Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Combinatorial constraints on novelty priors

Łukasz Konowałek

lukasz.konowalek@wum.edu.pl

Medical University of Warsaw

Research Article

Keywords: category formation, combinatorial model, innovation probability, category density, baseline modeling

Posted Date: May 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9646899/v1

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

Additional Declarations: No competing interests reported.

Combinatorial constraints on novelty prior

Author

Łukasz Konowałek

Affiliation

Department of Child Psychiatry
Medical University of Warsaw
Warsaw, Poland

Corresponding author

Łukasz Konowałek
Department of Child Psychiatry
Medical University of Warsaw
ul. Zwirki i Wigury 61
02-091 Warsaw, Poland
Email: lukasz.konowalek@wum.edu.pl

Keywords

category formation; combinatorial model; innovation probability; category density; baseline modeling

Statements and Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest/Competing interests. The author declares no competing interests.

Availability of data and material. All datasets used in this study are publicly available. The GUM corpus is available at https://github.com/amir-zeldes/gum. The Kinetics-400 dataset is available at https://github.com/cvdfoundation/kinetics-dataset. The Switchboard Dialog Act corpus is available at https://web.stanford.edu/~jurafsky/ws97.

Code availability. Analysis scripts sufficient to reproduce the published results are included in the project archive.

Authors' contributions. Łukasz Konowałek: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Combinatorial constraints on novelty prior

Abstract

Novelty detection in online category formation is often modeled as partition extension: each new observation either joins an existing category or initiates a new one. Bayesian species-sampling models, such as the Chinese Restaurant Process and the Pitman–Yor process, formalize this problem through prior-predictive rules over partitions. But how much of the resulting novelty bias reflects the specific prior, and how much follows from the generic geometry of partition space itself? We derive a combinatorial structural score from Stirling-number contrasts and use it as a prior-like predictor of innovation. Across lexical, referential, dialogue-act, and perceptual event-boundary datasets, this score explained a substantial portion of the fitted Pitman–Yor logit surface–not the raw innovation decisions themselves: 68–79% using the combinatorial score alone, and 78–98% when combined with a finite-size correction. In held-out prediction of observed innovation decisions, the resulting model performed competitively with type–token, CRP, and Pitman–Yor alternatives, with the clearest Pitman–Yor advantage appearing in dialogue acts. These results suggest that part of the apparent prior over innovation in species-sampling models reflects domain-general combinatorial geometry, while deviations from this geometry identify additional sequential or representational structure.

Keywords: category formation, combinatorial model, innovation

1. Introduction

Online processing requires continual decisions about novelty. As experience unfolds, each new observation must be related to what has already been encountered: it may be assigned to an existing category, episode, referent, action class, or event, or it may require the creation of a new one. This problem arises across domains, from lexical processing and discourse tracking to action perception and event segmentation. In each case, novelty detection is not merely the detection of an outlier. It is a decision about how an actively growing partition of experience should be extended.

Machine learning approaches to novelty detection typically define novelty operationally, for example in terms of distance from known observations, low density under a learned distribution, high surprisal, or reconstruction error (Pimentel et al., 2014). These approaches are useful for identifying observations that depart from a learned model, but they do not by themselves explain how a system should update the inventory of categories during online processing. A category system is not only a classifier over known labels; it is also a partition that can grow.

Bayesian nonparametric models provide one of the clearest formal treatments of this problem. In species-sampling models, including the Chinese Restaurant Process and the Pitman–Yor process, each new observation either joins an existing block of a partition or initiates a new block (Pitman, 2006; Teh, 2006). This framework has been influential in cognitive modeling because it provides a principled way to assign probability to previously un-

seen categories. In particular, Navarro and Kemp (2017) showed that people are sensitive not only to similarity, but also to the distribution of previous observations across categories when deciding whether an item belongs to "none of the above." On this view, novelty detection is naturally modeled as prior-predictive partition extension.

However, the psychological interpretation of such priors is not straightforward. In the Pitman–Yor process, the probability of opening a new category is determined by a strength parameter, a discount parameter, the number of observations, and the number of occupied categories. These parameters are often interpreted as capturing biases toward innovation, reinforcement, or heavy-tailed category growth. Yet random partition models are themselves constrained by the combinatorial structure of partition space. The connection between exchangeable partition priors and Stirling-type combinatorics is well established in probability theory (Gnedin & Pitman, 2006; Pitman, 2006). Closely related statistical work uses generalized Stirling-number asymptotics to approximate Gibbs partition weights and discovery probabilities (Cerquetti, 2012), and recent Bayesian nonparametric work has used Stirling-number structure to construct priors over latent partitions or to model sequential discoveries directly (Zito et al., 2023, 2024). More recent work on random partition models has similarly emphasized that properties such as rich-get-richer behavior, cluster-size balance, and prior distributions over the number of clusters can follow from general structural assumptions rather than from domain-specific mechanisms alone (Lee & Sang, 2022).

This raises a question for cognitive modeling. When a Pitman–Yor or CRP-style model predicts a particular novelty rate, how much of that predic-

tion reflects a substantive psychological assumption about novelty, and how much reflects the generic geometry of the partition space? Put differently, before interpreting prior parameters as cognitive biases, we need a baseline for what partition extension looks like under the combinatorial constraints of the problem itself.

Here we ask how much of the Pitman–Yor prior-predictive novelty bias can be recovered from a simple combinatorial score. The score is derived from contrasts between Stirling numbers of the second kind, which count the number of ways to partition N observations into K non-empty categories. It is not proposed as a replacement for species-sampling priors. Rather, it is used as a diagnostic tool: it allows us to ask how much of the predictive surface of Pitman–Yor-style novelty models can be recovered from partition geometry alone, and how much remains as residual structure requiring additional explanation. The contribution is therefore diagnostic rather than generative: the goal is to interpret what existing novelty priors are capturing, not to introduce a new complete process model of category formation.

We evaluate this idea in two ways. First, we compare the combinatorial score with fitted Pitman–Yor predictions to quantify the portion of prior-predictive novelty bias that can be explained by partition geometry. Second, we test whether a calibrated version of this score predicts innovation in naturalistic online streams, including lexical types, coreference chains, dialogue acts, and perceptual event boundaries. These datasets differ in their representational content, but they share a common formal structure: observations arrive sequentially and are assigned either to an existing equivalence class or to a newly created one.

Our aim is therefore not to show that combinatorics replaces Bayesian nonparametric modeling. Instead, we ask what part of Bayesian nonparametric novelty models is already implied by the geometry of the hypothesis space. The answer matters for theory: if a large fraction of Pitman–Yor novelty bias is recovered from a simple partition-geometric score, then fitted prior parameters should not be interpreted as pure measures of psychological novelty preference. They also absorb generic constraints imposed by the combinatorics of partition growth.

2. Pitman-Yor process vs. combinatorics

The Pitman–Yor process is the natural point of comparison because it already treats novelty as partition extension. After the random discrete measure is marginalized out, the process induces a sequential predictive rule over the current partition. If N observations have been assigned to K occupied blocks with block sizes $n_1, \ldots, n_K$, then the next observation joins an existing block j with probability

$$P(y_{N+1} = j \mid \Pi_N) = \frac{n_j - d}{\theta + N},$$

and opens a new block with probability

$$P(y_{N+1} = \text{new} \mid \Pi_N) = \frac{\theta + dK}{\theta + N}.$$

Here θ is the strength parameter and d is the discount parameter. The Chinese Restaurant Process is recovered when $d = 0$ (Pitman, 2006; Teh, 2006). In psychological applications, this rule is attractive because it gives an explicit probability that the next item belongs to "none of the above" (Navarro & Kemp, 2017).

