Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model

Francesca Basini[1] Vasiliki Tsouli Ioannis Ntzoufras[2] Nial Friel[3,4*]

[1]Department of Mathematics, University of Warwick, UK
[2] Department of Statistics, Athens University of Economics and Business, Greece
[3]School of Mathematics and Statistics, University College Dublin, Dublin 4, Ireland
[4] Insight Centre for Data Analytics, University College Dublin, Dublin 4,Ireland

July 20, 2021

Abstract

Competitive balance is a desirable feature in any professional sports league and encapsulates the notion that there is unpredictability in the outcome of games as opposed to an imbalanced league in which the outcome of some games are more predictable than others, for example, when an apparent strong team plays against a weak team. In this paper, we develop a model-based clustering approach to provide an assessment of the balance between teams in a league. We propose a novel Bayesian model to represent the results of a football season as a dense network with nodes identified by teams and categorical edges representing the outcome of each game. The resulting stochastic block model facilitates the probabilistic clustering of teams to assess whether there are competitive imbalances in a league. A key question then is to assess the uncertainty around the number of clusters or blocks and consequently estimation of the partition or allocation of teams to blocks. To do this, we develop an MCMC algorithm that allows the joint estimation of the number of blocks and the allocation of teams to blocks. We apply our model to each season in the English premier league from $1978/79$ to $2019/20$. A key finding of this analysis is evidence which suggests a structural change from a reasonably balanced league to a two-tier league which occurred around the early 2000's.

Keywords Bayesian statistics · sports analytics · Markov Chain Monte Carlo · stochastic block model · competitive balance

1 Introduction

Uncertainty in the outcome in sporting events would, on the face of it, appear to be a key ingredient in a successful sporting league. If the outcome of a sporting contest is very predictable it is reasonable to anticipate that this may diminish the contest from the point of view of the spectator. This captures the notion that competitive balance in a sporting league should be, in some sense, where the result of any match is quite unpredictable. The study of competitive balance is one which has interested researchers for some time, particularly in fields as diverse as economics to sports science. In economics, there is much interest to investigate relationships between the competitiveness of a sporting league and the distribution of revenue within the league. For example, Brandes and Franck (2007) study the effect which competitiveness of the Bundesliga in Germany has on attendance while (Manasis, Ntzoufras, and Reade 2021) study the effect of competitiveness for several European leagues; see also (Penn and Berridge 2019) and (Plumley, Ramchandani, and Wilson 2018) who present a perspective in the context of the English Premier League. More generally, competitive balance is of vital interest and used as an argument by sports governing bodies to ensure that, for example, revenue streams are allocated proportionately. In fact, in the context of soccer, that was partly the reason for introducing the UEFA Financial Fair Play Regulations (FFP) to limit excessive spending and investments by club owners together with the intention to ensure that clubs budgets are regulated, reducing debts from season to season and creating profit. The new regulations were agreed in 2009 and put in practice three years later.

*Address for correspondance: nial.friel@ucd.ie

Despite the strong interest in competitive balance, it has largely remained quite a vague concept. Hence, there is not a universally accepted measure or index of competitive balance. In the literature of sports analytics, several authors have proposed statistical measures of competitive balance. Most of these indices are based on intuition about the notion of competitiveness. These are typically descriptive statistics based on summarising the spread of points or wins achieved by each team at the end of the league (Evans 2014) and references therein. For example, the Herfindahl–Hirschman index of competitive balance is adapted from the Herfindahl-Hirschman index which is used to measure the spread of market share by firms in a given industry. Other definitions are based on the idea of entropy among many others. We refer the reader to (Manasis et al. 2013) for a statistical perspective on this topic; see also in (Evans 2014) for a comprehensive review of this literature. A thorough search of the related methodology reveals very few publications with competitive balance indexes which are based on solid statistical models or techniques. Some exceptions include the approach of (Koning 2000) based on multinomial probit type of model with response the game outcome (win, draw, loss), the index introduced by (Haan et al. 2008) based on a regression model for the goal difference and the autoregressive win percentage by (Vrooman 1995) which is based on an auto-regressive model. Further, model based approaches are the variance decomposition measure (Eckard 1998) and CBR (Humphreys 2002) which are based on ANOVA type models. Finally, two applications of Markov models provided measures of competitive balance of sports leagues: (a) the statistical test of theoretical and actual transitional probabilities which allows for the testing of a wide range of hypothesis regarding competitive balance relating to strata of a league structure (Koop 2004) and (b) a 'Gini type' single statistic measure of the competitive balance of a league system (Buzzacchi et al. 2003).

