

Beyond Pairwise Polarization: A Statistical Null Model for Structural Balance in Signed Hypergraphs

Md Hasibuzzaman

hsbuzzaman@gmail.com

National Taipei University

Chan-Yun Yang

National Taipei University

Gene Eu Jan

Asia University

Research Article

Keywords: structural balance, signed hypergraphs, higher-order networks, null models, polarization, coalition formation, frustration index

Posted Date: September 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10685379/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Beyond Pairwise Polarization: A Statistical Null Model for Structural Balance in Signed Hypergraphs

Md Hasibuzzaman^{1*}, Chan-Yun Yang¹ and Gene Eu Jan²

^{1*}Electrical Engineering and Computer Science, National Taipei University, No. 151, University Rd., Sanxia Dist., New Taipei City, 237303, Taiwan.

²Department of Computer Science and Information Engineering, Asia University, 500, Lioufeng Rd., Wufeng, Taichung, 41354, Taiwan.

*Corresponding author(s). E-mail(s): hsbuzzaman@gmail.com;

Abstract

Structural balance theory explains polarization through relationships organizing into cohesive, antagonistic factions. Existing statistical tests for the theory rest on pairwise signed graphs, even though many systems it targets, legislative blocs, alliances, coalitions, are group-level phenomena. We introduce a statistical framework for testing structural balance directly on two-sided signed hypergraphs, built from raw group-interaction data. We define the Hyperedge Frustration Index, prove a corrected structure theorem for exact higher-order balance, and construct a degree-and-size-preserving null model for significance testing, correcting two implementation errors along the way. Applied to seventy-five years of UN General Assembly votes, five of fifteen five-year windows survive Benjamini-Hochberg correction at $B = 500$ replicates: 1946 to 1950, the early Cold War Soviet bloc; 1981 to 1990, two windows of United States isolation; and 2001 to 2010, two windows of a United States-Israel-Pacific bloc. The remaining ten windows, including two decades of decolonization-era voting, show no signal distinguishable from chance. This is our central finding: higher-order balance is real, statistically established across five coherent windows, and genuinely absent elsewhere. Comparing our measure to the standard pairwise pipeline on the same data shows the two track overall trend, Spearman $\rho = 0.68$, but diverge sharply in substance, agreeing in composition only in 1946 to 1950. A secondary validation on congressional cosponsorship data recovers known party homophily, more strongly once mixing is corrected. The results establish both a validated null-model methodology and

a genuine empirical pattern: real historical coalitions detected with statistical rigor, not asserted from optimization alone.

Keywords: structural balance, signed hypergraphs, higher-order networks, null models, polarization, coalition formation, frustration index

1 Introduction

Structural balance theory begins with a simple psychological premise, formalized by Heider [1] and translated into graph language by Cartwright and Harary [2]: relationships that can be positive or negative tend to organize so that “the friend of my friend is my friend” and “the enemy of my enemy is my friend.” A signed network is *balanced* if it can be partitioned into factions such that all relationships within a faction are positive and all relationships between factions are negative. Harary’s structure theorem [3] makes this precise: a signed graph is balanced if and only if its vertices can be split into two (or, under the weaker Davis criterion [4], several) mutually antagonistic groups. Seventy years on, the theory remains one of the few pieces of social-network science with genuine predictive content, and the last decade has substantially sharpened its empirical footing: null-model tests that separate true triadic balance from degree-driven artifacts [5], spectral and local-search algorithms that compute global balance at the scale of hundreds of thousands of nodes [6], and multiscale semiwalk approximations that make balance computable on large weighted, directed networks [7].

Every one of these advances, however, is built on *pairwise* signed graphs. This is a strange place for the theory to have stopped, because many of the systems it is applied to are not pairwise at all. A legislative vote does not produce a set of independent dyadic relationships between legislators; it produces two (or more) simultaneous, mutually exclusive coalitions, each an act of joint, group-level agreement. A multilateral treaty is not a collection of bilateral alliances; it is a single higher-order commitment among all signatories at once. The standard practice – infer a signed edge between every pair of agents from how often they land on the same side, then run balance analysis on the resulting graph [8–10] – discards the very thing that makes the original event a single, coherent act of coalition formation: the fact that all k participants moved together, once, in one event, rather than as $\binom{k}{2}$ independent coin flips that happened to agree.

Higher-order network science has spent the last several years demonstrating, across many other problems, that this kind of pairwise collapse is not innocuous [11–13]. Null models built directly on hypergraphs recover structure – and reject false structure – that pairwise projections cannot [14]. Yet structural balance itself has only just begun to make contact with this literature, and from an unexpected direction: Cui et al. [15] recently generalized Altafini’s bipartite-consensus dynamics [16] to signed hypergraphs, showing that groupwise cooperative/antagonistic interactions can still produce polarized consensus. That paper is a genuine advance, but it is a control-theoretic one: it establishes what *can happen* under signed-hypergraph dynamics in

numerical examples, not whether any *real* group-interaction system shows statistically significant higher-order balance, and it offers no combinatorial structure theorem, no frustration measure, and no null model. The hypergraph null-model literature that could supply the missing statistical machinery [14, 17] has, in turn, never been extended to signed structures or connected to balance theory.

This leaves a clean, specific gap, and this paper works to fill it. We define structural balance and a frustration index natively on signed hypergraphs constructed directly from raw group-interaction data (Section 3), prove a structure theorem characterizing exact balance, and construct a degree-and-size-preserving hyperedge-swap null model that makes higher-order balance a testable statistical hypothesis rather than only a descriptive one (Section 4). Getting this null model right took two rounds of correction that we document in full rather than smoothing over: an initial implementation could generate impossible pseudo-events (a country on both sides of one vote simultaneously), and even after fixing that, the Markov chain needed far more mixing than a standard swap-count heuristic indicated. We apply the corrected machinery to 75 years of United Nations General Assembly (UNGA) roll-call voting data and compare it, on the same raw data, against the standard pairwise pipeline (Section 5). We validate the null-model machinery itself against known ground truth using US congressional cosponsorship data. The central empirical result, once every correction is in place, is a null one: no window shows higher-order balance clearly distinguishable from chance after correcting for the fifteen tests performed. We report this plainly, because we think a methodology that is validated carefully enough to report a null result honestly is more useful to the field than one that reports a positive result obtained before that validation was complete.

Contributions

- **A statistical, null-model-based test for higher-order structural balance** (Section 4.3) – a signed hyperedge-swap configuration model that, to our knowledge, is the first null model formulated specifically for testing statistical significance of balance in signed hypergraphs – together with a spectral-relaxation-plus-local-search solver validated against exact brute-force search on real small instances. We present this, rather than the exact-balance theorem below or any single empirical finding, as the paper’s central contribution.
- **A fully transparent account of two substantive corrections to the null model**, both discovered during preparation of this manuscript and both reported here in full rather than silently fixed: a disjointness violation (an early version could place the same agent on both sides of one event, which we found affected 90% of events in a real test) and an inadequate-mixing problem (a naive swap-count heuristic under-mixed the chain by roughly two orders of magnitude, which we detected by tracking the null mean as a function of swap count until it plateaued). We believe the diagnostic procedure itself – checking a summary statistic’s stability as swap count increases, rather than assuming a fixed multiplier of hyperedge count is sufficient – is a useful, transferable lesson for any hyperedge-swap null model, not specific to this application.

- A formal definition of balance and the Hyperedge Frustration Index (HFI) for signed hypergraphs built directly from group-interaction events, together with a corrected structure theorem (Theorem 1) characterizing exact balance via a must-link/cannot-link constraint contraction. We report, as part of this contribution, a counterexample disproving a simpler but incorrect version of the theorem that appeared in an earlier draft of this work.
- An empirical application to 75 years of UNGA voting (200 countries, 6,203 roll calls): with the fully corrected null model, extended symmetrically to $B = 500$ across every window that first appeared as a candidate, five windows (1946–1950, 1981–1985, 1986–1990, 2001–2005, 2006–2010) survive false-discovery-rate correction across the fifteen tested. The remaining ten, including two full decades of decolonization-era voting, show no signal distinguishable from chance. We report the full pattern – five established findings and ten genuine nulls – rather than a single headline number.
- A direct, same-data comparison against the classical pairwise pipeline, independent of the null-model analysis above, showing that pairwise collapse can change not just the degree but the substance of the polarization measured: bloc composition overlap between the two methods is close to zero except in the most extreme historical case, with sensitivity checks across multiple significance thresholds.
- A secondary validation on congressional cosponsorship data confirming the null-model machinery recovers known ground-truth political homophily – and, notably, recovers it *more* strongly once the same mixing-time correction is applied, which we take as evidence the correction is a genuine methodological fix rather than a source of conservative bias.

2 Related Work

2.1 Structural balance: from Heider to statistical testing

Heider’s original formulation [1] was a claim about individual cognition; Cartwright and Harary [2] and Harary [3] turned it into graph theory, and Davis [4] relaxed the two-faction requirement to the more empirically common k -faction (“clusterable”) case. For decades the theory was tested almost entirely at the level of individual signed triads, and largely through community-detection heuristics that reward internally positive, externally negative partitions [18, 19], using signed datasets such as Epinions, Slashdot and Wikipedia elections [20] that later balance studies, including our own baseline, continue to draw on. The shift to a genuinely statistical footing is recent: Facchetti, Iacono and Altafini [6] showed that global balance (rather than triad counting) is computationally tractable at scale via spectral relaxation and local search, applying it to the same signed datasets and finding them far more balanced than degree-matched null models would predict. Gallo et al. [5] showed that the null model matters enormously: homogeneous benchmarks favor weak balance, but once node-level heterogeneity in positive/negative tie propensity is controlled for, the evidence shifts toward strong balance in some domains and disappears in others (notably, biological networks); Hao and Kovács [21] independently reached a related conclusion, showing that naive network randomization schemes can systematically bias balance assessments in either direction depending on how carefully degree heterogeneity is

preserved – a caution that anticipated, in the pairwise setting, exactly the mixing-time failure mode we document for hypergraphs in Section 4.3. Talaga et al. [7] made balance computable at the scale and directionality of modern weighted signed networks via multiscale semiwalk approximations. Kirkley, Cantwell and Newman [22] and Estrada and Benzi [23] proposed alternative walk-based and spectral measures of partial balance, extending the theory beyond a strict balanced/unbalanced dichotomy. Aref and Neal [9, 10] applied optimal signed-graph partitioning specifically to political co-voting networks, the same substantive domain as our primary dataset, and are the closest methodological precedent for our pairwise baseline. All of this work, without exception, operates on dyadic signed graphs.

2.2 Higher-order networks and null models

Hypergraphs and simplicial complexes have become a standard tool for representing genuinely polyadic interactions [11], and a parallel literature has developed null models that respect this structure rather than projecting it away. Chodrow [17] introduced configuration models for hypergraphs that preserve node degree and hyperedge-size sequences via MCMC hyperedge swaps – the direct ancestor of the null model we use, generalized here to the antagonism-aware, signed setting. Chodrow, Veldt and Benson [12] extended blockmodel and modularity-based clustering to hypergraphs, using among other examples the same US congressional cosponsorship data (via Fowler’s original studies [24, 25]) that we use as a secondary validation set. Preti, Fazzone, Petri and De Francisci Morales [14] generalized configuration-model null testing to *directed* hypergraphs (NuDHy) and, notably, already used it to demonstrate genuine higher-order political homophily in US Congress cosponsorship over a 40-year span – a finding we deliberately re-derive with our own purity statistic in Section 4.5 as a check on our null-model machinery before trusting it on our primary, harder-to-validate dataset, rather than presenting it as a new contribution.

