THE BRADLEY–TERRY STOCHASTIC BLOCK MODEL

BY LAPO SANTI[1,a], NIAL FRIEL[2,b]

[1] *School of Mathematics and Statistics, University College Dublin, Dublin, Ireland,* [a] *lapo.santi@ucdconnect.ie*

[2] *Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland,* [b] *nial.friel@ucd.ie*

The Bradley–Terry model is widely used for the analysis of pairwise comparison data and, in essence, produces a ranking of the items under comparison. We embed the Bradley–Terry model within a stochastic block model, allowing items to cluster. The resulting Bradley–Terry SBM (BT–SBM) ranks clusters so that items within a cluster share the same tied rank. We develop a fully Bayesian specification in which all quantities—the number of blocks, their strengths, and item assignments—are jointly learned via a fast Gibbs sampler derived through a Thurstonian data augmentation. Despite its efficiency, the sampler yields coherent and interpretable posterior summaries for all model components. Our motivating application analyzes men's tennis results from ATP tournaments over the seasons 2000–2022. We find that the top 100 players can be broadly partitioned into three or four tiers in most seasons. Moreover, the size of the strongest tier was small from the mid-2000s to 2018 and has increased since, providing evidence that men's tennis has become more competitive in recent years.

1. Introduction. Rankings are ubiquitous. From institutional league tables in education and health to online reviews, sports tournaments and app ratings, ordered lists routinely inform decisions, allocate resources, and shape reputations. Their appeal lies in their simplicity: complex information about each item is distilled into a single number. Yet this simplicity can be misleading, creating a false sense of precision and granularity that the data often do not support. As Goldstein and Spiegelhalter (1996) famously cautioned, there are *"quantifiable uncertainties which place inherent limitations on the precision with which institutions can be compared"*–a critique that extends beyond schools and hospitals to rankings of all kinds.

Nowhere are these tensions more evident than in competitive sports, where rankings are both economically consequential and statistically fragile. In professional tennis – the substantive focus of this paper–rankings are central. A seemingly minor drop in position can alter tournament seedings, jeopardize endorsement deals, affect a player's self-image, and ultimately determine whether a career remains sustainable (Schöttl et al., 2025). Rankings depend on players' performances across tournaments, but it remains unclear to what extent a drop in rank reflects a genuine performance gap.

Ranking lists in many contexts, particularly sports, are typically strict, in the sense that if two players are selected at random, one player will always be ranked higher than the other. In so doing, such a ranking produces a strict complete ordering of the items (for example, players) under consideration and implies that tied ranks are not permitted.

In professional tennis, this assumption can be particularly brittle. Two players separated by a single rank may never have competed head-to-head, may have followed markedly different tournament schedules, or may have been subject to different quality of opponents.

A strict complete ordering nonetheless treats these noisy and context-dependent differences as meaningful, potentially overstating separation where none is credibly supported by the data.

Keywords and phrases: Bradley–Terry model, Stochastic block model, Ranking data, Bayesian inference, Gibbs sampling, Tennis analytics.

The limitations of strict rankings have only recently begun to be addressed in Bayesian approaches (Pearce and Erosheva, 2025; Piancastelli and Friel, 2024), which go beyond strict total orders by accommodating tied ranks of items.

Building on this line of work, our objective is to develop a model that clusters items that share the same rank while imposing a strict ordering only on the cluster labels. This yields a coarser ranking, but it is often a more faithful representation of uncertainty: rather than forcing spurious pairwise distinctions, players of similar abilities may be assigned to the same cluster.

To operationalize this idea, we take the Bradley–Terry (BT) model (Bradley and Terry, 1952) as our starting point. The BT framework assigns each player a latent strength (or ability) parameter.

Sorting these parameters yields a ranking of players, thereby inducing a strict total order. In other words, the BT model inherits the very limitation we aim to overcome: it treats even negligible differences in estimated strengths as meaningful, producing a strictly ordered list of players.

To relax this assumption, we extend the BT framework by embedding it with the Stochastic Block Model (SBM) (Holland, Laskey and Leinhardt, 1983; Nowicki and Snijders, 2001). This combination preserves the well-known BT likelihood while introducing a latent block structure that clusters players into groups of similar strength. The resulting model replaces spurious distinctions with a clustered ordering in which only the clusters themselves are strictly ranked.

This formulation, which combines pairwise comparison modelling with latent clustering, directly motivates the following methodological contributions: (i) *End-to-end probabilistic analysis of pairwise-data networks:* we model pairwise outcomes as a directed comparison network and make *ranking itself a random object*. The BT likelihood is retained, but ranking is transferred from individual items to *ordered blocks*. The result is a fully generative Bayesian model in which posterior samples propagate uncertainty through every layer–from block membership to block strengths and even the number of blocks–so that summaries and decisions are *probabilistic by design* rather than point-estimates; (ii) *Finite yet data-driven number of blocks:* a nonparametric Gibbs-type (Gnedin) prior partitions items into a random but finite number of blocks, with reinforcement dynamics well suited to bounded, tournament-like populations such as seasonal sports leagues.

This prior combines parsimony with heavy-tailed flexibility and lets K, the number of clusters, be learned from the data; (iii) *Fully conjugate inference with variable dimension:* a Thurstonian data augmentation (Caron and Doucet, 2012) yields closed-form full conditionals and a fast single-site Gibbs sampler that *jointly* learns the number of blocks, their strengths, and player memberships–traversing a variable-dimensional posterior *without* reversible-jump or split–merge moves, with per-iteration cost $\mathcal{O}(|E| + nK)$, where E is the total number of observed interactions; and (iv) *Uncertainty-aware posterior summaries:* from the same posterior draws, we obtain fully probabilistic and internally coherent summaries that directly address applied questions–the probability that a player belongs to a given tier, the distribution over K, odds-interpretable block strengths, and entropy-based measures of competitive balance with credible intervals. These posterior quantities are designed for reporting, comparison, and principled decision-making.

Empirically, we apply the BT–SBM to twenty-three seasons (2000/21 to 2022/23) of the results of matches from the *Association of Tennis Professionals* (ATP), focusing on the top $n = 105$ players each year. The posterior summaries reveal a clear and mutually reinforcing narrative: during the so-called big four era (when the players, Federer, Nadal, Djokovic, and Murray dominated), from roughly 2008–2018, the elite field contracts and stabilises, producing a sharp concentration of strength, reflected in an small number of players estimated to be assigned to the strongest cluster of players. While after 2018, the distribution of posterior

memberships broadens, signalling the rise of a more open competitive landscape. Beyond interpretability, the BT–SBM also improves predictive accuracy relative to the standard Bradley–Terry model,

confirming the practical value of block-ranking as a robust, data-supported alternative to noisy individual rankings.

Taken together, our results make the case for moving beyond strict complete orderings toward clustered, uncertainty-aware rankings that offer a more interpretable and empirically faithful view of pairwise data.

2. Mens APT results data. The substantive application which motivates our work concerns mens professional tennis. Specifically, we consider results of matches from the *Association of Tennis Professionals* (ATP) men's tour[1]. The APT men's tour comprises a series of tournaments. A feature of the APT series is that the performance of each player in tournament is then used to determine an APT rank of each player. As highlighted above, this is a strict complete ranking. We focus on an analysis of matches over 23 seasons from $2000/2001$ through to $2022/2023$, where we use the 2017/2018 season as an illustrative example. Throughout this study period we fixed the number of players each season to $n = 105$, to ensure comparability across seasons.

For each season, we record the outcome of each match as an ordered pair between two players. In particular, let w_{ij} denote the number of times that player i defeated player j, for all $(i, j) \in E$, where $E = \{(i, j) : n_{ij} > 0\}$. While n_{ij} denotes the total number of times players i and j played each other that season.

We collect these results into a matrix $\mathbf{W}$ of size $n \times n$. Note that the matrix is not symmetric: if two players met n_{ij} times, then $w_{ji} = n_{ij} - w_{ij}$. We define the match count matrix $\mathbf{N} = \mathbf{W} + \mathbf{W}^\top$, so that its (i, j)th entry is given by $n_{ij} = w_{ij} + w_{ji}$, $i, j = 1, \ldots, n$.

By construction, $\mathbf{N}$ is symmetric and $n_{ii} = 0$ for all i.

Figure 1 presents a graphical representation of $\mathbf{W}$, with players ordered along both axes according to their ATP ranking at the end of the season 2018. Each cell encodes the number of wins of the row player against the column opponent, using a colour gradient. Dark green indicates missing matches, white means no wins, and the light-green to orange palette represents increasing win counts. On the right, we show the total number of matches played by each player.

Several structural features emerge from an inspection of Figure 1. The matrix is sparse – most pairs of players never meet – and this sparsity reflects the well-known "winners-play-more" effect of knockout tournaments: lower-ranked players (in the bottom-right corner) face each other less often, while top players (top-left) meet frequently. The matrix is also skew-symmetric along the main diagonal. In the top-right corner, matches predominantly result in wins for higher-ranked players; in the bottom-left, we mainly observe losses. Along the diagonal band, by contrast, outcomes tend to be more balanced, with reciprocal results and local cycles that are inconsistent with the notion of a strict global ordering.

Just as important is what the pairwise comparison matrix fails to capture. Each cell simply counts wins but ignores contextual factors: the timing of the match (early vs. late season), the surface (clay, grass, hard), the physical condition of the player (injuries, fatigue), or situational aspects (home advantage, retirements). These structural features – sparsity, skew-symmetry, and local inconsistencies – motivate moving beyond a strict one-dimensional ordering to a model that explicitly accommodates uncertainty, allowing for tied or clustered ranks when the data do not credibly support fine distinctions.

[1] See https://www.atptour.com/ for further details.

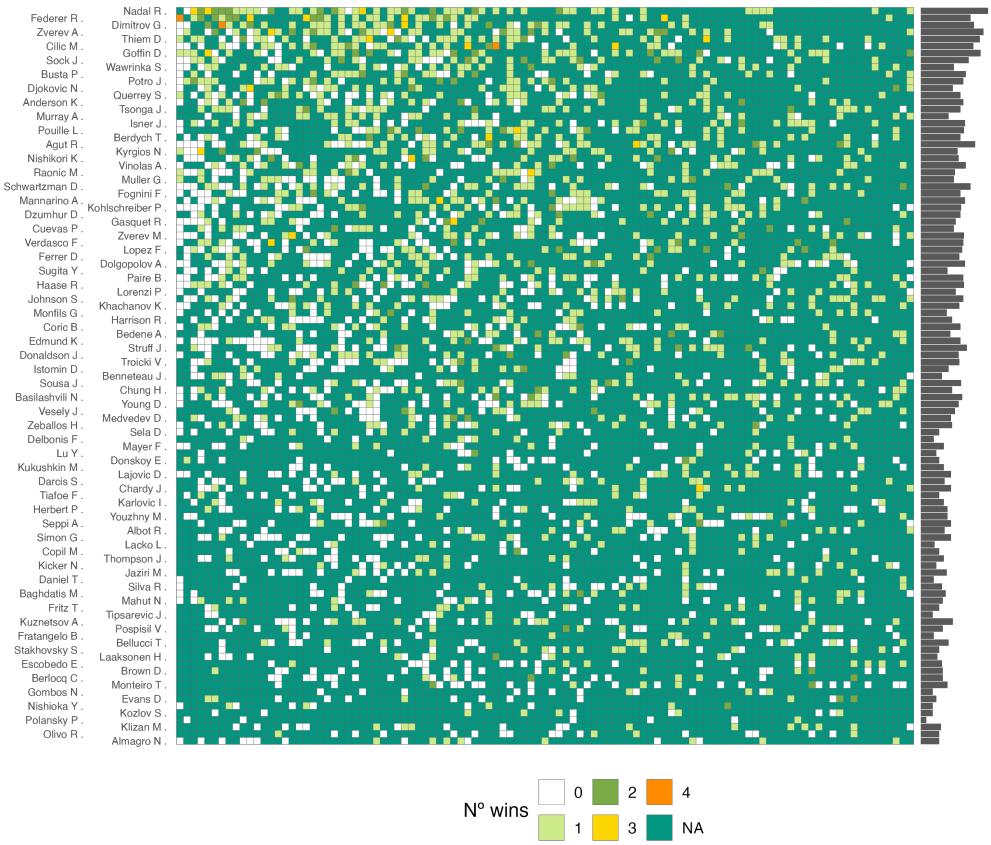


Fig 1: The match outcome matrix $\mathbf{W}$ for the ATP 2017-2018 season represented as an adjacency matrix where the players are ordered by their end-of-season ATP ranking. Each cell represents the number of wins of the row player against the column player. Dark green denotes missing data; white indicates no wins; yellow to orange indicates increasing win counts. Right margin bars show total matches played. The names on the y-axis are staggered across two columns to improve readability.

3. Literature Review. We briefly review key contributions that inform our work. At the core of our framework lies the Bradley-Terry (BT) model, which assigns the probability that item i defeats item j as

$$p(i \text{ beats } j \mid \boldsymbol{\lambda}) = \frac{\lambda_i}{\lambda_i + \lambda_j}, \qquad \lambda_i, \lambda_j > 0 \quad \text{for all } (i,j) \in E, \tag{1}$$

where $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_n\}$ denotes the vector of item-specific strength or ability parameters for the n items under comparison. The BT model remains a cornerstone of paired-comparison modelling and has inspired a rich body of work (Caron and Doucet, 2012; Glickman, 2008; Seymour et al., 2020; Wainer, 2023; Whelan, 2017; Croon and Luijkx, 1993).

Ordering the parameters of $\boldsymbol{\lambda}$ yields an implicit ranking: if $\lambda_i > \lambda_j$, then the probability that i defeats j exceeds 0.5, and sorting the strengths produces a complete linear order of the items. The classical BT model therefore enforces a strict ranking through the latent strengths. Such order is characterized by the property of linear stochastic transitivity (LST): if $\lambda_i > \lambda_j$ and $\lambda_j > \lambda_k$, then necessarily $\lambda_i > \lambda_k$, so all items can be placed on a single latent strength scale.

While convenient, LST is implausible in practice, as it assumes that outcomes of a match depend solely on the strength parameters of both player. In tennis, however, performances are shaped by multiple interacting factors, as discussed in Section 2. These sources of variability generate ranking-inconsistent outcomes, casting doubt on the adequacy of a representation of strength on a per-player basis.

To capture uncertainty and heterogeneity in the data, we move beyond strict one-to-one rankings by allowing players to be grouped into clusters. In our setting, players in the same cluster share the same strength parameter $\boldsymbol{\lambda}$. This implies that if two players belong to the same cluster, the probability of one defeating the other is exactly 0.5. Clustering thus provides a principled way to model rank-indifference: groups of players are indistinguishable in terms of estimated strength, while different clusters can still be ordered according to their common $\boldsymbol{\lambda}$.

Several authors have extended BT models to incorporate clustering. Wu et al. (2015) propose a mixture BT model to capture population heterogeneity, grouping items with distinct preference patterns, though without explicitly modeling rank indifference. Hsiao and Wang (2021) consider a "Cluster-BT" framework in which items are partitioned into latent clusters and within-cluster comparisons follow the BT model, while inter-cluster comparisons are treated as random/noise (i.e. non-informative). Spearing et al. (2021) introduced a BT-based model that clusters both items and interactions (e.g., games or matchups) to account for inconsistencies in outcomes, in the context of basketball data. In contrast, our focus is strictly on item-level clustering: we aim to discover interpretable transitive structures by estimating a partition of tied-ranked players. Pearce and Erosheva (2025) developed a rank-clustering approach in the BT family that fuses nearby strengths using a spike-and-slab prior. While their model shares our motivation of relaxing strict complete orderings, it treats clustering as a form of shrinkage on strength parameters.

In contrast to these papers, we adopt a fully generative Bayesian perspective that partitions items into latent clusters of comparable strength. Within each cluster, players are treated as rank-indifferent, while clusters themselves are strictly ordered. This structure balances interpretability with flexibility: it acknowledges the limits of LST at the player level while retaining a coherent transitive hierarchy at the cluster level. Recovering this clustered ordering and quantifying its uncertainty is the central goal of our model.

4. Embedding the BT model within an SBM. Competitions, whether in football, chess, or tennis, are inherently relational systems where hidden hierarchies and clusters emerge from repeated encounters. SBMs have been successfully employed to uncover such latent structures in various sports. Notably, Basini et al. (2023) apply SBMs to football leagues to assess competitive balance, while Vaca-Ramírez and Peixoto (2022) review applications to chess tournaments and American football. In a similar spirit, we employ SBMs to model tennis competitions. Building on this line of work, we introduce the Bradley–Terry Stochastic Block Model (BT–SBM), which combines the Bradley–Terry framework for pairwise comparisons with the SBM for clustering relational data. Before detailing the model, we briefly review the main features of SBMs.

4.1. *Overview of SBMs.* The SBM provides a flexible probabilistic framework for latent clustering in networks. Each node $i \in \{1, \dots, n$ is assumed to belong to an unobserved block $x_i \in \{1, \dots, K\}$, and the presence of an edge between nodes depends solely on their block memberships – a property known as *stochastic equivalence*. We first describe the canonical SBM (Nowicki and Snijders, 2001), before reviewing some recent approaches in this area. We begin by letting $\mathbf{x} = (x_1, \dots, x_n)$ denote the latent block assignments, y_{ij}, the binary outcome of whether an edge is present or not is modelled as,

$$y_{ij} \mid \mathbf{x}, \boldsymbol{\theta} \sim \text{Bernoulli}(\theta_{x_i x_j}),$$

where $\boldsymbol{\theta} = (\theta_{rs})_{r,s=1}^{K}$ is a $K \times K$ block–connectivity matrix. Where θ_{rs} represents the probability of an edge between any node in block r and any node in block s. Each element θ_{rs} is typically assigned an independent $\mathrm{Beta}(a, b)$ prior. The assignment vector $\mathbf{x}$ is modelled via block mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, drawn from a Dirichlet prior, $\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$. Marginalizing over $\boldsymbol{\pi}$ induces a Dirichlet–multinomial distribution on $\mathbf{x}$, favouring balanced cluster allocations when the hyperparameters α_k are symmetric.

Recent developments in Bayesian nonparametrics have introduced the overarching class of partition priors for $\mathbf{x}$, known as Gibbs–type priors, which encompass the traditional Dirichlet–multinomial, as outlined above, as a special case (Pitman, 1996; De Blasi et al., 2015). Examples include the Gnedin process (Gnedin and Pitman, 2006), along with the well-known Dirichlet process (Ferguson, 1973) and the Pitman–Yor process (Pitman and Yor, 1997). A key virtue of Gibbs–type priors is that they allow the number of occupied clusters K to be data-driven while offering fine control over the dynamics of reinforcement (how cluster sizes grow) and innovation (how new clusters are created) (De Blasi et al., 2015; De Blasi, Lijoi and Pruenster, 2013).