This predictive rule is not separate from combinatorics. The Pitman–Yor process is part of a broader family of exchangeable random partitions whose probabilities can be described by exchangeable partition probability functions. In the Gibbs partition framework, the probability of a partition factors into weights associated with block sizes and weights associated with the total number of observations and blocks. The consistency conditions for such partitions are governed by generalized Stirling triangles (Gnedin & Pitman, 2006; Pitman, 2006). Thus, the connection between Pitman–Yor processes and Stirling-type combinatorics is not new. The question pursued here is narrower: when a fitted Pitman–Yor model produces a novelty surface over (N, K), how much of that surface can be recovered from a simple combinatorial contrast that does not contain the Pitman–Yor parameters?

We begin with the unweighted counting case. The number of ways to partition N labeled observations into K non-empty, unlabeled blocks is given by the Stirling number of the second kind, $S(N, K)$. These numbers satisfy

$$S(N+1, K) = KS(N, K) + S(N, K-1).$$

The recurrence separates two combinatorial contributions to a partition of $N+1$ observations into K blocks. The term $KS(N, K)$ counts extensions in which the new observation is absorbed into one of K already existing blocks. The term $S(N, K-1)$ counts extensions in which the new observation is a singleton block, while the previous N observations formed only $K-1$ blocks. This recurrence can be given a probabilistic interpretation if one samples uniformly from partitions with a fixed final number of blocks. However, that is a retrospective distribution over completed partitions, not the Pitman–Yor prior-predictive question of whether an active partition with K blocks should

extend to $K + 1$ blocks.

We therefore use the recurrence not as a literal online probability, but as a source of a structural contrast. The contrast compares the combinatorial contribution associated with singleton creation to the contribution associated with absorption into the current K-block state:

$$x_{\text{comb}}(N, K) = \log S(N, K - 1) - \log S(N, K) - \log K. \tag{1}$$

This quantity is a log-combinatorial score, not a new exchangeable partition prior. Larger values indicate that, under the counting geometry of set partitions, the current state is closer to singleton-generating structure than to absorption-generating structure. In the empirical analyses below, the score is always evaluated on the pre-decision state (N_t, K_t) of the online sequence.

For empirical use, the exact Stirling expression is inconvenient because the relevant counts grow very rapidly. Using a saddle-point approximation for large N and K (Temme, 1993), the structural score can be written as a function of structural density, $\rho = N/K$, plus a finite-size correction:

$$x_{\text{comb}}(N, K) = g(\rho) + O\left(\frac{1}{N}\right). \tag{2}$$

The function $g(\rho)$ is defined in Appendix A. Intuitively, it captures the density-dependent pressure toward absorption: as more observations are packed into fewer categories, further innovation becomes less likely. The approximation is used for $\rho > 1$. On the boundary $\rho = 1$, where all previous observations occupy distinct categories, we use the exact Stirling contrast

$$x_{\text{comb}}(N, N) = \log \frac{N - 1}{2},$$

obtained from $S(N, N) = 1$ and $S(N, N - 1) = \binom{N}{2}$. States with $N \leq 1$ do not yield a finite contrast and are omitted from model fitting.

This score allows two related analyses. The first is a diagnostic analysis of the Pitman–Yor surface itself. For each dataset, we fit the Pitman–Yor parameters and compute the corresponding prior-predictive logit

$$\eta_{\mathrm{PY}}(N, K) = \operatorname{logit}\left(\frac{\theta + dK}{\theta + N}\right).$$

We then ask how much of this fitted surface can be recovered from the combinatorial score alone:

$$\eta_{\mathrm{PY}}(N, K) = \alpha_0 + \alpha_1 x_{\mathrm{comb}}(N, K) + \varepsilon,$$

and how much is recovered when the leading finite-size correction is also included:

$$\eta_{\mathrm{PY}}(N, K) = \alpha_0 + \alpha_1 x_{\mathrm{comb}}(N, K) + \alpha_2 \frac{1}{N} + \varepsilon.$$

This analysis does not claim that the combinatorial score is the Pitman–Yor process in another notation. Instead, it asks whether the fitted Pitman–Yor novelty bias is largely aligned with the geometry of partition space, or whether it contains substantial residual structure beyond that geometry.

The second analysis treats innovation decisions in the data directly. Real category systems may contain categories that are more salient, more accessible, or more attractive than others. We therefore map the structural score to observed online decisions using a calibrated logit model:

$$\operatorname{logit} P(y_{t+1} = 1 \mid N_t, K_t) = \beta_0 + \beta_1 x_{\mathrm{comb}}(N_t, K_t) + \beta_2 \frac{1}{N_t}.$$

For empirical tractability we use the approximation $x_{\mathrm{comb}}(N, K) \approx g(\rho)$ for $\rho > 1$ and the exact boundary value for $\rho = 1$, with the finite-size correction represented by $1/N$. Here, β_0 captures average category attractiveness, β_1

captures sensitivity to the structural score, and β_2 captures the finite-size correction. The weighting assumptions and the derivation of this form are given in Appendix B.

3. Method

3.1. Data and online partition states

The empirical analyses used lexical and coreference annotations from the Georgetown University Multilayer corpus (GUM) (Zeldes, 2017), dialogue-act annotations from the Switchboard Dialog Act corpus (SWDA) (Jurafsky et al., 1997), and perceptual event-boundary annotations for Kinetics-400 clips (Kay et al., 2017; Shou et al., 2021). Table 1 summarizes the operationalization used in each dataset.

Dataset	Observation	Category	Sequence unit	Observations	Sequences	Mean innovation
GUM-LEX	word token	lemma	document	267,799	281	0.34
GUM-COREF	mention	entity chain	document	75,416	281	0.52
SWDA	utterance	dialogue act	conversation	212,428	1,115	0.14
KINETICS	sampled frame	event segment	annotator–clip sequence	2,351,717	92,607	0.19

Table 1: Datasets and operational definitions used in the empirical analyses.

For each sequence, observations were processed in their original order. The predictors for decision $t + 1$ were computed only from the pre-decision state: the number of previous observations N_t, the number of active categories K_t, structural density $\rho_t = N_t/K_t$, and $1/N_t$. The outcome $y_{t+1} = 1$ indicated innovation; $y_{t+1} = 0$ indicated absorption.

3.2. Analyses

The first analysis decomposed fitted Pitman–Yor predictions. We estimated the Pitman–Yor parameters on training sequences, computed the fitted prior-predictive logits on held-out states, and regressed those logits on $x_{\mathrm{comb}}(N, K)$, first alone and then with $1/N$ added.

The second analysis evaluated prediction of observed innovation decisions. All models were trained and tested on the same sequence-level splits, so the same document, conversation, or annotator–clip sequence did not appear in both training and test sets. We compared the combinatorial model with type–token, CRP, and Pitman–Yor alternatives using held-out mean log loss and AUC. Calibration was assessed by binning held-out observations by structural density ρ and comparing observed and predicted innovation rates within each bin. For SWDA, we additionally ran a diagnostic speaker-state analysis in which the active partition was computed separately for the A and B sides of each conversation. This analysis tested the sensitivity of all models to the definition of the active conversational state.