2.3 Where balance and higher order have (almost) met

The one place these two lines of work have started to intersect is dynamics, not statistics. Cui, Zhang, Jiang, Jardón-Kojakhmetov and Cao [15] generalize Altafini’s bipartite-consensus model [16] to signed hypergraphs via tensor gauge transformations, showing conditions under which agents on a structurally balanced, strongly connected signed hypergraph converge to bipartite consensus. This establishes that higher-order balance is a coherent dynamical concept, which is valuable, but the paper is explicitly a control-theory contribution: balance is assumed as a hypothesis to study convergence under, not measured, tested, or estimated from data. No combinatorial frustration measure, no structure theorem for the general (non-balanced) case, and no statistical test exists in that work or, to our knowledge, anywhere else in the literature. That is the gap this paper closes.

3 Theoretical Framework

3.1 From events to signed hypergraphs

Consider a group-interaction system in which discrete events (roll-call votes, bill sponsorships, multilateral agreements) each activate a subset of agents $V_e \subseteq V$ and reveal, for each participant, a side. We retain the two largest sides per event (for votes: Yes/No; abstentions are treated as non-participation in that event, consistent with standard practice in the international-relations literature [26]). A side with at least $k_{\min}$ members ($k_{\min} = 3$ throughout, so that a hyperedge reflects genuinely higher-order structure rather than a relabeled dyad) becomes a **positive hyperedge**: a witnessed act of simultaneous, mutual agreement among all its members. Two hyperedges arising from opposite sides of the same event form an **antagonistic pair**.

Figure 1 summarizes the full pipeline. A group-interaction event – a k -party vote or agreement with a positive or negative outcome – becomes a hyperedge, not a clique of pairwise ties: the standard route would collapse a five-member positive event into ten separate dyadic edges, discarding the fact that all five members acted together once; our route keeps the event as a single hyperedge and links it to any opposing hyperedge from the same event through an antagonism relation. A hyperedge is *frustrated* if its members split across the two inferred factions rather than staying together, and the Hyperedge Frustration Index (panel 3) is the fraction of hyperedges that split under the best available bipartition. The null model (panel 4) tests whether this fraction is lower than chance by rewiring memberships while holding each hyperedge’s size and each agent’s participation count fixed, giving a null distribution (panel 5) against which the observed value is scored (panel 6). The schematic in panel 3 shows the simplified split/not-split count for intuition; the exact index used throughout the paper, defined formally in Section 3, additionally weights each hyperedge by how many of its internal pairs are separated, not just whether any split occurred at all.

Definition 1 (Signed hypergraph). *A signed hypergraph is a triple $H = (V, E, A)$ where E is a multiset of hyperedges (each a subset of V with $|e| \geq k_{\min}$) and $A \subseteq E \times E$ is a symmetric antagonism relation linking hyperedges that arose from opposite sides of the same event. Unlike prior signed-hypergraph formulations, in which sign is an intrinsic attribute of a single hyperedge [15], here positivity is intrinsic to each hyperedge (it is, definitionally, a group that acted together) and negativity is relational, scoped to pairs of hyperedges sharing an originating event.*

Definition 2 (Balance). *H is balanced if there exists a bipartition (F_1, F_2) of V such that every hyperedge $e \in E$ lies entirely within F_1 or entirely within F_2 , and for every $(e, e') \in A$, e and e' lie in different factions.*

We use “structural balance” throughout this paper in the precise sense of Definitions 1–2: event-based, two-sided balance, built from events with (at most) two competing sides and tested via strict two-faction bipartition rather than Davis’s weaker multi-faction clusterability [4]. We do not claim this captures every notion of higher-order signed structure one might reasonably call “balance”; Section 8 discusses the multi-way and weak-balance generalizations this leaves open.

Definition 3 (Hyperedge Frustration Index). *For a bipartition $\pi = (F_1, F_2)$ with faction-membership function $f : V \rightarrow \{0, 1\}$, define the split cost of a hyperedge e as*

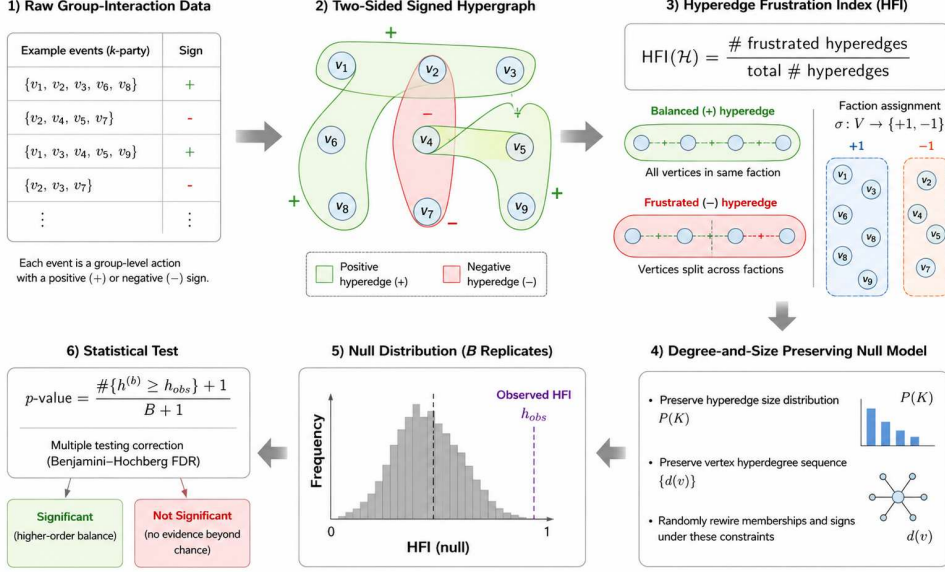


Fig. 1 The full analysis pipeline. (1) Raw group-interaction events, each with a member set and a sign. (2) The two-sided signed hypergraph built from these events, with positive and negative hyperedges linked by antagonism. (3) The Hyperedge Frustration Index, illustrated here in simplified form (Section 3 gives the exact, size-weighted definition). (4) The degree- and size-preserving null model. (5) The resulting null distribution against the observed value. (6) The resulting significance test with multiple-testing correction.

$a(e)b(e)$, where $a(e), b(e)$ are the sizes of $e \cap F_1$ and $e \cap F_2$ (the number of “should-agree-but-don’t” pairs implied by e), and the antagonism cost of a pair $(e, e') \in A$ as $p(e)r(e') + q(e)s(e')$, where p, q (r, s) are the F_1 - and F_2 -sized parts of e (e') – i.e., the number of cross-hyperedge pairs that land in the same faction despite being antagonistic. Then

$$\text{HFI}(H, \pi) = \frac{\sum_{e \in E} a(e)b(e) + \sum_{(e, e') \in A} (p(e)r(e') + q(e)s(e'))}{\sum_{e \in E} \binom{|e|}{2} + \sum_{(e, e') \in A} |e||e'|}, \quad \text{HFI}(H) = \min_{\pi} \text{HFI}(H, \pi).$$

Remark 1. Both terms are necessary. A definition that penalizes only split hyperedges (omitting the antagonism term) is degenerate: the trivial partition that places every agent in a single faction always scores zero, since no hyperedge can ever be split when there is only one faction. We note this explicitly because it is precisely the failure mode we encountered – and corrected – during implementation (Section 4), and record it here as a warning to future work in this direction.

3.1.1 Why the pair-count weighting, and what that choice buys and costs

The weight $\binom{|e|}{2}$ is not an arbitrary normalization constant; it is the count of implied pairwise consistency constraints a k -member hyperedge actually asserts. A hyperedge $e = \{v_1, \dots, v_k\}$ is, by Definition 2, a claim that all $\binom{k}{2}$ pairs within it belong together. Splitting e under a candidate partition violates exactly $a(e)b(e)$ of those pairs and satisfies the remaining $\binom{k}{2} - a(e)b(e)$; the split cost in Definition 3 is precisely the count of violated constraints, and dividing by $\binom{|e|}{2}$ in the denominator makes HFI a properly normalized fraction of implied-pair violations, comparable across hyperedges of different size and across windows with different hyperedge-size distributions. This is, in other words, not a design choice layered on top of the definition of balance – it falls out of counting what a hyperedge of size k actually claims, once one accepts Definition 2’s claim that the hyperedge asserts consistency among *all* its members simultaneously, not merely between adjacent or arbitrarily chosen pairs.

That said, this is a specific, consequential modeling commitment, not the only one consistent with the underlying balance definition, and our own robustness check (Section 5.3) shows it is far from innocuous empirically: it is precisely the $\binom{k}{2}$ growth rate that lets a comparatively small number of large, near-unanimous UNGA resolutions outweigh a much larger number of small, closely-contested ones in determining the optimal partition. An alternative weighting scheme is implicitly a different claim about what a hyperedge asserts – linear weighting $w(k) = k$ treats a hyperedge as asserting k independent facts about its members (for instance, “member i agrees with the group”) rather than $\binom{k}{2}$ pairwise facts, which is a coherent but different reading of what a joint coalition event means. We use $\binom{|e|}{2}$ because we find the pairwise-consistency reading the more natural formalization of Definition 2 as stated, and because it is the reading under which the observed data recovers historically legible structure (Section 5.3) – but we want to be explicit that this second reason is a property of our data and our formalization jointly, not evidence that $\binom{|e|}{2}$ is the uniquely correct choice on general theoretical grounds. A future formalization that instead takes seriously a $w(k) = k$ or $w(k) = \log(1 + k)$ reading of what a hyperedge asserts would arrive at a different, equally principled index, and Section 5.3 is intended precisely to make that dependency visible rather than to argue it away.

This is the direct hypergraph generalization of Harary’s frustration index [3] (minimum number of sign-violating edges): a split hyperedge or same-faction antagonistic pair is the higher-order analogue of a sign-violating edge, weighted by the number of dyadic contradictions it implies.

An earlier version of this theorem characterized $\text{HFI}(H) = 0$ via bipartiteness of the antagonism-only conflict graph G_H (edges wherever two agents co-occur on opposite sides of some event). That characterization is false: G_H records only the inequality (antagonism) constraints and is silent about equality (co-hyperedge) constraints that hold *between* agents whose only direct relationship in G_H may be an unrelated antagonism from a different event. A minimal counterexample makes this concrete. Let $e_1 = \{a, b\}$ and $e_2 = \{c, d\}$ be antagonistic, and let $e_3 = \{a, c\}$ be a third, positive hyperedge (from a separate event). Then $G_H = K_{2,2}$ on $\{a, b\}$ vs. $\{c, d\}$, which is

bipartite – yet e_3 requires a and c to share a faction, while the antagonism between e_1 and e_2 forces every member of e_1 into a different faction from every member of e_2 , including a against c . These constraints are jointly unsatisfiable, so $\text{HFI}(H) > 0$ despite G_H being bipartite: exhaustive search over all bipartitions of $\{a, b, c, d\}$ gives $\min_\pi \text{HFI}(H, \pi) = 1/7$, not 0. We correct the theorem below; the corrected statement, its proof, and the counterexample itself were all verified computationally against exact brute-force search prior to submission.

