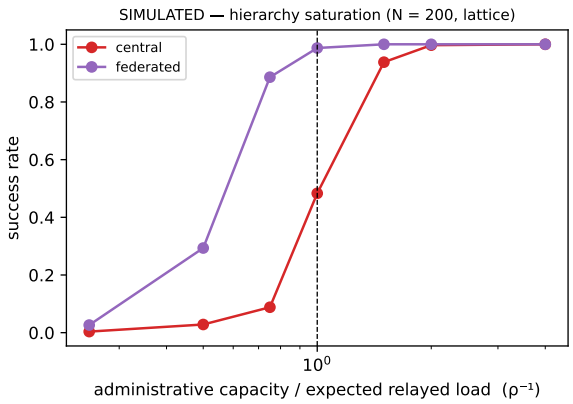
the one cell where the distribution is nearly uniform (exponent 0.5, entropy 5.78 bits) the Shannon code is longer than the six raw bits (6.24). As the remark after Theorem 1 notes, this is a property of the suboptimal Shannon code, not of coding: a Huffman code is never longer than the fixed-length code. We keep the Shannon code because it is the one Theorem 3 is stated for, and the sweep then reproduces that theorem too: after a shift in the plan distribution, a non-adapting code costs 1.90 extra bits per recipient against a divergence of 1.98 bits, within the one-bit rounding; a code that re-learns after ten periods pays half, an instantly adapting code almost nothing. Because decoding success does not depend on which plan was drawn, distribution shocks raise cost without lowering success in this model (Table 23), which is a modelling choice and not a finding. Link removal lowers success by 0.07–0.08 under records and the hierarchies at the reference budget and by almost nothing under the flat regimes; the latter is partly a floor effect, since at bandwidth 2 the flat regimes are already failing half of their tasks before the shock (the replication package reports the decomposition by bandwidth). The budget sweep traces the saturation of Theorem 5 and the routing bound of Theorem 4 together (Table 24, Figure 8): central success is 0.03 at a budget of half the load, 0.48 at a budget equal to the load, and 0.94 at one and a half times; federated success, with the same budgets, is 0.29, 0.99 and 1.00, because its root carries a tenth of the central root's load.

Table 23. [SIMULATED] Shock responses by regime: difference-in-differences of success and of bits per task (post minus pre, shocked minus unshocked run of the same cell and seed); standard error of the mean in parentheses; achieved divergence for distribution shocks (non-uniform cells only).

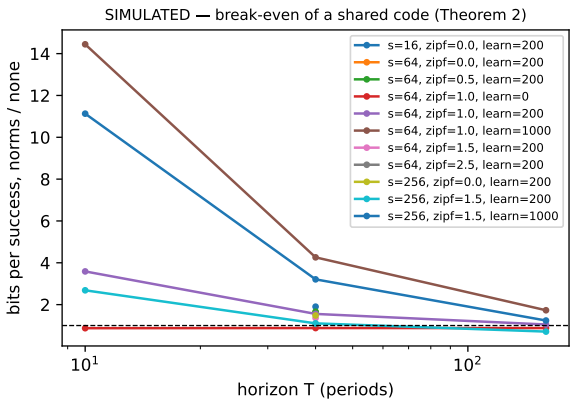
Shock	Regime	Budget	Δ success (SE)	Δ bits/task	$D_{KL}(q\|\|p)$
distribution	central	0.5	+0.000 (0.000)	+6.286	1.030
distribution	central	2.0	+0.000 (0.000)	+6.286	1.030
distribution	federated	0.5	+0.000 (0.000)	+4.224	1.030
distribution	federated	2.0	+0.000 (0.000)	+4.224	1.030
distribution	none	2.0	+0.000 (0.000)	+0.000	1.030
distribution	norms	2.0	+0.000 (0.000)	+1.031	1.030
distribution	records	2.0	+0.000 (0.000)	+1.049	1.030
network	central	0.5	+0.000 (0.000)	+0.031	0.000
network	central	2.0	-0.073 (0.006)	+0.031	0.000
network	federated	0.5	-0.019 (0.003)	+0.565	0.000
network	federated	2.0	-0.075 (0.006)	+0.565	0.000
network	none	2.0	-0.004 (0.001)	+0.000	0.000
network	norms	2.0	-0.004 (0.001)	+0.001	0.000
network	records	2.0	-0.070 (0.005)	-0.003	0.000

Table 24. [SIMULATED] Hierarchy saturation ($N = 200$, lattice): success, latency failures, mean latency and root load against the administrative budget multiple, at equal total budgets for the two designs.

Regime	Budget	Success	Latency failures	Latency	Root load
central	0.25	0.004	0.996	2.98	4748
central	0.50	0.028	0.972	2.62	3286
central	0.75	0.088	0.912	2.49	1823
central	1.00	0.483	0.517	2.21	418
central	1.50	0.938	0.062	1.33	302
central	2.00	0.997	0.003	1.09	302
central	4.00	1.000	0.000	1.00	302
federated	0.25	0.026	0.974	2.59	464
federated	0.50	0.293	0.707	2.27	119
federated	0.75	0.886	0.114	1.46	44
federated	1.00	0.987	0.013	1.16	42
federated	1.50	1.000	0.000	1.01	42
federated	2.00	1.000	0.000	1.00	42
federated	4.00	1.000	0.000	1.00	42



(a) Saturation of hierarchies.



(b) Break-even of a shared code.

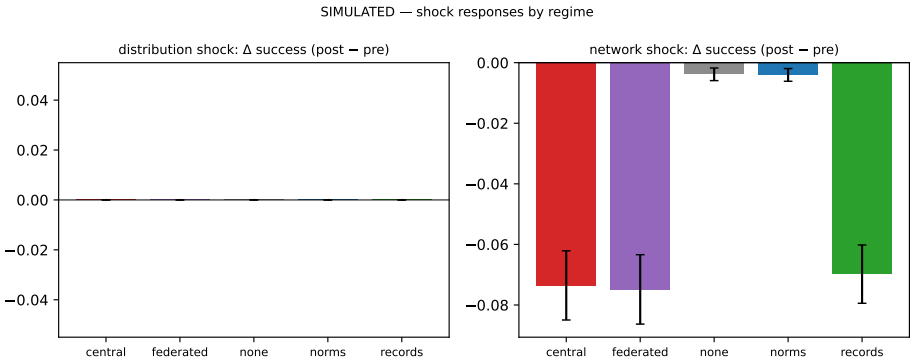


Figure 8. [SIMULATED] Success against the administrative budget multiple (left), ratio of bits per success with and without a shared code against the horizon (right), and difference-in-differences shock responses (bottom).

Table 25. [SIMULATED] Ablations ($N = 200$, lattice, eight seeds) and theorem checks.