Finally, we conducted a simulation-based recovery check. We sampled states (N, K), generated binary innovation outcomes from the calibrated combinatorial model with fixed parameters, and refit the model to test parameter recovery.

4. Results

4.1. Pitman–Yor predictions are partly geometric

The primary analysis asked how much of the fitted Pitman–Yor novelty surface could be recovered from the combinatorial structure of the same on-

line states. The answer was substantial. Across datasets, the combinatorial score alone explained between 68% and 79% of the variance in the fitted Pitman–Yor prior-predictive logit surface (Table 2). Adding the finite-size term increased the explained variance to between 78% and 98%.

This pattern is important because the decomposition did not fit the observed innovation decisions directly. It first fit the Pitman–Yor process and then asked whether the resulting Pitman–Yor predictor could be reconstructed from partition geometry. The high R^2 values therefore show that much of the functional shape of the fitted Pitman–Yor novelty bias is aligned with the geometry of set partitions. The result is strongest in GUM-COREF, where x_{comb} and $1/N$ together explained 98% of the fitted Pitman–Yor surface, but the same qualitative pattern appeared in all four domains. Figure 1 shows where this diagnostic reconstruction should be placed in the manuscript.

Dataset	R^2_x	$R^2_{x+1/N}$	Comb. LL	Δ PY	Δ +PY
GUM-lex	0.682	0.782	0.578	0.001	-0.001
GUM-coref	0.775	0.980	0.667	0.005	0.000
SWDA	0.788	0.932	0.346	-0.004	-0.004
Kinetics	0.762	0.906	0.485	0.013	-0.001

Table 2: Decomposition of Pitman–Yor predictions into combinatorial geometry on the primary held-out split used for the diagnostic analysis. R^2_x reports how much of the fitted Pitman–Yor logit surface is explained by the combinatorial score alone; $R^2_{x+1/N}$ adds the finite-size term. Comb. LL is the held-out mean log loss of the combinatorial model. Δ PY is the direct Pitman–Yor mean log loss minus the combinatorial mean log loss; Δ +PY is the log-loss change after adding the fitted Pitman–Yor logit as an additional predictor to the combinatorial model. Negative values indicate improvement over the combinatorial model.

[Insert Figure 1 about here]

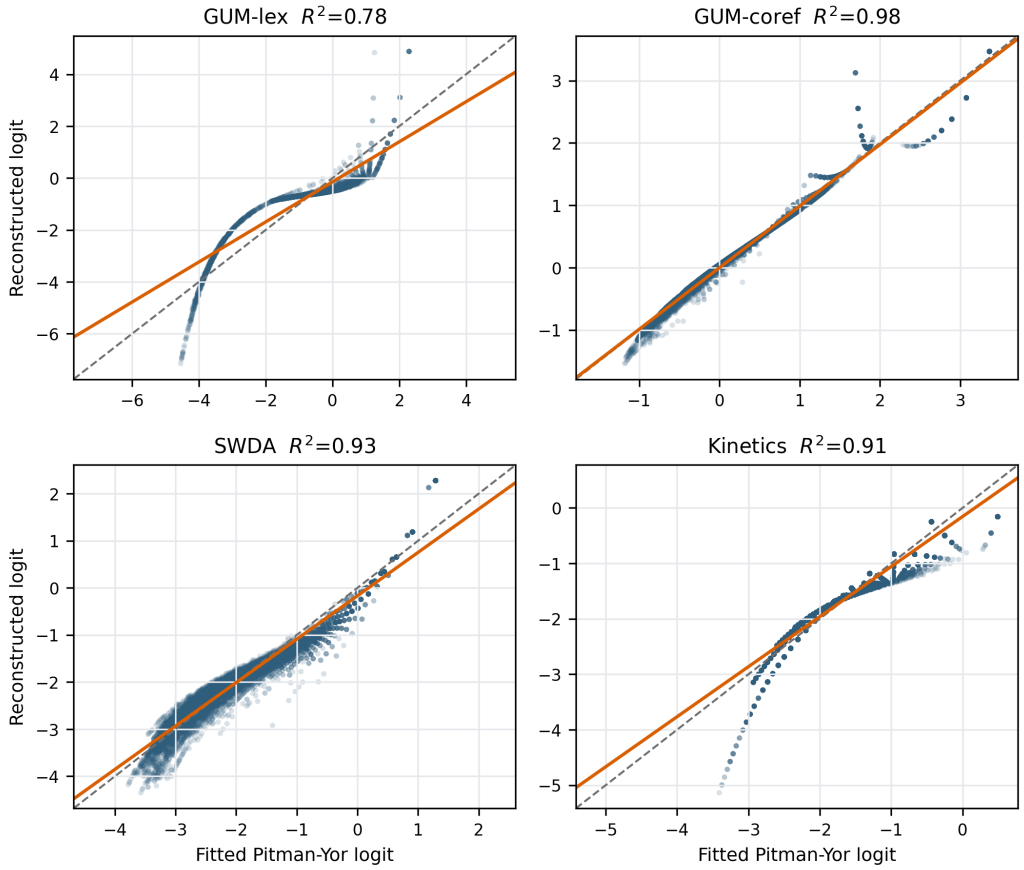


Figure 1: Diagnostic reconstruction of fitted Pitman–Yor predictions from combinatorial geometry. Each blue point is a held-out pre-decision partition state (N, K). The horizontal axis gives the prior-predictive logit assigned to that state by the fitted Pitman–Yor model; the vertical axis gives the corresponding value reconstructed from x_{comb} and $1/N$. The dashed line marks perfect reconstruction of Pitman–Yor predictions, and the solid line shows the empirical linear fit. The reported R^2 values quantify how much of the fitted Pitman–Yor predictive surface is recovered from combinatorial geometry.

The nested predictive comparison sharpened this interpretation. In GUM-

lex, GUM-coref, and Kinetics, adding the fitted Pitman–Yor logit to the combinatorial model changed held-out mean log loss by at most 0.001. In these datasets, the Pitman–Yor predictor contributed little once the combinatorial score and finite-size term were already included. SWDA was the clear exception: the Pitman–Yor term improved mean log loss by about 0.004. This residual is small in absolute terms, but it is theoretically informative because dialogue acts are organized by turn-taking and local pragmatic structure, not only by the density of the active category partition.

The main result is therefore not that the combinatorial score replaces the Pitman–Yor process. Rather, it shows that a large part of the fitted Pitman–Yor novelty surface is already recoverable from a low-dimensional description of partition geometry. The remaining deviations, most clearly visible in SWDA, mark the places where additional structure in the online state may be needed.

4.2. Prediction of naturalistic innovation

The combinatorial model was then evaluated as a direct predictor of observed online innovation decisions. It achieved the best held-out mean log loss in three of the four domains (Table 3). It substantially outperformed CRP in GUM-lex, GUM-coref, and Kinetics, and improved on TTR and Pitman–Yor in the same three datasets. The main exception was again SWDA. When dialogue was treated as a single conversation-level partition, Pitman–Yor and CRP achieved slightly lower held-out mean log loss and higher AUC than the combinatorial model.