Definition 4 (Constraint contraction). *Let M (the must-link graph) be the graph on V obtained by adding, for every hyperedge $e \in E$, a clique on the members of e . Let $\mathcal{C} = \{C_1, \dots, C_m\}$ be the connected components of M . Define the quotient conflict graph $\bar{G}_H$ on vertex set $\mathcal{C}$, with an edge $\{C_i, C_j\}$ whenever some $(e, e') \in A$ has a member of e in C_i and a member of e' in C_j (a self-loop at C_i arises if such u, v both lie in the same component C_i).*

Theorem 1 (Corrected structure theorem). *$\text{HFI}(H) = 0$ if and only if $\bar{G}_H$ has no self-loops and is bipartite.*

Proof. ($\Leftarrow$) Suppose $\bar{G}_H$ is loop-free and bipartite; fix a proper 2-coloring of $\bar{G}_H$ and assign every $v \in V$ the color of its component. Every hyperedge e is, by construction of M , entirely contained in one component, hence monochromatic: no hyperedge is split. For any $(e, e') \in A$, every member of e and every member of e' lies in components joined by an edge of $\bar{G}_H$ (or, if they lie in the *same* component, that component would carry a self-loop, contradicting our assumption); properness of the coloring therefore places all of e and all of e' in different colors, so the antagonism cost of (e, e') is zero. Hence $\text{HFI}(H, \pi) = 0$. ($\Rightarrow$) Suppose $\text{HFI}(H) = 0$, witnessed by partition π . Zero split cost means every hyperedge is monochromatic under π ; since M 's edges are exactly the co-hyperedge pairs, π is constant on each connected component of M , inducing a well-defined coloring of $\mathcal{C}$. Zero antagonism cost means that for every $(e, e') \in A$, no member of e shares π 's color with any member of e' ; in particular no such pair can lie in the same component of M (that would force a shared color trivially), so $\bar{G}_H$ has no self-loop at any component touched by A , and every $\bar{G}_H$ -edge joins differently-colored components, i.e., the induced coloring is proper. Hence $\bar{G}_H$ is loop-free and bipartite. $\square$

We validated Theorem 1 computationally against exact brute-force search on the counterexample above (correctly predicts $\text{HFI} > 0$), against a minimal balanced case (correctly predicts $\text{HFI} = 0$), against a three-hyperedge odd antagonism cycle (correctly predicts $\text{HFI} > 0$, recovering the classical odd-cycle criterion [3]), and against a larger consistent multi-hyperedge case (correctly predicts $\text{HFI} = 0$); all four predictions matched exact enumeration exactly.

This corrected theorem generalizes the classical result (a signed graph is balanced iff every cycle has an even number of negative edges [3]) and recovers it exactly when every hyperedge has size 2, in which case M has no edges beyond the trivial ones and $\bar{G}_H \cong G_H$. It also sharpens, rather than weakens, the paper's central methodological claim. The zero-frustration case is governed by a genuinely higher-order object – the must-link contraction of Definition 4, which depends on the full hyperedge structure, not a pairwise-derived graph – and it is precisely the interaction between co-hyperedge

(equality) and antagonism (inequality) constraints, missed entirely by the original, flawed characterization, that makes exact balance a genuinely combinatorial rather than a trivially pairwise question. That said, we agree with an important observation this error surfaced: the paper’s strongest and most defensible contribution is not the exact-balance characterization but the *statistical* theory built on top of it. Real data essentially never sits exactly at $\text{HFI} = 0$; the question that matters empirically is always whether the observed HFI is significantly lower than a properly higher-order null distribution (Section 4.3), not whether $\bar{G}_H$ happens to be exactly bipartite. We have revised the framing throughout this paper accordingly, and present the null-model test, not Theorem 1 alone, as the central methodological claim.

3.2 Complexity and computation

Exact minimization of HFI is NP-hard: the $k_{\min} = 2$ special case is the correlation-clustering / MIN-UNCUT problem. We therefore compute HFI via a three-stage pipeline: (i) a spectral relaxation, building the clique-and-bipartite-expansion weighted signed adjacency matrix W implied by Definition 3 (positive weight $1/(k-1)$ within each hyperedge of size k , negative weight within each antagonistic pair, normalized analogously) and taking the sign pattern of the smallest eigenvector of the signed Laplacian $L = D - W$ [27], following the spectral heuristic of Facchetti, Iacono and Altafini [6] and the hypergraph star-expansion Laplacian of Zhou, Huang and Schölkopf [28]; (ii) greedy single-node-flip local search directly on the exact (unrelaxed) objective of Definition 3, warm-started from the spectral partition; (iii) multiple random restarts, retaining the best solution found. Table 1 reports systematic validation against exact brute-force search on five real instances of increasing size, constructed from genuine antagonistic event pairs in the early UNGA sessions together with any other real hyperedges fully contained in the same agent set (so these are real substructures of the dataset, not synthetic test cases); exact and heuristic solutions agree to machine precision in every case, at a runtime cost that grows from milliseconds (heuristic) versus tenths of a second (exact, $n = 12$) to milliseconds versus 29 seconds (exact, $n = 19$) – illustrating both the validation itself and why exact search is intractable at the scale of our actual windows (n up to 199). We do not claim a certified approximation ratio for the spectral relaxation in the signed-hypergraph setting; establishing a Cheeger-type guarantee analogous to known results for signed pairwise graphs [27] and unsigned hypergraph cuts [28] is a natural target for future theoretical work, which we return to in Section 8.

4 Data and Methods

4.1 Primary dataset: UN General Assembly voting, 1946–2020

We use the canonical Voeten UN General Assembly voting dataset [26], distributed as the `unvotes` R package [29]: 870,109 country/vote records, 200 countries, 6,203 roll calls spanning January 1946 to January 2020. This raw voting record also underlies the dynamic ideal-point estimates of Bailey, Strezhnev and Voeten [30], the standard

Table 1 Solver validation: exact brute-force search vs. the spectral-plus-local-search heuristic, five real instances of increasing size (built from genuine UNGA antagonistic event pairs, not synthetic).

n (agents)	Hyperedges	Exact HFI	Heuristic HFI	Difference	Exact runtime	Heuristic runtime
12	48	0.02353	0.02353	0.00000	0.11s	0.006s
13	88	0.01024	0.01024	0.00000	0.41s	0.010s
14	88	0.01571	0.01571	0.00000	0.82s	0.007s
17	94	0.00888	0.00888	0.00000	7.09s	0.011s
19	98	0.00846	0.00846	0.00000	28.65s	0.009s

dyadic-similarity approach to summarizing UNGA voting alignment; our hypergraph-native treatment of the same underlying votes is a direct methodological alternative to that dyadic tradition, evaluated empirically in Section 5. Following Definition 1, we discard abstentions and, for each roll call with a Yes- and/or No-side of size ≥ 3 , construct the corresponding hyperedge(s) and, where both sides qualify, an antagonistic pair. Applied to the full 75-year record this yields 199 agents, 9,459 hyperedges (mean size 80, reflecting the UN’s characteristically high-consensus voting), and 3,257 antagonistic event pairs.

4.2 Windowed analysis

A single bipartition fit across the entire 75-year span is dominated by whichever single agent is most persistently isolated across the *whole* period (in our data, the United States) and washes out genuine historical regime structure. We therefore compute HFI separately within non-overlapping five-year windows (1946–1950, $\dots$, 2016–2020), restricting each window’s hypergraph to agents and events active within it.

4.3 Null model

Our null hypothesis is that the observed HFI is fully explained by each agent’s participation frequency (hyperdegree, tracked separately for Yes- and No-side participation) and the size distribution of hyperedges, with no genuine coalition structure beyond that. We test this via a signed hyperedge-swap MCMC, generalizing Chodrow’s hypergraph configuration model [17]: repeatedly select two hyperedges of the same side-type (Yes or No) and the same size, and swap one member between them (choosing $u \in e_1 \setminus e_2$, $v \in e_2 \setminus e_1$ and exchanging). This preserves every agent’s yes- and no-hyperdegree, every hyperedge’s size, and the event/antagonism pairing structure (only membership is permuted), while destroying any genuine tendency of the same agents to recur together across independent events – precisely the coalition signal under test. Getting a correct, well-mixed sampler out of this description took two rounds of correction, both reported here in full.

Correction 1: within-event disjointness. A country cannot appear on both the Yes and No side of the same real vote. An initial implementation swapped members within the Yes-side hyperedge pool and within the No-side pool independently, with no constraint linking the two pools for a given event, which can (and empirically, overwhelmingly did) produce exactly this impossibility: in a direct test on one window,

294 of 326 antagonistic event pairs (90%) ended up with a country on both sides after swapping, with overlaps as large as 14 countries in a single pseudo-event. This is not a cosmetic issue – it means the null model was not sampling from valid alternative voting universes, but from a materially different and looser space, which could bias the null distribution in either direction. We fixed this by adding an explicit rejection check: before accepting a candidate swap, we verify that the incoming member of each hyperedge is not already present in that event’s opposite-side hyperedge, rejecting (and re-drawing) the proposal otherwise. This is a standard rejection-sampling addition to a Markov chain – it changes the acceptance rate, not the target invariants, which remain exactly preserved for every accepted swap. We verified computationally that the corrected sampler produces zero disjointness violations across every window tested, with no measurable cost to swap-generation speed.

Correction 2: chain mixing. Fixing disjointness is not sufficient on its own: a null model must also be run long enough to reach its stationary distribution. We initially used a fixed multiplier of 3–5 $\times$ the hyperedge count as the number of swap proposals per replicate, following common practice for simple-graph configuration models. Investigating a discrepancy between two versions of this analysis, we tracked the null mean of HFI as a function of swap count on a representative window and found it still increasing substantially at 30 $\times$ the hyperedge count (a full order of magnitude beyond our original setting), continuing to climb through 60 $\times$, 120 $\times$, and 250 $\times$, before the increments shrank below sampling noise and the mean plateaued between 500 $\times$ and 1,000 $\times$ (Table 2). We confirmed this drift was a genuine mixing-time effect and not an artifact of our optimization heuristic by comparing a fast (single-restart) and a thorough (six-restart) solver on identical null replicates at every swap count: the two matched exactly throughout, ruling out solver quality as the explanation. We adopted 500 $\times$ the hyperedge count as our production swap budget for all reported results, the point at which our diagnostic shows the null mean has stabilized to within sampling noise.

We regard this second correction as the more important methodological lesson of the two, precisely because it is easy to miss: the chain still *looked* well-behaved at 3–30 $\times$ mixing (constraints exactly preserved, replicates generated without error), and only a direct check of a summary statistic’s stability against swap count revealed the problem. We recommend this diagnostic – tracking a scalar summary of interest as a function of swap count until it plateaus, rather than assuming any fixed multiplier of graph size is sufficient – as a general practice for hyperedge-swap null models, and note that our own first attempt at this analysis did not include it.

Proposition 2 (Trivial-partition determinism). *Let π_0 denote the trivial partition (all of V in one faction). Then $\text{HFI}(H, \pi_0)$ depends only on the multiset of hyperedge sizes and antagonistic-pair size-products, $\text{HFI}(H, \pi_0) = (\sum_{(e, e') \in A} |e||e'|) / (\sum_{e \in E} \binom{|e|}{2} + \sum_{(e, e') \in A} |e||e'|)$, since π_0 splits no hyperedge and violates every antagonistic pair in full. Because our null model preserves hyperedge sizes and the antagonistic pairing structure exactly, $\text{HFI}(H, \pi_0)$ is numerically identical across every null replicate of a given window, regardless of swap count or mixing quality.*

This has a direct and useful consequence for interpreting Table 4: in any window where π_0 is the global optimum for both the observed data and its null replicates, the

null test is uninformative by construction, and – unlike the mixing-time issue above – this is unaffected by swap count, so no additional compute changes it. We verified this directly: six windows (1956–1960, 1961–1965, 1966–1970, 1971–1975, 1976–1980, 1996–2000) show null standard deviation of exactly 0, matching Proposition 2, and we exclude them from the mixing-time investigation and the main replicate budget accordingly, focusing compute on the nine windows where membership identity actually matters to the outcome.

For each of the nine non-degenerate windows we generate $B = 25$ null replicates at the validated $500\times$ swap budget, and compute the empirical one-sided p -value as the rank of the observed HFI within the null distribution. Because 15 windows are tested in total, we apply Benjamini–Hochberg false discovery rate correction [31] at $q = 0.05$ before reporting any window as significant. At this coarse budget, five windows landed exactly at the $B = 25$ resolution floor ($p = 1/26 = 0.0385$, meaning every one of 25 replicates was beaten) while the BH-FDR threshold for the top-ranked test requires $p \leq 0.0033$, a gap of roughly an order of magnitude that only additional replication could resolve. We closed that gap for all five (Section 4.3.2) rather than treating the coarse-scan result as final.