Ablation	Result
Reference (N=200, lattice, no shock): success / bits per success / welfare per agent	central: 0.997 / 82.8 / 42.7; federated: 1.000 / 53.3 / 43.7; none: 0.918 / 15.3 / 55.0; norms: 0.960 / 23.7 / 57.3; records: 1.000 / 24.8 / 59.3
Records without verification	success 0.960, bits per success 25.8, welfare 57.2
Norms, extra bits per task after a distribution shock by adaptation delay	delay 0: +0.03; delay 10: +1.92; delay 1000: +3.79 (KL = 1.98 bits)
Break-even (norms cheaper than none incl. set-up)	s=16,zipf=0.0,T=40,learn=200.0: no; s=64,zipf=0.0,T=40,learn=200.0: no; s=64,zipf=0.5,T=40,learn=200.0: no; s=64,zipf=1.0,T=10,learn=0.0: yes; s=64,zipf=1.0,T=10,learn=200.0: no; s=64,zipf=1.0,T=10,learn=1000.0: no; s=64,zipf=1.0,T=40,learn=0.0: yes; s=64,zipf=1.0,T=40,learn=200.0: no; s=64,zipf=1.0,T=40,learn=1000.0: no; s=64,zipf=1.0,T=160,learn=0.0: yes; s=64,zipf=1.0,T=160,learn=200.0: no; s=64,zipf=1.0,T=160,learn=1000.0: no; s=64,zipf=1.5,T=40,learn=200.0: no; s=64,zipf=2.5,T=40,learn=200.0: no; s=256,zipf=0.0,T=40,learn=200.0: no; s=256,zipf=1.5,T=10,learn=200.0: no; s=256,zipf=1.5,T=10,learn=1000.0: no; s=256,zipf=1.5,T=40,learn=200.0: no; s=256,zipf=1.5,T=40,learn=1000.0: no; s=256,zipf=1.5,T=160,learn=200.0: yes; s=256,zipf=1.5,T=160,learn=1000.0: no
Heterogeneity ε: success none vs norms	eps=0.05,none: 0.984; eps=0.05,norms: 0.960; eps=0.25,none: 0.918; eps=0.25,norms: 0.960; eps=0.6,none: 0.808; eps=0.6,norms: 0.960
Federated branching b: success / messages per task / root load	b=4: 0.93 / 8.86 / 48; b=8: 0.99 / 6.61 / 42; b=16: 1.00 / 6.27 / 209
Coding gain (norms): $H(p)$, L_{code}, L_{raw}	s=16,zipf=0.0: 4.00, 4.00, 4; s=64,zipf=0.0: 6.00, 6.00, 6; s=64,zipf=0.5: 5.78, 6.24, 6; s=64,zipf=1.0: 4.86, 5.44, 6; s=64,zipf=1.5: 3.44, 4.00, 6; s=64,zipf=2.5: 1.45, 1.96, 6; s=256,zipf=0.0: 8.00, 8.00, 8; s=256,zipf=1.5: 3.92, 4.44, 8
Theorem checks	T1 ($H \le L < H+1$): True; code shorter than raw whenever $H \le \log_2 s - 1$: True; T3 extra bits per recipient 1.90 vs KL 1.98

8.4. What the experiments say about the hypothesis

On the cost side, which is the side the model actually measures, H3 is supported for operating cost and refuted for total cost within short horizons: a code saves bits per task in proportion to the regularity of the task environment, and it repays its learning cost only when that regularity is large and the horizon long (Theorem 2). Records are the regime that wins most often, and they win by

replacing a round of acknowledgements with a lookup, not by compressing messages. Hierarchies buy routing at the price of relaying, and they fail abruptly when the budget falls below the load they relay; federation halves that load and is the only hierarchical design that survives a tight budget. On the success side, the ranking of regimes is the ranking of their error parameters: with heterogeneity at 0.6 the no-code regime succeeds in 81% of tasks against 96% with a code, with heterogeneity at 0.05 the ranking reverses (98% against 96%) because the code's own error rate is then the larger one. The fragility search therefore finds what the framework predicts and nothing else: institutions hurt when their fixed cost is not amortized, when a hierarchy is asked to relay more than it can process, and when there is too little regularity for a code to compress. These are the cases the historical literature on over-extended administrations and rigid bureaucracies describes, and none of them is measurable with the indicators available in Seshat.

9. Synthesis: theory, data, and rival explanations

9.1. What is observable and what is not

Table 26 sets the framework's mechanisms against the two kinds of evidence. The coding gain, the stale-code penalty and the routing bounds are measured in the simulation and are not observable in Seshat, which records whether writing, coins or courts existed and never how long a message was or how many tasks waited. The break-even condition is reflected in the historical record only through the association of capacity with scale (H1), which is consistent with institutions being built where they pay but equally consistent with the reverse. Backlog dynamics are observable nowhere in history. The one mechanism that the historical tests address, sustained overload leading to breakdown (Conjecture 2), is addressed only through the proxy P, and the proxy's claim to track utilization (Conjecture 1) is what the collinearity and placebo results undermine.

Table 26. Mechanisms of the framework against the two kinds of evidence.

Mechanism	Historical record ([EMPIRICAL])	Simulation ([SIMULATED])
Coding gain (Thm. 1)	Not observable: presence of codes, not message lengths	Reproduced in every cell; Shannon code longer than raw in one near-uniform cell
Break-even (Thm. 2)	Indirect: capacity rises with scale (H1)	Code repaid only with large concentrated plan space and long horizon
Stale code (Thm. 3)	Not observable	Extra bits 1.90 vs divergence 1.98
Routing and root load (Thm. 4)	Partially: administrative levels coded, loads not	At equal budgets, central collapses below budget multiple 1; federated keeps most of its success
Backlog (Thm. 5)	Not observable; P is a proxy under C1	By construction
Overload $\rightarrow$ breakdown (C2)	Tested through P: positive but not beyond components	Not modelled
Smooth vs threshold (C3)	No threshold in Seshat; negligible nonlinearity in modern panel	Convex latency in utilization

9.2. The model against simpler rivals

The pre-registration named the rivals and the results rank them. In Seshat, a scale-only model explains crisis severity as well as the mismatch model, in and out of sample; the placebo built from production infrastructure is not distinguishable from P; capacity alone does nothing, and the equal-and-opposite restriction is not rejected only because β_K is imprecise. The simplest reading is Johnson's: larger polities face more that can go wrong when a crisis strikes, and whether their administrations were "adequate" is not something the presence codes can tell us. In the modern panel, a capacity-only model explains instability onsets as well as or better than the mismatch model; so does development; population adds little. The simplest reading is the state-capacity literature's (Besley and Persson, 2009; Hanson and Sigman, 2021; Goldstone et al., 2009): states with impartial administrations, predictable law and fiscal and territorial reach experience fewer civil wars, coups and irregular transitions, and the mismatch index inherits that association with its sign reversed. In neither panel does the *difference* between scale and capacity carry information that its components do not.

The rival mechanisms named in the mandate fare as follows. Military competition: the warfare-technology index is included as a covariate in every specification and does not displace the scale

association; it is not itself associated with severity once scale is in the model. Environmental shocks: an epidemic accompanying the crisis is a covariate; the climate-mediated pathway of Hoyer et al. (2023) is not testable with Polaris alone. Geography: absorbed by region effects, and the leave-one-region-out results show no region driving the estimates. Resource base: the production-infrastructure index is a covariate and, used as a placebo, mimics capacity. Inequality and elite conflict: not measurable at the polity level in Polaris and only through regime-type proxies in V-Dem; the structural-demographic mechanism (Turchin, 2004; Turchin and Nefedov, 2009) remains an open rival that this paper cannot distinguish from scale. Cultural diffusion and selection: addressed by region and family controls for inference and by the crisis-selection caveat, not for identification. Coercion: the military professionalization index is a covariate in H1 and was not associated with severity.

9.3. Consistency with simulation-derived predictions

The simulation predicts that institutions lower operating cost per successful task (and raise success by the construction of their error parameters), which agrees with the positive association of capacity with scale and, in the modern panel, with the negative association of capacity with instability; agreement of this kind is weak evidence because the data cannot distinguish capacity enabling scale from scale demanding capacity. The simulation also predicts a sharp utilization effect—hierarchies fail when relayed load exceeds administrative capacity—whose historical analogue would be a threshold in the mismatch index. The historical tests find no threshold in Seshat and, in the modern panel, a nonlinearity that meets the pre-registered rule but is negligible in size. Either there is no utilization threshold in history, or the proxy does not measure utilization; the collinearity of the proxy's components makes the second reading at least as likely, and the paper cannot tell them apart. Disagreement is more informative than agreement here: it marks the precise place where better data would be decisive.