Dataset	Model	Test obs.	Mean log loss	Δ vs. comb.	AUC	Δ AUC
GUM-lex	Combinatorial	134,258	0.581	—	0.685	—
	TTR	134,258	0.582	0.001	0.685	0.000
	CRP	134,258	0.633	0.051	0.630	-0.056
	Pitman-Yor	134,258	0.582	0.001	0.685	-0.000
GUM-coref	Combinatorial	37,809	0.667	—	0.628	—
	TTR	37,809	0.668	0.001	0.628	-0.000
	CRP	37,809	0.747	0.080	0.546	-0.083
	Pitman-Yor	37,809	0.672	0.005	0.626	-0.002
SWDA	Combinatorial	105,325	0.346	—	0.732	—
	TTR	105,325	0.350	0.005	0.727	-0.006
	CRP	105,325	0.343	-0.003	0.739	0.006
	Pitman-Yor	105,325	0.342	-0.004	0.739	0.006
Kinetics	Combinatorial	1,129,974	0.485	—	0.557	—
	TTR	1,129,974	0.504	0.019	0.557	0.000
	CRP	1,129,974	0.522	0.037	0.524	-0.034
	Pitman-Yor	1,129,974	0.503	0.018	0.558	0.000

Table 3: Held-out predictive performance by dataset and model. Mean log loss is the primary comparison metric; lower values indicate better predictive performance. The Δ columns report each alternative model's deviation from the combinatorial baseline. Positive Δ mean log loss indicates worse performance than the combinatorial model; negative values indicate better performance.

We therefore asked whether the SWDA result was sensitive to how the active conversational state was defined. The main analysis pools both speakers into one conversation-level state. However, each side of a conversation may maintain a partially distinct local repertoire of dialogue acts. We recomputed the online partition separately for the A and B sides of each conversation and refit the same models on these speaker-specific states (Table 4).

State definition and model	Test obs.	Mean log loss	AUC
Conversation-level combinatorial	105,121	0.346	0.733
Conversation-level Pitman–Yor	105,121	0.342	0.739
Speaker-specific combinatorial	104,005	0.341	0.716
Speaker-specific combinatorial + side intercept	104,005	0.341	0.716
Speaker-specific Pitman–Yor	104,005	0.337	0.727
Speaker-specific combinatorial + side intercept + PY logit	104,005	0.336	0.730

Table 4: Diagnostic analysis of state definition in SWDA. Conversation-level states treat both speakers as contributing to a single active dialogue-act partition. Speaker-specific states compute the active partition separately for the A and B sides of each conversation.

This diagnostic used the subset of held-out observations for which the speaker-specific state could be defined, so the number of test observations differs slightly from the main model comparison in Table 3. The speaker-specific analysis improved the combinatorial model, reducing mean log loss from 0.346 to 0.341. However, the same state redefinition also improved Pitman–Yor, from 0.342 in the conversation-level analysis to 0.337 in the speaker-specific analysis. Adding a side intercept did not materially change the combinatorial result, suggesting that the gain came from computing separate active partitions rather than from a simple baseline difference between A and B sides. Crucially, the speaker-specific analysis did not absorb the species-sampling residual: Pitman–Yor remained better than the speaker-specific combinatorial model, and adding the Pitman–Yor logit further reduced mean log loss to 0.336. The diagnostic therefore shows that SWDA is sensitive to state definition, but it does not remove the need for additional conversational structure.

The fitted model also reproduced the overall level of innovation in the nat-

uralistic datasets. Across held-out observations, the mean predicted probability of innovation closely matched the empirical innovation rate: 0.340 predicted versus 0.336 observed in GUM-LEX, 0.519 versus 0.508 in GUM-COREF, 0.135 versus 0.134 in SWDA, and 0.191 versus 0.190 in KINETICS. Figure 2 shows the same comparison across bins of structural density.

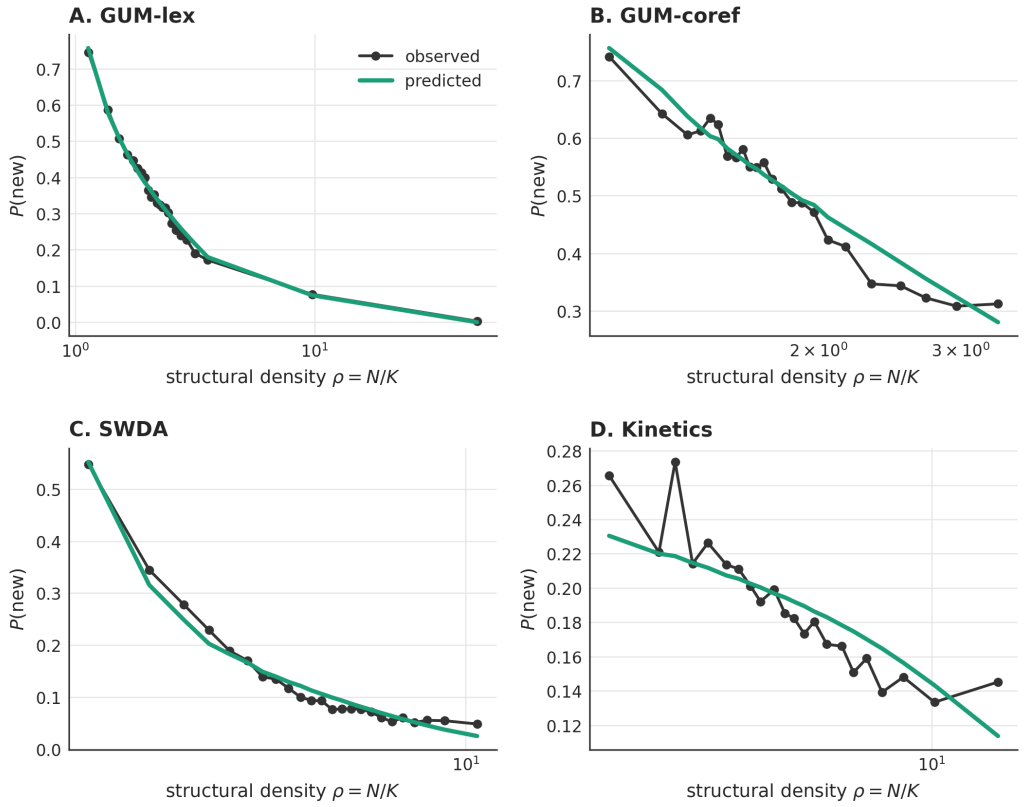


Figure 2: Calibration of naturalistic innovation rates. Points show observed innovation rates in held-out observations grouped by structural density ρ; lines show the corresponding mean predicted probability from the full combinatorial model in the same empirical bins. The plotted predictions therefore reflect the empirical distribution of both ρ and N within each bin, rather than a smooth function of ρ alone.

The calibration curves should be interpreted as calibration over the empirical distribution of online partition states, not as smooth response functions of ρ alone. This distinction is especially important for KINETICS, where short event sequences make both N and $\rho = N/K$ highly discrete; the apparent local irregularities in the predicted curve reflect the finite-size correction evaluated over these discrete states. In SWDA, the larger deviations are consistent with the diagnostic analysis above: dialogue-act innovation is sensitive to how the active conversational state is defined, but state definition alone does not explain the Pitman–Yor advantage. Figure 3 shows where the binwise performance comparison should be placed in the manuscript.