The notion of competitiveness in sports is multidimensional with many qualitative characteristics. This has generated a considerable number of different indexes and measures of competitiveness. One of the characteristics, well explained in the Economics literature is the the so-called three-dimensional factorization of competitive balance; see for example (Cairns 1987). These three dimensions are: a) the match uncertainty, which refers to a particular game, b) the seasonal dimension, which focuses on the relative quality of teams in the course of a particular season and c) the between-seasons dimension, which focuses on the relative quality of teams across seasons.

Another aspect that has recently considered by (Manasis et al. 2013) is the multi-levelled nature of European leagues offering multiple awards as opposed to the common single prize offered by North American ones. These multiple awards are related with qualification positions (of different quality) for playing at UEFA champions league, Europa league and to league positions leading to relegation. The attention of the economic analysis of competitive balance is the effect on the fans' behaviour, which is the longstanding "Uncertainty of Outcome Hypothesis" (Fort and Maxcy 2003).

The novel statistical model which we develop is based on the idea of a stochastic block model (SBM) which is popular in the analysis of relational network data (Nowicki and Snijders 2001). The central idea of an SBM is to partition the nodes of a network (or graph) into blocks (or clusters) so that nodes within a block tend to have the same probability of being connected by an edge and that this connection probability varies by block. We extend this framework to data arising from a football league. By analogy, we conceptually consider teams as being nodes of a complete graph, where every pair of nodes is connected by an edge and also that the relational edge between two teams is the categorical outcome, a win, draw or loss, resulting from when they play against each other. In particular and by analogy to a standard SBM, we aim to partition teams into blocks so that the outcome (a win, draw or loss) when two teams from a same block play each other follows the same multinomial distribution, the parameters of which vary according to the block label of each team. Effectively, teams within a block are considered to be *balanced* as the outcome for each game follows the same multinomial probability mass function. As such, an important aspect of our modelling framework is to infer both the number of blocks, but also to assign probability to the membership or allocation of each team to a block. In particular, if we infer that a single block is most likely, this suggests that the league is balanced. While additionally estimating the number of teams in the strongest block, integrating over the posterior uncertainty in the number of blocks, provides a further means to quantify the strength and quality of competitiveness. For example, if we find that a two block model has most support, but that the weakest block contains only two teams, then there is evidence that the league is quite competitive for the teams in the top block.

Hence, in this paper we offer an innovative implementation of an SBM for football (which can be used also for other competitive sports between two opponents). First of all, we should emphasize that the proposed method is more rigorous than other methods in the literature which only use the information from the final league table. Importantly, relational information of each individual game is considered instead. By considering the game specific outcome, we directly consider competitiveness in multiple dimensions: for each game separately and in terms of seasonal dimension. Moreover, the across season competitiveness can be also evaluated by analysing the main outcomes of SBM across different seasons. Our proposed method is solely based on a solid statistical model and not on intuition leading to multiple outcomes for measuring different characteristics of competitive balance. To be more specific:

1. Primarily, the number of groups specify whether the teams can be classified in different blocks/levels of competitiveness. This can be also related with the different levels of awards offered in a league. An SBM consisting of one group correspond to leagues where all teams have similar probabilities to win against each other implying a balanced league.

2. The size of the teams belonging in the top block/cluster is also an important index that can be evaluated in terms of competitiveness. Even in cases of two groups, a top-team block with many members indicate high competitiveness for the higher league positions. Note that as (Manasis et al. 2013) have indicated, the competitiveness for top league positions are more important for the fans than competition for bottom league positions (leading to relegation).

3. Finally, the persistence of specific teams appearing in the top-teams group and the variety of different teams appearing in the the top-teams group is an important measure of across-seasons competitiveness.