9.4. Scope conditions and failed predictions

Three statements survive, each with its scope. First, capacity and scale are positively associated across 317 historical polities, 579 ethnographic societies and 14,647 modern country-years, with the sign stable within regions and eras and over fifty-year differences; the direction is not identified. Second, among the 127 historical crises with a defined pre-crisis index, a larger polity's crisis is recorded with more adverse consequences, at about 0.42–0.44 log-odds per standard deviation of scale on the common sample; this is a statement about scale, not about mismatch. Third, modern states with higher administrative, legal, fiscal and territorial capacity experience fewer instability onsets in the following five years, at about -0.63 log-odds per standard deviation; this replicates

a known result and is as strongly related to past as to future instability. Five predictions failed or fell short. The mismatch index does not outperform its components out of sample in either panel. The restriction it imposes is rejected in the modern panel. The placebo index is not distinguishable from it in Seshat. The negative control in the modern panel passes by the letter but by less than one standard error. There is no threshold in Seshat and only a small nonlinearity in the modern panel; and the secondary polity-duration outcome runs against the hypothesis. By the rule written before the data were examined, the universal theory of coordination, institutions and civilizational complexity is not supported by the historical and modern evidence available in 2026, and the paper is written as the test that did not confirm it.

10. Limitations

The limitations were written down during the analysis and before this section; the full list is in the replication package. The ones that bear on how the results should be read are these.

Data. The CrisisDB outcome is crisis-selected (150 of 626 polities), so the Seshat test concerns the severity of recorded crises, not their occurrence. Only 6% of Polaris facts are dated, so pre-crisis capacity is in practice whole-polity capacity and temporal precedence is weak; a dated-only test was not estimable. Seshat capacity items record the existence of institutions, not their reach, quality or cost, and the demand side is scale only because the databank has no usable indicator of division of labour or dispersion. Institutional codes and crisis codes come from the same project and overlapping historiography, which permits circularity that the inferred-code and late-crisis sensitivities can only partly address. Seshat polity boundaries are periodization decisions, so polity duration is not an instability outcome. In the modern panel, the six capacity items correlate 0.75 with electoral democracy and are rated by experts who know what came later; population begins in 1820 for most countries and there is no dispersion measure; 39 V-Dem entities without Maddison population drop out. The Ethnographic Atlas is a cross-section with coarse ordinal community sizes (the top category is bounded at 50,000) and no outcome; jurisdictional levels are themselves a response to scale in the scalar-stress tradition (Johnson, 1982; Carneiro, 1967), so the H1 association there replicates a known size–hierarchy regularity rather than testing compression. The severity share aggregates fourteen weakly correlated consequences (Section 6.2).

Design. Nothing is identified causally. The modern negative control passes the pre-registered letter (past 0.44 against future 0.48 log-odds per SD) by a margin within one standard error; the Oster adjustment under proportional selection crosses zero, with the caveat that proportional selection is a strong assumption when one control (electoral democracy) correlates 0.75 with a component of the

regressor, and the robustness value is 0.07 against a partial R^2 of 0.006. The Seshat binary outcome was underpowered in advance. There are 42 regions and ten macro-regions, so cluster-robust inference is anti-conservative and alternative clusterings are reported. The verdict on H2 rests on a pre-registered out-of-sample rule against baselines, which is the right test of "beyond conventional baselines" but is also a demanding one at $n = 93$–127, and one of the baselines (scale and capacity entered separately) nests the mismatch model, so that P cannot do better than it in expectation; we report the fold-level results and the Wald test of the restriction so that readers who prefer a different rule can apply it. Several specification errors found by the adversarial reviews (a placebo that reparametrized the scale-only baseline, a consequence item used as a covariate, silently failed cross-validation folds, mislabelled rows) were corrected and are listed as amendments A-004 to A-007. The Equinox-2020 replication listed in the pre-registration was not executed (not blocked: the raw file is in the package, the parser was not written) and is recorded as such in the audit.

Theory. The theorems describe an idealized communication system in which capacity is exogenous; historically it is built in response to demand. The conjectures that link the framework to the proxies are exactly what the measurement results call into question: with scale and capacity correlated 0.84, the proxy P retains 17% of their joint variance, and a residual of that kind cannot be assumed to track utilization.

Simulation. Regimes are imposed; the model says nothing about emergence. The success ranking of regimes is the ranking of their decoding-error parameters (0.25 per unit of dialect distance without a code, 0.02 with a code, zero with records) and reverses when heterogeneity is low; only the cost side is measured rather than assumed. Bit and message accounting is stylized (one-bit acknowledgements, four-bit records, codewords repeated on every relay hop), and only within-simulation comparisons are meaningful. Decoding success does not depend on the plan drawn, so distribution shocks affect cost but not success by construction; link removal does not touch the administrative tree. The saturation of hierarchies follows from the budget multiple, a free parameter swept across its range; it is confirmed, not discovered, and its position depends on the "capacity follows load" allocation rule. Shannon rather than Huffman codes are used so that the stale-code penalty is comparable to the theorem; an optimal code would never be longer than fixed-length coding. Several of these choices were corrected after the computational review (amendment A-005).

Process. The study was executed autonomously by a language-model agent under a human mandate in a single session: the pre-registration, code, analyses and text were produced by the agent, with the mandate's prohibitions on fabrication, imputation of missing data as zero and unapproved spending enforced by the execution log and the replication package. The principal

investigator has not yet reviewed the code or the text at the time of writing; three adversarial reviews commissioned from independent reviewer personas and the responses to them are included in the package. Reference verification relied on Crossref metadata and primary-source abstracts; the chapter contents of Johnson (1982) and the published tables of Sng (2014) were not read.

11. Conclusion

We set out to state a universal theory of human coordination in a form that could be tested, and to test it. The statement survived: institutions as shared codes, records and routing structures that lower the communication cost per coordinated task are a coherent object, with five provable properties and a simulable cost structure, and the simulations show both the gains and the three ways institutions can hurt—unpaid set-up costs, saturated hierarchies, and task environments with too little regularity to compress. The test did not confirm the theory's empirical corollary. In the Seshat databank and in a two-century country-year panel, the mismatch between coordination demand and coordination capacity is associated with instability in the predicted direction, robustly across regions and eras, and yet it adds nothing identifiable to what scale alone (in history) or capacity alone (in modern states) already explain; a placebo does as well, past instability loads on the index as strongly as future instability, and the coefficient does not survive a standard adjustment for unobservables. Under the rule we committed to before looking, the universal framing is not supported, and we have written the narrower paper instead.

The result should be read for what it says about measurement as much as about the hypothesis. Scale and capacity are so tightly coupled across the historical record that their difference is a residual carrying one sixth of their joint variance; a mismatch index built from presence codes cannot track utilization in the sense the theory requires, and the data that would—message volumes, administrative workloads, backlogs, the reach rather than the existence of institutions—are not in any databank we know of. The one historical case in which such quantities have been reconstructed, Sng's late-imperial China, found the effect of distance on fiscal reach that the framework predicts. A credible next test would need either more cases of that kind, or a source of variation in demand or capacity unrelated to local instability, or a historical databank that codes institutional reach and administrative load rather than presence. Until then, the honest summary is that capacity matters, scale matters, and that whether their mismatch matters is a question the available evidence cannot answer.

Three things are offered for reuse: a pre-registration and measurement model for Seshat that treats missing as missing and carries ranges, dates and latent-score uncertainty through to inference; a validated bibliography and novelty matrix for the strands that converge on this hypothesis; and a factorial agent-based testbed in which the mechanisms are measurable. The replication package

reproduces every number in this paper from the raw downloads with one command.

References

Acemoglu, Daron, Simon Johnson, and James A Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 2001, *91* (5), 1369–1401.

AHMED, ALI T. and DAVID STASAVAGE, "Origins of Early Democracy," *American Political Science Review*, 2020, *114* (2), 502–518.

Alberti, Gianmarco, "Modeling Group Size and Scalar Stress by Logistic Regression from an Archaeological Perspective," *PLoS ONE*, 2014, *9* (3), e91510.