[Insert Figure 3 about here]

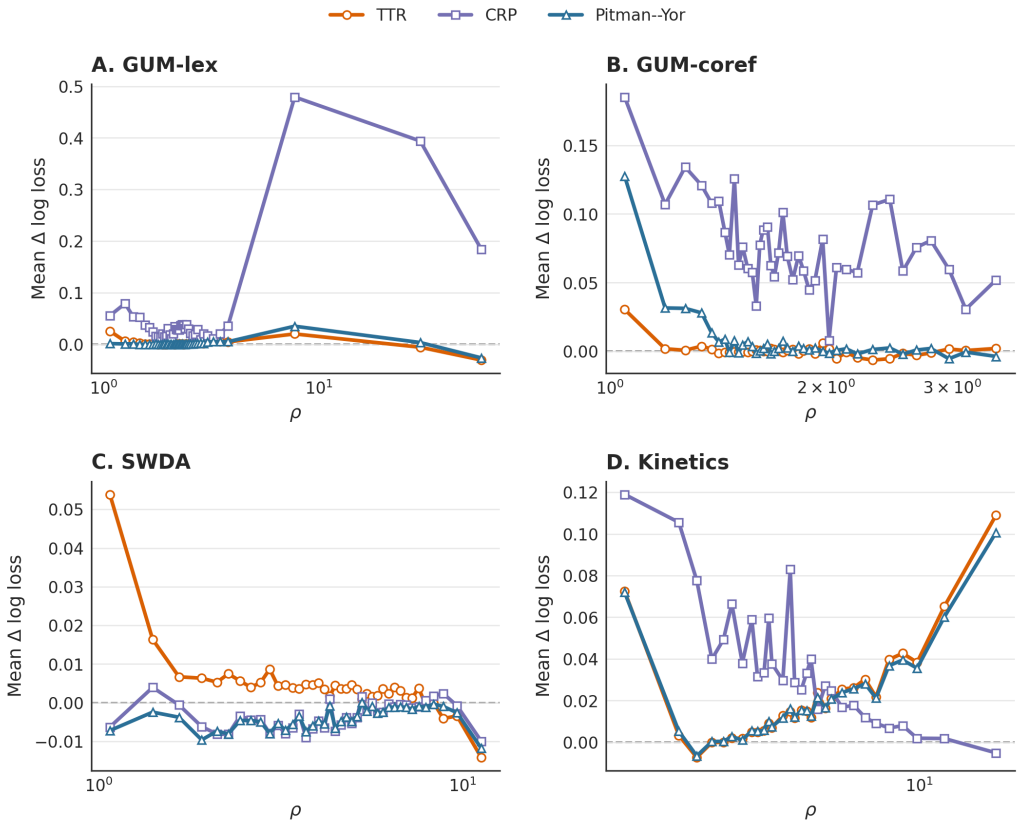


Figure 3: Mean held-out log-loss difference between each alternative model and the combinatorial baseline across quantile bins of structural density ρ. Positive values indicate lower mean log loss for the combinatorial baseline.

The binwise analysis confirms that the combinatorial advantage is regime-dependent rather than uniform. In KINETICS, the advantage is visible across broad regions of the observed density range. In GUM-LEX and GUM-COREF, the advantage is smaller but stable in aggregate. In SWDA, by contrast, the prior-predictive alternatives perform better overall in the conversation-level analysis. The speaker-specific diagnostic shows that all models benefit from a more local definition of the active partition, while the remaining

Pitman–Yor advantage points to additional conversational dependencies beyond partition density.

4.3. Simulation-based validation

Finally, the simulation-based validation checked whether the empirical model could recover known generating parameters. When data were generated from the full combinatorial model, the fitted model recovered the parameters closely: $\beta_0 = 2.00$ was recovered as 2.00, $\beta_1 = 0.80$ as 0.81, and $\beta_2 = -8.00$ as -7.02. The same simulation also showed that the generating model was distinguishable from the alternatives. The full model achieved the lowest held-out log loss, followed by the density-only combinatorial model, whereas TTR, CRP, and Pitman–Yor showed larger predictive losses. Thus, the finite-size correction can be estimated under known generating conditions, but in the empirical analyses it is treated as part of model specification rather than as a separate psychological effect. Figure 4 shows where the simulation-based validation should be placed in the manuscript.

[Insert Figure 4 about here]

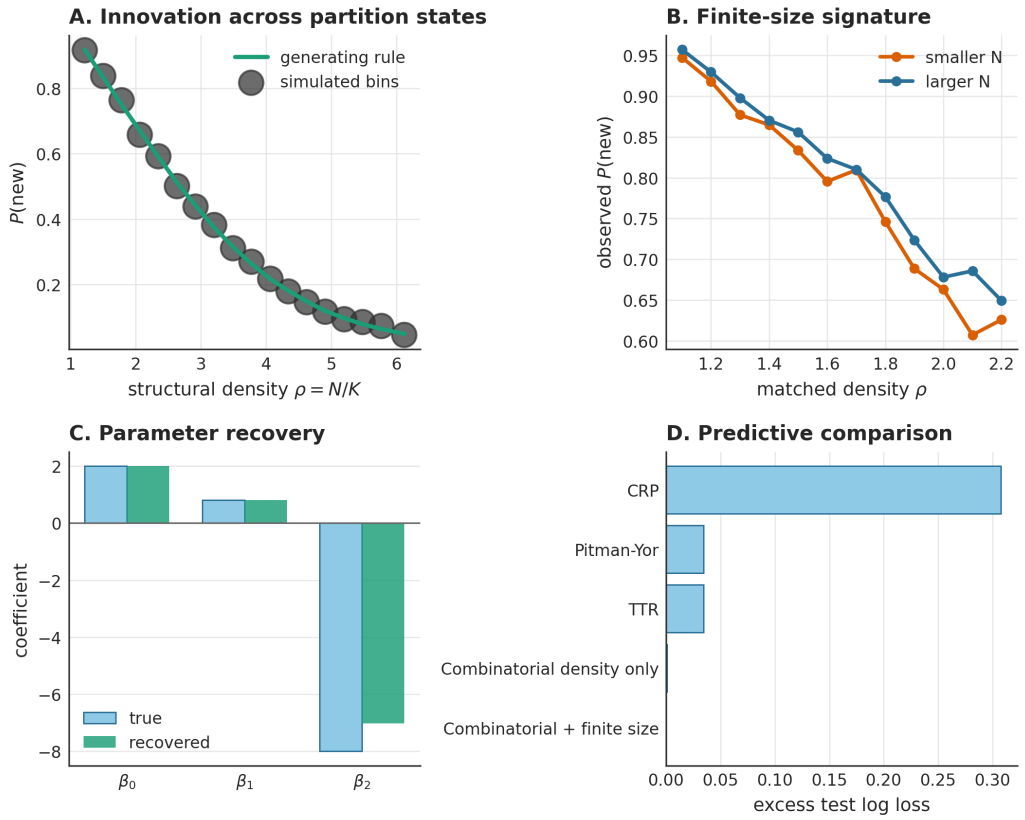


Figure 4: Simulation-based validation of the combinatorial model. Panel A shows the innovation curve generated by the model across partition states. Panel B shows the finite-size signature at matched structural density. Panel C compares true and recovered parameter values. Panel D compares held-out predictive performance against alternative models.