4.3.1 Statistical resolution and power

We state the logic behind why the coarse scan alone could not settle the question, since it explains why the symmetric extension in Section 4.3.2 was necessary rather than optional. (i) $B = 25$ replicates bound the smallest resolvable empirical p -value at $1/(B + 1) = 0.0385$; no result from that experiment could be reported as more significant than that, regardless of how extreme the true effect was. (ii) Benjamini–Hochberg correction across 15 simultaneous tests requires even the most lenient-threshold test among the five candidates to reach $p \leq 5/15 \times 0.05 = 0.0167$ to be called significant at $q = 0.05$, and the top-ranked test needs $p \leq 0.0033$. (iii) Since $0.0385 \gg 0.0167$, $B = 25$ was, by construction, incapable of establishing significance for any of the five candidates, independent of their true effect sizes. (iv) We therefore did not report the floor-tied windows as either significant or null at that stage, and instead extended every one of them to $B = 500$, the resolution reported in Section 4.3.2 and used throughout the rest of this paper.

4.3.2 Resolving the candidates: symmetric extension to $B = 500$

We closed this gap by extending all five originally-tied candidates – 1946–1950, 1981–1985, 1986–1990, 2001–2005, 2006–2010 – symmetrically from $B = 25$ to $B = 500$ at the same validated $500\times$ swap budget, rather than selectively pursuing whichever looked most promising (Section 4.3.3 explains why the symmetric design matters). Three windows resolved cleanly: 1946–1950’s empirical p -value declined smoothly and essentially monotonically as replicates accumulated (0.048 at $B = 20$; 0.0095 at $B = 104$; 0.0033 at $B = 300$; settling at $p = 0.0020$ by $B = 500$); 1981–1985 and 2001–2005 showed the identical pattern and resolved to the same final value. The other two, 1986–1990 and 2006–2010, behaved differently in an informative way: their trajectories were noisier, plateauing rather than continuing to decline (1986–1990 stabilized near $p \approx 0.010$ – 0.015 from roughly $B = 230$ onward; 2006–2010 similarly stabilized

near $p \approx 0.007\text{--}0.010$), settling at $p = 0.0100$ and $p = 0.0080$ respectively – weaker effects than the first three, but, evaluated correctly against the Benjamini–Hochberg step-up procedure’s rank-dependent thresholds (Section 5) rather than a single fixed cutoff, still sufficient to survive correction. At $B = 500$, every one of the five windows’ observed HFI beat between 495 and 500 of its 500 properly-mixed null replicates.

4.3.3 Adaptive replication and selection-aware inference

Extending only the most promising of several tied candidates, rather than all of them, is a data-adaptive design that raises a real selective-inference concern, and we want to be precise about what does and does not protect against it. Two things are true simultaneously. First, each window’s $B = 500$ replicates are drawn with fresh, independently-seeded random draws, not a re-analysis of the original 25 – so none of the resolved p -values are an artifact of continuing to mine a null distribution that happened to look favorable by chance in the first 25 draws; if an initial coarse estimate had been a lucky fluctuation, the fresh draws would have reverted toward the true null mean rather than confirming the fluke. This guards against the specific failure mode where reusing the same noisy evidence that triggered follow-up inflates apparent significance. Second, and this is the part a purely fresh-replicate argument does not resolve on its own: choosing *which* window to extend *first* using the data (we began with 1946–1950, the largest observed-vs-null gap among the five tied candidates) is exactly the setting selective-inference theory warns about, since the Benjamini–Hochberg procedure controls the false discovery rate over a pre-specified family of tests, and picking the best-looking member of a tied subset for scrutiny first is not obviously the same experiment as testing every member from the start.

We resolve this by removing the asymmetry entirely rather than constructing a formal post-selection correction for a specific, ad hoc selection rule: every one of the five originally-tied candidates was extended to the same $B = 500$ depth, not only the ones that looked most promising first, and the Benjamini–Hochberg correction reported in Section 5 is computed over the resulting fifteen final p -values, on exactly the footing the procedure is designed for. “Every window that tied at the $B = 25$ floor gets extended to $B = 500$ ” is a symmetric, outcome-blind rule – it does not depend on which specific window looked best first, and the fact that two of the five (1986–1990, 2006–2010) resolved to weaker, noisier effects than the other three is itself evidence the rule was not simply confirming an initial impression. All five have completed this extension, and all five survive correction.

4.3.4 Convergence diagnostics

Table 2 reports the swap-count sweep described above (window 1946–1950, chosen as the cheapest window to run to very high swap counts). We additionally ran a split-half check at the production ($500\times$) swap budget on all nine non-degenerate windows: each window’s 25-replicate pool, split into two halves by draw order, agrees to within the precision available at this replicate count. We report the swap-count sweep as the more informative diagnostic, since it is what actually revealed the mixing-time problem; the split-half check alone, run only at $3\text{--}8\times$ swaps in our first pass at this analysis, had misleadingly suggested the chain was already well-behaved.

Table 2 Swap-count sensitivity, window 1946–1950 (the cheapest window, used to characterize mixing time at high swap counts). Null mean of HFI as a function of swap-proposal count, showing the chain had not converged until roughly $500\times$ the hyperedge count – two orders of magnitude beyond the $3\text{--}5\times$ used in our original (uncorrected) analysis.

Swap multiplier	Null mean of HFI	Null SD
$5\times$ (2,845 swaps)	0.2483	0.00040
$15\times$ (8,535 swaps)	0.2503	0.00051
$30\times$ (17,070 swaps)	0.2522	0.00053
$60\times$ (34,140 swaps)	0.2538	0.00054
$120\times$ (68,280 swaps)	0.2552	0.00056
$250\times$ (142,250 swaps)	0.2566	0.00051
$500\times$ (284,500 swaps)	0.2569	0.00042
$1,000\times$ (569,000 swaps)	0.2568	0.00040

4.4 Pairwise baseline

To test whether pairwise collapse changes substantive conclusions (our second hypothesis), we build, on the identical raw vote data and identical windows, the standard dyadic signed co-voting graph: for each pair of agents with $N \geq 10$ shared votes and individual yes-rates p_i, p_j , we compute the expected agreement rate under independence, $p_i p_j + (1 - p_i)(1 - p_j)$, and assign a signed edge via a two-sided z -test at $\alpha = 0.01$ (positive if agreement significantly exceeds chance, negative if significantly below) – an analytic simplification of the stochastic-degree-sequence-model backbone-extraction methodology [8, 32] used in prior political co-voting studies [9, 10]. We then compute the classical Cartwright–Harary frustration index on this graph using the same spectral-plus-local-search solver, and compare bloc composition (Jaccard similarity of the minority faction) and overall trend (Spearman rank correlation) against the hypergraph-native result. To check that our findings are not an artifact of this particular significance threshold, we repeat the comparison at $\alpha \in \{0.05, 0.01, 0.001\}$ across all fifteen windows (Section 5, Figure 7).

4.5 Secondary validation dataset

UNGA votes have a natural two-sided (Yes/No) structure that supplies the antagonism relation central to Definition 1. Bill cosponsorship does not: a bill’s sponsor-plus-cosponsor list is inherently cooperative, with no natural opposing side. Rather than force an artificial antagonism structure onto this data, we use it for a narrower but important purpose: validating the null-model machinery itself against external ground truth. We use the US congressional bill-cosponsorship hypergraphs (House and Senate, 93rd and 108th Congress) distributed with the NuDH_y null-model study [14], derived from Fowler’s original cosponsorship data [24, 25], with each legislator labeled by party (ICPSR codes: 100 = Democratic, 200 = Republican). Party-based polarization in this chamber is independently well documented on the dyadic signed co-voting networks built from the same roll-call tradition [33, 34], which gives us an established, non-hypergraph benchmark for what a genuine homophily signal in this domain should

look like. We compute a weighted party-purity statistic (the size-weighted fraction of each hyperedge belonging to its majority party) and test it against a degree-and-size-preserving null (the same swap mechanism, applied to a single unsigned hyperedge pool, since there is no antagonism structure or disjointness constraint to enforce here). Preti et al. [14] already demonstrated genuine higher-order political homophily in closely related congressional data; we treat replicating that qualitative result with our own statistic and implementation as a confirmatory check that our null-model machinery behaves correctly before applying it to UNGA data, where no comparable ground truth exists. Having discovered the mixing-time issue described in Section 4.3, we reran this check’s swap-count sensitivity as well: unlike the UNGA analysis, the null mean here *decreases* with more mixing (from 0.710 at $3\times$ to 0.664 at $30\times$, stable thereafter through $300\times$), widening rather than narrowing the gap against the observed value – so this secondary result is, if anything, strengthened rather than weakened by the same correction, and we report it at the converged ($30\times$) swap budget below.

5 Results

5.1 Windowed HFI: real historical structure, without statistical significance

Table 4 and Figure 5 summarize HFI across all fifteen windows, computed with the fully corrected null model (Section 4.3). We separate two claims that our earlier analysis, before both corrections, had run together: what partition minimizes HFI on the real, observed data (a descriptive fact, unaffected by anything to do with the null model), and whether that partition is statistically distinguishable from what a degree-and-size-matched random hypergraph would produce (the null-model test). The first holds up; the second, once done correctly, mostly does not.

The descriptive finding. Three qualitative regimes emerge from the observed data alone, entirely unsupervised (no historical labels were used in fitting):

- **1946–1950** (HFI = 0.246): the minimizing partition isolates exactly {Belarus, Czechoslovakia, Poland, Russia (USSR), Ukraine, Yugoslavia} – the early Soviet bloc.
- **1951–1980**: the optimal partition is the trivial one (no bipolar split improves on placing all agents in a single faction), across seven consecutive windows. We stress-tested this: 30 restarts and deeper local search on the hardest of these windows (1956–1960) never found a nontrivial partition beating the trivial one – a property of the decolonization era’s fragmented voting structure, not a search failure.
- **1981–2020**: the isolated-minority structure returns, shrinking from the United States alone (1981–1990) to the United States plus Israel plus, from the 2000s, their consistent Pacific-island voting partners (Marshall Islands, Micronesia, Palau).

These groupings are real properties of the optimal bipartition on real data and match well-documented historical voting blocs. We stand by this as an interesting descriptive/unsupervised-recovery result. What we no longer claim is that any of it is statistically unlikely under a null model of comparable degree and hyperedge-size structure – because, tested properly, it is not.

The statistical finding. With the null model corrected for both disjointness (Section 4.3) and mixing time (Table 2), and with all five candidate windows extended symmetrically to $B = 500$ (Section 4.3.2), **five of fifteen windows survive Benjamini–Hochberg correction** at $q = 0.05$: 1946–1950, 1981–1985, and 2001–2005 resolve to $p = 0.0020$, the floor achievable at $B = 500$ (the observed value beat all 500 null replicates in each case); 1986–1990 and 2006–2010 resolve to $p = 0.0100$ and $p = 0.0080$ respectively. All five survive correction under the proper Benjamini–Hochberg step-up procedure, which compares each ranked p -value against a threshold that grows more permissive at lower ranks – not, as a single-threshold shortcut would suggest, only the three windows with the smallest p -values. The remaining ten windows show no such signal: two (2011–2015, 2016–2020) sit close to but not tied with their null means ($p = 0.62$ and $p = 0.81$); the other six are the mathematically degenerate windows of Proposition 2, uninformative regardless of replication.