4.2. *A Gnedin Prior for the Latent Partition.* We now replace the simpler Dirichlet–multinomial prior on $\mathbf{x}$ by a Gibbs–type prior, thereby linking smoothly with the earlier mention of mixing weights and allowing more flexibility in the number of clusters. Concretely, we posit

$$p(\mathbf{x}) = \psi_{n,K(\mathbf{x})} \prod_{k=1}^{K(\mathbf{x})} (1-\sigma)_{m_k - 1}, \quad \sigma < 1, \tag{2}$$

where $m_k = |\{i : x_i = k\}|$ is the size of block k, $K(\mathbf{x})$ is the number of unique labels in the vector $\mathbf{x}$, and $(a)_r$ denotes the ascending factorial (Pochhammer) $(a)_r = a(a+1)\cdots(a+r-1)$. From now on, we will simply refer to $K(\mathbf{x})$ as K for ease the notation, while aware that $\mathbf{x}$ encodes the number of clusters K.

The triangular recursion on ψ,

$$\psi_{n,K} = (n - \sigma K)\, \psi_{n+1,K} + \psi_{n+1,K+1}, \quad \psi_{1,1} = 1,$$

induces the predictive urn scheme (Lijoi, Mena and Prünster, 2007):

$$p(x_{n+1} = k \mid \mathbf{x}) \propto \begin{cases} \psi_{n+1,K}\,(m_k - \sigma), & k = 1, \ldots, K, \\ \psi_{n+1,K+1}, & k = K+1. \end{cases}$$

Within this class, we choose the *Gnedin model* $\mathbf{x} \sim p_{\mathrm{GN}}(\mathbf{x} \mid \gamma)$, corresponding to $\sigma = -1$ and a single parameter $\gamma \in (0, 1)$. In that case,

$$\psi_{n,K} = \frac{(\gamma)_{n-K} \prod_{k=1}^{K-1} (k^2 - \gamma k)}{\prod_{i=1}^{n-1} (i^2 + \gamma i)},$$

and the predictive rule simplifies to:

$$p(x_{n+1} = k \mid \mathbf{x}) \propto (m_k + 1)\,(n - K + \gamma), \qquad p(x_{n+1} = K+1 \mid \mathbf{x}) \propto K^2 - K\gamma. \tag{3}$$

In the context of our motivating application to tennis data, we favour the Gnedin prior, for several reasons which we now outline. In modelling latent player groups, the true number of blocks is unknown but plausibly finite and moderate (not diverging with n), because only a finite set of players competes in a season. Relative to the Dirichlet–multinomial discussed above, the Gnedin prior allows a *random yet almost surely finite* number of clusters K, thereby preserving flexibility (see, e.g., De Blasi et al., 2015; Legramanti et al., 2022).

In particular, Legramanti et al. (2022) emphasise that the Gnedin process generates random and finite number of groups. But also that it is subject to what they term a reinforcement process.

This aligns well with our tennis context, where "reinforcement" captures the idea that the prior favours the expansion of existing clusters, giving additional weight to groups that are already established, rather than continuously spawning new ones. Finally, in the same paper they observe that the prior on the *population* number of groups has mode at 1 and heavy tails, favouring parsimonious yet robust specifications of K.

Another useful property is that one can study the prior distribution of K in *closed analytic form*. Specialising the general Gibbs–type result (2) to the Gnedin case ($\sigma = -1$) yields the exact pmf

$$p(K = k \mid n, \gamma) = \binom{n}{k} \frac{(1-\gamma)_{k-1}\,(\gamma)_{n-k}}{(1+\gamma)_{n-1}}, \qquad k = 1, 2, \ldots, n.$$

Hence,

$$\mathbb{E}[K \mid n, \gamma] = \sum_{k=1}^{n} k\, p(K = k \mid n, \gamma) = \frac{\Gamma(n+1)\,\Gamma(1+\gamma)}{\Gamma(n+\gamma)}, \tag{4}$$

which coincides with the expression reported in Legramanti et al. (2022). In addition, the variance of K admits a closed form:

$$\mathrm{Var}(K \mid n, \gamma) = \mathbb{E}[K \mid n, \gamma]\,\big[n - \gamma(n-1)\big] - \mathbb{E}[K \mid n, \gamma]^2. \tag{5}$$

This compact expression shows how the variability of the number of clusters depends jointly on n and the hyperparameter γ. Lower values of γ inflate the term $n - \gamma(n-1)$, producing a heavier-tailed prior on K and thereby increasing the chance of allocating additional clusters even under moderate sample sizes. A more detailed discussion of the mean and the variance of K is reported in Appendix C.

From an empirical perspective, recent studies within the SBM/eSBM literature report strong performance of Gnedin–based specifications relative to other Gibbs–type competitors (Legramanti et al., 2022; Lu, Durante and Friel, 2025).

4.3. *Block-clustered BT likelihood.* Once the latent partition $\mathbf{x} \sim p_{\mathrm{GN}}(\mathbf{x} \mid \gamma)$ is drawn and the number of occupied blocks K determined, each block is associated with a single latent *strength* parameter λ_k. Intuitively, λ_k represents the overall competitive strength of players in block k, capturing how likely they are to prevail against members of other groups. The collection of block strengths thus forms a K-dimensional vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, which summarises the structure of the competition at the group level. To retain flexibility while maintaining conjugacy with the likelihood, we assign each λ_k an independent Gamma prior,

$$\lambda_k \mid \mathbf{x} \sim \Gamma(a, b), \tag{6}$$

a standard conjugate choice in Bradley–Terry or Plackett–Luce models (Gormley and Murphy, 2009-06; Guiver and Snelson, 2009). This choice accommodates a wide range of relative block strengths through its shape and rate parameters, while preserving interpretability: higher values of λ_k correspond to stronger groups. Conditional on $\mathbf{x}$ and $\boldsymbol{\lambda}$, the probability that player i beats j is

$$p(i \text{ beats } j \mid \mathbf{x}, \boldsymbol{\lambda}) = \frac{\lambda_{x_i}}{\lambda_{x_i} + \lambda_{x_j}},$$

and hence the full likelihood over all comparisons is

$$\mathcal{L}(\mathbf{W} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N}) = \prod_{i<j : n_{ij}>0} \left(\frac{\lambda_{x_i}}{\lambda_{x_i} + \lambda_{x_j}}\right)^{w_{ij}} \left(\frac{\lambda_{x_j}}{\lambda_{x_i} + \lambda_{x_j}}\right)^{n_{ij} - w_{ij}}. \tag{7}$$

4.4. *Data augmentation and conjugacy.* Based on the development of the BT-SBM as outlined so far, one could focus on the posterior distribution,

$$p\Big(\boldsymbol{\lambda}, \mathbf{x}, \mid \mathbf{W}, \mathbf{N}\Big) \propto \underbrace{\mathcal{L}(\mathbf{W} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N})}_{\text{BT- likelihood as in (7)}} \times \underbrace{p(\boldsymbol{\lambda} \mid \mathbf{x})}_{\text{Strengths' prior as in (6)}} \times \underbrace{p_{\mathrm{GN}}(\mathbf{x} \mid \gamma)}_{\text{Gnedin prior spec. of (2)}} .$$

However, posterior inference, via a standard Metropolis-within-Gibbs is not straightforward. In particular, this results in a slow mixing chain that converges slowly to the posterior distribution.

To address this issue, we follow the approach presented in Caron and Doucet (2012) that reinterprets each individual comparison as a continuous-time event, by introducing the latent variables:

$$Y_{k,i} \sim \mathrm{Exp}(\lambda_{x_i}), \quad Y_{k,j} \sim \mathrm{Exp}(\lambda_{x_j}), \quad k = 1, \ldots, n_{ij}.$$

In this formulation, often referred to as *Thurstonian*, player i "wins" the kth match when $Y_{k,i} < Y_{k,j}$, which yields $p(Y_{k,i} < Y_{k,j}) = \dfrac{\lambda_{x_i}}{\lambda_{x_i} + \lambda_{x_j}}$, coinciding with the usual BT probability that player i beats player j. We then define

$$Z_{ij} = \sum_{k=1}^{n_{ij}} \min\{Y_{k,i}, Y_{k,j}\} \;\sim\; \Gamma\big(n_{ij},\, \lambda_{x_i} + \lambda_{x_j}\big),$$

This results in a complete-data likelihood which is expressed as

$$\mathcal{L}(\mathbf{W}, \mathbf{Z} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N}) = \prod_{i<j : n_{ij}>0} \lambda_{x_i}^{w_{ij}}\, \lambda_{x_j}^{n_{ij}-w_{ij}}\, \frac{Z_{ij}^{n_{ij}-1}}{\Gamma(n_{ij})} \exp\big[-(\lambda_{x_i} + \lambda_{x_j}) Z_{ij}\big]. \tag{8}$$

The result of augmenting the likelihood with $\mathbf{Z}$ is that this allows one to develop a Gibbs sampler for all parameters in the model, as we will shortly outline. Moreover, the resulting MCMC algorithm is fast to run and offer the potential to scale well to larger datasets.

4.5. *Full Posterior Distribution.* Combining the above elements, we obtain the full posterior,

$$p\Big(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{Z} \mid \mathbf{W}, \mathbf{N}\Big) \propto \underbrace{\mathcal{L}(\mathbf{W}, \mathbf{Z} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N})}_{\text{Augmented likelihood as in (8)}} \times \underbrace{p(\boldsymbol{\lambda} \mid \mathbf{x})}_{\text{Strengths' prior as in (6)}} \times \underbrace{p_{\mathrm{GN}}(\mathbf{x} \mid \gamma)}_{\text{Gnedin prior spec. of (2)}} .$$

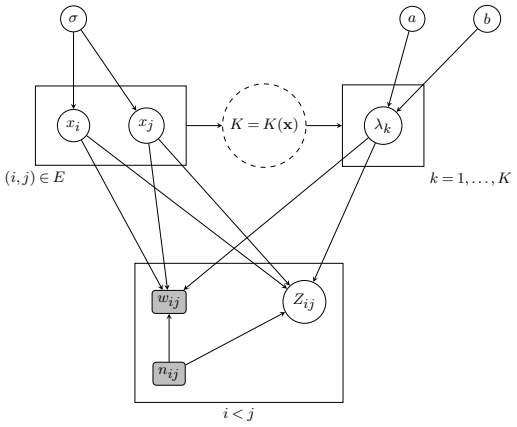
Substituting the full expressions for each term on the right hand side, and letting the proportionality sign absorb all the constants that do not depend on $\{\boldsymbol{\lambda}, \mathbf{x}, \mathbf{Z}\}$ yields,

$$p\Big(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{Z} \mid \mathbf{W}, \mathbf{N}\Big) \propto \prod_{1 \le i < j \le n :\, n_{ij}>0} \frac{\lambda_{x_i}^{w_{ij}}\, \lambda_{x_j}^{n_{ij}-w_{ij}}}{\Gamma(n_{ij})}\, Z_{ij}^{n_{ij}-1}\, \exp\Big\{-(\lambda_{x_i} + \lambda_{x_j}) Z_{ij}\Big\}$$

$$\times \prod_{k=1}^{K} \frac{b^a}{\Gamma(a)} \lambda_k^{a-1} \exp(-b\lambda_k) \times \psi_{n,K} \prod_{k=1}^{K} (1-\sigma)_{\,m_k - 1}. \tag{9}$$

This compact form is the basis for deriving the full conditional updates in Gibbs sampling or other posterior explorations.

In summary, Figure 2 provides a joint representation of the augmented BT–SBM model and its generative process.



Require: n, $\mathbf{N} = \{n_{ij}\}_{i<j}$, a, b, γ.
Initialize: $\{K, x_1\} \leftarrow 1, \lambda_1 \sim \Gamma(a, b)$;
for $i = 2$ **to** n **do**
  $x_i \sim \mathrm{GN}(\gamma)$ as in (3)
  if $x_i = K + 1$ **then**
    $K \leftarrow K + 1$; $\lambda_K \sim \Gamma(a, b)$
  end if
end for
for all pairs $i < j$ with $n_{ij} > 0$ **do**
  $w_{ij} \mid n_{ij}, \mathbf{x}, \boldsymbol{\lambda} \sim \mathrm{Bin}\left(n_{ij}, \frac{\lambda_{x_i}}{\lambda_{x_i} + \lambda_{x_j}}\right)$
end for

Fig 2: On the left, the directed acyclic graph (DAG) highlights the hierarchical structure of the model: hyperparameters (a, b) and σ) govern the latent cluster strengths $\boldsymbol{\lambda}$ and the item-specific allocations $\mathbf{x}$, respectively. In turn, these latter determine the distribution of the observed match outcomes $\mathbf{W} = (w_{ij})$ given the number of matches $\mathbf{N} = (n_{ij})$. The number of occupied blocks K is determined by $\mathbf{x}$, and K affect the size of the plate for λ_k. The grey-coloured boxes represent observed outcomes. On the right, the algorithm sketches the generative procedure required to sample from the BT-SBM. Items are allocated one at a time either to an existing block or to a newly created one according to the predictive distribution $p(x_{n+1} \mid \mathbf{x})$ in Eq. 3, while the number of occupied blocks K evolves accordingly. After block assignments and cluster strengths have been sampled, match outcomes are generated conditionally on $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{N})$. This procedure yields partitions with the characteristic features of the Gnedin prior–namely, a small modal number of clusters and a heavy-tailed distribution–and match outcomes reflecting K different tiers of strength.

5. Posterior inference and single-site Gibbs sampling. Having derived the full posterior distribution in Section 4.5, we now turn to inference. In particular, we employ a single-site Gibbs sampler which exploits the conditional conjugacy of the augmented model, which leads to closed-form full-conditional distributions for all variables in (9), which we now detail.

Auxiliary–variable update:. For every pair with at least one match ($n_{ij} > 0$),

$$Z_{ij} \mid \text{rest} \;\sim\; \Gamma\big(n_{ij},\, \lambda_{x_i} + \lambda_{x_j}\big)\,. \tag{10}$$

Block–strength update:. Let $I_k = \{i : x_i = k\}$ and set

$$w_i = \sum_{j \neq i} w_{ij}, \qquad Z_i = \sum_{j \neq i} Z_{ij}.$$

With a $\Gamma(a, b)$ prior, the full conditional is

$$\lambda_k \mid \text{rest} \;\sim\; \Gamma\big(a + \textstyle\sum_{i \in I_k} w_i,\; b + \sum_{i \in I_k} Z_i\big). \tag{11}$$

See Appendix A for further details of this derivation.

Block assignment update:. For each item i we sample the label x_i from its unnormalised full conditional

$$p(x_i = k \mid \text{rest}) \;\propto\; p(x_i = k \mid \mathbf{x}_{-i}) \;\times\; \frac{p\big(\mathbf{W} \mid x_i = k,\, \mathbf{x}_{-i}, \boldsymbol{\lambda},\, \mathbf{Z}\big)}{p\big(\mathbf{W}_{-i} \mid \mathbf{x}_{-i}, \boldsymbol{\lambda},\, \mathbf{Z}\big)}, \tag{12}$$

where the prior term follows the urn scheme shown in Sect.4.2. In the second term, we have the ratio where the denominator cancels out those pairs that do not involve i.

For convenience, we denote by $L_i(k)$ the likelihood factor associated with assigning item i to cluster k, i.e. the term multiplying the prior in (12), defined as:

$$L_i(k) \;:=\; \frac{p\big(\mathbf{W} \mid x_i = k,\, \mathbf{x}_{-i}, \boldsymbol{\lambda},\, \mathbf{Z}\big)}{p\big(\mathbf{W}_{-i} \mid \mathbf{x}_{-i}, \boldsymbol{\lambda},\, \mathbf{Z}\big)} \;=\; \lambda_k^{w_i}\, e^{-\lambda_k Z_i}, \text{for } k = 1, \dots K,$$

with $w_i = \sum_{j \neq i} w_{ij}$ and $Z_i = \sum_{j \neq i} Z_{ij}$. (See App. AA for details.)

For a new cluster $(k = K + 1)$, we follow Neal (2000)'s Algorithm 3, by integrating out the yet–to–be-sampled λ_{K+1} to obtain:

$$\begin{aligned} L_i(k) \;&=\; \int_0^\infty \lambda_k^{w_i}\, e^{-\lambda_k Z_i}\, \frac{b^a}{\Gamma(a)}\, \lambda_k^{a-1} e^{-b\lambda_k}\, d\lambda_k \\ &=\; \frac{b^a\, \Gamma(a + w_i)}{\Gamma(a)}\, (b + Z_i)^{-(a+w_i)} \quad \text{for } k = K + 1. \end{aligned} \tag{13}$$

Combining prior and likelihood terms gives

$$p(x_i = k \mid \text{rest}) \;\propto\; \begin{cases} (m_k + 1)(n - K + \gamma)\, L_i(k), & k = 1, \dots, K; \\ (K^2 - K\gamma)\, L_i(k), & k = K + 1. \end{cases} \tag{14}$$

Normalising over $k \in \{1, \dots, K + 1\}$ yields the sampling probabilities. If the draw selects the new label $K+1$, we set $x_i \leftarrow K + 1$, and then we sample

$$\lambda_{K+1} \sim \Gamma\big(a + w_i,\, b + Z_i\big),$$

and increment the cluster count $K \leftarrow K + 1$. Conversely, K decreases naturally whenever a cluster becomes empty. Note that the number of *occupied* clusters K is a random quantity, determined by the current configuration of $\mathbf{x}$. Unlike reversible-jump or split-merge samplers, this approach explores a variable-dimensional posterior space without explicit trans-dimensional proposals: clusters are created or removed naturally as labels are updated.

5.1. *Hyperparameter specification and scale alignment.* The behaviour of the model is chiefly governed by three hyperparameters: (a, b), which define the prior over block strengths, and γ, which controls the partition structure and the overall model complexity (see Sect. 4.2). Each block strength λ_k is assigned a $\Gamma(a, b)$ prior (Eq. (6)), where the shape parameter a and the rate b affect the $\boldsymbol{\lambda}$ mean and variance, and also affect the probability of sampling a new cluster (see Eq. 13). Choosing appropriate values for (a, b) is therefore essential, and we adopt a simple yet effective heuristic to fix them in a principled way.

This heuristic is motivated by several considerations, the first of which arises from the multiplicative invariance of the Bradley–Terry likelihood, which depends only on ratios such as

$$\frac{\lambda_{x_i}}{\lambda_{x_i} + \lambda_{x_j}} = \frac{c\,\lambda_{x_i}}{c\,\lambda_{x_i} + c\,\lambda_{x_j}}, \quad \text{for each } (i, j) \in E.$$

Therefore, the overall magnitude of $\boldsymbol{\lambda}$ cannot be identified from the data. To ensure identifiability, we normalise the strength vector $\boldsymbol{\lambda}$ at each iteration (step 5 in Alg. 1) through a standard procedure known as *global rescaling*. Specifically, we impose that the arithmetic mean of the log-strengths be zero:

$$\frac{1}{K} \sum_{k=1}^{K} \log \lambda_k = 0,$$

as working on the log-scale provides greater numerical stability[2]. As a consequence, the rescaled $\log \boldsymbol{\lambda}$ satisfies, a posteriori, $\mathbb{E}[\log \lambda_k \mid \mathbf{W}] = 0$. Therefore, it seems a natural choice to select hyperparameters (a, b) such that the prior mean of $\log \lambda_k$ aligns with the normalisation-induced zero expectation:

$$\mathbb{E}_{\text{prior}}[\log \lambda_k \mid a, b] = \mathbb{E}[\log \lambda_k \mid \mathbf{W}] = 0,$$

which, for $\lambda_k \sim \Gamma(a, b)$, is obtained by setting

$$b = \exp\{\psi(a)\},$$

where ψ is the derivative of the logarithm of the Gamma function, known as the *digamma* function (Soch et al., 2025).