5. Discussion

Taken together, the results show that the combinatorial score captures a large part of the fitted Pitman–Yor novelty surface, predicts naturalistic innovation rates competitively, and is statistically recoverable under known generating conditions. The main qualification is domain-dependence: the

combinatorial account works best when innovation is well described as category growth in an active partition, and less well when the observed sequence is strongly shaped by additional sequential structure.

5.1. Interpretation of Pitman–Yor parameters

The central interpretive consequence of these results concerns the Pitman–Yor parameters. In cognitive applications, the strength parameter θ and discount parameter d are routinely read as psychological quantities: a bias toward innovation, a tendency toward rich-get-richer reinforcement, or a preference for heavy-tailed category growth. The present analysis complicates that reading. If most of the fitted Pitman–Yor predictive surface is recoverable from a low-dimensional combinatorial score, then these parameters are not pure measures of novelty preference. They also absorb the constraints imposed by the geometry of the hypothesis space. A fitted Pitman–Yor prior should therefore be interpreted as a mixture of at least two components: a generic partition-geometric component and a residual component that may reflect domain-specific psychological structure. Researchers using species-sampling models to compare individuals, tasks, or linguistic domains should treat fitted parameter differences as upper bounds on psychological differences, not as direct estimates of them.

The calibrated slope β_1 gives a more specific handle on this. In the weighted derivation, $\beta_1 = 1 - c_1$, where c_1 describes how average category attractiveness varies with combinatorial density. A value near one therefore indicates that the mean absorption weight does not change systematically with $g(\rho)$ beyond the baseline partition geometry; it does not imply that all individual category weights are uniform. Values substantially above or

below one indicate systematic density-dependent reweighting as the partition grows denser. This parameter can therefore serve as a compact domain diagnostic rather than a nuisance calibration coefficient: it measures how far a particular sequence type departs from the equal-attractiveness assumption that underlies the unweighted combinatorial baseline.

5.2. Consequences for novelty models

The most direct consequence concerns the interpretation of "distributional information" in Navarro and Kemp (2017). Navarro and Kemp argue that novelty judgments depend on the frequency distribution over previously observed categories, not only on similarity. In their words, the models differ in "how they exploit the frequency distribution over categories." The present results separate two components of that distributional information. The first is coarse partition geometry: the number of observations N, the number of occupied categories K, and the density $\rho = N/K$. The second is the full frequency table $(n_1, \ldots, n_K)$. Our model addresses only the first component. It can speak to effects such as familiar addition and novel addition, where N and K change, but it cannot explain Navarro and Kemp's transfer effect, where N and K can remain fixed while the balance of category frequencies changes. That limitation is useful: it identifies exactly where hierarchical Bayesian machinery or item-level similarity has to enter.

This also clarifies the role of priors in the broader Bayesian program described by Griffiths et al. (2008). Bayesian cognitive models usually ask what prior knowledge a learner brings to a task and how that prior combines with data. Our analysis suggests that, for partition-extension problems, some of what appears as prior knowledge is already induced by the hypothesis space

itself. Before adding a psychologically meaningful prior over novelty, the partition space already imposes a structural gradient over (N, K). A novelty model should therefore distinguish between three things: the geometry of the hypothesis space, the substantive prior assumptions added by the modeler, and the likelihood or similarity information supplied by the stimulus.

The same distinction matters for Pitman–Yor language models. Goldwater et al. (2011) use the Pitman–Yor process as an adaptor that converts a generator over types into a token-frequency distribution, thereby explaining why type frequencies matter in language. Teh (2006) uses hierarchical Pitman–Yor processes to model context-specific word prediction and to connect Bayesian nonparametrics with Kneser–Ney smoothing. Our results do not challenge those interpretations. They suggest a more local diagnostic: part of the Pitman–Yor predictive behavior in a given domain may be recovered from partition geometry alone, while the residual marks what the adaptor, hierarchy, or context model is actually contributing. The SWDA result is the clearest example: partition geometry captures some novelty pressure, but the remaining Pitman–Yor advantage points to sequential and pragmatic structure that a density-only model cannot represent.

The practical implication is that novelty models should not report only whether a CRP, Pitman–Yor, or type–token model fits. They should also ask how much of that fit is explained by a geometry-only partition baseline. If the baseline accounts for most of the predictive structure, the model is mainly capturing generic partition growth. If it does not, the residual becomes theoretically informative: it points to category-size imbalance, similarity, context, hierarchy, or task-specific mechanisms.

5.3. Comparison to other accounts

The closest statistical relatives of the present analysis are Bayesian nonparametric accounts of Gibbs partitions and sequential discoveries. Cerquetti used asymptotic results for generalized Stirling numbers to approximate Gibbs partition weights and discovery probabilities in species-sampling problems (Cerquetti, 2012). This is close to the present work at the mathematical level, but the aim is different. In Cerquetti's work, Stirling asymptotics are used to approximate Bayesian nonparametric partition quantities. Here they are used diagnostically: the question is how much of an already fitted Pitman–Yor novelty surface can be recovered from a simple partition-geometric score.

Zito and colleagues develop the species-sampling tradition in a complementary direction. Their model of sequential discoveries directly specifies the probability of a new discovery and studies its asymptotic and finite-sample properties (Zito et al., 2023). In related work, the Stirling-gamma prior uses Stirling-number structure to define interpretable priors over latent partitions, especially through the induced distribution of the number of clusters (Zito et al., 2024). These models use combinatorics constructively, to build or estimate priors. The present analysis uses combinatorics in the reverse direction, to decompose the behavior of an existing prior-predictive model. The question is not which prior should be placed on partitions, but how much of a fitted prior can be explained by the geometry common to partitions.

A related interpretive caution is developed by Lee and Sang (2022). They show that the rich-get-richer behavior of many random partition models can follow from product-form exchangeability and projectivity, rather than from

an additional substantive assumption about cluster growth. Their target is balancedness: how probability mass is distributed over partitions with different cluster-size profiles. The present target is innovation: how much novelty bias is already determined by the coarse state (N, K). The shared lesson is that some apparently meaningful properties of partition models may reflect generic constraints of the partition framework itself.

The comparison with type–token models points in the same direction. Type–token approaches work because innovation rates depend strongly on the ratio of accumulated tokens to observed types, but they do not explain why this ratio is predictive. The combinatorial score offers a structural interpretation: the type–token relation is a coarse description of where the current state falls in partition space, and partition density determines the combinatorial pressure toward absorption versus innovation.

This interpretation also separates the present approach from adjacent-possible models of innovation. Urn models with triggering explain Heaps-like growth by making the possibility space endogenous: each novelty expands the set of future novelties (Rizzo et al., 2026; Tria et al., 2014, 2018). The present analysis does not model that expansion. It instead takes the current partition state as given and asks what novelty bias follows from its combinatorial geometry alone. The two approaches are complementary rather than competing: adjacent-possible models describe how hypothesis spaces grow; the combinatorial diagnostic describes what structure is already implied by a given active partition.