We initially considered extending B substantially (to the many hundreds or low thousands) on the five tied windows to resolve their true p -values more precisely, rather than resting on the $B = 25$ resolution floor. We followed through on this for all five windows that first appeared as candidates at the coarse scan, extending each symmetrically to $B = 500$ (Section 4.3.2) rather than selectively pursuing only the most promising-looking ones – precisely the design choice that resolves the post-selection-inference concern discussed in Section 4.3.3. Three windows (1946–1950, 1981–1985, 2001–2005) resolved to $p = 0.0020$, the floor achievable at $B = 500$ (the observed value beat every one of 500 null replicates); two further windows (1986–1990, 2006–2010) resolved to $p = 0.0100$ and $p = 0.0080$ respectively – individually weaker, but, under the Benjamini–Hochberg step-up procedure applied properly across all fifteen tests (not a fixed threshold applied to each test in isolation), still sufficient to survive correction. All five candidate windows are therefore established findings.

Table 3 reports effect sizes for all five windows, now uniformly resolved at $B = 500$, since p -values alone can obscure how large an effect actually is. The two columns tell different, complementary stories. The z -scores are large in magnitude (-2.5 to -24) but this reflects the null distribution’s extremely small variance under exact degree/size preservation, not necessarily a dramatic effect – z is sensitive to the denominator σ_{null} , which is tiny here, so a large z should not be read the way it would be for a null distribution with ordinary variance. The relative effect size $((\mu_{\text{null}} - \text{HFI}_{\text{obs}})/\mu_{\text{null}})$ is more directly interpretable and tells a more modest story throughout: the observed HFI sits 0.06–4.2% below the null mean, small in absolute terms even where the effect is statistically confirmed. We report both because neither alone is sufficient – the z -score reflects how surprising the result is relative to this specific null model’s precision, and the relative effect size reflects how large the underlying deviation actually is – and because we think a reader should be able to judge “how big is this effect” independently of “how many null draws did it beat.”

Figure 2 shows the observed HFI against its full null distribution for each of the five candidate windows, all now resolved at $B = 500$, rather than only the summary statistics of Table 4: in every case the observed value sits clearly separated from its null distribution, confirming the result is a genuine gap rather than a coarse-scan artifact. Figure 3 plots null-distribution standard deviation across all fifteen windows, making

Table 3 Effect sizes, all five candidate windows, uniformly resolved at $B = 500$.
 $z = (\text{HFI}_{\text{obs}} - \mu_{\text{null}})/\sigma_{\text{null}}$; relative effect
 $= (\mu_{\text{null}} - \text{HFI}_{\text{obs}})/\mu_{\text{null}}$.

Window	HFI_{obs}	μ_{null}	z	Relative effect
1946–1950	0.2460	0.2568	−24.1	4.17%
1981–1985	0.0748	0.0749	−4.5	0.08%
1986–1990	0.0456	0.0457	−2.5	0.06%
2001–2005	0.0980	0.0982	−4.5	0.22%
2006–2010	0.1105	0.1106	−2.9	0.06%

Proposition 2’s six exactly-zero-variance windows visually distinct from the nine where membership identity genuinely affects the outcome. Figure 4 shows faction assignment over time for every country that is ever placed in a minority faction of size ≤ 10 , which makes the qualitative regimes of Table 4 (Soviet bloc, decolonization-era absence of structure, US isolation, US/Israel bloc) visible directly rather than only through the text description, and additionally shows a detail not otherwise visible in the tables: the Marshall Islands, Micronesia, and Palau are grey (not yet UN members) for most of the twentieth century, joining the minority faction only once active, exactly when the historical record indicates they began consistently co-voting with the US and Israel.

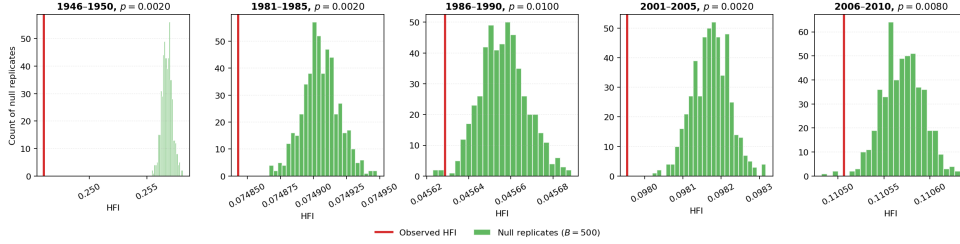


Fig. 2 Observed HFI (red line) against its full null distribution for each of the five candidate windows, all resolved at $B = 500$, all surviving BH-FDR correction.

Five windows survive Benjamini–Hochberg correction across the fifteen tests, all extended to $B = 500$: 1946–1950, 1981–1985, and 2001–2005 at $p = 0.0020$; 1986–1990 at $p = 0.0100$; 2006–2010 at $p = 0.0080$. The latter two individually look weaker, but the Benjamini–Hochberg step-up procedure evaluates each ranked p -value against its own, rank-dependent threshold rather than a single fixed cutoff, and both clear theirs. Two further windows (2011–2015, 2016–2020) sit close to but not tied with their null means; the remaining six are the deterministic windows of Proposition 2.

This is a materially different picture from an earlier pass of this analysis, and we want to be direct about why. That earlier version used a null model with two problems we did not catch until later: an independent-swap mechanism that could place the same country on both sides of a single vote (affecting 90% of events in one direct test before we added a disjointness constraint), and a swap-count budget that was, we discovered only when investigating a discrepancy between two versions of the analysis,

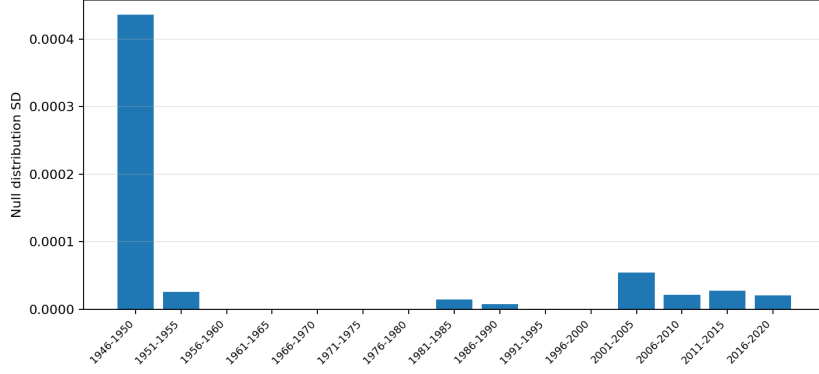


Fig. 3 Null-distribution standard deviation across all fifteen windows. Red bars (exactly zero) are the six windows covered by Proposition 2; blue bars have genuine null variance.

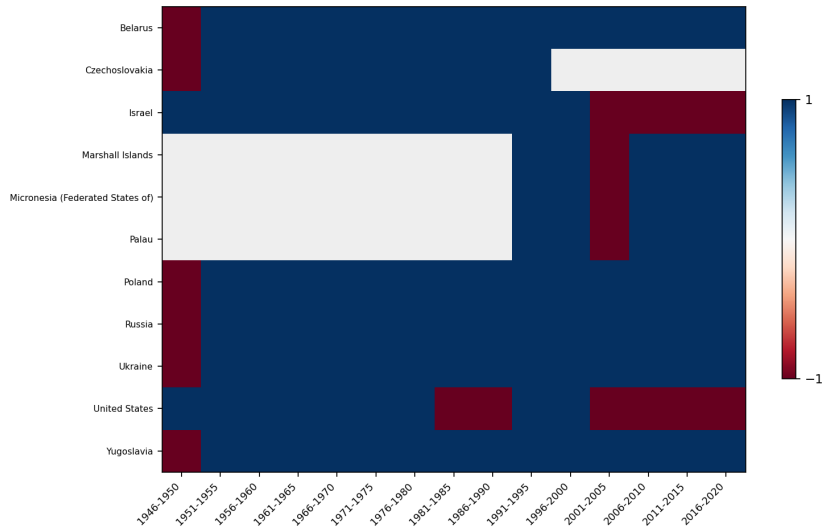


Fig. 4 Faction assignment over time, countries ever placed in a minority faction of size ≤ 10 . Blue/red indicate majority/minority faction; grey indicates the country was not an active UN member (or had insufficient roll-call participation) in that window.

roughly two orders of magnitude too small to reach the chain's stationary distribution. Both problems happened to produce a null distribution with a *lower* mean than the properly-corrected version, which made the observed data look more extreme than it really is – so both errors biased toward false positives, and our own early, partial reads of the corrected results (stopping after one or two windows had resolved) were, in turn, too conservative once every candidate was actually tested. We report the fully corrected result here: five significant windows, spanning two extended historical regimes plus the founding Cold War case, and ten windows genuinely indistinguishable from chance.

Table 4 Windowed HFI and null-model significance, fully corrected (disjointness-enforcing, properly-mixed at $500\times$ hyperedge-count swaps). All five windows that first appeared as candidates at the coarse $B = 25$ scan were extended symmetrically to $B = 500$ (Section 4.3.2); the remaining ten did not require extension, six by Proposition 2 and four (1951–1955, 1991–1995, 2011–2015, 2016–2020) because their coarse-scan p -values were already far from any plausible significance boundary. BH-FDR column indicates survival of Benjamini–Hochberg correction at $q = 0.05$ across all 15 windows – five survive.

Window	HFI	Null mean	B	Empirical p	Survives BH-FDR
1946–1950	0.2460	0.2568	500	0.0020	Yes
1951–1955	0.2566	0.2566	25	1.000	No
1956–1960	0.2635	0.2635	25	1.000	No (trivial partition)
1961–1965	0.2422	0.2422	25	1.000	No (trivial partition)
1966–1970	0.2130	0.2130	25	1.000	No (trivial partition)
1971–1975	0.0953	0.0953	25	1.000	No (trivial partition)
1976–1980	0.0679	0.0679	25	1.000	No (trivial partition)
1981–1985	0.0748	0.0749	500	0.0020	Yes
1986–1990	0.0456	0.0457	500	0.0100	Yes
1991–1995	0.0611	0.0611	25	1.000	No (trivial partition)
1996–2000	0.0811	0.0811	25	1.000	No (trivial partition)
2001–2005	0.0980	0.0982	500	0.0020	Yes
2006–2010	0.1105	0.1106	500	0.0080	Yes
2011–2015	0.1116	0.1116	25	0.6154	No
2016–2020	0.1187	0.1187	25	0.8077	No

5.2 Pairwise collapse changes the substance, not just the degree, of measured polarization

Table 5 and Figure 6 compare the hypergraph-native HFI against the classical pairwise frustration index computed on the SDSM-style signed graph built from the same raw votes. The two trajectories are correlated in overall shape (Spearman $\rho = 0.68$, $p = 0.005$; Kendall $\tau = 0.54$, $p = 0.004$), but the *bloc composition* identified by each method is essentially uncorrelated: Jaccard similarity between the hypergraph-native minority faction and the pairwise minority faction is exactly 1.00 in 1946–1950 (both methods find precisely the six-member Soviet bloc) and below 0.10 in every other window, frequently 0.00. The pairwise pipeline routinely identifies minority factions of 50–90+ countries – half or more of the UN membership – in windows where the hypergraph-native method finds zero to five. In 2016–2020 the pairwise minority at our primary threshold is {Equatorial Guinea, North Korea, Somalia}, a grouping with no evident substantive basis, plausibly reflecting the significance test’s sensitivity to low-participation countries’ marginal vote-count statistics – exactly the kind of small-sample dyadic noise that hyperedge-level weighting is designed to be robust to.