A second reason motivating this heuristic in choosing (a, b) lies in its computational convenience, as it stabilises the propensity to sample new clusters. In particular, substituting $b = \exp\{\psi(a)\}$ in Eq. (13) makes this probability less sensitive to variations in the scale of $\boldsymbol{\lambda}$. A detailed derivation and empirical validation of this argument are provided in Appendix B.

Once $b = \exp\{\psi(a)\}$ has been fixed, the remaining shape parameter a is the only degree of freedom and governs the variance of $\log \lambda_k$, given by

$$\mathrm{Var}(\log \lambda_k \mid a, b) = \psi_1(a),$$

where ψ_1 is the *trigamma* function, i.e., the derivative of the digamma function. From the relationship between a and $\mathrm{Var}(\log \lambda_k)$ examined in Appendix B, we find that values of $a \in [2, 4]$ provide a reasonable balance, allowing for a moderate variability among the λ_k's that can be quantified as:

$$\tau = \sqrt{\psi_1(a)} = \mathrm{SD}(\log \boldsymbol{\lambda}) \in [0.53, 0.80].$$

Having fixed (a, b) to control the distribution of block strengths, we now turn to the parameter γ, which directly controls the mean and the variance of K through the Gnedin prior (2). To adopt a conservative specification, we fix $\gamma = 0.8$, which for $n = 105$ yields $\mathbb{E}[K \mid \gamma = 0.8, n = 105] \approx 2$ and $\mathrm{Var}(K \mid \gamma = 0.8, n = 105) \approx 46$, computed using (4) and (5). This choice produces considerable shrinkage on K, as it discourages partitions with many clusters unless they are strongly supported by the data. However, the heavy right tail of the distribution still allows the sampler to explore different regions of the complex partition space.

Code validation and computational complexity. Before turning to real data, we validate both the statistical accuracy and the computational efficiency of the proposed algorithm. First, we evaluate its ability to recover the model parameters and overall structure through a simulation study (see App. D). The results indicate that the posterior samples obtained via Alg. 1 accurately recover both the true number of clusters (K) and the underlying partition structure ($\mathbf{x}$) when the data are generated from the BT–SBM, confirming the correctness of the inference procedure and the reliability of its implementation. Further details on the data-generating process, MCMC settings, and performance metrics are provided in Appendix D.

In addition to statistical validation, we assess the computational scaling of the algorithm. Each Gibbs sweep requires $\mathcal{O}(|E|)$ operations to update the latent variables Z_{ij}, $\mathcal{O}(nK)$ operations to update the cluster assignments, and $\mathcal{O}(K)$ operations to update the block intensities λ_k. Hence, the total cost per iteration scales as

$$\mathcal{O}(|E| + K(n+1)),$$

[2]This also allows one to interpret λ_k as the odds of beating a block of average strength (Newman, 2022). For instance, a block with win probability p_1 against the average block has $\lambda_k = p_1/(1-p_1)$, directly expressing strength in odds form.

which is effectively linear in $|E|$ in our setting, since the number of occupied blocks K is much smaller than the number of nodes ($K \ll n$, typically $K \leq 6$).

We verify this scaling empirically by simulating networks of increasing size n from a BT–SBM with $K = 5$ blocks (see App. D). We control the expected sparsity of each network through its *edge density*, defined as the ratio between the number of observed edges and the total number of possible directed edges, i.e. $\frac{|E|}{n(n-1)}$. Table 1 reports the total runtime (in minutes) required for 10,000 Gibbs iterations on each of the simulated networks:

n	$\lvert E\rvert$	Density	Total time (min)
100	4,950	0.50	0.35
500	24,950	0.10	2.06
1,000	49,950	0.05	4.91
5,000	249,950	0.01	123.07

TABLE 1

Empirical scaling of total runtime with network size for 10,000 Gibbs iterations. Networks are generated from a BT–SBM with $K = 5$ blocks and varying edge density. Computations were carried out on a MacBook Air with an Apple M1 processor and 8 GB of RAM. The observed increase in computational cost aligns with the expected linear dependence on $|E|$, confirming that the algorithm scales efficiently with network size and remains computationally tractable even for substantially larger graphs.

The speed results in Table 1 are encouraging, especially given that computational efficiency was not a primary focus of this work. There remains substantial room for improvement–for instance, by porting the `R` code to `C++` or using more efficient operations. To conclude, Algorithm 1 summarizes the steps described above that, together, constitute the conjugate single-site Gibbs with augmentation we employ to perform inference for our BT–SBM model.

In the next section, we review methods to summarize the posterior samples; we then illustrate their performance of the BT-SBM in a simulation study, before finally turning to the real data application.

6. Point Estimates and Uncertainty Quantification. Our MCMC sampler yields posterior samples of $\mathbf{x}$ and $\boldsymbol{\lambda}$, which we use to approximate the posterior distribution and derive summary statistics. After collecting the samples $\{\mathbf{x}^{(t)}, \boldsymbol{\lambda}^{(t)}, K^{(t)}\}_{t=1}^{T}$, we discard the first B iterations as burn-in to eliminate the non-stationary phase of the chain. Again, we emphasise, that $K^{(t)}$ is number of occupied clusters in $\mathbf{x}^{(t)}$ and hence it can be easily obtained from the latter.

6.1. *Identifiability.* The retained samples suffer from two well-known identifiability issues: *label switching* and *scale invariance*. We address both below.

Label switching.. The model likelihood is invariant under permutations of cluster labels. That is, for any permutation π of the labels,

$$\mathcal{L}(\mathbf{W} \mid \mathbf{x}, \boldsymbol{\lambda}, \mathbf{N}) = \mathcal{L}(\mathbf{W} \mid \mathbf{x}_\pi, \boldsymbol{\lambda}_\pi, \mathbf{N}),$$

where $\mathbf{x}_\pi$ and $\boldsymbol{\lambda}_\pi$ denote the permuted assignment vector and cluster strengths, respectively. As a result, the posterior distribution is also invariant under label permutations, and the MCMC sampler may visit the same partition structure under different labellings across iterations. This phenomenon is known as the *label switching problem*, and it makes direct posterior summaries of $\mathbf{x}$ or $\boldsymbol{\lambda}$ meaningless unless the samples are relabelled consistently.

To address this, we relabel each sample by sorting the blocks in decreasing order of strength: the block with the largest $\lambda_k^{(t)}$ is assigned label 1, the second largest label 2, and so on. This post-processing aligns all samples to a common labelling convention. The procedure is detailed in Algorithm 2.

Algorithm 1 Conjugate single-site Gibbs with augmentation for the BT–SBM

Require: Initialization at $t = 0$: $\{x_i^{(0)} = i\}_{i=1}^n$, $K^{(0)} = n$ (overdispersed), and $\lambda_k^{(0)} \sim \Gamma(a, b)$ independently.

1: **for** $t = 0, 1, \ldots, T-1$ **do**

2: **Update** Z_{ij}**:** For each pair (i, j) with $n_{ij} > 0$, sample

$$Z_{ij}^{(t+1)} \mid \boldsymbol{\lambda}^{(t)}, \mathbf{x}^{(t)}, \mathbf{N} \sim \Gamma\Big(n_{ij},\, \lambda^{(t)}_{x_i^{(t)}} + \lambda^{(t)}_{x_j^{(t)}}\Big).$$

3: **Update** λ_k**:** For each occupied cluster k, sample

$$\lambda_k^{(t+1)} \mid \mathbf{Z}^{(t+1)}, \mathbf{x}^{(t)}, \mathbf{W}, \mathbf{N} \sim \Gamma\Big(a + \textstyle\sum_{i \in I_k^{(t)}} w_i,\; b + \sum_{i \in I_k^{(t)}} Z_i^{(t+1)}\Big).$$

4: **Update** x_i**:** For each $i = 1, \ldots, n$,

1. Remove i from its current cluster; update $m^{(t)}_{-i,k}$. If that cluster becomes empty, remove it and re-index the clusters.
2. Compute the unnormalized probabilities
$$p\Big(x_i^{(t+1)} = k \,\Big|\, \mathbf{x}_{-i}, \boldsymbol{\lambda}^{(t+1)}, \mathbf{Z}^{(t+1)}, \mathbf{W}, \mathbf{N}\Big), \quad k = 1, \ldots, K^{(t)} + 1,$$
according to (14).
3. Normalize and sample a new value for x_i.
4. If $x_i^{(t+1)} = K^{(t)} + 1$, sample
$$\lambda^{(t+1)}_{K^{(t)}+1} \mid \mathbf{Z}^{(t+1)}, \mathbf{W}, \mathbf{N} \sim \Gamma\Big(a + w_i,\; b + Z_i^{(t+1)}\Big),$$
and set $K^{(t+1)} \leftarrow K^{(t)} + 1$.

5: **Global rescaling** To enforce the scale constraint $\prod_{k=1}^{K^{(t+1)}} \log \lambda_k^{(t+1)} = 0$, compute

$$\log g^{(t+1)} \;=\; \frac{1}{K^{(t+1)}} \sum_{k=1}^{K^{(t+1)}} \log \lambda_k^{(t+1)}$$

and set $\log \lambda_k^{(t+1)} \leftarrow \log \lambda_k^{(t+1)} - \log g^{(t+1)}$ for all occupied k.

6: **end for**

7: **return** Collect $\{\, (\mathbf{x}^{(t)}, \boldsymbol{\lambda}^{(t)}, K^{(t)}) \,\}_{t=1}^T$ from the posterior, discarding the first B iterations.

Algorithm 2 Label-switching correction

Require: $\{\mathbf{x}^{(t)}, \boldsymbol{\lambda}^{(t)}\}_{t=B+1}^T$

for $t = 1, \ldots, T$ **do**

Extract the set of occupied clusters and their $\lambda_k^{(t)}$ values.

Sort the $\lambda_k^{(t)}$ in decreasing order.

Relabel clusters by rank in the sorted list (i.e., largest $\lambda_k^{(t)}$ becomes label 1, etc.).

end for

return $\{\mathbf{x}_\pi^{(t)}, \boldsymbol{\lambda}_\pi^{(t)}\}_{t=B+1}^T$

Scale invariance.. The BT–SBM likelihood is also invariant under multiplicative rescaling of the block strengths, and this issue is addressed above (see Sect. 5.1).

6.2. *Summarizing Posterior Samples.* We summarize three key quantities from the posterior: (i) the partition $\mathbf{x}$, (ii) the cluster strengths $\boldsymbol{\lambda}$, and (iii) the number of occupied clusters K. Below, we outline how we extract point estimates and quantify posterior uncertainty for each.

Partition point estimates.. To summarize the posterior distribution over the $\{\mathbf{x}^{(t)}\}_{t=B+1}^{T}$ samples, we seek a consensus partition $\widehat{\mathbf{x}}$ that best represents the clustering structure. A principled and widely used approach is to minimize the expected *Variation of Information (VI)* (Meilă, 2007), an information-theoretic metric defined as:

$$\mathrm{VI}(A,B) = H(A) + H(B) - 2\,I(A,B), \tag{15}$$

where $H(\cdot)$ is the Shannon entropy of a partition, and $I(A,B)$ is the mutual information between two partitions A and B:

$$H(A) = -\sum_{a} \frac{n_a}{n} \log\left(\frac{n_a}{n}\right), \qquad I(A,B) = \sum_{a,b} \frac{m_{ab}}{n} \log\left(\frac{m_{ab}\cdot n}{n_a \cdot n_b}\right), \tag{16}$$

where n_a is the size of cluster a under partition A, n_b the size of cluster b under B, and m_{ab} the number of shared items between clusters a and b.

Entropy is a measure of the uncertainty of a single partition, while mutual information captures the overlap between two. The Variation of Information (VI) is a metric on the space of partitions (Meilă, 2007) and has become a standard tool in Bayesian clustering (Wade and Ghahramani, 2018; Rastelli and Friel, 2018; Legramanti et al., 2022). The consensus partition $\widehat{\mathbf{x}}$ is defined as the minimizer of the posterior expected VI:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathbb{E}\big[\mathrm{VI}(\mathbf{x}, \mathbf{x}_{\pi}^{(t)})\big] \approx \arg\min_{x} \frac{1}{T-B} \sum_{t=B+1}^{T} \mathrm{VI}(\mathbf{x}, \mathbf{x}_{\pi}^{(t)}).$$

where each sample $\mathbf{x}_{\pi}^{(t)}$ is relabelled to account for label switching, as described above. This estimator, efficiently implemented in `R` through the `mcclust` and `mcclust.ext` packages (Wade and Ghahramani, 2018), tends to penalize the formation of small clusters, yielding a parsimonious summary partition.

A partition estimate summarises the full posterior sample $\{\mathbf{x}^{(t)}\}_{t=B+1}^{T}$ into a single labelling. To fully exploit the richness of the posterior, it is therefore crucial to assess and appropriately represent the *uncertainty* surrounding $\widehat{\mathbf{x}}$. A common approach is to compute the co-clustering probability between pairs of nodes and represent it as a heatmap, i.e., the similarity or co-clustering matrix (Legramanti et al., 2022). However, such heatmaps can be difficult to interpret and, as noted by Wade and Ghahramani (2018), they tend to understate posterior uncertainty.

An appealing alternative proposed by Wade and Ghahramani (2018) is the construction of a *credible ball*–the analogue of a credible interval in the high-dimensional space of partitions–which captures the range of plausible clusterings around the point estimate $\widehat{\mathbf{x}}$. Partitions on the edge (or surface) of this ball provide an intuitive sense of alternative groupings that remain well supported by the posterior distribution.

Define an $\varepsilon-$ball around $\widehat{\mathbf{x}}$ of size ε as

$$B_{\varepsilon}(\widehat{\mathbf{x}}) := \big\{\mathbf{x} : \mathrm{VI}(\mathbf{x}, \widehat{\mathbf{x}}) \le \varepsilon\big\}.$$

Then the posterior probability of $B_{\varepsilon}(\widehat{\mathbf{x}})$ can be expressed as

$$P\big(B_{\varepsilon}(\widehat{\mathbf{x}}) \mid \text{data}\big) \;=\; E\big[\mathbb{I}\{\mathrm{VI}(\mathbf{x}, \widehat{\mathbf{x}}) \le \varepsilon\} \mid \text{data}\big] \;\approx\; \frac{1}{T-B} \sum_{t=B+1}^{T} \mathbb{I}\big\{\mathrm{VI}(\mathbf{x}^{(t)}, \widehat{\mathbf{x}}) \le \varepsilon\big\},$$

where $\mathbb{I}\{\mathrm{VI}(\mathbf{x}^{(t)}, \widehat{\mathbf{x}}) \le \varepsilon\} = 1$ if $\mathbf{x}^{(t)}$ lies within a ε VI-distance from $\widehat{\mathbf{x}}$. The $(1-\alpha)$ *credible ball* around $\widehat{\mathbf{x}}$ is then defined as the smallest metric ball, $B_{\varepsilon^*}(\widehat{\mathbf{x}})$, where

$$\varepsilon^* = \inf\big\{\varepsilon \ge 0 : P\big(B_{\varepsilon}(\widehat{\mathbf{x}}) \mid \text{data}\big) \ge 1-\alpha\big\}.$$

In our application, we set $\alpha = 0.05$, yielding a 95% credible ball. Intuitively, this region contains the set of partitions that are *sufficiently close* to the optimal partition $\widehat{\mathbf{x}}$ under the VI metric to collectively account for 95% of the posterior probability. A small ε^* indicates a concentrated posterior, reflecting strong confidence in the estimated clustering.

Once the credible ball is constructed, it is often insightful to examine its *surface*, that is, the set of partitions $\mathbf{x}$ such that $\mathrm{VI}(\mathbf{x}, \widehat{\mathbf{x}}) = \varepsilon^*$. Among these, Wade and Ghahramani (2018) identify three representative types:

- **Vertical upper bounds**: surface partitions with the *fewest* clusters, corresponding to the coarsest partition within the 95% credible ball and maximally distant from $\widehat{\mathbf{x}}$;
- **Vertical lower bounds**: surface partitions with the *largest* number of clusters, corresponding to the finest partition within the 95% credible ball and maximally distant from $\widehat{\mathbf{x}}$;
- **Horizontal bounds**: surface partitions that are maximally distant from $\widehat{\mathbf{x}}$, irrespective of the number of clusters.

Inspecting these surface partitions provides an intuitive view of how posterior uncertainty distributes across alternative but still plausible clusterings. The computation of credible balls and their corresponding bounds is implemented in the `R` package `mcclust` (Wade and Ghahramani, 2018).

Number of clusters K and model complexity. The credible ball can also be used to quantify the uncertainty surrounding the number of clusters K. Specifically, we define $\widehat{K}^{(\mathrm{VI})}$ as the number of groups in the partition point estimate $\widehat{\mathbf{x}}$ obtained under the VI loss, and denote by K_{ub} and K_{lb} the numbers of clusters associated with the *vertical upper* (coarsest) and *vertical lower* (finest) bounds, respectively (Wade and Ghahramani, 2018). The resulting range,

$$\widehat{K}^{(\mathrm{VI})}_{[K_{\mathrm{ub}},\, K_{\mathrm{lb}}]},$$

provides a compact measure of the variability of K within the 95% credible region around $\widehat{\mathbf{x}}$.

These quantities complement the direct inspection of the number of occupied clusters K across MCMC samples. We approximate its posterior distribution by collecting $\{K^{(t)} = K(\mathbf{x}^{(t)})\}_{t=B+1}^{T}$ and summarizing its empirical frequencies. As a point estimate, we report the posterior mode

$$\widehat{K} := \mathrm{mode}\{K^{(t)}\},$$

interpreted as the most probable number of clusters supported by the posterior under the Gnedin prior, together with its 95% credible interval.

These two estimators of K, ($\widehat{K}$ and $\widehat{K}^{(\mathrm{VI})}$), along with their associated uncertainty measures, are not theoretically guaranteed to coincide. In practice, however, we observe substantial agreement between them across most seasons, with occasional discrepancies reflecting the distinct nature of the two summaries.

Strength estimates.. Given the normalized and relabelled MCMC samples $\boldsymbol{\lambda}_{\pi}^{(t)}$, we summarize the strength of each block k by its posterior mean:

$$\widehat{\lambda}_k = \frac{1}{T-B} \sum_{t=B+1}^{T} \lambda_{\pi,k}^{(t)}, \quad \text{for } k = 1, \ldots, K.$$

7. Analysis of the last 23 years of men's tennis. We now present a detailed analysis of the 2017/2018 ATP men's tennis season, highlighting the main outputs of our model. We focus on this season because of the substantial uncertainty regarding the number of clusters. This case study also serves to illustrate the BT–SBM more concretely in the context of real competition data.

We subsequently repeat the analysis across all seasons from 2000/2001 to 2022/2023 (23 seasons in total), a period spanning the rise, peak, and eventual decline of the dominance exerted by Federer, Nadal, Djokovic, and Murray—the quartet commonly referred to as the "Big Four."