Angeletos, George-Marios and Alessandro Pavan, "Efficient Use of Information and Social Value of Information," *Econometrica*, 2007, *75* (4), 1103–1142.

Arrow, Kenneth J., *The Limits of Organization*, New York: W. W. Norton, 1974.

Axtell, Robert L., Joshua M. Epstein, Jeffrey S. Dean, George J. Gumerman, Alan C. Swedlund, Jason Harburger, Shubha Chakravarty, Ross Hammond, Jon Parker, and Miles Parker, "Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley," *Proceedings of the National Academy of Sciences*, 2002, *99* (suppl_3), 7275–7279.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker, "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable," *American Journal of Political Science*, 1998, *42* (4), 1260.

Beheim, Bret, Quentin D. Atkinson, Joseph Bulbulia, Will Gervais, Russell D. Gray, Joseph Henrich, Martin Lang, M. Willis Monroe, Michael Muthukrishna, Ara Norenzayan, Benjamin Grant Purzycki, Azim Shariff, Edward Slingerland, Rachel Spicer, and Aiyana K. Willard, "Treatment of missing data determined conclusions regarding moralizing gods," *Nature*, 2021, *595* (7866), E29–E34.

Beniger, James R., *The Control Revolution: Technological and Economic Origins of the Information Society*, Cambridge, MA: Harvard University Press, 1986.

Besley, Timothy and Torsten Persson, "The Origins of State Capacity: Property Rights, Taxation, and Politics," *American Economic Review*, 2009, *99* (4), 1218–1244.

_ and _, *Pillars of Prosperity: The Political Economics of Development Clusters*, Princeton University Press, 2011.

Bettencourt, Luís M. A., "The Origins of Scaling in Cities," *Science*, 2013, *340* (6139), 1438–1441.

_ , José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B. West, "Growth, innovation, scaling, and the pace of life in cities," *Proceedings of the National Academy of Sciences*, 2007, *104* (17), 7301–7306.

Blanton, Richard and Lane Fargher, *Collective Action in the Formation of Pre-Modern States*, Springer New York, 2008.

BLAYDES, LISA and ERIC CHANEY, "The Feudal Revolution and Europe's Rise: Political Divergence of the Christian West and the Muslim World before 1500 CE," *American Political Science Review*, 2013, *107* (1), 16–34.

Bollen, Kenneth A., "Structural Equation Models with Observed Variables," 1989.

Bolt, Jutta and Jan Luiten van Zanden, "Maddison-style estimates of the evolution of the world economy: A new 2023 update," *Journal of Economic Surveys*, 2024, *39* (2), 631–671.

Bolton, P. and M. Dewatripont, "The Firm as a Communication Network," *The Quarterly Journal of Economics*, 1994, *109* (4), 809–839.

Borcan, Oana, Ola Olsson, and Louis Putterman, "State history and economic development: evidence from six millennia," *Journal of Economic Growth*, 2017, *23* (1), 1–40.

Box-Steffensmeier, Janet M. and Bradford S. Jones, *Event History Modeling: A Guide for Social Scientists*, Cambridge University Press, 2004.

Boyd, Robert and Peter J. Richerson, *Culture and the Evolutionary Process*, Chicago: University of Chicago Press, 1985.

_ and _ , "The evolution of reciprocity in sizable groups," *Journal of Theoretical Biology*, 1988, *132* (3), 337–356.

Brambor, Thomas, Agustín Goenaga, Johannes Lindvall, and Jan Teorell, "The Lay of the Land: Information Capacity and the Modern State," *Comparative Political Studies*, 2019, *53* (2), 175–213.

Butzer, Karl W., "Collapse, environment, and society," *Proceedings of the National Academy of Sciences*, 2012, *109* (10), 3632–3639.

Calvó-Armengol, Antoni, Joan de Martí, and Andrea Prat, "Communication and influence: Communication and influence," *Theoretical Economics*, 2015, *10* (2), 649–690.

Cameron, A. Colin and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

Carneiro, Robert L., "On the Relationship between Size of Population and Complexity of Social Organization," *Southwestern Journal of Anthropology*, 1967, *23* (3), 234–243.

Carter, David B. and Curtis S. Signorino, "Back to the Future: Modeling Time Dependence in Binary Data," *Political Analysis*, 2010, *18* (3), 271–292.

CEDERMAN, LARS-ERIK, "Modeling the Size of Wars: From Billiard Balls to Sandpiles," *American Political Science Review*, 2003, *97* (1), 135–150.

Chandler, Alfred D., *The Visible Hand: The Managerial Revolution in American Business*, Cambridge, MA: Harvard University Press, 1977.

Chwe, Michael Suk-Young, "Communication and Coordination in Social Networks," *Review of Economic Studies*, 2000, *67* (1), 1–16.

_ , *Rational Ritual: Culture, Coordination, and Common Knowledge*, Princeton: Princeton University Press, 2001.

Cinelli, Carlos and Chad Hazlett, "Making Sense of Sensitivity: Extending Omitted Variable Bias," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2019, *82* (1), 39–67.

Cline, Eric H., *1177 B.C.: The Year Civilization Collapsed: The Year Civilization Collapsed*, Princeton University Press, 2014.

Coase, R. H., "The Nature of the Firm," *Economica*, 1937, *4* (16), 386–405.

Conley, T.G., "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Coppedge, Michael, John Gerring, Adam Glynn, Carl Henrik Knutsen, Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Fernando Bizzarro, Joshua Krusell, Matthew Maguire, Kyle Marquardt, Kelly McCann, Valeriya Mechkova, Farhad Miri, Josefine Pernes, Jeffrey Staton, Natalia Stepanova, Eitan Tzelgov, and Yi ting Wang, *Varieties of Democracy: Measuring Two Centuries of Political Change*, Cambridge University Press, 2020.

_ , _ , **Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell et al.**, "V-Dem Country-Year Dataset v16," Varieties of Democracy (V-Dem) Project, https://www.v-dem.net/ 2026. Obtained from the official vdemdata R package (commit f4dd26922e, 2026-03-17); accessed 2026-11-01.

Cover, Thomas M. and Joy A. Thomas, *Elements of Information Theory*, Wiley, 2005.

Cox, D. R., "Regression Models and Life-Tables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1972, *34* (2), 187–202.

Crawford, Vincent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), 1431.

Cremer, J., L. Garicano, and A. Prat, "Language and the Theory of the Firm," *The Quarterly Journal of Economics*, 2007, *122* (1), 373–407.

Currie, Thomas E., Simon J. Greenhill, Russell D. Gray, Toshikazu Hasegawa, and Ruth Mace, "Rise and fall of political complexity in island South-East Asia and the Pacific," *Nature*, 2010, *467* (7317), 801–804.

Daniels, Bryan C., David C. Krakauer, and Jessica C. Flack, "Control of finite critical behaviour in a small-scale social system," *Nature Communications*, 2017, *8* (1).

Dedeo, Simon, "Major Transitions in Political Order," in "From Matter to Life," Cambridge University Press, 2017, pp. 393–428.

Dessein, Wouter, "Authority and Communication in Organizations," *Review of Economic Studies*, 2002, *69* (4), 811–838.

DEWAN, TORUN and DAVID P. MYATT, "The Qualities of Leadership: Direction, Communication, and Obfuscation," *American Political Science Review*, 2008, *102* (3), 351–368.

Dewatripont, Mathias and Jean Tirole, "Modes of Communication," *Journal of Political Economy*, 2005, *113* (6), 1217–1238.

Dincecco, Mark, *State Capacity and Economic Development: Present and Past*, Cambridge University Press, 2017.

_ **and Yuhua Wang**, "Violent Conflict and Political Development Over the Long Run: China Versus Europe," *Annual Review of Political Science*, 2018, *21* (1), 341–358.

Dow, Malcolm M., "Galton's Problem as Multiple Network Autocorrelation Effects: Cultural Trait Transmission and Ecological Constraint," *Cross-Cultural Research*, 2007, *41* (4), 336–363.