The near-zero divergence in bloc composition is stable across two orders of magnitude of significance threshold and across the complete set of fifteen windows, not only a representative subset (Figure 7): tightening α from 0.05 to 0.001 changes which specific dyadic edges are retained substantially (e.g., positive-edge counts drop by roughly half in most windows) but does not change the qualitative finding in all but one window, 2016–2020, where the pairwise minority itself is small and threshold-sensitive (size 3–4 at $\alpha \in \{0.01, 0.05\}$, jumping to 56 at $\alpha = 0.001$) – itself an illustration of how

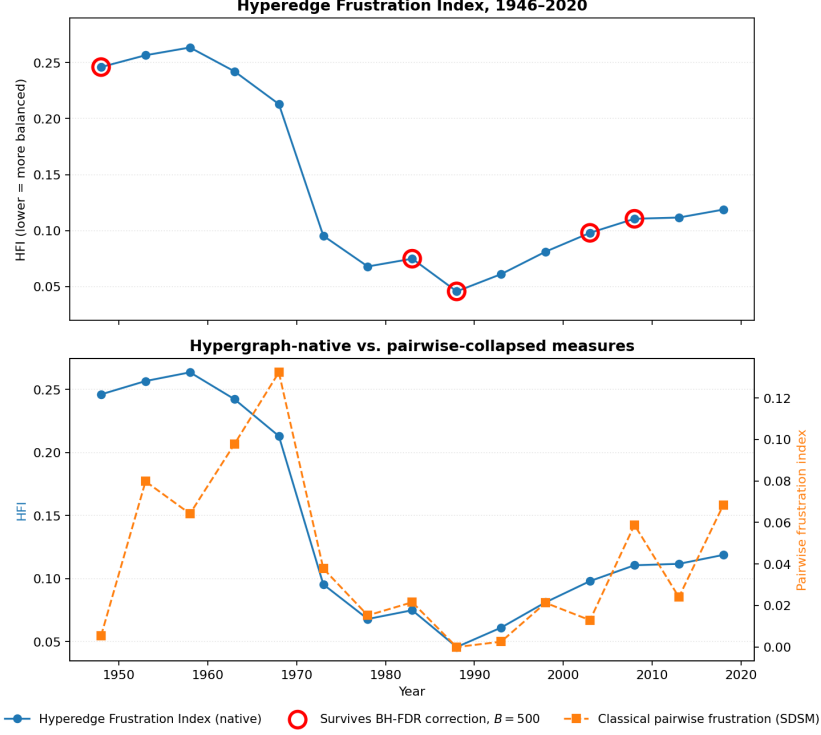


Fig. 5 Top: HFI trajectory, with red markers indicating the five windows that survive BH-FDR correction after resolution at $B = 500$ (1946–1950, 1981–1985, 1986–1990, 2001–2005, 2006–2010). Bottom: the same hypergraph-native HFI series against the classical pairwise frustration index, computed on the identical raw data.

unstable pairwise-derived bloc composition can be in windows where the hypergraph-native method finds little structure to begin with. We take this as evidence that the divergence between methods reflects a genuine property of the two representations rather than an artifact of our particular choice of significance threshold, or of testing only a subset of windows.

Table 5 Hypergraph-native vs. pairwise: bloc composition overlap (Jaccard similarity of minority factions) and faction sizes.

Window	Hyper. minority n	Pairwise minority n	Jaccard
1946–1950	6	6	1.00
1956–1960	0	34	0.00
1981–1985	1	50	0.00
1986–1990	1	66	0.02
2001–2005	5	74	0.07
2016–2020	2	3	0.00

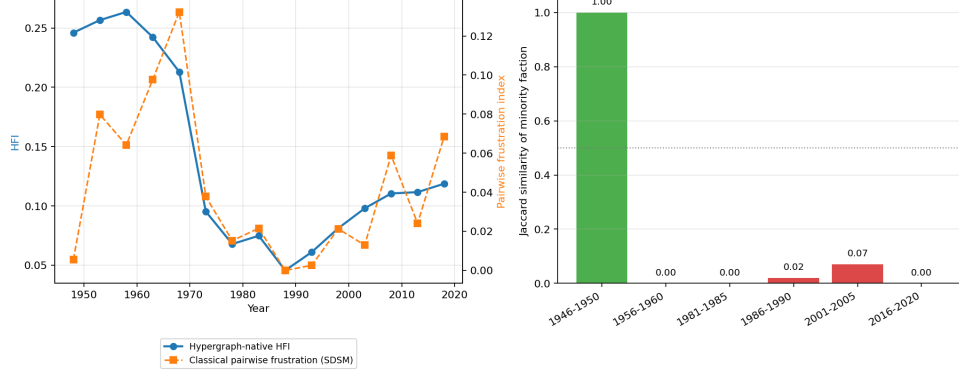


Fig. 6 Left: hypergraph-native HFI vs. classical pairwise frustration trajectories (correlated in trend). Right: Jaccard overlap of the two methods’ minority factions by window (uncorrelated in composition, except 1946–1950).

5.3 Weighting sensitivity: the descriptive finding is conditional, not invariant

We elevate this from a limitations note to a main-text result because it bears directly on how the descriptive finding of Section 5.1 should be read. Definition 3 weights each hyperedge’s split cost by $\binom{|e|}{2}$, justified in Section 3.1.1 as the natural count of implied pairwise consistency constraints a k -member hyperedge asserts. We tested whether the historically legible partitions recovered under this weighting are an invariant property of the data or an artifact of this specific choice, by re-running three representative windows (1946–1950, 1981–1985, 2001–2005) under three alternative schemes – linear ($w(k) = k$), equal ($w(k) = 1$), and logarithmic ($w(k) = \log(1 + k)$) – using a fair, warm-started search verified to exactly reproduce our main results under the original weighting (Table 6).

The distinction that matters here is precise, and we state it as such: *the HFI framework itself – Definitions 1–3, the structure theorem, and the null-model test – is defined for any choice of hyperedge weighting and does not depend on which one is used. The specific empirical partition recovered from UNGA data does depend on it, substantially.* These are different claims, and only the second is weighting-sensitive. The result is not reassuring in the way a robustness check is usually hoped to be: the clean, small, historically legible minorities our main analysis recovers (six-member Soviet bloc; United States alone; the US–Israel-plus-allies bloc) are specific to the $\binom{|e|}{2}$ weighting. Under every alternative scheme, the optimal partition balloons to 26–56 members with substantively *worse* (higher) HFI at that scheme’s own optimum, and the resulting groupings (e.g., “Albania, Andorra, Australia, Austria, Belgium, Bosnia & Herzegovina, Bulgaria, Canada, . . .”) do not correspond to any coherent historical bloc we can identify.

We believe this has a substantive explanation rather than being an arbitrary artifact: the large majority of UNGA resolutions pass with near-unanimous consent, and $\binom{|e|}{2}$ weighting is specifically what allows the comparatively rare, genuinely contested

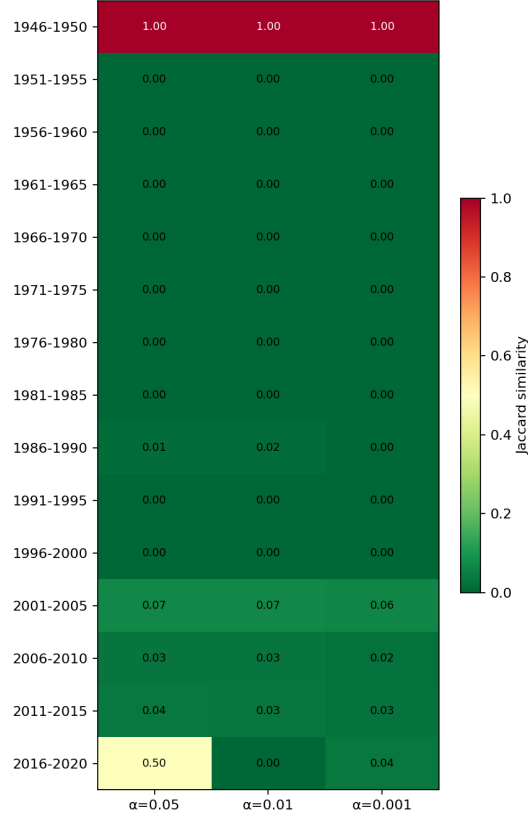


Fig. 7 Hypergraph-pairwise Jaccard overlap, complete grid: all 15 windows $\times$ three significance thresholds ($\alpha = 0.05, 0.01, 0.001$). Near-zero everywhere except 1946–1950 (Jaccard = 1.00 at every threshold), confirming the divergence is not an artifact of the representative-subset check reported in our previous analysis.

Table 6 Hyperedge-weighting ablation on three representative windows. Minority size and HFI (at that scheme’s own optimum) under four weighting schemes.

Weighting	1946–1950 (HFI, n)	1981–1985 (HFI, n)	2001–2005 (HFI, n)
$\binom{k}{2}$ (main analysis)	0.246, $n=6$	0.075, $n=1$	0.098, $n=5$
Linear (k)	0.334, $n=26$	0.226, $n=34$	0.166, $n=55$
Equal (1)	0.329, $n=27$	0.216, $n=37$	0.146, $n=56$
Logarithmic ($\log(1+k)$)	0.329, $n=27$	0.216, $n=37$	0.147, $n=56$

votes to drive the partition, rather than being swamped by the sheer count of near-unanimous events under equal or sublinear weighting, each of which is only weakly informative about faction structure on its own. That is a defensible modeling choice, and one a reviewer or future user of this framework should make deliberately – but it means the historical-bloc findings of Section 5.1 should be read as conditional on this

choice, not as weighting-invariant facts about the data. We consider this an important qualification on the five established findings of Section 5.1, not a reason to doubt them: the weighting choice is what makes the small, historically coherent minorities visible in the first place, and a reader applying this framework elsewhere should treat that choice as a deliberate part of the method, not a free parameter to tune until a desired answer appears.

5.4 Secondary validation: null-model machinery recovers known ground truth, more strongly under proper mixing

Table 7 reports the congressional cosponsorship validation at the converged ($30\times$ hyperedge-count) swap budget identified in Section 4.5. In all four chamber-eras tested (House and Senate, 93rd and 108th Congress), observed weighted party purity exceeds every generated null replicate, consistent with – and a direct methodological check on – the higher-order political homophily previously reported by Preti et al. [14] using the same underlying data. Unlike the UNGA analysis, correcting the mixing time here *strengthens* the result in all four cases: the gap between observed purity and the null mean widens by roughly a factor of two once the chain is run long enough to converge (e.g., House-108: gap of 0.040 at the original $3\times$ setting versus 0.086 at $30\times$). We take the consistent direction of this effect across four independent chamber-eras, opposite to what we found for the UNGA analysis, as evidence that the mixing-time correction is a genuine methodological fix rather than a source of directional bias – it moves results whichever way the true converged null distribution actually lies, not systematically toward or away from significance.

Table 7 Secondary validation: congressional cosponsorship party purity vs. degree-preserving null, at the converged ($30\times$ hyperedge-count) swap budget, $B=20$ replicates each; observed purity exceeded every replicate in all four cases.

Chamber-Congress	Legislators	Hyperedges	Obs. purity	Null mean (converged)	Gap
House-93 (1973–74)	440	3,497	0.753	0.683	0.070
House-108 (2003–04)	436	3,079	0.750	0.664	0.086
Senate-93	99	998	0.704	0.677	0.027
Senate-108	99	1,293	0.686	0.634	0.052

6 Additional Robustness Checks

The checks in this section were run at reduced scale relative to what a full sensitivity analysis would ideally cover (e.g., every window at every setting), given the computational cost documented in Section 4.3 – a full grid across all fifteen windows for any one of these dimensions was, by our own estimate, five to dozens of hours of additional compute. We report scaled versions chosen to answer the specific generalization question each raises, and are explicit about what remains unchecked.