This era profoundly shaped men's tennis, and our model captures its evolution through changes in the inferred competitive structure over time. For each season, our MCMC algorithm (described in Sect. 5) was independently run for 30,000 iterations, discarding the first 10,000 as burn-in. Fitting the model to all 23 seasons under these specifications required approximately 35 minutes in total. We used a Gnedin prior for $\mathbf{x}$ with hyperparameter $\gamma = 0.8$, and a Gamma prior for the rate parameters with shape $a = 2$ and rate $b = \exp\{\psi(a)\} = 1.526$.

The label switching correction algorithm (Alg. 2) was subsequently applied. All computations, regarding both the simulation study and the application to the real data, were implemented in `R`, and the associated code and data are available at: https://github.com/laposanti/BT-SBM.

7.1. *The 2017/2018 ATP men's Season.* We examine in detail the 2017/2018 ATP season. Table 2 reports the posterior probabilities for different numbers of blocks under the Gnedin-type prior. The posterior slightly favours a four-block model ($\widehat{K} = 4$), although the 95% credible interval on K spans configurations between 3 and 7 clusters.

Season	2	3	4	5	6	7	8
2017/2018	–	0.259	0.315	0.216	0.110	0.056	0.025

TABLE 2

Posterior distribution of the number of clusters (K) inferred for the 2017/2018 ATP men's season under the Gnedin-type prior. Each entry reports the posterior probability of a given K, with the most probable configuration ($K = 4$) highlighted in orange. Although the four-block model achieves the highest support ($p(K = 4 \mid \mathbf{W}) = 0.315$), the posterior mass remains substantial across $K \in \{3, \dots, 7\}$, indicating non-negligible uncertainty about the true number of latent groups.

Given this diffuse posterior, we next identify a representative point estimate. We obtain an estimated partition $\widehat{\mathbf{x}}$ via VI loss (Sect. 6.2) and reorder the data accordingly, as displayed in Figure 3. Because the VI loss tends to penalise small clusters, it produces a parsimonious summary ($\widehat{K}^{(\mathrm{VI})} = 3 < \widehat{K} = 4$). In this case, the VI point estimate with $\widehat{K}^{(\mathrm{VI})} = 3_{[3,11]}$ reveals a coarse but interpretable block structure that divides players into three main groups: a narrow elite, a compact middle section, and a broad lower tier of weaker competitors.

The top group isolates *Nadal* and *Federer*, reflecting the marked competitive gap of the 2017–2018 season. This dominance is clearly visible in the top-right corner of the reordered adjacency matrix, where dark-green pixels correspond to head-to-head winning probabilities close to one. Moving across the matrix, the bottom-left corner tells the complementary story: its prevalence of white pixels indicates that players in the weakest block rarely beat those above them.

Along the main diagonal, the three squared blocks correspond to within-block matchups. These regions are almost uniform in colour, consistent with the assumption that for players belonging to the same block the probability of victory is roughly 0.5, a consequence of sharing the same strength parameter ($\hat{\lambda}_i = \hat{\lambda}_j \iff x_i = x_j$). Taken together, these visual patterns reveal a sharply hierarchical yet clustered structure: a small elite dominates, a compact group of contenders follows closely, and a large set of statistically similar players forms the broad lower tier.

Although the VI estimate provides a concise summary, the posterior retains non-negligible uncertainty, supporting partitions up to $K = 11$ within the 95% credible ball. The *vertical*

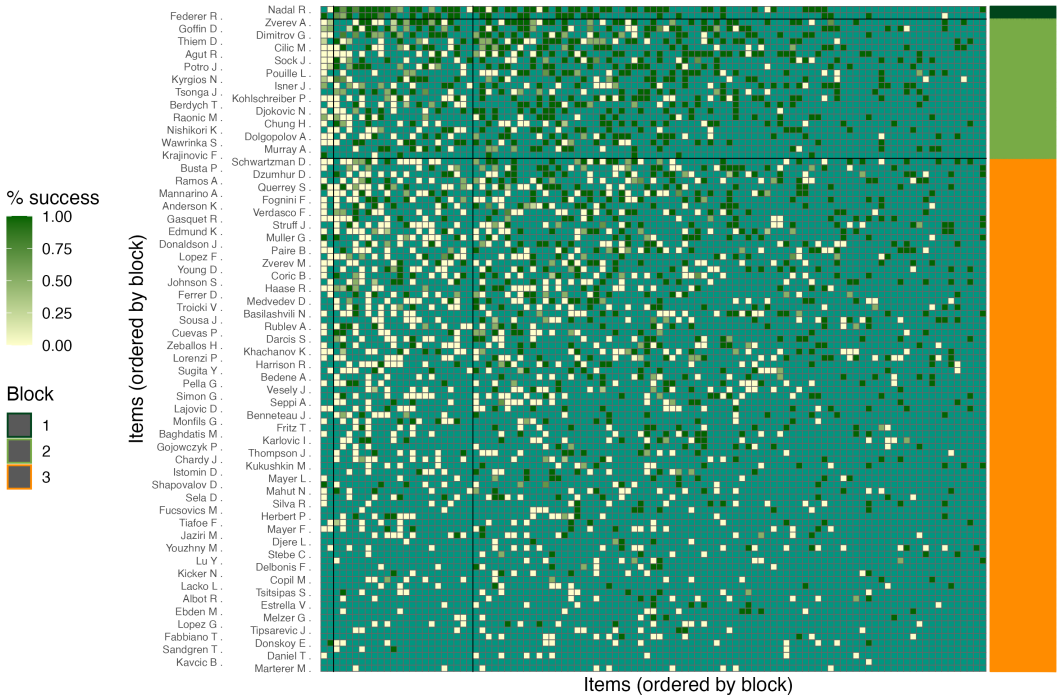


Fig 3: Reordered adjacency matrix of ATP matches during the 2017/2018 season, obtained from the estimated partition $\hat{\mathbf{x}}$ with $K = 3$. Rows and columns correspond to players, ordered first by inferred cluster membership (from 1 = strongest to 3 = weakest) and then by their marginal win proportions. Coloured bars on the left mark the three inferred clusters (dark-green, light green, and orange). Each pixel represents the empirical proportion of wins of the row player against the column player, with a white-to-green gradient indicating increasing success rates for the row player. The uniform colour pattern along the main diagonal reveals strong within-cluster balance. In contrast, the top-right corner is dominated by darker shades, corresponding to encounters between top-block and weakest-block players, where stronger opponents prevail almost systematically, as expected. Conversely, the bottom-left corner is mostly white, reflecting matches in which the weakest blocks concede to the elite group. Players' names are staggered across two columns to improve readability. Nadal corresponds to the first row, Federer to the second, Zverev to the third, and so on.

upper, *vertical lower*, and *horizontal* bounds partitions are reported in Fig. 4, which illustrates how uncertainty propagates across the hierarchy.[3] The coarsest partition (vertical upper bound) nearly coincides with the VI estimate, confirming that the VI criterion favours fewer and broader blocks. At the opposite extreme, the finest configuration (vertical lower bound) supports higher resolutions: the middle and weakest tiers split into six and four sub-blocks, respectively. Between these extremes, the horizontal bound provides an intermediate $K = 6$ solution with more balanced block sizes.

[3]Magnified versions of these configurations are provided in Appendix E.

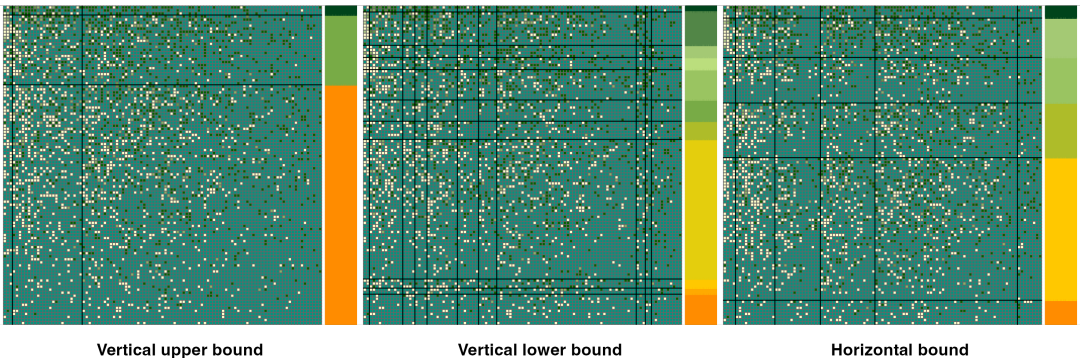


Fig 4: Reordered adjacency matrix of the 2017–2018 ATP season. From left to right, the matrix is reordered according to the vertical upper bound, vertical lower bound, and horizontal bound partitions. Magnified views of these plots are shown in Appendix E. A detailed guide to interpreting this type of graph is provided in Fig. 3.

Compared with the $K = 3$ configuration, these higher-resolution partitions introduce new blocks primarily in two regions: between block 2 and block 3, and at the bottom of block 3. The elite group remains unchanged, while the middle and lower parts of the hierarchy are refined by the addition of further layers. This indicates that the global order of strengths is stable, whereas the granularity of the lower tiers remains more flexible. Depending on the chosen resolution, these additional groups can either be kept separate or merged into broader tiers, yielding more parsimonious summaries. In this sense, the credible ball spans a continuum of plausible resolutions–from coarse to fine–rather than representing competing hierarchies of the data.

At the limit, the standard Bradley–Terry model corresponds to the extreme case in which every player forms their own block–a strict global order that enforces sharp distinctions even where the data provide little evidence for them. Our block-based approach, by contrast, preserves the stable clustered structure while explicitly accounting for the uncertainty arising from group separation.

Uncertainty in cluster assignments. We can examine this uncertainty more directly by looking at the posterior probabilities of cluster membership–that is, how confidently each player is assigned to a block. Figure 5 displays $p(x_i = k \mid \mathbf{W})$, restricted to draws with $K^{(t)} = 4$ to isolate allocation uncertainty from uncertainty in K and to illustrate how a fourth block could also be supported. Players are ordered as in Figure 3 to facilitate direct comparison with the point-estimate adjacency matrix above.

Two main factors drive uncertainty in cluster allocation: limited data and discontinuous performance. The first mainly affects players at the bottom of the hierarchy, for whom we observe few matches; as data become sparse, posterior assignments become more diffuse. In contrast, the top tier–where data are abundant–shows highly concentrated memberships ($p(x_i = k \mid \mathbf{W}) \approx 1$).

For players between blocks 1 and 2–where the network is dense–uncertainty instead arises from irregular or exceptional seasons. Examples include fast-rising competitors such as *Filip Krajinović*, whose 2017 Paris Masters final marked a sudden ascent, and established players like *Novak Djokovic* and *Juan Martín del Potro*, whose injury-interrupted seasons blur their position within the hierarchy.

Finally, allocation uncertainty is highest between blocks 2 and 3 and at the bottom of block 3, where players are hardest to separate with confidence. These players could plausibly form an additional, distinct group, reinforcing the patterns observed in the vertical lower and horizontal partitions described above.

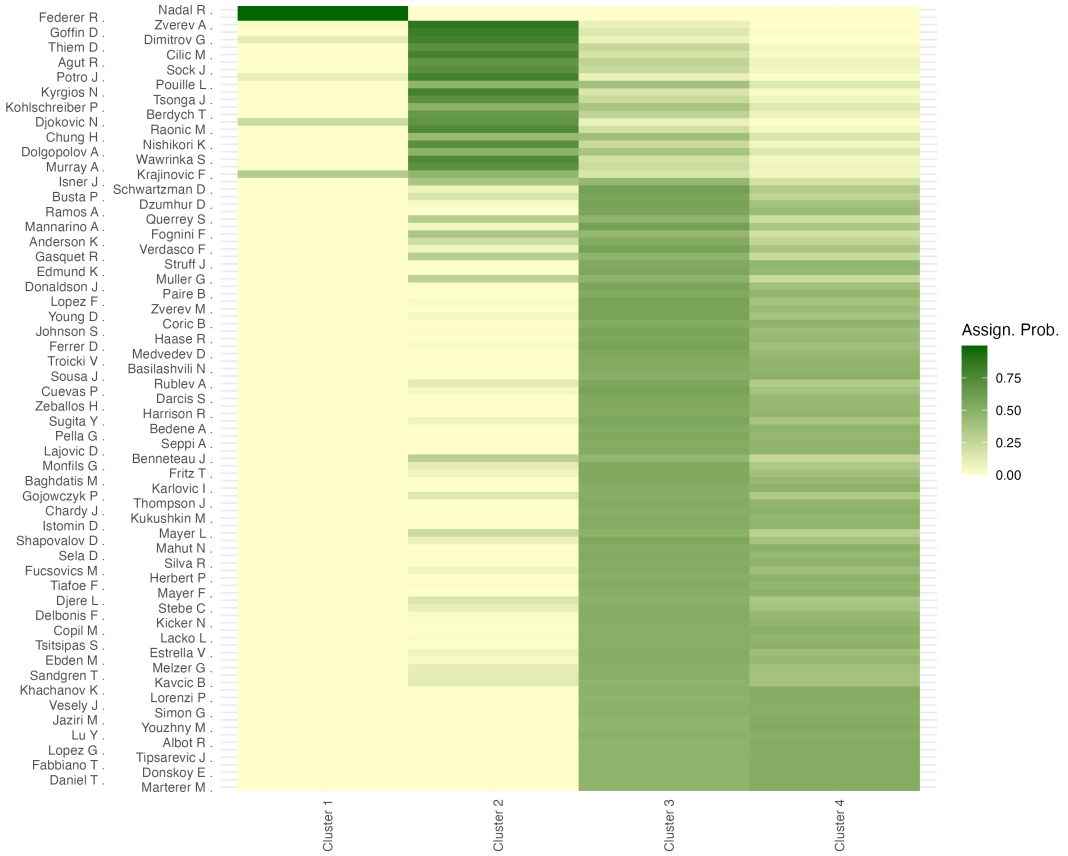


Fig 5: Posterior probabilities of cluster membership $p(x_i = k \mid \mathbf{W})$ for each player in the 2017/2018 ATP season, conditioned on MCMC draws with $K = 4$. Each row corresponds to a player and each column to an inferred block label (14). Players are ordered as in Figure 3, and darker shades indicate higher posterior probabilities. Players at the top and bottom of the hierarchy show near-certain assignments, whereas uncertainty concentrates between blocks 2 and 3 and at the bottom of block 3, where several players display diffuse membership–compatible also with block 4. This pattern suggests that a subset of these players could plausibly form an additional, well-supported fourth block, consistent with the posterior evidence on K. Names are staggered across two columns for readability.

Player-level strength uncertainty. The uncertainty visible in cluster assignments is mirrored in the posterior distribution of player-specific strengths $\tilde{\lambda}_i$. Figure 6 reports posterior means and 95% HPD intervals from relabelled draws (Algorithm 2), with

$$\tilde{\lambda}_i^{(t)} := \lambda^{(t)}_{\pi, x_i^{(t)}} \quad \text{for } i \in 1, \ldots, n, \tag{17}$$

filtered to $K^{(t)} = 3$. Elite players (Nadal, Federer, Zverev) display large $\tilde{\lambda}_i$ values and tight credible intervals; the weakest block is equally well separated at the lower end. In contrast, mid-tier players exhibit broad and overlapping HPD intervals–precisely where the posterior mass is most diffuse and clustering least stable.

Taken together, the credible-ball partitions (Figs. 4), the uncertainty analyses in Figures 5 and 6 highlight a key aspect of the posterior structure. A stable elite group is followed by a more unstable block of players that can split into smaller sub-groups depending on the resolution of the partition. In this middle region, block assignments and strength estimates are

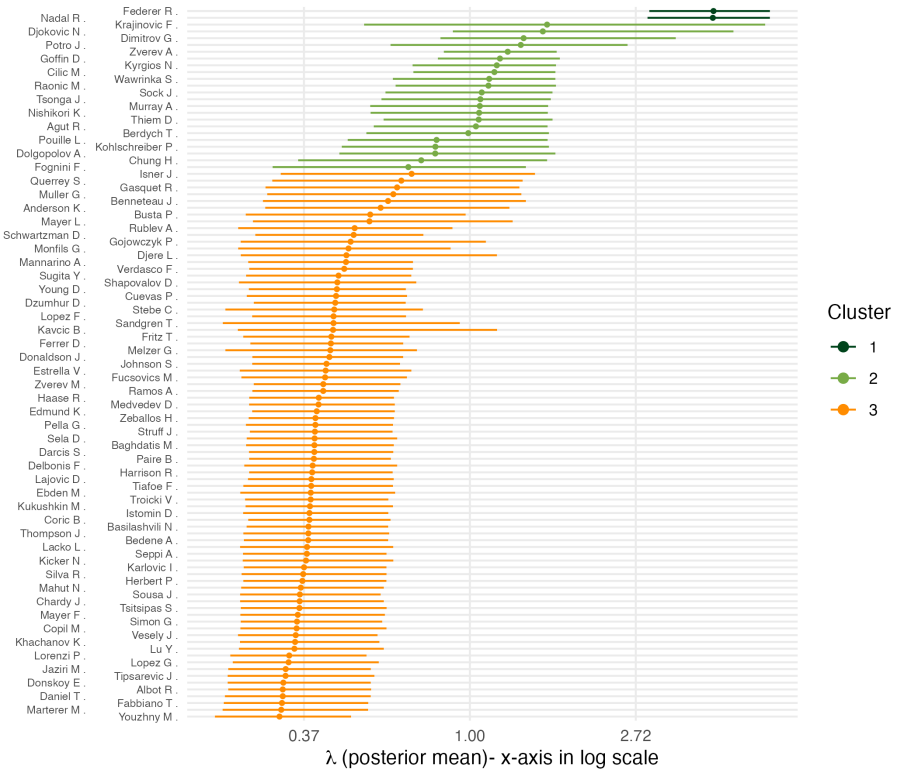


Fig 6: Posterior estimates of player-specific strength parameters $\tilde{\lambda}_i$ for the 2017/2018 ATP season, conditioned on MCMC draws with $K^{(t)} = 3$. Players are ordered by their posterior mean strength, with 95% HPD credible intervals. Elite players are sharply separated, whereas mid-ranked competitors exhibit wide overlaps, mirroring the assignment uncertainty in Figure 5. This alignment underscores how the posteriors uncertainty concentrates in the middle region of the hierarchy and at the the borders between blocks.

inherently harder to pin down, and such uncertainty is well-accounted in the BT-SBM, which is robust to the variability of tennis competition. In this sense, the block-based representation offers a more faithful summary: it preserves the clear dominance relations at the top and bottom while acknowledging the ambiguity that naturally arises in the middle of the hierarchy.

7.2. *BT–SBM analysis: all seasons* $2000/01$–$2022/23$. We now apply our model independently across seasons from the early 2000s to the early 2020s–a period spanning the rise, dominance, and gradual decline of the so-called *Big Four*. This era is marked by a remarkably stable elite group of players that captured the lion's share of major titles, profoundly shaping the competitive structure of the men's tour. In the seasons immediately preceding the COVID-19 disruption, we observe the fading of this dominance and the emergence of a new, more fragmented cluster of elite contenders.