Dunbar, R. I. M., "Coevolution of neocortical size, group size and language in humans," *Behavioral and Brain Sciences*, 1993, *16* (4), 681–694.

D'Arcy, Michelle and Marina Nistotskaya, "State First, Then Democracy: Using Cadastral Records to Explain Governmental Performance in Public Goods Provision," *Governance*, 2016, *30* (2), 193–209.

Eisenstadt, Shmuel N., *The Political Systems of Empires*, New York: Free Press, 1963.

Ellison, Glenn, "Learning, Local Interaction, and Coordination," *Econometrica*, 1993, *61* (5), 1047.

Epstein, Joshua M. and Robert L. Axtell, *Growing Artificial Societies: Social Science from the Bottom Up*, The MIT Press, 1996.

FEARON, JAMES D. and DAVID D. LAITIN, "Ethnicity, Insurgency, and Civil War," *American Political Science Review*, 2003, *97* (01), 75–90.

Feinman, Gary M., "Size, Complexity, and Organizational Variation: A Comparative Approach," *Cross-Cultural Research*, 2010, *45* (1), 37–58.

Fernández-Villaverde, Jesús, Mark Koyama, Youhong Lin, and Tuan-Hwee Sng, "The Fractured-Land Hypothesis," *The Quarterly Journal of Economics*, 2023, *138* (2), 1173–1231.

Flack, Jessica C., "Multiple time-scales and the developmental dynamics of social systems," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2012, *367* (1597), 1802–1810.

_ , "Coarse-graining as a downward causation mechanism," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2017, *375* (2109), 20160338.

Flannery, Kent V., "The Cultural Evolution of Civilizations," *Annual Review of Ecology and Systematics*, 1972, *3* (1), 399–426.

François, Pieter, J.G. Manning, Harvey Whitehouse, Rob Brennan, Thomas Currie, Kevin Feeney, and Peter Turchin, "A Macroscope for Global History: Seshat Global History Databank, a methodological overview," *Digital Humanities Quarterly*, 2016, *10* (4).

Fukuyama, Francis, *The Origins of Political Order: From Prehuman Times to the French Revolution*, New York: Farrar, Straus and Giroux, 2011.

Galeotti, Andrea, Christian Ghiglino, and Francesco Squintani, "Strategic information transmission networks," *Journal of Economic Theory*, 2013, *148* (5), 1751–1769.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 2000, *108* (5), 874–904.

Gavrilets, Sergey, Edgar A. Duenez-Guzman, and Michael D. Vose, "Dynamics of Alliance Formation and the Egalitarian Revolution," *PLoS ONE*, 2008, *3* (10), e3293.

Gelman, Andrew and Eric Loken, "The Statistical Crisis in Science," *American Scientist*, 2014, *102* (6), 460.

Gennaioli, Nicola and Hans-Joachim Voth, "State Capacity and Military Conflict," *The Review of Economic Studies*, 2015, *82* (4), 1409–1448.

Goldstone, Jack A., *Revolution and Rebellion in the Early Modern World*, Berkeley: University of California Press, 1991.

_ , Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward, "A Global Model for Forecasting Political Instability," *American Journal of Political Science*, 2009, *54* (1), 190–208.

Graeber, David and David Wengrow, *The Dawn of Everything: A New History of Humanity*, New York: Farrar, Straus and Giroux, 2021.

Greif, Avner, "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition," *American Economic Review*, 1993, *83* (3), 525–548.

_ , *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*, Cambridge University Press, 2006.

HAGENBACH, JEANNE and FRÉDÉRIC KOESSLER, "Strategic Communication Networks," *Review of Economic Studies*, 2009, *77* (3), 1072–1099.

Hamilton, Marcus J, Bruce T Milne, Robert S Walker, Oskar Burger, and James H Brown, "The complex structure of hunter–gatherer social networks," *Proceedings of the Royal Society B: Biological Sciences*, 2007, *274* (1622), 2195–2203.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank, "Glottolog 5.3," Max Planck Institute for Evolutionary Anthropology, https://glottolog.org 2026. CLDF languages table, accessed 2026-11-01.

Hanson, Jonathan K. and Rachel Sigman, "Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research," *The Journal of Politics*, 2021, *83* (4), 1495–1510.

Hayek, Friedrich A., "The Use of Knowledge in Society," *American Economic Review*, 1945, *35* (4), 519–530.

Hegre, Håvard, "Toward a Democratic Civil Peace? Democracy, Political Change, and Civil War, 1816–1992," *American Political Science Review*, 2001, *95* (1), 33–48.

Henrich, Joseph, "Cultural group selection, coevolutionary processes and large-scale cooperation," *Journal of Economic Behavior & Organization*, 2004, *53* (1), 3–35.

Hidalgo, César, *Why Information Grows: The Evolution of Order, from Atoms to Economies*, New York: Basic Books, 2015.

HOFFMAN, PHILIP T., *Why Did Europe Conquer the World?*, Princeton University Press, 2015.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant, *Applied Logistic Regression*, Wiley, 2013.

Hoyer, Daniel, James S. Bennett, Jenny Reddish, Samantha Holder, Robert Howard, Majid Benam, Jill Levine, Francis Ludlow, Gary Feinman, and Peter Turchin, "Navigating polycrisis: long-run socio-cultural factors shape response to changing climate," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2023, *378* (1889).

_ , Samantha Holder, James S. Bennett, Pieter François, Harvey Whitehouse, Alan Covey, Gary Feinman, Andrey Korotayev, Vadim Vustiuzhanin, Johannes Preiser-Kapeller, Kathryn Bard, Jenny Levine, Jennifer Reddish, Georg Orlandi, Rachel Ainsworth, and Peter Turchin, "All Crises are Unhappy in their Own Way: The Role of Societal Instability in Shaping the Past," 2024. SocArXiv preprint, DOI 10.31235/osf.io/rk4gd; no journal version found as of November 2026.

Innis, Harold A., *Empire and Communications*, Oxford: Oxford University Press, 1950.

Jackson, Matthew O., *Social and Economic Networks*, Princeton University Press, 2008.

Johnson, Gregory A., "Information Sources and the Development of Decision-Making Organizations," in Charles L. Redman et al., eds., *Social Archeology: Beyond Subsistence and Dating*, New York: Academic Press, 1978, pp. 87–112.

_ , "Organizational Structure and Scalar Stress," in Colin Renfrew, Michael J. Rowlands, and Barbara Abbott Segraves, eds., *Theory and Explanation in Archaeology: The Southampton Conference*, New York: Academic Press, 1982, pp. 389–421.

Johnson, Noel D. and Mark Koyama, "States and economic growth: Capacity and constraints," *Explorations in Economic History*, 2017, *64*, 1–20.

Kandori, Michihiro, George J. Mailath, and Rafael Rob, "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 1993, *61* (1), 29.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale, and Michael C. Gavin, "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity," *PLOS ONE*, 2016, *11* (7), e0158391.

Ko, Chiu Yu, Mark Koyama, and Tuan-Hwee Sng, "UNIFIED CHINA AND DIVIDED EUROPE," *International Economic Review*, 2018, *59* (1), 285–327.

Kohler, Timothy A. and Michael E. Smith, eds, *Ten Thousand Years of Inequality: The Archaeology of Wealth Differences*, Tucson: University of Arizona Press, 2018.

_ , _ , Amy Bogaard, Gary M. Feinman, Christian E. Peterson, Alleen Betzenhauser, Matthew Pailes, Elizabeth C. Stone, Anna Marie Prentiss, Timothy J. Dennehy, Laura J. Ellyson, Linda M. Nicholas, Ronald K. Faulseit, Amy Styring, Jade Whitlam, Mattia Fochesato, Thomas A. Foor, and Samuel Bowles, "Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica," *Nature*, 2017, *551* (7682), 619–622.