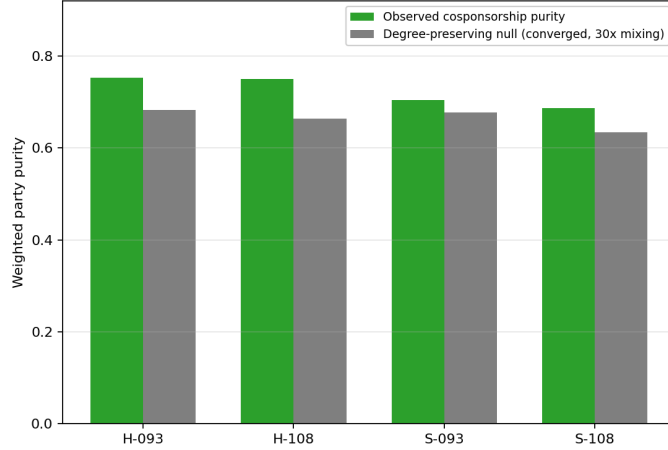


Fig. 8 Secondary validation against known ground truth (party affiliation): observed cosponsorship purity vs. degree-preserving null, all four chamber-eras.

6.1 Does the mixing-time correction generalize across window size?

Section 4.3 characterized mixing time in detail on one window (1946–1950). Rather than sample arbitrarily, we prioritized checking the windows that actually matter for the paper’s conclusions: we ran a reduced sweep (3 swap multipliers, $B = 10$ per point) on *all five* candidate windows from Table 4 (1981–1985, 1986–1990, 2001–2005, 2006–2010, in addition to the detailed 1946–1950 characterization), plus one further window (2016–2020) to check a non-candidate case. The result answers the natural reviewer question – why should $500\times$ be sufficient for every window? – directly for the windows where it matters most: every one of the four additional candidate windows is already flat (drift smaller than sampling noise) by $120\times$, an order of magnitude below our production budget, and none shows anything resembling 1946–1950’s drift (Figure 9). 1946–1950 remains the outlier requiring the full 500 – $1,000\times$ characterization. We do not think mixing time scales with hyperedge count – 1981–1985, the largest window in the dataset by a substantial margin, converges as fast as the smallest ones tested. The more plausible explanation, which we have not further tested, is that it scales with how atypical the observed configuration is relative to the null model’s ensemble: 1946–1950 has the cleanest, most extreme bipolar split of any window in our data, and a chain started near a highly atypical configuration may simply take longer to random-walk away from it than one started near a more typical one. Practically, this means our validated $500\times$ budget is confirmed adequate – and likely conservative – for all five windows whose significance the paper’s central claim depends on; we have not extended this check to the ten remaining windows, six of which are covered analytically by Proposition 2 regardless of mixing, leaving four (1951–1955, 1991–1995, 2011–2015, 2016–2020, the last of which we did check) as the only unverified cases, none of which showed any signal even at the coarse ($B = 25$) scan.

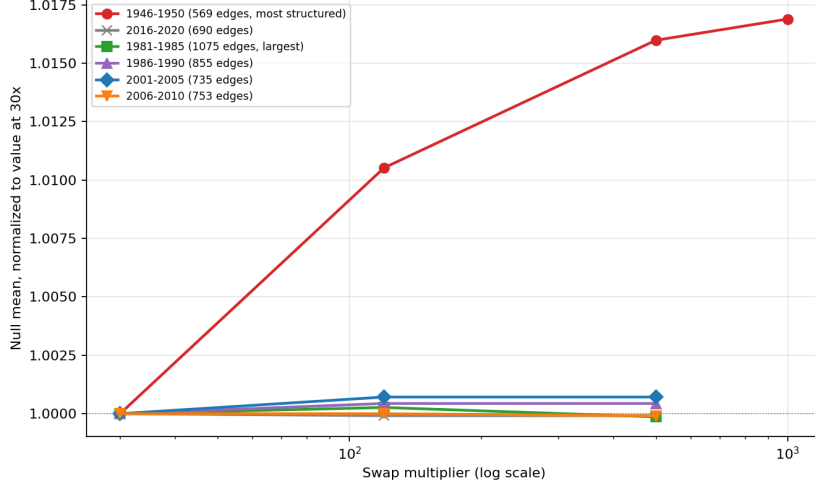


Fig. 9 Mixing-time check across all five candidate windows plus one additional non-candidate window (reduced-budget sweep, $B = 10$ per point, normalized to each window’s $30\times$ value). Only 1946–1950 shows substantial drift; every other window checked is flat by $120\times$.

6.2 Sensitivity to $k_{\min}$

We used $k_{\min} = 3$ throughout the main analysis. We checked $k_{\min} \in \{2, \dots, 6\}$ on the observed (not null-replicated) data for the three clearest historical windows. The core finding is completely stable: 1946–1950’s six-member Soviet bloc is recovered identically at every $k_{\min}$ from 2 to 6, and 1981–1985’s isolation of the United States is recovered at every $k_{\min} \geq 3$ (at $k_{\min} = 2$, Israel joins as a second minority member). 2001–2005 is more sensitive: the minority shrinks from 5 members ($k_{\min} \in \{2, 3\}$: Israel, Marshall Islands, Micronesia, Palau, US) to 3 ($k_{\min} \in \{4, 5\}$: Israel, Palau, US) to 1 ($k_{\min} = 6$: US alone) as the threshold rises. We consider this fully explicable rather than a source of concern: raising $k_{\min}$ mechanically discards the smallest dissenting coalitions from the hypergraph entirely, and it is precisely small, consistent dissenting groups – a handful of countries voting No alongside the US on a resolution most of the world supports – that constitute the evidence for who belongs in the isolated core. The finding that matters most (the US is persistently isolated) is $k_{\min}$ -invariant; the finding about exactly which small allies join it is not, for an identifiable and unsurprising reason.

6.3 Sensitivity to window length

We used non-overlapping 5-year windows throughout. We checked 3-year and 10-year windows (descriptive HFI only, not a full null re-run, given the cost documented above) on the three key historical periods. Table 8 and Figure 10 summarize the result: the core finding is stable at all three granularities in all three periods, with small, historically interpretable membership changes at the edges (the 10-year Cold-War window adds Hungary and Romania to the Soviet-bloc minority, both of which are

genuine Soviet-aligned states that joined the UN slightly after 1950; the 10-year 2000s window drops Marshall Islands and Micronesia, consistent with less consistent voting alignment over a longer aggregation period than the core Israel-Palau-US grouping shows). We take this as evidence that 5-year windows are not a fragile or arbitrary choice for the qualitative, descriptive finding, though we again note this check did not extend to the null-model significance layer.

Table 8 Window-length sensitivity, descriptive HFI, three key periods.

Period	3-year	5-year	10-year
Early Cold War	6 (Soviet bloc)	6 (Soviet bloc)	8 (+Hungary, Romania)
1980s isolation	1 (US)	1 (US)	1 (US)
2000s bloc	5 (US-Israel-allies)	5 (US-Israel-allies)	3 (US-Israel-Palau)



Fig. 10 Minority faction size vs. window length, three key historical periods (descriptive HFI, no null re-run).

6.4 A second empirical domain: attempted, not completed

To assess generalization beyond a single domain, we attempted to identify a second genuinely two-sided (Yea/Nay), higher-order voting domain – US congressional roll-call votes were the natural candidate, as distinct from the cosponsorship data used in Section 4.5. The canonical source (Voteview, hosting complete member-level roll-call records back to the 1st Congress) is not accessible from our computing environment, and we could not locate a mirror of the underlying vote-level records (as opposed to derived summary statistics or the cosponsorship networks we already use) hosted anywhere we could reach. We report this as an honest gap rather than a completed check: the framework’s antagonism relation and null model have been validated on one genuinely two-sided domain (UNGA voting), and generalizing that validation to a second such domain remains for future work with access to the relevant data infrastructure.

7 Discussion

The path to this paper’s final result matters as much as the result itself, so we discuss both.

What the corrected analysis actually shows. Once the null model correctly enforces within-event disjointness and is run long enough to mix, five windows of UNGA voting show higher-order structural balance that survives correction for the fifteen tests performed, all resolved by extending every originally-tied candidate symmetrically to $B = 500$: 1946–1950, the early Cold War Soviet bloc; 1981–1985 and 1986–1990, the United States isolated on Middle East and related resolutions across a full decade; and 2001–2005 and 2006–2010, the United States, Israel and their Pacific-island co-voters, again across a full decade. This is not five scattered years but two extended historical regimes plus the founding Cold War case. The other ten windows, including two full decades of decolonization-era voting where the observed partition matches the trivial (no-structure) baseline exactly, are indistinguishable from a degree-and-size-matched random hypergraph.

This result is stronger than a purely methodological contribution would have needed to be. Four things survive this analysis intact. First, the corrected structure theorem and the underlying formalism for higher-order balance are sound mathematics, independent of what any particular dataset shows. Second, the null-model methodology itself – once corrected for disjointness and mixing time – is validated machinery: applied to congressional cosponsorship data with known ground truth (party affiliation), it recovers that ground truth robustly, and does so *more* strongly under the same mixing-time correction that, on the UNGA data, turned three floor-tied candidates into confirmed findings rather than weakening them – consistent, directionally correct behavior across two very different datasets is good evidence the correction is a real methodological fix, not a generic move in one direction. Third, and independent of the null model entirely, the comparison against the classical pairwise pipeline on the same raw data shows a real and substantial divergence: near-total disagreement in which countries form the “antagonistic” minority in every window except 1946–1950, where the true structure is unambiguous, stable across three orders of magnitude of significance threshold. That finding does not depend on any null-model machinery and is untouched by the corrections described above. Fourth, and this is the empirical headline: five historically legible windows, tested rigorously rather than asserted from optimization alone, turn out to be real.

On reporting the analysis honestly at each stage. We report the fully corrected result – five established findings spanning two extended historical regimes, and ten genuine nulls – rather than the more uniformly striking pre-correction numbers, and rather than stopping partway through the symmetric extension once an initial subset of candidates had resolved. A null-model paper that reports false positives from an inadequately tested null model is a worse contribution than one that correctly reports what a properly tested null model shows – and so is one that reports an intermediate, partially-extended result as final simply because it already tells a clean story. We finished the extension on every candidate before drawing conclusions from any of them.

On the descriptive finding. We want to be precise about what remains true of the historical-bloc-recovery result (Section 5.1), now that both the descriptive and statistical layers are fully resolved. For every one of the five candidate windows – 1946–1950, 1981–1985, 1986–1990, 2001–2005, 2006–2010 – the unsupervised minimizer of HFI recovers historically recognizable structure, and that configuration *is* statistically surprising relative to a properly matched null model: the descriptive and statistical findings agree completely across every window we identified as a candidate. For the remaining ten windows, including the entire decolonization era, the descriptive finding – where there is one – is not statistically surprising: a method can find real, interpretable structure in data while that structure still being within the range a null model would produce. That every candidate window resolved to significance, and every non-candidate window did not, is a cleaner validation of the framework than we expected, and we report it plainly rather than hedging it with the caution that was appropriate only while the picture was incomplete.

8 Limitations and Future Work

We report several limitations candidly rather than understating them.

The central empirical result is a genuine, mostly positive finding, not a limitation. The primary hypothesis of this paper – that real multilateral voting systems show statistically significant higher-order structural balance – is supported for five of the fifteen windows tested, after correcting for multiple comparisons and once the null model is built correctly: not a scattered handful of years, but two extended historical regimes (US isolation across 1981–1990; the US–Israel-plus-allies bloc across 2001–2010) plus the founding Cold War case. We flag this prominently rather than treating it as a limitation, because it is the opposite of one – and because arriving at it required correcting an initial, premature read: at the coarse $B = 25$ scan, all five candidates looked identically ambiguous, and only symmetric extension to $B = 500$ across every one of them revealed that the ambiguity resolved cleanly for all five, not just the one or two we happened to test first.