7.2.1. *Model comparison: BT–SBM vs. standard BT.* Before exploring the temporal evolution of player clusters, we first ask whether the BT–SBM provides a tangible improvement over the standard Bradley–Terry (BT) model in terms of out-of-sample predictive accuracy and model fit.

In the standard BT model, each player i has an individual strength $\lambda_i > 0$. With independent priors $\lambda_i \sim \Gamma(a, b)$ and the identifiability constraint $\prod_i \lambda_i = 1$, the posterior is

$$p_{\text{BT}}(\boldsymbol{\lambda} \mid \mathbf{W}) \propto \left[\prod_{i=1}^{n} \lambda_i^{a-1} e^{-b\lambda_i}\right] \times \prod_{1 \le i < j \le n} \left(\frac{\lambda_i}{\lambda_i + \lambda_j}\right)^{w_{ij}} \left(\frac{\lambda_j}{\lambda_i + \lambda_j}\right)^{w_{ji}}, \qquad \text{subject to } \prod_{i=1}^{n} \lambda_i = 1.$$

Regarding the BT-SBM, its corresponding posterior $p(\boldsymbol{\lambda}_{\text{BT-SBM}}, \mathbf{x} \mid \mathbf{W})$ is defined in (9).

To assess predictive performance across the two models $\mathcal{M} \in \{\text{BT}, \text{BT-SBM}\}$, we perform leave-one-out cross-validation using Pareto-smoothed importance sampling (PSIS–LOO) (Vehtari, Gelman and Gabry, 2017).

For each directed edge $(i, j) \in E$ with at least one recorded match, we evaluate the model's ability to predict the outcome w_{ij} when it is held out from the data. That is, for each edge, model $\mathcal{M}$ is effectively fit on $\mathbf{W}_{-ij}$—the data matrix excluding w_{ij}—and predictive accuracy is assessed via the log posterior predictive density:

$$\text{lpd}_{ij,\mathcal{M}} = \log\big[\, p_{\mathcal{M}}\big(w_{ij} \mid \mathbf{W}_{-ij}\big)\,\big].$$

In practice, PSIS–LOO approximates this by reweighting posterior samples $\Theta_{\mathcal{M}}^{(t)} \sim p_{\mathcal{M}}(\Theta_{\mathcal{M}} \mid \mathbf{W})$ using importance weights that account for the omission of w_{ij}, yielding

$$\widehat{\text{lpd}}_{ij,\mathcal{M}} = \sum_{t=B+1}^{T} r_{ij,\mathcal{M}}^{(t)} \log\big[\, p_{\mathcal{M}}\big(w_{ij} \mid \Theta_{\mathcal{M}}^{(t)}\big)\,\big],$$

where $r_{ij,\mathcal{M}}^{(t)}$ are normalised PSIS importance weights.

Summing over all relevant edges yields the expected log pointwise predictive density (ELPD) for model $\mathcal{M}$,

$$\widehat{\text{ELPD}}_{\mathcal{M}} = \sum_{(i,j) \in E} \widehat{\text{lpd}}_{ij,\mathcal{M}},$$

which serves as a measure of out-of-sample predictive performance. Higher ELPD values indicate better generalisation and model fit. We computed $\widehat{\text{ELPD}}_{\mathcal{M}}$ for both models in each of the 23 seasons (2000–2022) and summarised their difference as

$$\Delta\text{ELPD} = \widehat{\text{ELPD}}_{\text{BT-SBM}} - \widehat{\text{ELPD}}_{\text{BT}},$$

with one–standard–error bounds $\text{SE}_\Delta = \sqrt{\text{SE}_{\text{BT-SBM}}^2 + \text{SE}_{\text{BT}}^2}/2$ obtained from the PSIS diagnostics.

The clustered BT–SBM outperformed the classical BT in every season. The median gain was 22.99 ELPD points (median SE_Δ = 12.1), and in approximately 87% of seasons the improvement exceeded one SE, providing moderate to strong evidence in favour of the clustered specification (see Fig. 7).

Statistic	Min	Median	Mean	Max
ΔELPD	11.17	22.52	21.99	35.49

Proportion of seasons with $\Delta\text{ELPD} > \text{SE}_\Delta$: 0.87

TABLE 3

Season-aggregated model comparison metrics. The table reports summary statistics of the predictive gain (ΔELPD) of the clustered BT–SBM over the classical BT model across all seasons. Values above zero indicate that the clustered BT–SBM provides a better out-of-sample predictive fit. In 87% of the seasons, the predictive improvement exceeded one standard error, further supporting the robustness of the clustering formulation.

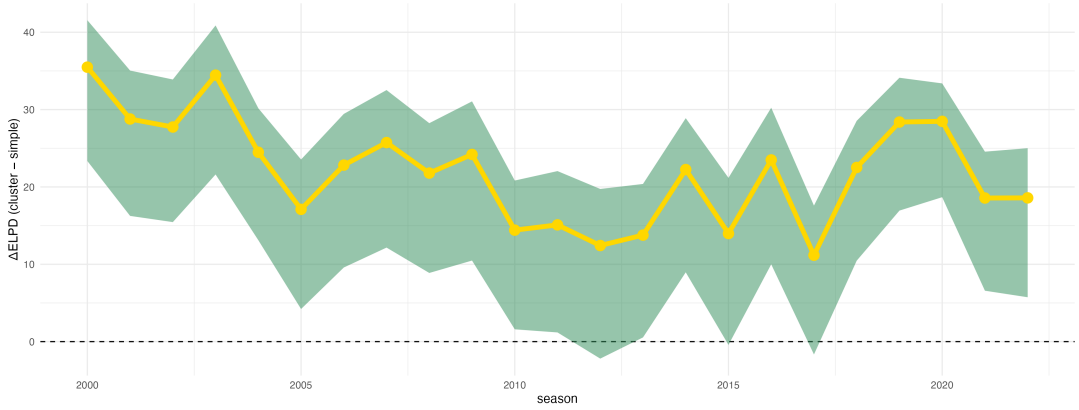


Fig 7: Predictive gain of the clustered BT–SBM over the classical BT model for each season, expressed as the difference in expected log predictive density (ΔELPD). The yellow line shows the mean ΔELPD across seasons, with shaded bands denoting $\pm 1\,\mathrm{SE}_\Delta$ obtained from PSIS–LOO. Positive values indicate superior predictive performance of the BT–SBM relative to the standard BT model. The consistently positive trend highlights the improved fit achieved by incorporating block structure into the model.

7.2.2. *Results' analysis.* We now turn to the core output of our longitudinal analysis, which examines how the inferred cluster structure evolves across seasons. Table 4 reports the posterior probabilities associated with different values of K, with the modal number of blocks for each year highlighted using a consistent colour palette adopted throughout this section[4]. A few broad patterns stand out in the evolution of cluster structure over time.

Early 2000s.. Between 2000 and 2005, the three-block solution is consistently dominant (green cells in Tab. 4), suggesting a relatively fluid and competitive upper tier. During this phase, multiple players rotated through the top cluster from one season to the next, reflecting a more open contest among elite players.

Figure 8 supports this reading by showing the size of the top block across seasons. In the early 2000s, the top cluster is noticeably larger, indicating a broader elite. This is a direct signal of competitiveness at the top: a more inclusive top tier implies less dominance by a fixed few.

Mid 2000s to 2015.. Starting in 2006, the model increasingly favours four-block solutions, and the top cluster size shrinks abruptly, marking the emergence and consolidation of the Big Four – a tightly held elite with little rotation. The big four start to consistently dominate outcomes, and this is reflected both in the reduced top-cluster size and player-level posterior inclusion probabilities, which are examined in Figure 9.

Here, each point represents the posterior probability that a given player belonged to the top block in a given season. During the early 2000s, top-cluster inclusion is more diffuse: several players exhibit intermediate probabilities, suggesting a more permeable elite. But in the following decade, a small number of players begin to dominate the top cluster with probabilities consistently near 1.

The sparsity of intermediate probabilities during this era reinforces this picture: most players were either clearly in or out of the top block, with little in-between. These fluctuations

[4]The colour scale has been manually chosen: a green tone, reminiscent of grass courts, denotes $K = 3$, whereas an orange shade, akin to the clay tournaments, represents $K = 4$.

Season \ $\widehat{K}$	2	3	4	5	6	7	8
2000/2001	0.120	0.510	0.243	0.078	0.030	0.011	0.005
2001/2002	0.013	0.565	0.269	0.100	0.035	0.013	0.003
2002/2003	0.335	0.370	0.156	0.076	0.037	0.014	0.007
2003/2004	–	0.520	0.303	0.115	0.039	0.014	0.006
2004/2005	0.002	0.512	0.288	0.122	0.048	0.018	0.007
2005/2006	–	0.218	0.379	0.241	0.097	0.041	0.015
2006/2007	–	0.017	0.299	0.285	0.179	0.107	0.061
2007/2008	–	0.534	0.285	0.113	0.040	0.016	0.006
2008/2009	–	0.172	0.372	0.238	0.120	0.057	0.026
2009/2010	–	0.051	0.433	0.300	0.124	0.055	0.022
2010/2011	–	0.278	0.322	0.201	0.103	0.051	0.024
2011/2012	–	0.151	0.382	0.252	0.120	0.055	0.024
2012/2013	–	0.023	0.312	0.289	0.184	0.105	0.048
2013/2014	–	0.021	0.365	0.303	0.176	0.080	0.035
2014/2015	–	0.024	0.418	0.305	0.160	0.061	0.023
2015/2016	–	0.021	0.504	0.282	0.116	0.049	0.019
2016/2017	–	0.082	0.396	0.278	0.133	0.066	0.029
2017/2018	–	0.259	0.315	0.216	0.110	0.056	0.025
2018/2019	–	0.346	0.332	0.175	0.086	0.035	0.016
2019/2020	0.010	0.501	0.299	0.122	0.044	0.016	0.005
2020/2021	0.352	0.371	0.165	0.072	0.024	0.009	0.005
2021/2022	–	0.248	0.341	0.210	0.110	0.054	0.021
2022/2023	0.003	0.387	0.314	0.156	0.077	0.038	0.014

TABLE 4

Posterior probabilities of the number of clusters (K) across ATP men's seasons (2000/2001–2022/2023), inferred under the Gnedin-type prior. Each row corresponds to a season, and each column to a candidate value of K. The modal number of clusters for each year is highlighted using a consistent colour palette: green shades denote $K = 3$, and orange shades denote $K = 4$. Early seasons display stronger support for $K = 3$, indicating a relatively fluid elite tier, while later years increasingly favour $K = 4$, reflecting the emergence of a more stratified competitive hierarchy during the Big Four era. More recent seasons seem to support once again the three block solution.

across both block size and individual inclusion probabilities offer a consistent and data-driven narrative of the Big Four era.

From 2018 to recent years. After 2018, the pattern loosens again: more players appear in the middle range of top-cluster inclusion probabilities, and the elite cluster begins to broaden. Taken together, Table 4, Figure 8, and Figure 9 reinforce a coherent narrative. The Big Four era is characterized by both a structural contraction of the elite cluster and consequent subtle expansions that are non-trivial to detect or quantify. In the next section, we introduce a summary indicator designed to capture and formalize this dynamic.

7.3. *Measuring Competitive Balance.* To quantify changes in competitive balance over time, we measure the Shannon entropy of the composition of the top block (block 1) at each MCMC iteration t. Let $m_1^{(t)}$ denote the number of players assigned to block 1 at iteration t, and $m_k^{(t)}$ the number in block k, for $k = 1, \ldots, K^{(t)}$. We define the proportions $p_k^{(t)} = \frac{m_k^{(t)}}{\sum_{\ell=1}^{K^{(t)}} m_\ell^{(t)}}$ for $k = 2, \ldots, K^{(t)}$. The entropy of the top-block composition is then

$$H^{(t)} = -\sum_{k=1}^{K^{(t)}} p_k^{(t)} \log p_k^{(t)}. \tag{18}$$

Low entropy ($H^{(t)} \approx 0$) indicates concentrated dominance by block 1, whereas high entropy reflects a more even distribution of players across blocks. For interpretability across iterations

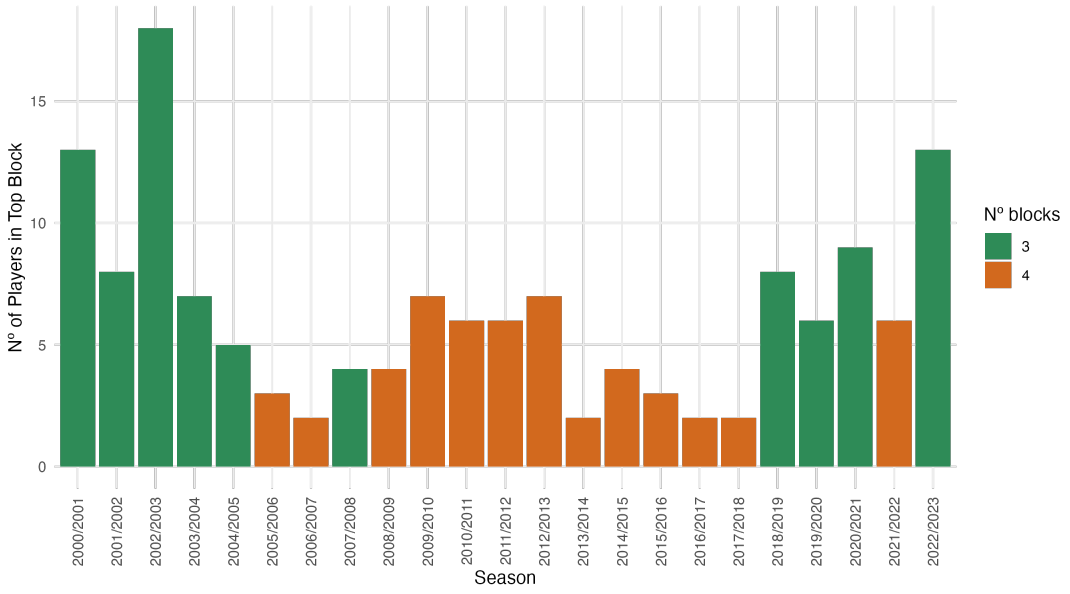


Fig 8: Estimated number of players assigned to the top cluster in each ATP season during the 2000/2001–2022/2023 ATP seasons. The top block is substantially larger in the early 2000s, indicating a broader and more competitive elite group. From 2006 onward, the size of the top cluster shrinks markedly, coinciding with the rise and consolidation of the Big Four, who dominated the upper tier with little rotation. In 2018, the top block begins to expand again, signalling the end of that concentrated dominance and the re-emergence of a more open competition structure.

with varying $K^{(t)}$, we report the normalized entropy $\widetilde{H}^{(t)} = \frac{H^{(t)}}{\log K^{(t)}} \in [0,1]$, which equals 0 under complete dominance and 1 when cluster sizes are uniform.

Figure 10 displays the posterior mean normalized entropy over time, with 95% credible intervals. The trajectory closely mirrors our earlier findings: competitive balance was highest in the early 2000s, collapsed during the Big Four era, and resurged after 2018. This metric offers a compact yet powerful diagnostic of competitive balance–and can be readily applied to other sports or domains to assess the degree of concentration or uncertainty in competitive hierarchies.

8. Conclusion and Future Directions. This work has introduced a Bayesian SBM for pairwise competitive outcomes, grounded in the Bradley–Terry framework and enriched by a nonparametric prior over partitions. At its core, the BT–SBM imposes that players in the same block have identical strength, i.e., $\lambda_i = \lambda_j \iff x_i = x_j$. This assumption induces a clustered ordering where co-clustered players share the same position in the ranking. This structure offers a principled and interpretable summary of competitive hierarchies by blending notions of similarity and dominance.

Our model was motivated by and applied to two decades of ATP men's tennis. The analysis uncovered interpretable temporal trends – most notably the emergence, consolidation, and eventual loosening of a dominant elite (the so-called Big Four). The results consistently supported a clustered ranking of players, where the top tiers showed stable membership and sharp separations, while the middle and lower tiers exhibited more diffuse and uncertain boundaries. The model's ability to express uncertainty in both strength and cluster assignment represents a clear methodological advantage over classical approaches.

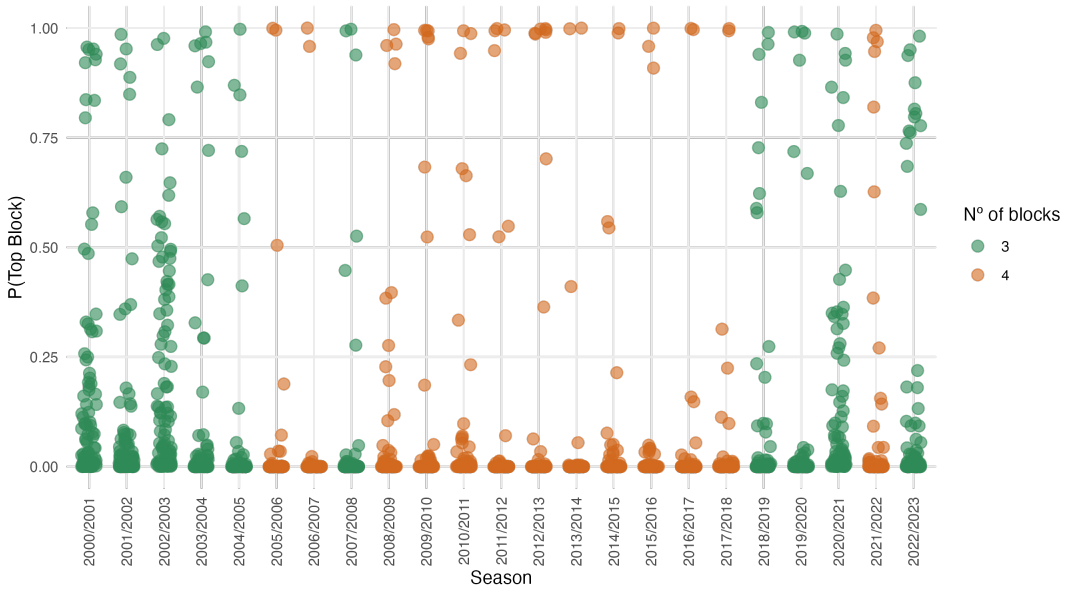


Fig 9: Posterior probability of belonging to the top cluster, by player and season, during the 2000/2001–2022/2023 ATP seasons. Each point represents a player's posterior inclusion probability in the top block for a given season; points are horizontally jittered to avoid overlap. Early seasons display a diffuse pattern with many players showing intermediate probabilities, consistent with a fluid and competitive elite. From the mid-2000s to 2018, only a small group—the Big Four—maintain probabilities close to one, reflecting near-constant dominance and reduced permeability of the top cluster. Around 2010, only a few points appear, offering limited evidence of a transition compared to the clearer regimes before and after. After 2020, intermediate probabilities become more frequent again, suggesting the re-emergence of a broader and more contested elite.