Krakauer, David, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nihat Ay, "The information theory of individuality," *Theory in Biosciences*, 2020, *139* (2), 209–223.

Kushilevitz, Eyal, "Communication Complexity," in "Advances in Computers," Elsevier, 1997, pp. 331–360.

Lee, Melissa M. and Nan Zhang, "Legibility and the Informational Foundations of State Capacity," *The Journal of Politics*, 2017, *79* (1), 118–132.

Lewis, David K., *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press, 1969.

Lindley, D. V., "The theory of queues with a single server," *Mathematical Proceedings of the Cambridge Philosophical Society*, 1952, *48* (2), 277–289.

Lobo, Jose, Luis MA Bettencourt, Michael E Smith, and Scott Ortman, "Settlement scaling theory: Bridging the study of ancient and contemporary urban systems," *Urban Studies*, 2019, *57* (4), 731–747.

Loynes, R. M., "The stability of a queue with non-independent inter-arrival and service times," *Mathematical Proceedings of the Cambridge Philosophical Society*, 1962, *58* (3), 497–520.

Mace, Ruth, Mark Pagel, John R. Bowen, Keith F. Otterbein, Mark Ridley, Thomas Schweizer, and Eckart Voland, "The Comparative Method in Anthropology [and Comments and Reply]," *Current Anthropology*, 1994, *35* (5), 549–564.

Macy, Michael W. and Robert Willer, "From Factors to Actors: Computational Sociology and Agent-Based Modeling," *Annual Review of Sociology*, 2002, *28* (1), 143–166.

Mann, Michael, "The autonomous power of the state: its origins, mechanisms and results," *European Journal of Sociology*, 1984, *25* (2), 185–213.

_ , *The Sources of Social Power*, Cambridge University Press, 1986.

Marschak, Jacob and Roy Radner, *Economic Theory of Teams*, New Haven: Yale University Press, 1972.

Marzen, Sarah E. and Simon DeDeo, "The evolution of lossy compression," *Journal of The Royal Society Interface*, 2017, *14* (130), 20170166.

Mayshar, Joram, Omer Moav, and Luigi Pascali, "The Origin of the State: Land Productivity or Appropriability?," *Journal of Political Economy*, 2022, *130* (4), 1091–1144.

McNeill, William H., *The Pursuit of Power: Technology, Armed Force, and Society since A.D. 1000*, University of Chicago Press, 1982.

Middleton, Guy D., *Understanding Collapse: Ancient History and Modern Myths*, Cambridge University Press, 2017.

Milgrom, Paul R., Douglass C. North, and Barry R. Weingast*, "THE ROLE OF INSTITUTIONS IN THE REVIVAL OF TRADE: THE LAW MERCHANT, PRIVATE JUDGES, AND THE CHAMPAGNE FAIRS," *Economics & Politics*, 1990, *2* (1), 1–23.

Miranda, Lux and Jacob Freeman, "The two types of society: Computationally revealing recurrent social formations and their evolutionary trajectories," *PLOS ONE*, 2020, *15* (5), e0232609.

Morris, Stephen, "Contagion," *Review of Economic Studies*, 2000, *67* (1), 57–78.

_ **and Hyun Song Shin**, "Social Value of Public Information," *American Economic Review*, 2002, *92* (5), 1521–1534.

Mount, Kenneth and Stanley Reiter, "The informational size of message spaces," *Journal of Economic Theory*, 1974, *8* (2), 161–192.

Murdock, George Peter, "Ethnographic Atlas: A Summary," *Ethnology*, 1967, *6* (2), 109.

NAROLL, RAOUL, "A Preliminary Index of Social Development," *American Anthropologist*, 1956, *58* (4), 687–715.

Nisan, Noam and Ilya Segal, "The communication requirements of efficient allocations and supporting prices," *Journal of Economic Theory*, 2006, *129* (1), 192–224.

North, Douglass C., *Institutions, Institutional Change and Economic Performance*, Cambridge University Press, 1990.

North, Douglass C, "Institutions," *Journal of Economic Perspectives*, 1991, *5* (1), 97–112.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor, "The preregistration revolution," *Proceedings of the National Academy of Sciences*, 2018, *115* (11), 2600–2606.

Olson, Mancur, *The Logic of Collective Action: Public Goods and the Theory of Groups, With a New Preface and Appendix*, Harvard University Press, 1965.

Ortman, Scott G., Andrew H. F. Cabaniss, Jennie O. Sturm, and Luís M. A. Bettencourt, "The Pre-History of Urban Scaling," *PLoS ONE*, 2014, *9* (2), e87902.

Oster, Emily, "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 2017, *37* (2), 187–204.

Ostrom, Elinor, *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press, 1990.

_ , "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," *American Economic Review*, 2010, *100* (3), 641–672.

Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi ting Wang, Juraj Medzihorsky, Joshua Krusell, Farhad Miri, and Johannes von Roemer, "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data," V-Dem Working Paper 21, Varieties of Democracy Institute 2018.

Peregrine, Peter N., "Social Resilience to Climate-Related Disasters in Ancient Societies: A Test of Two Hypotheses," *Weather, Climate, and Society*, 2018, *10* (1), 145–161.

Peregrine, Peter Neal, "Political participation and long-term resilience in pre-Columbian societies," *Disaster Prevention and Management: An International Journal*, 2017, *26* (3), 314–329.

Radner, Roy, "The Organization of Decentralized Information Processing," *Econometrica*, 1993, *61* (5), 1109.

Richerson, Peter, Ryan Baldini, Adrian V. Bell, Kathryn Demps, Karl Frost, Vicken Hillis, Sarah Mathew, Emily K. Newton, Nicole Naar, Lesley Newson, Cody Ross, Paul E. Smaldino, Timothy M. Waring, and Matthew Zefferman, "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence," *Behavioral and Brain Sciences*, 2014, *39*.

Rubinstein, Ariel, "The Electronic Mail Game: Strategic Behavior under "Almost Common Knowledge"," *American Economic Review*, 1989, *79* (3), 385–391.

Samejima, Fumiko, "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometrika*, 1969, *34* (S1), 1–97.

SCHEIDEL, WALTER, *Escape from Rome: The Failure of Empire and the Road to Prosperity*, Princeton University Press, 2019.

Schelling, Thomas C., *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, 1960.

Scott, James C., *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*, New Haven: Yale University Press, 1998.

SCOTT, JAMES C., *Against the Grain: A Deep History of the Earliest States*, Yale University Press, 2017.

Segal, Ilya, "The communication requirements of social choice rules and supporting budget sets," *Journal of Economic Theory*, 2007, *136* (1), 341–378.

Seshat: Global History Databank, "Polaris-2026 data release (repository build_polaris_dataset, commit 55ca5fbc2b)," https://seshatdatabank.info/data/ 2026. Accessed 2026-11-01.

Shannon, C. E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948, *27* (3), 379–423.

Sheehan, Oliver, Joseph Watts, Russell D. Gray, and Quentin D. Atkinson, "Coevolution of landesque capital intensive agriculture and sociopolitical hierarchy," *Proceedings of the National Academy of Sciences*, 2018, *115* (14), 3628–3633.

Shin, Jaeweon, Michael Holton Price, David H. Wolpert, Hajime Shimao, Brendan Tracey, and Timothy A. Kohler, "Scale and information-processing thresholds in Holocene social evolution," *Nature Communications*, 2020, *11* (1).

Simon, Herbert A., "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics*, 1955, *69* (1), 99.

_ , "The Architecture of Complexity," *Proceedings of the American Philosophical Society*, 1962, *106* (6), 467–482.

Skyrms, Brian, *Signals: Evolution, Learning, and Information*, Oxford University PressOxford, 2010.