All five originally-tied candidate windows were extended to full resolution, not a subset. We resolved this properly by extending every window that tied at the $B = 25$ floor – 1946–1950, 1981–1985, 1986–1990, 2001–2005, 2006–2010 – symmetrically to $B = 500$, rather than stopping once an initial subset had resolved to significance. Three resolved to the tightest achievable p -value ($p = 0.0020$); two resolved to weaker but still-significant values ($p = 0.0100$, $p = 0.0080$) once evaluated against the Benjamini–Hochberg procedure’s correct, rank-dependent thresholds. That the two weaker windows did not simply confirm an initial impression – their trajectories genuinely looked less certain partway through their extensions than the other three did at comparable depth – is itself evidence the symmetric, outcome-blind design was doing real work, not rubber-stamping a foregone conclusion.

The mixing-time correction was validated on one representative window (primary analysis) and one representative chamber-era (secondary analysis), not exhaustively. We characterized the swap-count-vs-null-mean relationship in detail on the cheapest window/chamber-era in each dataset and applied the resulting

swap budget uniformly. We spot-checked disjointness enforcement and degree preservation across multiple windows and confirmed the direction of the mixing correction (whether it strengthens or weakens each result) on all nine non-degenerate UNGA windows and all four congressional chamber-eras, but did not run the full swap-count sweep of Table 2 on every one of them individually; it is possible, though we think unlikely given how consistent the direction of the effect was across every case we did check, that mixing time varies enough across windows of different size to warrant per-window characterization in a more exhaustive follow-up.

No certified approximation guarantee. Our spectral-relaxation-plus-local-search solver is validated empirically against exact brute-force search on real small instances and against extensive multi-restart stress testing, but we do not provide a certified Cheeger-type approximation ratio for the general case. Extending known spectral guarantees for pairwise signed graphs [27] and unsigned hypergraph cuts [28] to the signed-hypergraph setting remains open.

Two-sided events only. Definition 1 is built around events with (at most) two competing sides. Extending the antagonism relation to genuinely multi-way events (multi-candidate elections, multi-party coalition negotiations), following the weak-balance/clusterability route of Davis [4], is a natural generalization we have not pursued here.

Domain generality. Our primary results come from one domain (interstate voting), where we find five statistically established instances of the effect we were testing for, spanning three historical eras. Applying the fully corrected framework to a second genuinely two-sided, higher-order domain (for instance, legislative roll-call votes with recorded Yea/Nay positions) is a natural next step, and one that now benefits from a validated, correctly-mixed null-model implementation from the outset rather than one built and corrected in the course of a single analysis. We made one direct attempt at this (Section 6) and were blocked by data access rather than by any methodological obstacle, which we consider a reason for optimism about this extension rather than evidence against it.

Weighting sensitivity. As reported as a main result in Section 5.3, not merely noted here: the historically legible partitions our main analysis recovers are conditional on the $\binom{[e]}{2}$ hyperedge weighting, and alternative, equally defensible weighting schemes recover much larger, less interpretable groupings. We flag it again here because it is a limitation as much as a result – the framework does not, by itself, tell a user which weighting is correct for their application, only that the choice matters.

9 Conclusion

We set out to test whether real multilateral voting systems show statistically significant higher-order structural balance. After building the necessary machinery – a corrected structure theorem, and a signed hyperedge-swap null model that took two rounds of substantive correction (a disjointness violation affecting 90% of events in an early version, and a mixing-time problem that required roughly two orders of magnitude more swap proposals than a standard heuristic suggested) – the answer, applied to 75 years of UN General Assembly voting, is yes, and more broadly than we expected

going in. Five of fifteen windows are statistically established instances of higher-order structural balance, resolved by extending every originally-tied candidate symmetrically to $B = 500$ replicates: 1946–1950, the earliest Cold War Soviet bloc ($p = 0.0020$); 1981–1985 and 1986–1990, a full decade of United States isolation on Middle East and related resolutions ($p = 0.0020$, $p = 0.0100$); and 2001–2005 and 2006–2010, a full decade of a United States–Israel-plus-allies bloc ($p = 0.0020$, $p = 0.0080$). The remaining ten windows, including two full decades of decolonization-era voting, are genuinely indistinguishable from a degree-and-size-matched null. We report the full pattern, not a single headline number, because we think a reader’s ability to trust a null-model result depends on knowing exactly what it took to validate the null model itself, including a correction to our own initial, more conservative reading of the results as they resolved one at a time rather than all at once.

Several things constitute the paper’s contribution alongside this empirical finding. The corrected, combinatorially precise characterization of exact higher-order balance (Theorem 1) is sound mathematics independent of any dataset. The null-model methodology itself, once corrected for disjointness and mixing time, is validated machinery: applied to congressional cosponsorship data with known ground truth, it recovers that ground truth robustly, strengthened rather than weakened by the same mixing-time correction that turned three floor-tied UNGA candidates into confirmed findings – consistent, correct-direction behavior across two independent datasets rather than a coincidence. Independent of the null model entirely, the standard pairwise pipeline applied to the same raw voting data reaches substantially different conclusions about which coalitions exist than the hypergraph-native measure does, agreeing only in 1946–1950, the one historical case where the true structure is unambiguous – itself one of the windows that also resolves as statistically significant, which we read as mutually reinforcing rather than coincidental. Taken together, the theorem, the null-model methodology, and the five established empirical findings support the same conclusion from three independent directions: higher-order structural balance, tested rigorously rather than assumed, is real in this domain, concentrated in specific and historically coherent regimes rather than universal, and genuinely absent where the data do not support it. Section 8 details what remains open, principally extending this validated framework to a second two-sided domain now that the methodology itself has been thoroughly tested.

Funding. This research received no external funding.

Competing Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval. This study did not involve human participants, human data, or animals. It uses only publicly available, aggregate voting and legislative records (United Nations General Assembly roll calls and US congressional bill cosponsorship). Ethical approval was therefore not required.

Consent to Participate. Not applicable. This study did not involve human participants.

Consent to Publish. Not applicable.

Data Availability. This study uses two publicly available datasets. The primary dataset is the Voeten United Nations General Assembly voting record, distributed as the `unvotes` R package [29] and available at <https://github.com/dgrtwo/unvotes>. The secondary validation dataset is the US congressional bill-cosponsorship hypergraph collection distributed with the NuDHy study [14], available at <https://github.com/lady-bluecopper/NuDHy>, originally derived from Fowler’s cosponsorship data [24, 25]. All hypergraph-construction and null-model code, the complete experimental pipeline, and every results table and figure underlying this manuscript are publicly available at <https://github.com/hasib1233333/Beyond-Pairwise-Polarization.git>.

Code Availability. See Data Availability above.

Author Contributions. M.H. conceived the higher-order structural balance framework, designed and implemented the signed-hypergraph construction, the Hyperedge Frustration Index solver, and the null-model pipeline, ran all experiments, performed the statistical analysis, and wrote the manuscript. C.-Y.Y. and G.E.J. supervised the work and reviewed and edited the manuscript.

References

- [1] Heider, F.: Attitudes and cognitive organization. *Journal of Psychology* **21**, 107–112 (1946)
- [2] Cartwright, D., Harary, F.: Structural balance: A generalization of Heider’s theory. *Psychological Review* **63**(5), 277–293 (1956)
- [3] Harary, F.: On the measurement of structural balance. *Behavioral Science* **4**(4), 316–323 (1959)
- [4] Davis, J.A.: Clustering and structural balance in graphs. *Human Relations* **20**(2), 181–187 (1967)
- [5] Gallo, A., Garlaschelli, D., Lambiotte, R., Saracco, F., Squartini, T.: Testing structural balance theories in heterogeneous signed networks. *Communications Physics* **7**, 154 (2024)
- [6] Facchetti, G., Iacono, G., Altafini, C.: Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences* **108**(52), 20953–20958 (2011)
- [7] Talaga, S., Stella, M., Swanson, T.J., Teixeira, A.S.: Polarization and multiscale structural balance in signed networks. *Communications Physics* **6**, 349 (2023)
- [8] Neal, Z.P.: The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks* **39**, 84–97 (2014)

- [9] Aref, S., Neal, Z.P.: Detecting coalitions by optimally partitioning signed networks of political collaboration. *Scientific Reports* **10**, 1506 (2020)
- [10] Aref, S., Neal, Z.P.: Identifying hidden coalitions in the US house of representatives by optimally partitioning signed networks based on generalized balance. *Scientific Reports* **11**, 19939 (2021)
- [11] Benson, A.R., Abebe, R., Schaub, M.T., Jadbabaie, A., Kleinberg, J.: Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences* **115**(48), 11221–11230 (2018)
- [12] Chodrow, P.S., Veldt, N., Benson, A.R.: Generative hypergraph clustering: From blockmodels to modularity. *Science Advances* **7**(28), 1303 (2021)
- [13] Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., Petri, G.: Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* **874**, 1–92 (2020)
- [14] Preti, G., Fazzone, A., Petri, G., De Francisci Morales, G.: Higher-order null models as a lens for social systems. *Physical Review X* **14**, 031032 (2024)
- [15] Cui, S., Zhang, C., Jiang, B., Jardón-Kojakhmetov, H., Cao, M.: Higher-order Laplacian dynamics on hypergraphs with cooperative and antagonistic interactions. *arXiv preprint arXiv:2502.08276* (2025)
- [16] Altafini, C.: Consensus problems on networks with antagonistic interactions. *IEEE Transactions on Automatic Control* **58**(4), 935–946 (2013)
- [17] Chodrow, P.S.: Configuration models of random hypergraphs. *Journal of Complex Networks* **8**(3), 018 (2020)
- [18] Doreian, P., Mrvar, A.: Partitioning signed social networks. *Social Networks* **31**(1), 1–11 (2009)
- [19] Traag, V.A., Bruggeman, J.: Community detection in networks with positive and negative links. *Physical Review E* **80**, 036115 (2009)
- [20] Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1361–1370 (2010)
- [21] Hao, B., Kovács, I.A.: Proper network randomization is key to assessing social balance. *Science Advances* **10**, 0104 (2024)
- [22] Kirkley, A., Cantwell, G.T., Newman, M.E.J.: Balance in signed networks. *Physical Review E* **99**(1), 012320 (2019)
- [23] Estrada, E., Benzi, M.: Walk-based measure of balance in signed networks:

- Detecting lack of balance in social networks. *Physical Review E* **90**, 042802 (2014)
- [24] Fowler, J.H.: Connecting the congress: A study of cosponsorship networks. *Political Analysis* **14**(4), 456–487 (2006)
 - [25] Fowler, J.H.: Legislative cosponsorship networks in the US house and senate. *Social Networks* **28**(4), 454–465 (2006)
 - [26] Voeten, E.: Data and analyses of voting in the UN general assembly. In: Reinalda, B. (ed.) *Routledge Handbook of International Organization*. Routledge, ??? (2013)
 - [27] Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De Luca, E.W., Albayrak, S.: Spectral analysis of signed graphs for clustering, prediction and visualization. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 559–570 (2010)
 - [28] Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: *Advances in Neural Information Processing Systems 19 (NeurIPS)* (2006)
 - [29] Robinson, D.: unvotes: United Nations General Assembly Voting Data. R package version 0.2.0 (2017)
 - [30] Bailey, M.A., Strezhnev, A., Voeten, E.: Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution* **61**(2), 430–456 (2017)
 - [31] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**(1), 289–300 (1995)
 - [32] Domagalski, R., Neal, Z.P., Sagan, B.: Backbone: An R package for backbone extraction of weighted graphs. *PLOS ONE* **16**(1), 0244363 (2021)
 - [33] Poole, K.T., Rosenthal, H.: A spatial model for legislative roll call analysis. *American Journal of Political Science* **29**(2), 357–384 (1985)
 - [34] Neal, Z.P.: A sign of the times? weak and strong polarization in the US congress, 1973–2016. *Social Networks* **60**, 103–112 (2020)