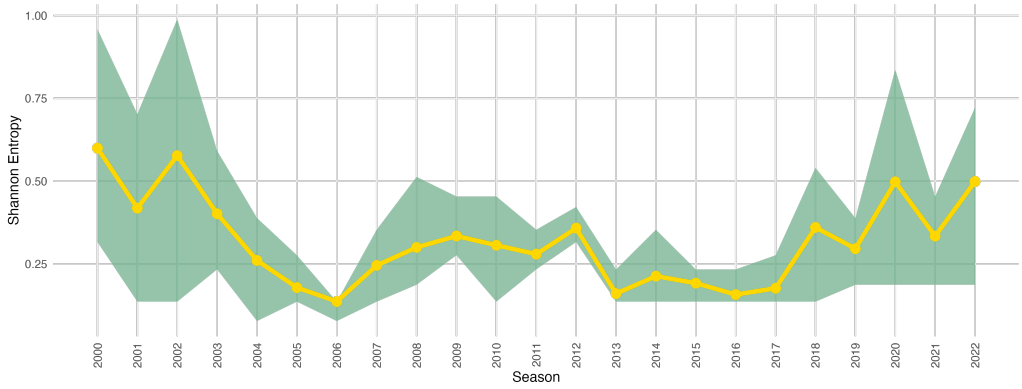


Fig 10: Posterior mean of the Shannon entropy of top-cluster composition across seasons, with 95% credible intervals. Entropy quantifies the degree of competitive balance: high values indicate a more contested situation, with less predictable outcomes, and in turn, a more competitive sport; low values reflect concentrated dominance by few players. The temporal trajectory reveals three phases: high entropy in the early 2000s (a fluid competitive landscape), a sharp decline during the Big Four era (2006–2018) marking strong concentration at the top, and a post-2019 rebound indicating renewed balance and greater turnover among elite players.

At the same time, the work acknowledges the limitations already highlighted in Section 2. Pairwise outcomes are sparse, sometimes inconsistent, and flatten away contextual information such as playing surface, player condition, or tournament setting. Moreover, the Bradley-Terry framework assumes linear stochastic transitivity, effectively reducing competition to a one-dimensional skill scale. Our formulation partially addresses these issues by introducing rank-indifference and latent clustering, but many dimensions of competitive heterogeneity remain unmodeled.

From a modelling standpoint, several extensions emerge as natural and valuable. First, our analysis treated each season independently. A richer formulation would incorporate *temporal dependence* in both cluster allocations $\mathbf{x}$ and strength parameters $\boldsymbol{\lambda}$, enabling smooth evolution of player skill and cluster structure. Hidden Markov models, Gaussian processes, or autoregressive nonparametric priors (e.g., dynamic Dirichlet processes) are promising approaches.

Second, the current model captures heterogeneity entirely through latent structure. Incorporating *covariates*–both at the edge level (e.g., surface type, tournament category, home advantage) and node level (e.g., age, handedness, ranking)–would address the present inability to account for observed factors that clearly shape outcomes.

Third, the assumption of linear stochastic transitivity remains restrictive. Competitive systems frequently display intransitivities or context dependence. Models that relax this assumption, for instance through weak or stochastic transitivity constraints (Lee and Chen, 2025; Spearing et al., 2021), offer a natural direction.

Fourth, the current model conditions on the observed match counts n_{ij}, focusing solely on modelling win outcomes $w_{ij} \mid n_{ij}$. In professional tennis, however, the frequency of matches between players is itself an informative and dynamic process. High-ranked players tend to play more matches – and against stronger opponents – which creates a feedback loop between exposure and success that is not captured by the present framework. A more comprehensive model would treat n_{ij} as endogenous to the competitive process.

Finally, although our empirical focus was tennis, the BT–SBM framework is broadly applicable. It can be employed in any setting wherever pairwise directed comparisons are observed – ranging from animal dominance in ecology (Baldassarre et al., 2023), to political preferences (Loewen, Rubenson and Spirling, 2012), transportation modelling (Hatzinger and Mazanec, 2007), and performance evaluation in AI systems (Chiang et al., 2024). Notably, this approach provides a natural diagnostic for *competition uncertainty* across time or contexts, as shown by our use of entropy to quantify elite concentration.

In summary, the BT–SBM constitutes a first step toward a richer understanding of competitive structures. Its combination of interpretability, uncertainty quantification, and probabilistic coherence makes it a compelling tool for analysing systems of competition. At the same time, it opens several avenues for future work, in particular the incorporation of covariates, temporal dynamics, and more flexible transitivity assumptions.

Code and Reproducibility. All code to reproduce the analyses, figures, and tables is openly available. We separate the *analysis workflow* and the *general-purpose package* as follows:

- **Reproducibility repository** (`BT-SBM-Bradley-Terry-Stochastic-Block-Model`): github.com/laposanti/BT-SBM-Bradley-Terry-Stochastic-Block-Model
- **R package** (`BTSBM`): laposanti.github.io/BTSBM/

Acknowledgments. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

Funding. This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6049. The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number 12/RC/2289_P2.

REFERENCES

BALDASSARRE, A., DUSSELDORP, E., D'AMBROSIO, A. et al. (2023). The Bradley–Terry Regression Trunk Approach for Modeling Preference Data with Small Trees. *Psychometrika* **88** 1443–1465. https://doi.org/10.1007/s11336-022-09882-6

BASINI, F., TSOULI, V., NTZOUFRAS, I. and FRIEL, N. (2023). Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model. **186** 530–556. https://doi.org/10.1093/jrsssa/qnad007

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs: the method of paired comparisons. **39** 324–345. https://doi.org/10.2307/2334029

CARON, F. and DOUCET, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. **21** 174–196. https://doi.org/10.1080/10618600.2012.638220

CHIANG, W.-L., ZHENG, L., SHENG, Y., ANGELOPOULOS, A. N., LI, T., LI, D., ZHU, B., ZHANG, H., JORDAN, M. I., GONZALEZ, J. E. and STOICA, I. (2024). Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*. *ICML'24*. JMLR.org. https://doi.org/10.48550/ARXIV.2403.04132

CROON, M. A. and LUIJKX, R. (1993). Latent structure models for ranking data. In *Probability models and statistical analyses for ranking data*, (M. A. Fligner, J. S. Verducci, J. Berger, S. Fienberg, J. Gani, K. Krickeberg, I. Olkin and B. Singer, eds.) **80** 53–74. Springer New York. https://doi.org/10.1007/978-1-4612-2738-0_4

DE BLASI, P., LIJOI, A. and PRUENSTER, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica* **23** 1299–1321. https://doi.org/10.5705/ss.2012.047

DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. and RUGGIERO, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 212–229. https://doi.org/10.1109/TPAMI.2013.217

FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability* **23** 1721-1754. https://doi.org/10.1214/12-AAP843

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230. https://doi.org/10.1214/aos/1176342360

GLICKMAN, M. E. (2008). Bayesian locally optimal design of knockout tournaments. **138** 2117–2127. https://doi.org/10.1016/j.jspi.2007.09.007

GNEDIN, A. (2010). A Species Sampling Model with Finitely Many Types. *Electronic Communications in Probability* **15** 79–88. https://doi.org/10.1214/ECP.v15-1532

GNEDIN, A. and PITMAN, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138** 5674–5685. https://doi.org/10.1007/s10958-006-0335-z

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **159** 385–443.

GORMLEY, I. C. and MURPHY, T. B. (2009-06). A grade of membership model for rank data. **4** 265–295. https://doi.org/10.1214/09-ba410

GUIVER, J. and SNELSON, E. (2009). Bayesian inference for Plackett–Luce ranking models. 377–384. https://doi.org/10.1145/1553374.1553423

HATZINGER, R. and MAZANEC, J. A. (2007). Measuring the part worth of the mode of transport in a trip package: An extended Bradley–Terry model for paired-comparison conjoint data. *Journal of Business Research* **60** 1290–1302. https://doi.org/10.1016/j.jbusres.2007.04.010

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

HSIAO, C. and WANG, I. (2021). Joint clustering and ranking from heterogeneous pairwise comparisons. IEEE international symposium on information theory. In *2021 IEEE international symposium on information theory (ISIT)* 2036–2041. IEEE. https://doi.org/10.1109/isit45174.2021.9517936

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. **2** 193–218. https://doi.org/10.1007/bf01908075

LEE, S. M. and CHEN, Y. (2025). Pairwise Comparisons without Stochastic Transitivity: Model, Theory and Applications. https://doi.org/10.48550/arXiv.2501.07437

LEGRAMANTI, S., RIGON, T., DURANTE, D. and DUNSON, D. B. (2022). Extended stochastic block models with application to criminal networks. **16** 2369-2395. https://doi.org/10.1214/21-aoas1595

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 715–740. https://doi.org/10.1111/j.1467-9868.2007.00609.x MR2370077

LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability* **18**. https://doi.org/10.1214/07-AAP495

LOEWEN, P. J., RUBENSON, D. and SPIRLING, A. (2012). Testing the power of arguments in referendums: a Bradley–Terry approach. *Electoral Studies* **31** 212-221. Special Symposium: Germany's Federal Election September 2009. https://doi.org/10.1016/j.electstud.2011.07.003

LU, C., DURANTE, D. and FRIEL, N. (2025). Zero-inflated stochastic block modelling of efficiency-security trade-offs in weighted criminal networks. https://doi.org/10.1093/jrsssa/qnaf029

MEILĂ, M. (2007). Comparing clusterings—an information based distance. **98** 873–895. https://doi.org/10.1016/j.jmva.2006.11.013

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9** 249–265. https://doi.org/10.2307/1390653

NEWMAN, M. E. (2022). Ranking with multiple types of pairwise comparisons. **478**. https://doi.org/10.1098/rspa.2022.0517

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. **96** 1077–1087. https://doi.org/10.1198/016214501753208735

PEARCE, M. and EROSHEVA, E. A. (2025). Bayesian rank-clustering. 1–49. https://doi.org/10.1017/psy.2025.10014

PIANCASTELLI, L. S. C. and FRIEL, N. (2024). The clustered Mallows model. **35**. https://doi.org/10.1007/s11222-024-10555-w

PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. *Lecture Notes–Monograph Series* **30** 245–267. https://doi.org/10.1214/lnms/1215453576

PITMAN, J. (2006). *Combinatorial Stochastic Processes* **1875**. Springer-Verlag, Berlin/Heidelberg. https://doi.org/10.1007/b11601500

PITMAN, J. and YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25** 855–900. https://doi.org/10.1214/aop/1024404422

RASTELLI, R. and FRIEL, N. (2018). Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing* **28** 1169–1186. https://doi.org/10.1007/s11222-017-9786-y

SCHÖTTL, K., KEINER, M., METZ, V. and KAINZ, P. (2025). The Financial Break Even in Professional Tennis: From Which Ranking Can You Afford Your Life from Professional Tennis? Manuscript under review. https://doi.org/10.2139/ssrn.5206641

SEYMOUR, R. G., SIRL, D., PRESTON, S., DRYDEN, I. L., ELLIS, M. J. A., PERRAT, B. and GOULDING, J. (2020). The Bayesian spatial Bradley–Terry model: urban deprivation modeling in Tanzania. **71** 288–308. https://doi.org/10.48550/arXiv.2010.14128

SOCH, J., THE BOOK OF STATISTICAL PROOFS, SARITAŞ, K., MAJA, MONTICONE, P., FAULKENBERRY, T. J., MARTIN, O. A., KIPNIS, A., BALKUS, S., LFKDLFDLK, ALLEFELD, C., ATZE, H., KNAPP, A., MCINERNEY, C. D., LO4DING00, OHAN, V., AMVOSK and MAXGROZO (2025). StatProofBook/StatProofBook.github.io: StatProofBook 2024. Zenodo. https://doi.org/10.5281/ZENODO.4305949

SPEARING, J., TAWN, J., IRONS, D. and PAULDEN, T. (2021). Modelling intransitivity in pairwise comparisons with application to baseball data. **32** 1383–1392. https://doi.org/10.1080/10618600.2023.2177299

VACA-RAMÍREZ, F. and PEIXOTO, T. P. (2022). Systematic assessment of the quality of fit of the stochastic block model for empirical networks. *Phys. Rev. E* **105** 054311. https://doi.org/10.1103/PhysRevE.105.054311

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing* **27** 1413–1432. https://doi.org/10.48550/arXiv.1507.04544

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis* **13** 559-626. https://doi.org/10.1214/17-ba1073

WAINER, J. (2023). A Bayesian Bradley–Terry model to compare multiple ML algorithms on multiple data sets. *Journal of Machine Learning Research* **24** 1–34. https://doi.org/10.5555/3648699.3649040

WHELAN, J. T. (2017). Prior distributions for the Bradley–Terry model of paired comparisons. https://doi.org/10.48550/arXiv.1712.05311

WU, R., XU, J., SRIKANT, R., MASSOULIÉ, L., LELARGE, M. and HAJEK, B. (2015). Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* 449-450. ACM. https://doi.org/10.1145/2835776.2835787

APPENDIX A: DERIVATIONS OF FULL CONDITIONALS

In this appendix, we derive the full conditionals used in the Gibbs sampler for our augmented Bradley–Terry model with block structure. We exploit the latent variables $Z_{ij} \sim \Gamma(n_{ij}, \lambda_{x_i} +$

λ_{x_j}) as described in the augmentation scheme of Eq. (8). We aim to derive the full conditional distributions for the block strengths λ_k and the cluster assignments x_i, based on the joint augmented likelihood:

$$\mathcal{L}(\mathbf{W}, \mathbf{Z} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N}) \propto \prod_{i<j} \lambda_{x_i}^{w_{ij}} \lambda_{x_j}^{n_{ij}-w_{ij}} \cdot Z_{ij}^{n_{ij}-1} \exp\left\{-(\lambda_{x_i} + \lambda_{x_j}) Z_{ij}\right\}. \tag{A.1}$$

Constants not involving $\boldsymbol{\lambda}$, $\mathbf{x}$, or $\mathbf{Z}$ are omitted throughout.

A.1. Full conditionals for $\boldsymbol{\lambda}$. We isolate the terms in A.1 involving λ_{x_i}:

$$\ell(\mathbf{W}, \mathbf{Z} \mid \boldsymbol{\lambda}, \mathbf{x}, \mathbf{N}) \propto \underbrace{\sum_{i \neq j} w_{ij} \log \lambda_{x_i}}_{\text{(I)}} - \underbrace{\sum_{i<j} (\lambda_{x_i} + \lambda_{x_j}) Z_{ij}}_{\text{(II)}}.$$

We rewrite each part in terms of individual-level quantities.

A.1.1. *Term (I).*. Since $\log \lambda_{x_i}$ is constant in j, we factor it out:

$$\sum_{i \neq j} w_{ij} \log \lambda_{x_i} = \sum_i \left(\sum_{j \neq i} w_{ij} \right) \log \lambda_{x_i} = \sum_i w_i \log \lambda_{x_i},$$

where $w_i = \sum_{j \neq i} w_{ij}$ is the total number of wins by individual i.

A.1.2. *Term (II).*. Each pair (i, j) contributes $\lambda_{x_i} Z_{ij} + \lambda_{x_j} Z_{ij}$, so the sum becomes:

$$\sum_{i<j} (\lambda_{x_i} + \lambda_{x_j}) Z_{ij} = \sum_i \lambda_{x_i} \sum_{j \neq i} Z_{ij} = \sum_i \lambda_{x_i} Z_i,$$

where $Z_i = \sum_{j \neq i} Z_{ij}$.

A.1.3. *Final Form.*. The complete-data log-likelihood, up to constants, becomes:

$$\ell(\mathbf{W}, \mathbf{Z} \mid \boldsymbol{\lambda}, \mathbf{x}) \propto \sum_i \left[w_i \log \lambda_{x_i} - \lambda_{x_i} Z_i \right].$$

Now, recall the prior $\lambda_k \sim \Gamma(a, b)$. Collecting the terms in the likelihood by block k, let $I_k = \{i : x_i = k\}$. The contribution of λ_k becomes:

$$\ell(\lambda_k \mid \cdot) \propto \lambda_k^{\sum_{i \in I_k} w_i} \exp\left\{ -\lambda_k \sum_{i \in I_k} Z_i \right\}.$$

Combining it with the Gamma prior:

$$p(\lambda_k \mid \mathbf{W}, \boldsymbol{Z}, \mathbf{x}) \propto \lambda_k^{a-1+\sum_{i \in I_k} w_i} \exp\left\{ -\lambda_k (b + \sum_{i \in I_k} Z_i) \right\},$$

we obtain as a posterior:

$$\lambda_k \mid \mathbf{W}, \boldsymbol{Z}, \mathbf{x} \sim \Gamma\left(a + \sum_{i \in I_k} w_i,\ b + \sum_{i \in I_k} Z_i \right).$$

A.2. Full Conditional for x_i. We now derive the conditional probability of assigning player i to block $k \in \{1, \ldots, K\} \cup \{K+1\}$. Suppose i is temporarily removed from its current block.

A.2.1. *Existing Block $k \in \{1, \dots, K\}$.*. Split the product in A.1 into terms involving i and those that do not:

$$\mathcal{L} = C(\mathbf{W}) \left[\prod_{j \neq i} \underbrace{\Lambda_{ij}(k)}_{\text{depends on } k} \right] \left[\prod_{\substack{p<q \\ p,q \neq i}} \Lambda_{pq} \right],$$

where $C(\mathbf{W})$ collects terms independent of $\mathbf{x}$ and $\boldsymbol{\lambda}$, and the second product is constant with respect to k. For $j > i$, one has

$$\Lambda_{ij}(k) = \lambda_k^{w_{ij}} \lambda_{x_j}^{n_{ij}-w_{ij}} Z_{ij}^{n_{ij}-1} e^{-\lambda_k Z_{ij}} e^{-\lambda_{x_j} Z_{ij}}.$$

For $j < i$, the roles of w_{ij} and $n_{ij} - w_{ij}$ swap, but the exponent of λ_k still contributes either w_{ij} or $n_{ij} - w_{ij}$. Summing over all $j \neq i$ therefore yields

$$\prod_{j \neq i} \Lambda_{ij}(k) \propto \lambda_k^{w_i} \exp(-\lambda_k Z_i).$$

The prior assignment probability from the Gibbs-type prior is

$$\Pr(x_i = k \mid \mathbf{x}_{-i}) \propto \psi_{n,K}(m_{-i,k} - \sigma),$$

where $m_{-i,k}$ is the size of cluster k excluding node i. Hence, the full conditional for $x_i = k$ becomes

$$p(x_i = k \mid \cdot) \propto \psi_{n,K}(m_{-i,k} - \sigma) \lambda_k^{w_i} \exp(-\lambda_k Z_i).$$

A.2.2. *New Block $k = \{K + 1\}$.*. Here, x_i is assigned to a new cluster. We integrate λ_k out:

$$\int_0^\infty \lambda_k^{w_i} e^{-\lambda_k Z_i} \cdot \frac{b^a}{\Gamma(a)} \lambda_k^{a-1} e^{-b\lambda_k} d\lambda_k = \frac{b^a \Gamma(a + w_i)}{\Gamma(a)} (b + Z_i)^{-(a+w_i)} \quad \text{for } k = K + 1.$$

The prior for a new cluster is $\psi_{n,K+1}$, so the full conditional becomes:

$$p(x_i = K + 1 \mid \cdot) \propto \psi_{n,K+1} \cdot \frac{b^a \Gamma(a + w_i)}{\Gamma(a)} (b + Z_i)^{-(a+w_i)}.$$

Together, these define the unnormalized probabilities for $x_i \in \{1, \dots, K + 1\}$, which are normalized to yield the full conditional distribution over block assignments.