Slingerland, Edward, M. Willis Monroe, Brenton Sullivan, Robyn Faith Walsh, Daniel Veidlinger, Ben Alderson-Day et al., "Historians Respond to Whitehouse et al. (2019), "Complex Societies Precede Moralizing Gods Throughout World History"," 2020. PsyArXiv preprint, DOI 10.31234/osf.io/2amjz; published in Journal of Cognitive Historiography.

Sng, Tuan-Hwee, "Size and dynastic decline: The principal-agent problem in late imperial China, 1700–1850," *Explorations in Economic History*, 2014, *54*, 107–127.

Soifer, Hillel, "State Infrastructural Power: Approaches to Conceptualization and Measurement," *Studies in Comparative International Development*, 2008, *43* (3-4), 231–251.

Stanish, Charles, *The Evolution of Human Co-operation: Ritual and Social Complexity in Stateless Societies*, Cambridge University Press, 2017.

STASAVAGE, DAVID, "When Distance Mattered: Geographic Scale and the Development of European Representative Assemblies," *American Political Science Review*, 2010, *104* (4), 625–643.

Stasavage, David, *The Decline and Rise of Democracy: A Global History from Antiquity to Today*, Princeton University Press, 2020.

Tainter, Joseph A., *The Collapse of Complex Societies*, Cambridge: Cambridge University Press, 1988.

Tilly, Charles, *Coercion, Capital, and European States, AD 990–1992*, Oxford: Blackwell, 1992.

Treier, Shawn and Simon Jackman, "Democracy as a Latent Variable," *American Journal of Political Science*, 2008, *52* (1), 201–217.

Trigger, Bruce G., *Understanding Early Civilizations: A Comparative Study*, Cambridge University Press, 2003.

Turchin, Peter, *Historical Dynamics: Why States Rise and Fall: Why States Rise and Fall*, Princeton University Press, 2004.

_ , *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth*, Chaplin, CT: Beresta Books, 2016.

_ , *End Times: Elites, Counter-Elites, and the Path of Political Disintegration*, New York: Penguin Press, 2023.

_ **and Sergey A. Nefedov**, *Secular Cycles*, Princeton University Press, 2009.

_ **and Sergey Gavrilets**, "Evolution of Complex Hierarchical Societies," *Social Evolution & History*, 2009, *8* (2), 167–198.

_ **, Harvey Whitehouse, Sergey Gavrilets, Daniel Hoyer, Pieter François, James S. Bennett, Kevin C. Feeney, Peter Peregrine, Gary Feinman, Andrey Korotayev, Nikolay Kradin, Jill Levine, Jenny Reddish, Enrico Cioni, Romain Wacziarg, Gavin Mendel-Gleason, and Majid Benam**, "Disentangling the evolutionary drivers of social complexity: A comprehensive test of hypotheses," *Science Advances*, 2022, *8* (25).

_ **, Rob Brennan, Thomas Currie, Kevin Feeney, Pieter Francois, Daniel Hoyer, Joseph Manning, Arkadiusz Marciniak, Daniel Mullins, Alessio Palmisano, Peter Peregrine, Edward A.L. Turner, and Harvey Whitehouse**, "Seshat: The Global History Databank," *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, 2015, *6* (1).

_ **, Thomas E. Currie, Harvey Whitehouse, Pieter François, Kevin Feeney, Daniel Mullins, Daniel Hoyer, Christina Collins, Stephanie Grohmann, Patrick Savage, Gavin Mendel-Gleason, Edward Turner, Agathe Dupeyron, Enrico Cioni, Jenny Reddish, Jill Levine, Greine Jordan, Eva Brandl, Alice Williams, Rudolf Cesaretti, Marta Krueger, Alessandro Ceccarelli, Joe Figliulo-Rosswurm, Po-Ju Tuan, Peter Peregrine, Arkadiusz Marciniak, Johannes Preiser-Kapeller, Nikolay Kradin, Andrey Korotayev, Alessio Palmisano, David Baker, Julye Bidmead, Peter Bol, David Christian, Connie Cook, Alan Covey, Gary Feinman, Árni Daníel Júlíusson, Axel Kristinsson, John Miksic, Ruth Mostern, Cameron Petrie, Peter Rudiak-Gould, Barend ter Haar, Vesna Wallace, Victor Mair, Liye Xie, John Baines, Elizabeth Bridges, Joseph Manning, Bruce Lockhart, Amy Bogaard, and Charles Spencer**, "Quantitative historical analysis uncovers a single dimension of complexity that structures global

variation in human social organization," *Proceedings of the National Academy of Sciences*, 2017, *115* (2).

Watts, Duncan J. and Steven H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, 1998, *393* (6684), 440–442.

Weber, Max, *Economy and Society: An Outline of Interpretive Sociology*, Berkeley: University of California Press, 1978. Edited by G. Roth and C. Wittich; first published 1922.

Whitehouse, Harvey, Pieter François, Patrick E. Savage, Daniel Hoyer, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, Jennifer Larson, John Baines, Barend ter Haar, Alan Covey, and Peter Turchin, "Testing the Big Gods hypothesis with global historical data: a review and "retake"," *Religion, Brain & Behavior*, 2022, *13* (2), 124–166.

_ , _ , _ , Thomas E. Currie, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, Robert M. Ross, Jennifer Larson, John Baines, Barend ter Haar, Alan Covey, and Peter Turchin, "RETRACTED ARTICLE: Complex societies precede moralizing gods throughout world history," *Nature*, 2019, *568* (7751), 226–229.

_ , _ , _ , _ , _ , _ , _ , _ , _ , _ , _ , _ , _ , and _ , "Retraction Note: Complex societies precede moralizing gods throughout world history," *Nature*, 2021, *595* (7866), 320–320.

Williamson, Oliver E., *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Wolpert, David, Eric Libby, Joshua A. Grochow, and Simon Dedeo, "The Many Faces of State Space Compression," in "From Matter to Life," Cambridge University Press, 2017, pp. 199–243.

WRIGHT, HENRY T. and GREGORY A. JOHNSON, "Population, Exchange, and Early State Formation in Southwestern Iran," *American Anthropologist*, 1975, *77* (2), 267–289.

Yao, Andrew Chi-Chih, "Some complexity questions related to distributive computing(Preliminary Report)," in "Proceedings of the eleventh annual ACM symposium on Theory of computing - STOC '79" STOC '79 ACM Press 1979, pp. 209–213.

Yates, JoAnne, *Control through Communication: The Rise of System in American Management*, Baltimore: Johns Hopkins University Press, 1989.

Yoffee, Norman, *Myths of the Archaic State: Evolution of the Earliest Cities, States, and Civilizations*, Cambridge University Press, 2005.

Young, H. Peyton, "The Evolution of Conventions," *Econometrica*, 1993, *61* (1), 57.

Zandt, Timothy Van, "Real-Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents," *Review of Economic Studies*, 1999, *66* (3), 633–658.

Appendices

A. Pre-registration summary

The analysis plan (`research/preregistration.md`, SHA-256 `ecda8804e17f9317...`) was frozen on 2026-11-01 before any association between a predictor and an outcome was computed. It fixes: the five hypotheses and their falsifiers; the units (polity, society, country-year) and the reasons for them (93.7% of Polaris facts undated; CrisisDB crisis-selected); the disjoint indicator sets for D and K listed in Section 4.3; the outcomes Y_1 (severity share), Y_{1b}, Y_2 (duration), Y_3 (succession), Y_4–Y_6 (modern events) and the composite; the five-year horizon; the models of Section 5; the baselines and the 80% fold rule; the Holm correction over three primary tests; the a-priori power computation and its decision rule; the robustness set; and the rule for narrowing the claim. Seven pre-analysis decisions (D-001 to D-007) and seven post-analysis amendments (A-001 to A-007: fold-internal preprocessing check; corrections and additions made after the adversarial reviews, listed in Section 10 and in the response-to-reviewers document) are recorded in `research/decisions.md`. The single deviation from the mandate's candidate formula (K standardized rather than logged, because it is a bounded share that can be zero) was recorded before analysis. The use of decade rather than year effects in the forward-chaining folds, so that test years have a fitted level, is noted in the code and in Section 5.4.