The clearest failure of the combinatorial model is SWDA. Dialogue-act sequences are not merely growing inventories of reusable types: they are

structured by turn-taking, adjacency pairs, and pragmatic licensing constraints that determine which act is locally appropriate regardless of how many types have appeared so far. The speaker-state diagnostic confirms that the definition of the active state matters: computing partitions separately for the A and B sides improved prediction. But this was not a solution to the species-sampling residual, because Pitman–Yor improved under the same speaker-specific split and remained better than the combinatorial model. The residual should therefore be interpreted as evidence for conversational dependencies beyond partition density, such as preceding dialogue acts, adjacency-pair structure, or local pragmatic role.

6. Limitations

Three limitations deserve note. First, the combinatorial score is not an exchangeable partition prior and should not be treated as a literal online probability derived from Stirling's recurrence; it is a structural predictor that requires empirical calibration. Second, the finite-size term $1/N$ is a correction to the asymptotic approximation rather than an independently motivated psychological mechanism, and it should be treated as such. Third, the datasets used here do not support stable participant-level parameter estimates; individual differences in β_0, β_1, and β_2 remain an open question that would require repeated within-subject measurements.

Data Availability

All datasets used in this study are publicly available. The GUM corpus is available at https://github.com/amir-zeldes/gum. The Kinetics-400

dataset is available at https://github.com/cvdfoundation/kinetics-dataset. The Switchboard Dialog Act corpus is available at https://web.stanford.edu/~jurafsky/ws97. Analysis scripts sufficient to reproduce the published results are included in the project archive.

Author contributions (CRediT)

Łukasz Konowałek: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used OpenAI Codex to assist with language editing, code refactoring, and manuscript organization. All generated material was reviewed, edited, and verified by the author, who takes full responsibility for the final content of the manuscript.

Acknowledgements

The author thanks Nicholas Chater and Thomas L. Griffiths for helpful comments that improved the clarity of the manuscript.

References

Cerquetti, A. (2012). Stirling's approximations for exchangeable gibbs weights.

Gnedin, A., & Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles [Originally published in *Zapiski Nauchnykh Seminarov POMI*, 325, 2005]. *Journal of Mathematical Sciences*, *138*(3), 5674–5685. https://doi.org/10.1007/s10958-006-0335-z

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12*, 2335–2382. http://jmlr.org

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 59–100). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.006

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13* (tech. rep. No. 97-02). University of Colorado, Boulder Institute of Cognitive Science. Boulder, CO.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. https://arxiv.org/abs/1705.06950

Lee, C. J., & Sang, H. (2022). Why the rich get richer? on the balancedness of random partition models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 12521–12541, Vol. 162). PMLR. https://proceedings.mlr.press/v162/lee22j.html

Navarro, D. J., & Kemp, C. (2017). None of the above: A bayesian account of the detection of novel categories. *Psychological Review*, *124*(5), 643–677. https://doi.org/10.1037/rev0000077

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Pitman, J. (2006). *Combinatorial stochastic processes* (Vol. 1875). Springer. https://doi.org/10.1007/b11601500

Rizzo, L., Lee, E. D., & Kertész, J. (2026). Adjacent possible innovation dynamics on local optima networks.

Shou, Z., Wang, W., & Li, Z. (2021). Generic event boundary detection: A benchmark for event segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman–yor processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Temme, N. M. (1993). Asymptotic estimates of stirling numbers. *Studies in Applied Mathematics*, *89*(3), 233–243. https://doi.org/10.1002/sapm1993893233

Tria, F., Loreto, V., & Servedio, V. D. P. (2018). Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. *Entropy*, *20*(10), 752. https://doi.org/10.3390/e20100752

Tria, F., Loreto, V., Servedio, V. D. P., & Strogatz, S. H. (2014). The dynamics of correlated novelties. *Scientific Reports*, *4*, 5890. https://doi.org/10.1038/srep05890

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612. https://doi.org/10.1007/s10579-016-9343-x

Zito, A., Rigon, T., & Dunson, D. B. (2024). Bayesian nonparametric modeling of latent partitions via stirling-gamma priors. *Bayesian Analysis*. https://doi.org/10.1214/24-BA1463

Zito, A., Rigon, T., Ovaskainen, O., & Dunson, D. B. (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association*, *118*(544), 2521–2532. https://doi.org/10.1080/01621459.2022.2060835

Appendix A. Saddle–Point Asymptotics of the Stirling Numbers

Throughout the appendix we work with $\nu = K/N = 1/\rho$ for convenience.

From the generating function $(e^x - 1)^K/K! = \sum_{N \geq K} S(N, K)\, x^N/N!$ one obtains the contour integral representation

$$S(N, K) = \frac{N!}{K!} \frac{1}{2\pi i} \int_{\mathcal{C}} \exp\Big(\phi(x)\Big) \frac{dx}{x}, \qquad \phi(x) = -N \log x + K \log(e^x - 1).$$

The saddle point $x_0(\nu)$ is the unique positive solution of

$$\nu = \frac{1 - e^{-x_0}}{x_0}, \tag{A.1}$$

which is strictly decreasing in $\nu \in (0, 1)$.

Taking logarithms of the one-term saddle-point approximation of Temme, 1993, eq. 2.9 gives

$$\log S(N, K) = N\Phi(\nu) + C(N, K) + O\Big(\frac{1}{N}\Big), \tag{A.2}$$

where the rate function is

$$\Phi(\nu) = -\log x_0(\nu) + \nu \log(e^{x_0(\nu)} - 1), \tag{A.3}$$

and the sublinear correction term is

$$\begin{aligned} C(N,K) = \log \frac{N!}{K!(N-K)!} + \log(N-K)! - \tfrac{1}{2}\log\frac{(N-K)}{2\pi} \\ + \log f(t_0) + O\Big(\frac{1}{N}\Big), \end{aligned} \tag{A.4}$$

where the first two terms combine exactly, yielding the more compact form used in the derivation:

$$C(N,K) = \log \frac{N!}{K!} + \tfrac{1}{2}\log\frac{2\pi}{N-K} + \log f(t_0) + O\Big(\frac{1}{N}\Big). \tag{A.5}$$

Here $t_0 = (N-K)/K$, and

$$f(t_0) = \sqrt{\frac{t_0}{(1+t_0)(x_0 - t_0)}},$$

which is real and positive since $x_0 > t_0$ holds for all $\nu \in (0,1)$, as follows directly from the saddle equation (A.1) and $t_0 = (1-\nu)/\nu$.

The approximation (A.2) is uniformly valid in ν bounded away from 0 and 1; relative errors for moderate N are reported in Table A.5.

Let $\psi(x;\nu) = \phi(x)/N = -\log x + \nu \log(e^x - 1)$ denote the normalized phase function, so that the saddle equation is $\partial_x \psi(x_0;\nu) = 0$. By the envelope theorem,

$$\Phi'(\nu) = \frac{\partial \psi}{\partial \nu}\bigg|_{x=x_0} = \log(e^{x_0(\nu)} - 1). \tag{A.6}$$

Appendix A.1. Asymptotic Logit

We work in the joint asymptotic regime $N, K \to \infty$ with $\nu = K/N$ held fixed. The derivation proceeds in three steps.

Table A.5: Relative error of the Temme approximation (A.2) for $\log S(N,K)$.