APPENDIX B: SCALE ALIGNMENT AND PRIOR STANDARD DEVIATION

In this appendix section, we provide a detailed motivation for the heuristic choice of fixing the Gamma prior's rate as $b = \exp\{\psi(a)\}$ and for choosing the shape parameter a in an interpretable way. Block strengths are modelled as $\lambda_k \sim \text{Gamma}(a, b)$ – using the shape–rate parametrization. Due to identifiability concerns (see Sect. 6.1), the sampler renormalizes the occupied blocks at each iteration so that their log-strengths have arithmetic mean equal to zero, i.e.

$$\frac{1}{K} \sum_{k=1}^{K} \log \lambda_k = 0 \qquad \Longrightarrow \qquad \mathbb{E}[\log \lambda_k \mid \mathbf{W}, \mathbf{x}] = 0.$$

Given this zero-induced expectation, it becomes natural to define the hyperprior so that the a-priori expected value of $\log \lambda_k$ is zero as well.

This moment-matching construction can be interpreted as a simple *hyperprior heuristic*. Rather than placing the prior on the raw block strengths λ_k, we work on $\log \lambda_k$, and then we match the prior and posterior means:

$$\mathbb{E}[\log \lambda_k \mid a, b, \mathbf{x}] \;=\; \mathbb{E}[\log \lambda_k \mid \mathbf{W}, a, b, \mathbf{x}] \;=\; 0.$$

From a modelling perspective, this alignment ensures that the prior and the likelihood operate on the same scale. From a computational standpoint, it prevents systematic up- or down-scaling in the creation of new clusters–an effect that would otherwise depend on arbitrary choices of a and b.

Formally, for $\lambda_k \sim \Gamma(a, b)$,

$$\mathbb{E}[\log \lambda_k \mid a, b] = \psi(a) - \log b,$$

where $\psi(a)$ is the *digamma function*, i.e. the derivative of $\log \Gamma(a)$ (see Soch et al., 2025). By setting this expectation to zero, we define b_0 as

$$b_0 := \exp\{\psi(a)\},$$

thereby ensuring that the prior mean of $\log \lambda_k$ is consistent with the global scale imposed by the sampler.

To study the role of this choice in the propensity to create new clusters (see Eq. (13)), define the bias $\delta = \psi(a) - \log b$, the discrepancy between the a-priori expectation and the normalization-induced (a-posteriori) expectation, which is zero. We want to see how the likelihood contribution to the creation of a new cluster (Eq. (13)) changes with respect to δ. Therefore, we compute

$$\begin{aligned}\Delta_i(\delta) &:= \log L_i(K+1 \mid a, b) - \log L_i(K+1 \mid a, b_0) \\ &\quad - a\,\delta - (a + w_i)\Big[\log\big(e^{\psi(a)-\delta} + Z_i\big) - \log\big(e^{\psi(a)} + Z_i\big)\Big] \quad \text{where } b_0 = \exp\psi(a).\end{aligned}$$

Differentiating gives

$$\Delta_i'(\delta) \;=\; -a \;+\; (a + w_i)\frac{b}{b + Z_i}, \qquad b \;=\; e^{\psi(a)-\delta},$$

so that $-a < \Delta_i'(\delta) < w_i$. Here $w_i = \sum_{j \neq i} w_{ij}$ denotes the total number of successes for item i, and Z_i is the aggregate augmentation mass attached to i. Specifically, with n_{ij} matches between i and j and the Gamma augmentation, we have

$$Z_{ij} \,\big|\, n_{ij}, \lambda_{x_i}, \lambda_{x_j} \;\sim\; \Gamma\big(n_{ij},\, \lambda_{x_i} + \lambda_{x_j}\big)\,,$$

as defined in Eq. (10). We report it for convenience also:

$$Z_i \;=\; \sum_{j \neq i} Z_{ij}, \qquad \mathbb{E}[Z_i \,|\, \boldsymbol{\lambda}, \mathbf{x}] \;=\; \sum_{j \neq i} \frac{n_{ij}}{\lambda_{x_i} + \lambda_{x_j}}.$$

Thus Z_i scales with the total number of matches played by item i (larger when i has faced many opponents) and is down-weighted when i or its opponents have large λ-values. Intuitively, Z_i can be interpreted as an "strength-adjusted" measure of the sample size associated with i: more matches and weaker opponents yield a larger Z_i.

The sign of $\Delta_i'(\delta)$ reveals two distinct regimes, clearly visible in Figures 11–12. When Z_i is small, $\frac{b}{b+Z_i} \approx 1$ and $\Delta_i'(\delta) \approx w_i > 0$: increasing δ (i.e. choosing $b < e^{\psi(a)}$) *inflates* the odds of creating a new cluster, making such items more likely to initiate new blocks. This effect appears in Figure 11 as the upward shift of the curves for low-Z_i items under the misaligned

prior (in this specific case $b = 1$, but other b values would yield the same effect). When Z_i is large, $\frac{b}{b+Z_i} \approx 0$ and $\Delta_i'(\delta) \approx -a < 0$: increasing δ then *deflates* the new-cluster odds, as seen in the lower placement of the high-Z_i curves in the same figure. Hence, any $\delta \neq 0$ induces a Z_i-dependent tilt in the new-cluster likelihood, while the aligned choice $\delta = 0$ (Figure 12) removes this effect entirely, causing all curves to stabilize along the same monotone and downward sloping trajectory.

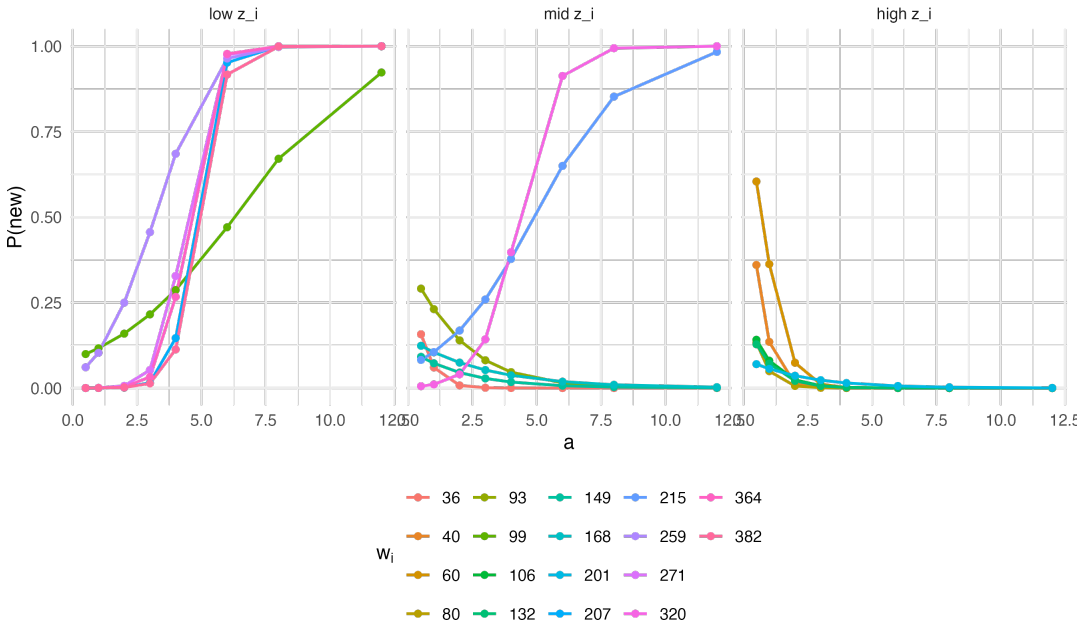


Fig 11: How the probability of creating a new cluster varies with a in the misaligned case ($b = 1$, $\delta = \psi(a)$). The three panels correspond to empirical tertiles of Z_i: low ($Z_i \leq \widehat{q}_{1/3}$), mid ($\widehat{q}_{1/3} < Z_i \leq \widehat{q}_{2/3}$), and high ($Z_i > \widehat{q}_{2/3}$), where $\widehat{q}_{1/3}$ and $\widehat{q}_{2/3}$ are the sample 33% and 66% quantiles of $\{Z_i\}_{i=1}^n$. For small Z_i, the evidence for a new cluster is systematically inflated (left panel), whereas for large Z_i it is deflated (right panel). Hence, setting $b = 1$ in the prior for $\log \lambda$ induces a non-linear, Z-dependent effect on the probability of creating a new block.

B.1. Prior standard deviation on log–λ. Once the mean of $\log \lambda_k$ is fixed, the remaining degree of freedom is its *standard deviation* of $\log \lambda_k$ which, for $\lambda_k \sim \text{Gamma}(a, b)$, corresponds to:

$$\text{Var}[\log \lambda_k] = \psi_1(a), \qquad \tau \equiv \text{SD}(\log \lambda_k) = \sqrt{\psi_1(a)},$$

where $\psi_1(a)$ is the *trigamma function*, the derivative of the digamma (Soch et al., 2025, Eq. 5.15.1). Hence, the shape parameter a controls the prior heterogeneity on $\log \lambda_k$: large a (small τ) yields concentrated priors with nearly homogeneous block strengths, while small a (large τ) allows for more heterogeneous blocks.

We therefore treat τ as an intuitive hyperparameter controlling the prior heterogeneity of block strengths. The mapping between a and τ follows from inverting the following expression:

$$\psi_1(a) = \tau^2 \qquad \Longrightarrow \qquad a = \psi_1^{-1}(\tau^2).$$

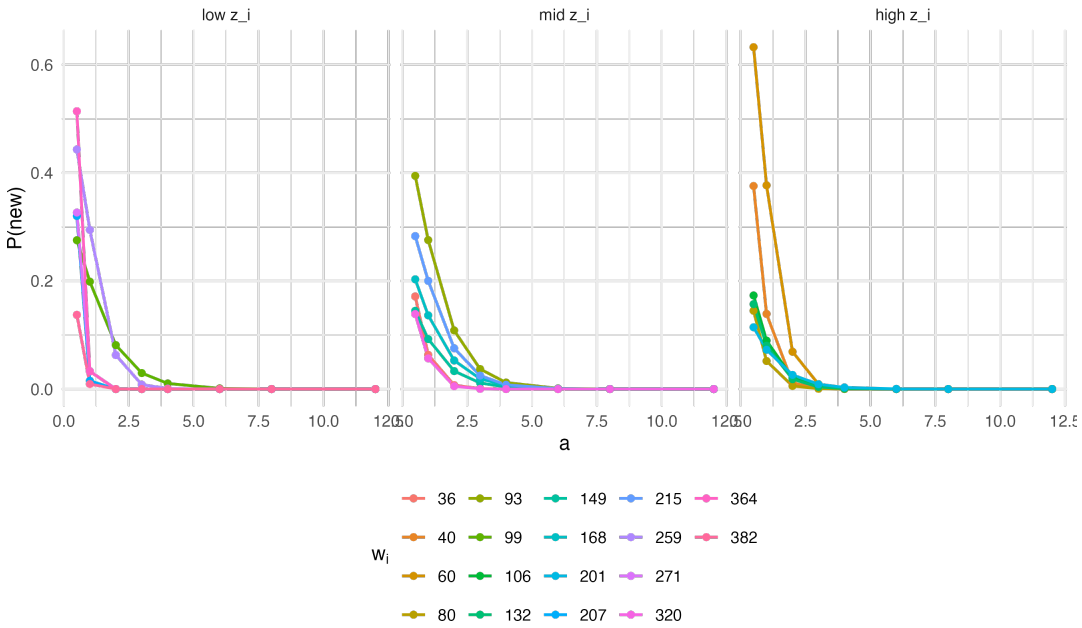


Fig 12: How the probability of creating a new cluster varies with a in the aligned case ($b_0 = \exp\{\psi(a)\}$, $\delta = 0$). The three panels correspond to empirical tertiles of Z_i: low ($Z_i \le \widehat{q}_{1/3}$), mid ($\widehat{q}_{1/3} < Z_i \le \widehat{q}_{2/3}$), and high ($Z_i > \widehat{q}_{2/3}$), where $\widehat{q}_{1/3}$ and $\widehat{q}_{2/3}$ are the sample 33% and 66% quantiles of $\{Z_i\}_{i=1}^n$. In all panels, the prior alignment removes the Z-dependent bias in the probability of creating a new cluster, producing consistent behaviour across Z values.

Representative values are shown in Table 5.

a	$\tau = \mathrm{SD}(\log \lambda_k)$	$\mathrm{SD}(\lambda_k) = \sqrt{a}/b$
6	0.43	0.44
5	0.47	0.50
4	0.53	0.57
3	0.63	0.69
2	0.80	0.93
1	1.28	1.78

TABLE 5

Relationship between the Gamma shape parameter a and the corresponding standard deviation of $\log \lambda_k$, denoted by $\tau = \sqrt{\psi_1(a)}$. The second column reports τ, which measures prior variability on the log scale. The rightmost column shows the standard deviation of λ_k under the scale alignment $b = \exp\{\psi(a)\}$, illustrating how dispersion on the log scale translates into actual variability in the block strengths.

In practice, moderate values such as $\tau \in [0.5, 0.8]$ (i.e. $a \in [4, 2]$) provide a good balance between flexibility and stability. From Table 5, these correspond to a standard deviation of $\mathrm{SD}(\lambda_k) \approx 0.6 - 0.9$ under the aligned prior $b = \exp \psi(a)$, meaning that block strengths typically vary within roughly a 50%–100% range around their geometric mean. Thus, τ acts as a single interpretable control knob for the prior heterogeneity of λ_k, while the alignment $b = \exp \psi(a)$ ensures this variability remains invariant to the scale of λ_k, identifiable, and well behaved with respect to the auxiliary varible Z .

APPENDIX C: CLOSED-FORM MEAN AND VARIANCE OF K UNDER THE GNEDIN MODEL

In this appendix, we report the closed-form expressions for the mean and the variance of K under the Gnedin prior, where K denotes the *sample* number of clusters (that is, the number of groups represented in the observed data). The calculations below are a routine specialization of standard manipulations for Gibbs–type partitions (Favaro, Lijoi and Prünster, 2013). We work in the finite–type case ($\sigma = -1$) known as the Gnedin model, for which the number of clusters has the explicit pmf

$$p(K = k \mid n, \gamma) = \binom{n}{k} \frac{(1-\gamma)_{k-1}\,(\gamma)_{n-k}}{(1+\gamma)_{n-1}}, \qquad k = 1, \ldots, n,$$

as reported in Eq. (3), with $(a)_m = \Gamma(a+m)/\Gamma(a)$ the rising Pochhammer symbol; see Gnedin (2010, Sec. 6, Eq. (9)). As in the general Gibbs–type literature, we combine (i) index–shift identities for binomials, $k\binom{n}{k} = n\binom{n-1}{k-1}$ and $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$, with (ii) expressing expectations in factorial form, and (iii) repeated applications of the Chu–Vandermonde (binomial–Pochhammer) identity

$$\sum_{r=0}^{N} \binom{N}{r} (a)_r (b)_{N-r} = (a+b)_N,$$

exactly in the manner used in the BNP literature to collapse partition sums; see Lijoi, Prünster and Walker (2008, App. A.2, Lemma A.1) for a precise statement and its use in Gibbs–type derivations, and Favaro, Lijoi and Prünster (2013) for related generalizations and the use of steps (i)–(iii). Nothing essentially new is given in this $\sigma = -1$ setting: the same general off–the–shelf identities in (Favaro, Lijoi and Prünster, 2013) lead directly to the closed–form mean and variance reported below, and we include the explicit formulas only for completeness and ease of reference.

C.1. First factorial moment and mean. Using the identity $k\binom{n}{k} = n\binom{n-1}{k-1}$, we write:

$$\mathbb{E}[K] = \sum_{k=1}^{n} k\, p(K = k \mid n, \gamma) = \frac{n}{(1+\gamma)_{n-1}} \sum_{k=1}^{n} \binom{n-1}{k-1} (1-\gamma)_{k-1} (\gamma)_{(n-1)-(k-1)}.$$

Setting $r = k-1$ and applying the Chu–Vandermonde identity with $a = 1-\gamma$, $b = \gamma$, and $N = n-1$, we obtain

$$\mathbb{E}[K] = \frac{n}{(1+\gamma)_{n-1}} (1)_{n-1} = \frac{\Gamma(n+1)\,\Gamma(1+\gamma)}{\Gamma(n+\gamma)}.$$

This expression coincides with the mean used for prior specification in recent applications of the Gnedin model (e.g. Legramanti et al., 2022), and can be viewed as the $\sigma = -1$ specialization of the general Gibbs–type expectation derived via factorial–moment expansions in Favaro, Lijoi and Prünster (2013).

C.2. Second factorial moment. Proceeding analogously and using $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$, we can derive the second factorial moment as:

$$\mathbb{E}[K(K-1)] = \frac{n(n-1)}{(1+\gamma)_{n-1}} \sum_{k=2}^{n} \binom{n-2}{k-2} (1-\gamma)_{k-1} (\gamma)_{n-k}.$$

Letting $r = k-2$ and factoring out one term from the rising factorial, $(1-\gamma)_{k-1} = (1-\gamma)(2-\gamma)_{k-2}$, we obtain

$$\mathbb{E}[K(K-1)] = \frac{n(n-1)(1-\gamma)}{(1+\gamma)_{n-1}} \sum_{r=0}^{n-2} \binom{n-2}{r} (2-\gamma)_r (\gamma)_{(n-2)-r}.$$

Applying the Chu–Vandermonde identity again, now with $a = 2-\gamma$, $b = \gamma$, and $N = n-2$, yields

$$\mathbb{E}[K(K-1)] = \frac{n(n-1)(1-\gamma)\,(n-1)!}{(1+\gamma)_{n-1}}.$$

This calculation mirrors the repeated use of the Vandermonde–Pochhammer convolution adopted by Lijoi, Prünster and Walker (2008) to simplify factorial–moment sums in Gibbs–type models, here specialized to the finite–type case $\sigma = -1$.

Variance. Ordinary and factorial moments are related through $\mathbb{E}[K^2] = \mathbb{E}[K(K-1)] + \mathbb{E}[K]$ (see, e.g., Pitman, 2006; Favaro, Lijoi and Prünster, 2013), so the variance follows immediately:

$$\mathrm{Var}(K) = \frac{n(n-1)(1-\gamma)\,(n-1)!}{(1+\gamma)_{n-1}} + \frac{\Gamma(n+1)\Gamma(1+\gamma)}{\Gamma(n+\gamma)} - \left(\frac{\Gamma(n+1)\Gamma(1+\gamma)}{\Gamma(n+\gamma)}\right)^2.$$

C.2.1. *Specification for the present context.*. For the empirical setting considered here, with $n = 105$ and $\gamma = 0.8$, we have

$$\mathbb{E}[K] = \frac{\Gamma(106)\Gamma(1.8)}{\Gamma(105.8)} \approx 2.36, \qquad \mathrm{Var}(K) \approx 45.95.$$

Thus, while the prior mean is about two to three clusters, the variance is quite large, yielding a heavy–tailed distribution that allows for additional clusters when supported by the data. This behaviour reflects the role of γ in the partition prior: smaller values of γ inflate the variability of K, favouring richer partitions, whereas larger values concentrate the prior on more parsimonious configurations (see also Gnedin, 2010; Favaro, Lijoi and Prünster, 2013; De Blasi et al., 2015).