B. Data dictionary (abridged)

Seshat facts (56,274 rows): sheet, section, region, polity, variable, value (presence codes P/IP $\to$ 1, A/IA $\to$ 0, disputed $\to$ 0.5, SU/blank $\to$ missing; numeric midpoints with bounds), date interval, flags for dated, inferred, disputed, unparsed, uncertain. **Polity aggregates**: mean, min, max, mean log, bound extremes, counts of facts, dated facts, inferred and disputed facts, per polity and variable; a pre-crisis version per CrisisDB record. **Polity metadata**: region, macro-region (ten groups fixed in `configs/regions.yaml`), era block (six), start and end years, midpoint, duration, censoring at 1900, centralization, relationships to preceding and succeeding entities, matched successor and its relationship, discontinuity indicator. **CrisisDB**: record, polity, start and end years, the 21 consequence codes, the number of instability items coded, the number present, the severity share, the binary severe indicator (defined when the five severe items are all coded), the reform

share and the epidemic flag. **Ethnographic Atlas**: society, coordinates, Glottocode and Glottolog family, region, focal year, EA030–EA034, EA042, EA066–EA068, EA070, EA072–EA073, EA090, EA005, EA028, EA113, community-size midpoint, population, jurisdictional levels. **Modern panel**: V-Dem identifiers and year, six capacity items, electoral democracy, regime identifiers and end types, civil war, coups, population and GDP per capita from Maddison (ISO3 merge; 39 unmatched V-Dem entities listed in the crosswalk file), derived event indicators and windows. The full dictionary with counts is `data/metadata/data_dictionary.md`.

C. Additional measurement detail

Item parameters of the two-parameter model are listed in `results/empirical/measurement_irt_item_parameters.csv`. Five record-keeping items (written record, script, lists and tables, calendar, practical literature) reach the discrimination ceiling because they are coded almost jointly in Seshat; the latent score is therefore dominated by the writing cluster, which is why the transparent block mean is the primary index. Figure 9 compares item difficulties estimated on the Old World and on the Americas and Oceania. Figure 10 shows the mismatch index over time by macro-region and the raw relation between pre-crisis mismatch and crisis severity.

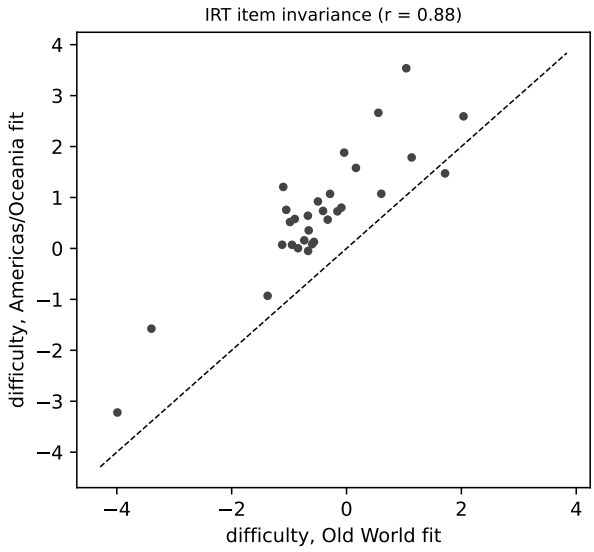


Figure 9. Item difficulties from two-parameter models fitted separately on Old-World polities and on the Americas and Oceania.

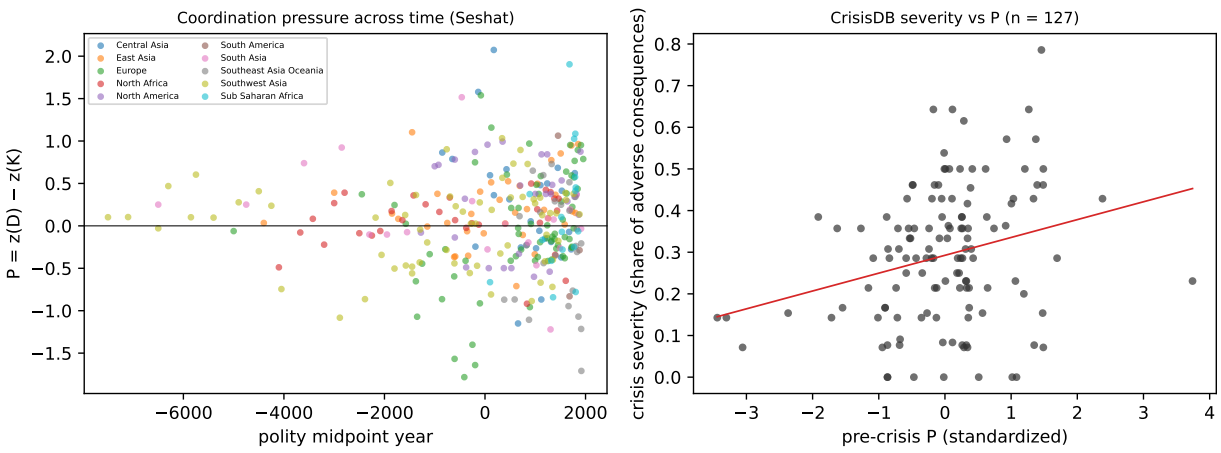


Figure 10. Left: mismatch index P against polity midpoint year, by macro-region. Right: crisis severity against pre-crisis P for the 127 CrisisDB records with a defined index, with the least-squares line.

D. Simulation scaling table

Table 27. [SIMULATED] Means by regime and N over all main-design cells without shocks (administrative capacity ratio 2).

Regime	N	Success	Msgs/task	Bits/task	Bits/success (set-up)	Welfare/agent	Gini
central	50	0.932	8.00	50.1	68.0	43.5	0.149
central	200	0.932	12.00	73.1	93.7	39.4	0.158
central	800	0.932	16.00	96.3	118.9	38.6	0.160
federated	50	0.934	6.81	43.2	60.5	43.9	0.147
federated	200	0.936	8.06	50.4	69.4	40.3	0.157
federated	800	0.936	9.34	57.9	78.0	39.9	0.158
none	50	0.520	4.00	14.0	62.4	30.7	0.224
none	200	0.524	4.00	14.0	59.1	31.0	0.224
none	800	0.525	4.00	14.0	58.1	30.9	0.222
norms	50	0.552	4.00	13.5	98.4	32.3	0.222
norms	200	0.555	4.00	13.5	93.1	32.5	0.220
norms	800	0.556	4.00	13.5	92.0	32.3	0.219
records	50	0.957	2.00	15.5	26.8	56.5	0.133
records	200	0.957	2.00	15.5	26.8	56.6	0.134
records	800	0.958	2.00	15.5	26.8	56.5	0.134

E. Replication guide

The package is a Python 3.12 project with pinned dependencies. From a clean clone:

```
uv venv -p 3.12 .venv && uv pip install -p .venv/bin/python -e .
make fetch     # 9 raw sources, SHA-256 verified
make data      # three panels + diagnostics
make measure   # D, K, P, IRT, invariance, power
make econ      # H1, H2, H4, H5, robustness
make sim       # 6,048 + 584 runs and analysis
make figures   # all tables and figures
make paper     # latexmk -> paper/main.pdf
make test      # unit and integration tests
```

Seeds are fixed (20261101 for the estimation code, 0–7 for the simulations); raw files are not redistributed because of the D-PLACE and V-Dem licences, but the registry allows byte-level verification of a fresh download. The execution log (results/logs/execution_log.md) records every command and every correction made during the study, and the final audit (results/logs/final_research_audit.md) lists completed, failed and blocked tasks.