N	K	$\log S(N,K)$ (exact)	Relative error
20	5	27.342	1.0×10^{-4}
20	10	29.409	8.6×10^{-5}
50	15	106.991	1.1×10^{-5}
50	25	98.717	1.0×10^{-5}
100	30	264.367	2.3×10^{-6}
100	50	234.022	2.1×10^{-6}
200	60	627.981	4.8×10^{-7}

Step 1: difference of the leading term. By a first-order Taylor expansion of Φ in ν,

$$N\Phi\Big(\nu - \tfrac{1}{N}\Big) - N\Phi(\nu) = -\Phi'(\nu) + O\Big(\frac{1}{N}\Big) = -\log(e^{x_0(\nu)} - 1) + O\Big(\frac{1}{N}\Big), \quad \text{(A.7)}$$

where the $O(1/N)$ remainder is $\Phi''(\nu)/(2N) + O(1/N^2)$.

Step 2: difference of the correction term. The dominant contribution to $C(N,K)$ in (A.4) is $\log(N!/K!)$. When K decreases by one,

$$\log \frac{N!}{(K-1)!} - \log \frac{N!}{K!} = \log K.$$

The remaining terms in $C(N,K)$ change by $O(1/N)$: the term $\frac{1}{2}\log(2\pi/(N-K))$ changes by $-1/(2(N-K)) = O(1/N)$, which requires $N - K = \Theta(N)$, i.e. ν bounded away from 1, and $\log f(t_0)$ changes by $O(1/N)$ since t_0 and x_0 vary continuously with ν. Therefore,

$$C(N, K-1) - C(N,K) = \log K + O\Big(\frac{1}{N}\Big). \quad \text{(A.8)}$$

Table A.6 confirms that the residual $[C(N, K-1) - C(N, K)] - \log K$ is $O(1/N)$.

Table A.6: Verification of (A.8): the residual $[C(N, K-1) - C(N, K)] - \log K$ converges to zero at rate $O(1/N)$. Here $C(N, K-1) - C(N, K)$ is computed as $[\log S(N, K-1) - \log S(N, K)] - [N\Phi((K-1)/N) - N\Phi(K/N)]$.

ν	N	K	$C(N, K-1) - C(N, K)$	$\log K$	Residual $\times N$
0.2	100	20	2.99225	2.99573	−0.348
0.2	200	40	3.68707	3.68888	−0.362
0.2	500	100	4.60443	4.60517	−0.369
0.2	1000	200	5.29795	5.29832	−0.372
0.2	2000	400	5.99128	5.99146	−0.373
0.3	100	30	3.39425	3.40120	−0.695
0.3	200	60	4.09083	4.09434	−0.703
0.3	500	150	5.00922	5.01064	−0.707
0.3	1000	300	5.70307	5.70378	−0.709
0.3	2000	600	6.39657	6.39693	−0.710

Step 3: assembling the structural score. Combining Steps 1 and 2,

$$\log S(N, K-1) - \log S(N, K) = -\log(e^{x_0(\nu)} - 1) + \log K + O\Big(\frac{1}{N}\Big).$$

Substituting into the exact score (1),

$$\begin{aligned} x_{\text{comb}}(N, K) &= \Big[\log S(N, K-1) - \log S(N, K)\Big] - \log K \\ &= -\log(e^{x_0(\nu)} - 1) + \log K - \log K + O\Big(\frac{1}{N}\Big). \end{aligned}$$

The two $\log K$ terms cancel exactly, leaving

$$x_{\text{comb}}(N, K) = g(\rho) + O\Big(\frac{1}{N}\Big), \qquad g(\rho) = -\log\Big(e^{x_0(1/\rho)} - 1\Big), \tag{A.9}$$

where $\rho = N/K$. The cancellation is exact: the $+\log K$ in the difference of log-Stirling numbers comes from the $\log(N!/K!)$ prefactor in Temme's formula, and the $-\log K$ comes from the definition of the score in (1).

Table A.7 confirms that $N \times (x_{\text{comb}} - g(\rho))$ converges to a finite constant, establishing the $O(1/N)$ rate empirically.

Table A.7: Convergence of the exact combinatorial score to $g(\rho)$. The column $N \times \varepsilon$ confirms the $O(1/N)$ error rate.

ρ	N	K	x_{comb} (exact)	$g(\rho)$	$N \times \varepsilon$
2	50	25	−1.4572	−1.3665	−4.54
2	100	50	−1.4118	−1.3665	−4.53
2	200	100	−1.3891	−1.3665	−4.52
2	500	250	−1.3755	−1.3665	−4.52
5	50	10	−5.2409	−4.9581	−14.14
5	100	20	−5.0953	−4.9581	−13.72
5	200	40	−5.0257	−4.9581	−13.52
5	500	100	−4.9849	−4.9581	−13.40

Appendix B. Weighted partitions and the empirical model

The combinatorial baseline derived in Appendix A assumes that all equivalence classes are equally likely targets of absorption. This corresponds to uniform weights $a_k = 1$ for all categories, yielding the structural score

$$x_{\text{comb}}(N, K) = g(\rho) + O\Big(\frac{1}{N}\Big).$$

We now relax this assumption by allowing categories to carry non–uniform weights. Let each existing category $k \in \{1, \ldots, K\}$ be associated with a

positive weight

$$a_k(\rho) = w_k \exp\Big(c_0 + c_1 g(\rho)\Big), \qquad w_k > 0,$$

where $g(\rho)$ is the combinatorial baseline function and $\rho = N/K$ is the structural density.

The total number of absorption configurations is then proportional to the total absorption mass

$$A(N,K) = \sum_{k=1}^{K} a_k(\rho).$$

We assume that the category–specific weights $\{w_k\}$ are regular in the sense that their empirical mean converges,

$$\frac{1}{K}\sum_{k=1}^{K} w_k \to \bar{w} \in (0,\infty) \qquad \text{as } K \to \infty,$$

and that the weights do not depend on N in a finely tuned way that would cancel subleading finite–size terms.

Under these assumptions, the total absorption mass satisfies

$$A(N,K) \sim K\,\bar{w}\,\exp\Big(c_0 + c_1 g(\rho)\Big), \qquad K \to \infty,$$

with residual corrections of order $O(1/K)$, which are of order $O(1/N)$ at fixed ρ.

Substituting this expression into the weighted combinatorial contrast,

$$x^{(w)}_{\mathrm{comb}}(N,K) = \log S(N,K-1) - \log S(N,K) - \log A(N,K),$$

and using the saddle–point expansion from Appendix A, we obtain

$$x^{(w)}_{\mathrm{comb}}(N,K) \approx g(\rho) - \Big(\log \bar{w} + c_0 + c_1 g(\rho)\Big) + \beta_2\,\frac{1}{N},$$

where the $1/N$ term arises from the subleading corrections in the Stirling asymptotics and vanishes in the infinite–sample limit.

Rearranging terms yields

$$x^{(w)}_{\mathrm{comb}}(N, K) \approx \beta_0 + \beta_1 g(\rho) + \beta_2 \frac{1}{N},$$

with

$$\beta_0 = -\log \bar{w} - c_0, \qquad \beta_1 = 1 - c_1.$$

Thus, allowing for weighted categories yields a natural extension of the combinatorial baseline in which the structural score depends on density through $g(\rho)$, modulated by average category attractiveness and a finite-size correction arising from subleading asymptotic terms.