APPENDIX D: SIMULATION STUDY

To validate the proposed model, we conduct a simulation study designed to test whether the BT–SBM can recover both the true number of blocks and the latent partition under a correctly specified data-generating process. The experiment mirrors the empirical ATP setting while allowing full control over the underlying cluster structure.

D.1. Data Generation. We generate synthetic datasets to assess the model's ability to recover the true number of clusters K^* and partition $\mathbf{x}^*$. The number of items is fixed to $n = 150$, and the target number of clusters varies as $K^* \in \{3, \ldots, 10\}$. For each target K^*, we generate a structured but interpretable dataset according to the following steps.

First, we simulate the edge structure to reproduce the sparsity observed in the real data: for each unordered pair (i, j), an edge is included with probability 0.5; otherwise, the pair is excluded. This ensures that roughly half of the potential matches are realized, producing a network density comparable to that of the empirical ATP season.

Second, we induce clusters of approximately equal size by repeating the true labels sequentially, $x_i^* = (1, 2, \ldots, K^*, 1, 2, \ldots, K^*, \ldots)$ up to $n = 150$. This yields equally-sized clusters allowing us to isolate the model's clustering accuracy from block-imbalance effects.

Third, we assign block strengths deterministically as equally spaced values $\lambda_k^* \in [0.1, 3]$, $k = 1, \ldots, K^*$, ensuring clear yet realistic differences between cluster-level abilities.

Finally, conditional on the generated cluster structure, we sample the number of matches between connected pairs from a $\mathrm{Poisson}(5)$ distribution, and simulate match outcomes w_{ij} under the Bradley–Terry- SBM likelihood.

Algorithm 3 Synthetic data generation for the BT–SBM. Edges match real-data sparsity; clusters are balanced; block strengths are equally spaced for clear separation.

Require $n = 105$, target $K^* \geq 2$
for all pairs (i, j) with $1 \leq i < j \leq n$ **do**
  Draw $e_{ij} \sim \mathrm{Bernoulli}(0.5)$
  if $e_{ij} = 1$ **then**
    Draw $n_{ij} \sim \mathrm{Poisson}(5)$
  else
    $n_{ij} \leftarrow 0$
  end if
end for
Assign cluster labels $x_i^* = (1, 2, \ldots, K^*, 1, 2, \ldots)$ truncated at $n = 105$
Set $\lambda_k^* \leftarrow 0.1 + (k-1)\,\dfrac{3 - 0.1}{K^* - 1}$ for $k = 1, \ldots, K^*$
for all pairs (i, j) with $n_{ij} > 0$ **do**
  Draw $w_{ij} \sim \mathrm{Binomial}\left(n_{ij},\ \dfrac{\lambda_{x_i^*}}{\lambda_{x_i^*} + \lambda_{x_j^*}}\right)$
end for

D.2. Recovery of the true number of clusters. For each $K^* \in \{3, \ldots, 10\}$, we generate 23 replicates and fit the model using 30,000 MCMC iterations with $\gamma = 4$. The posterior distribution of K is summarized by its posterior mode $\widehat{K}$, as reported in Table 6.

TABLE 6
Cross–tabulation of the estimated number of clusters $\widehat{K}$ (rows) versus the true number K^* (columns). Diagonal entries represent exact recovery; off–diagonal entries correspond to near misses.

$\widehat{K} \setminus K^*$	3	4	5	6	7	8	9	10
3	23	-	-	-	-	-	-	-
4	-	23	-	-	-	-	-	-
5	-	-	23	-	-	-	-	-
6	-	-	-	23	-	-	-	-
7	-	-	-	-	23	-	-	-
8	-	-	-	-	-	22	1	-
9	-	-	-	-	-	4	18	1
10	-	-	-	-	-	1	13	9

The results show that for $K^* \leq 9$, the model almost always recovers the correct number of clusters. For $K^* = 10$, the posterior tends to slightly underestimate K, typically returning $\widehat{K} = 9$. This mild shrinkage is expected: as the number of blocks grows, the effective sample size per block decreases, leading to occasional merging of clusters during posterior inference. Nevertheless, in all scenarios, the true K^* remains within the 95% posterior credible interval, indicating good posterior calibration.

D.3. Recovery of the true partition. We next assess how well the model recovers the true partition $\mathbf{x}^*$. Posterior samples $\mathbf{x}^{(t)}$ are relabelled to address label-switching using Algorithm 2, and summarized by the VI loss estimator, as reported in Sect. 6. Clustering accuracy is evaluated via the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), which ranges from 0 (no agreement) to 1 (perfect recovery).

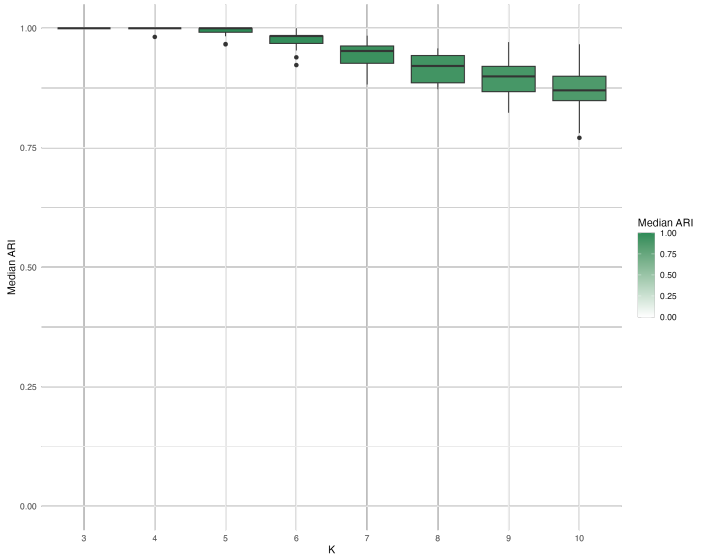


Fig 13: Adjusted Rand Index (ARI) across 23 replicates for each true number of clusters K^*. Each boxplot shows the distribution of ARI scores for a fixed K^*.

Figure 13 shows consistently high recovery accuracy. For $K^* \leq 9$, median ARI values exceed 0.9, with narrow dispersion across replicates. For $K^* = 10$, performance slightly decreases (median ARI ≈ 0.85), mainly due to the smaller sample size per block. Even in these cases, the estimated partitions remain close to the ground truth, and the VI loss estimator provides smooth, stable summaries of posterior clustering uncertainty.

Overall, the simulation study confirms that the BT–SBM equipped with a Gnedin prior reliably recovers both the number of clusters and the underlying partition structure when the model is correctly specified.

APPENDIX E: BOUNDARY PARTITIONS ON THE CREDIBLE BALL SURFACE

To provide a clearer view of posterior uncertainty around the estimated partition $\widehat{\mathbf{x}}$, we report the magnified components of Figure 4, that is, the surface partitions of the 95% credible ball defined in Section 6. Following Wade and Ghahramani (2018), these *boundary partitions* correspond to clusterings that are maximally distant from $\widehat{\mathbf{x}}$ under the VI loss, while still lying within the credible region. The distance employed here is the VI distance, which we omit to repeat for brevity. We display the three bounds reported in Figure 4:

1. *Vertical upper bound*, reported in Fig. 14, with its $K = 3$ blocks, is the most distant partition with the least amount of blocks (what we call the *coarsest partition*) within the 95% credible ball;
2. *Vertical lower bound*, reported in Fig. 15, showcases $K = 11$ blocks and represents the most distant partition with the maximum number of blocks (the *finest partition*) within the credible region;

3. *Horizontal bound*, reported in Fig. 16 with $K = 6$, is the most distant partition within the credible ball irrespective of K.

Together they characterize and narrow down the posterior uncertainty distribution with respect to the inferred partition $\widehat{\mathbf{x}}$

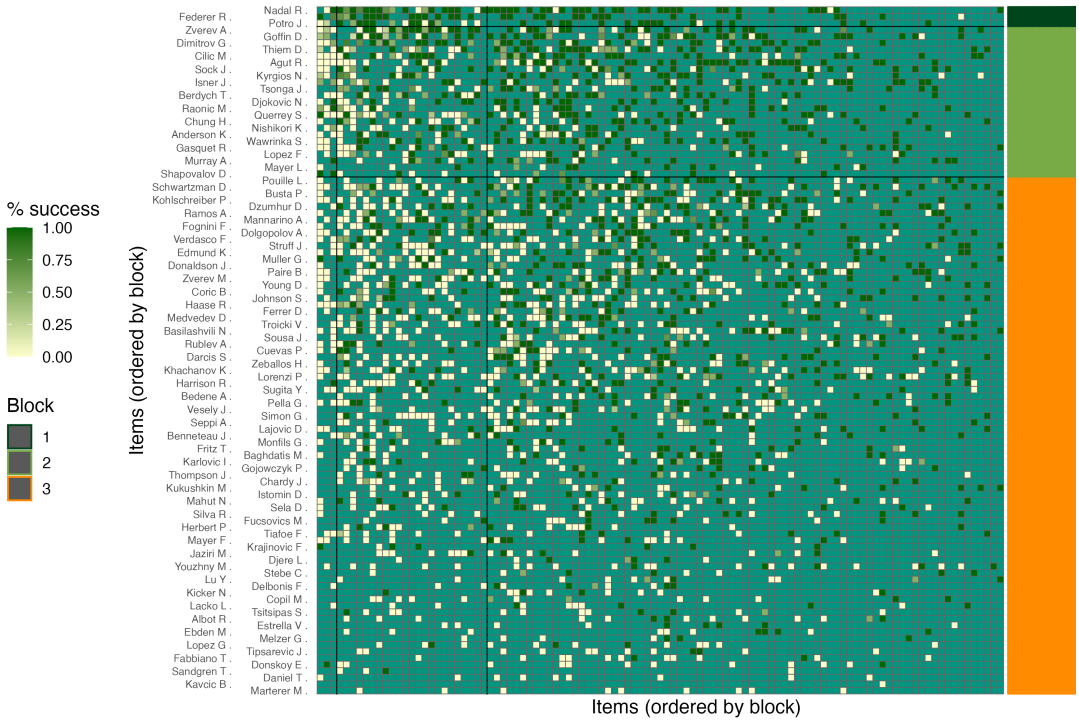


Fig 14: Adjacency matrix of the 2017/2018 season reordered according to the vertical upper bound ($K = 11$) of the 95% credible ball around $\widehat{\mathbf{x}}$. This is the coarsest partition within the credible ball that is maximally distant from $\widehat{\mathbf{x}}$. The only difference with respect to Fig. 3, is the fact that the top block now includes also *del Potro*, while other blocks remain identical. For more details about how to read this figure, check the caption of Fig. 3.

E.1. Description and interpretation. We begin by outlining what the three configurations share, and then describe where they differ. Across all of them, the elite block is sharply defined: *Nadal* and *Federer* appear in every configuration, while *Djokovic*, *Wawrinka*, and occasionally *del Potro* join them depending on the resolution. These players also display the highest assignment uncertainty in Fig. 5, indicating that the model hesitates only over the precise composition of the elite, not its existence. At the opposite end of the hierarchy, the weakest block is always cohesive and clearly separated. Between these two extremes lies a consistent group of contenders striving to enter the elite. In sum, the basic $K = 3$ tiered structure–elite, contenders, and weaker players–remains visible across all partitions, which mainly differ in how finely these three blocks are subdivided.

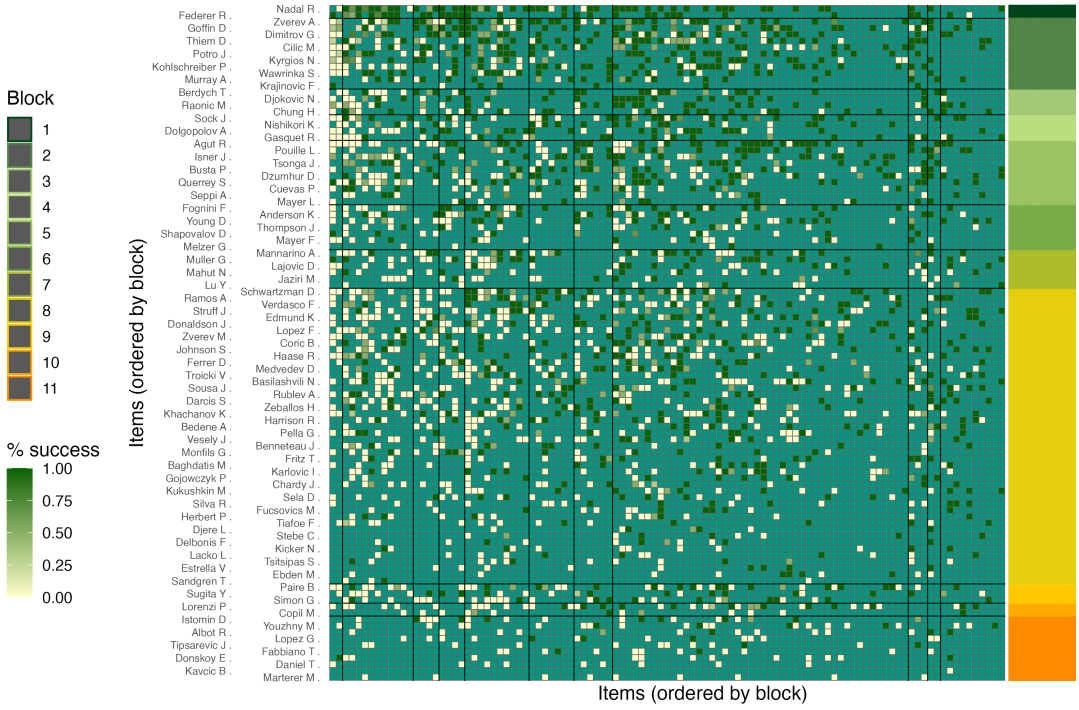


Fig 15: Adjacency matrix of the 2017/2018 season reordered according to the vertical lower bound ($K = 11$) of the 95% credible ball around $\widehat{\mathbf{x}}$. This is the finest partition within the credible ball, where both mid-tier players and weaker players are split into several smaller groups, capturing more subtle distinctions in playing strength. The elite block remains narrow–still dominated by *Nadal* and *Federer*–while the weakest players continue to cluster at the bottom into a large block. The main change is a dense fragmentation of the middle, where the large mid-tier group in the point estimate divides into six sub-blocks, and a further differentiation at the lower end, where additional blocks emerge to represent varying strength levels among the weaker players. For more details on how to read this figure, see the caption of Fig. 3.

Beyond this variability in *resolution*, the main differences among boundary partitions occur near the interfaces between the three principal blocks. Uncertainty concentrates at these transition regions – between the elite and contenders, and between the mid-tier and the weakest group–while the cores of each block remain stable.

For example, the *horizontal bound* (Fig. 16, $K = 6$) introduces additional blocks precisely at these interfaces: parts of the upper mid-tier and of the weakest group detach to form two small intermediate clusters. In the *vertical lower bound* (Fig. 15, $K = 11$), this pattern becomes more pronounced. The weakest block in the point estimate splits into several subgroups–two minor and one more substantial–capturing finer distinctions among the least successful players. Yet even in these high-resolution configurations, the same three main regions are still identifiable, merely refined by additional layers at their borders.

E.2. Summary and take-away. The boundary partitions confirm a stable *tiered structure* of player strengths, while revealing uncertainty in how finely those tiers should be divided.

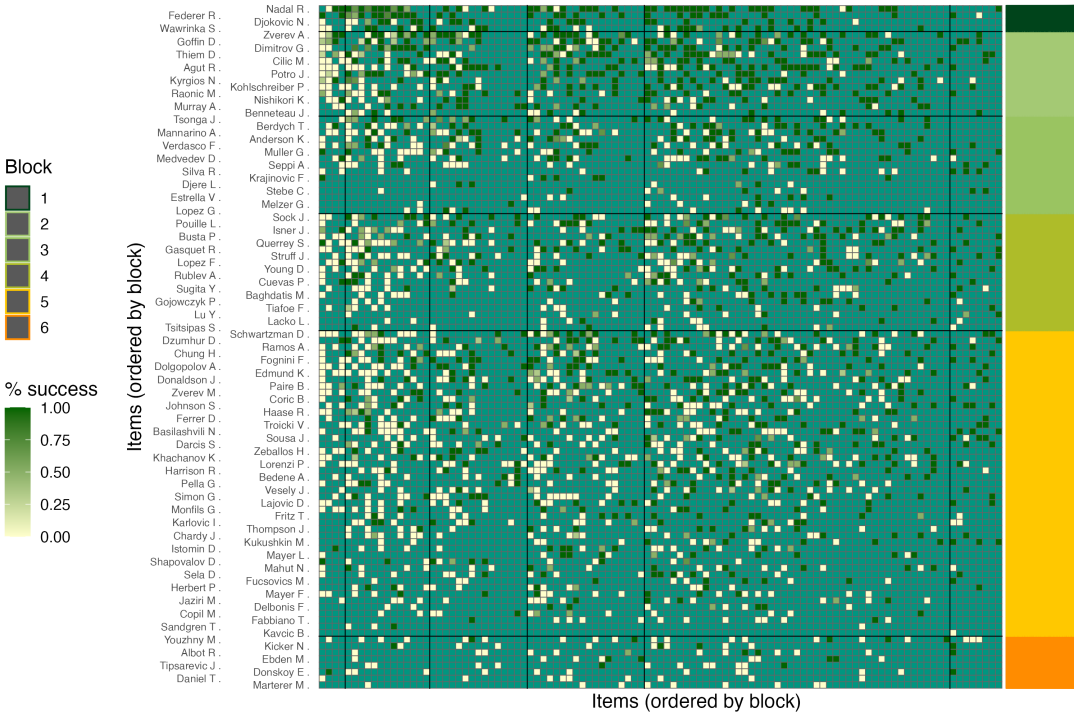


Fig 16: Adjacency matrix of the 2017/2018 season reordered according to the horizontal bound ($K = 6$) of the 95% credible ball around $\widehat{\mathbf{x}}$. This configuration doubles the number of clusters relative to the VI estimate, producing a more even subdivision into blocks. The elite block slightly expands to include *Djokovic* and *Wawrinka*, while the overall cluster sizes become more balanced. As in the other configurations, the main differences arise at the boundaries between blocks, where most of the posterior uncertainty concentrates. For more details on how to read this figure, see the caption of Fig. 3.

As the resolution increases, new blocks appear only at the boundaries, not in the core of the existing ones. This supports our modelling choice of *tiered ranks*: the tiers themselves capture the robust features of the posterior, while fine-grained allocations across neighbouring blocks reflect quasi-arbitrary choices of resolution. In short, the model identifies who belongs to which broad level of the hierarchy–the tiered order is clear–but the exact placement of the boundaries remains uncertain, and that uncertainty is an integral part of the data-generating process.