Finally, it is important to note that our proposed SBM method evaluates the overall competitiveness of a league and not the competition for the overall champion each season which is very important from the point of view of fans; see (Manasis et al. 2013) for a related discussion and empirical evidence.

Our motivation for this paper was specifically to analyse the English Premier League to try to shed light on the question of whether this league has become more imbalanced over time. The popular opinion seems to be that it is gradually becoming more and more imbalanced especially since the end of the old first division in 1991 and inception of the Premier League. Moreover, it is commonplace to see discussion in the media about the *big-six* teams, usually referring to the teams, Arsenal, Chelsea, Liverpool, Manchester City, Manchester United and Tottenham Hotspur, as having emerged in the recent past. Yet, it is not immediately clear how grounded in reality this notion actually is. To the best of our knowledge, there has been no statistical modelling framework in place to answer these and other questions. Our aim is to fill this gap.

As such, we apply our model to over 40 seasons of the old English first division/Premier league. Our findings are that the league was relatively balanced until the early part of the 2000s, but that it has become quite imbalanced since then. Over the first half of our study from $1978/79$ to around $2002/03$ we find many seasons for which a single block model is preferred and that in seasons where a two block is estimated to have highest posterior probability, that there is often considerable support for a single block model. Additionally, we find that typically the number of teams in the strongest block is often large, again suggesting that the league was relatively balanced for the majority of seasons prior to 2000. This is in contrast with our findings for the second half of the study period. In particular, we find that since the $2003/04$ season a two block model is typically best supported by the data and that membership to each block has become quite sharp or polarised, in the sense that the probability of allocation of any team to the strongest block is either very high or very low. Moreover, since $2003/04$ the number of teams in the top block varies between 2 and 8 teams (with the exception of only two seasons, $2010/11$ and $2015/16$). While since $2009/10$ the strongest block most often contains 6 teams and is relatively stable in terms of the membership of teams in this block, lending some support to the notion of a *big-six* groups of team in this period.

This paper is structured as follows. We begin in Section 2 by providing a brief summary of two measures of competitive balance and indicate their shortcoming. We present the form of the data which we analyse in Section 3 pointing out that the data can be considered as an adjacency matrix of a complete network. Section 4 develops the novel stochastic block model for this type of data. While Section 5 outlines an MCMC algorithm which we use to sample from our Bayesian model. The substantive application of our model to 42 seasons spanning the old English first division through to the present day Premier League is presented in Section 6. The paper concludes with Section 7.

2 Statistical approaches to assess competitive balance

In this section we present and discuss two indicative popular indices used to measure specific aspects of competitiveness in sports leagues and then we proceed with our proposed method. The first index is the Herfindahl–Hirschman index (HHI) which was originally used to measure the dominance of a firm or company to a specific industrial field and, hence, it is an indicator of the amount of competition among them (Hirschman 1945; Herfindahl 1950; Hirschman 1964). HHI naturally lends itself to the context of competitive balance in sports (Owen et al. 2007) since some of the basic notions are in common with company dominance in industry.

The second measure we explore is relative entropy which is an asymmetrical measure of similarity between the relative frequencies of the data and a theoretical distribution. Here the relative frequencies are replaced by the proportion of points won by each team and this is compared with the proportion of points expected in a perfectly balanced league (which corresponds to that of a uniform distribution where each team has the same amount of points).

Both of these measures are representative of a group of competitive balance measures that focus on the spread of points or wins at the end of the season. Hence all the relevant information is taken from the final league table; see at (Manasis et al. 2013; Evans 2014) for a comprehensive review of relative measures. Note that the relative entropy can be also adopted to represent game-by-game uncertainty but to the best of our knowledge it has not been used in this context in the competitive balance literature.

Here, we introduce a sophisticated statistical approach based on network modelling in order to analyse the competitiveness between teams in a football league. Our proposed method accounts for individual games and automatically identifies groups/clusters of different strength. Single block leagues are indicative of high competitiveness where every team can beat every opponent. This is, intuitively, a seminal characteristic of the Premier league for specific seasons according to the perspective of football fans and sports media. On the other hand, two block leagues may indicate a two-stage competition between teams of different strength (but not always – this depends on which and how many teams are separated from the main body of the league and the posterior uncertainty in the number of blocks).

2.1 Herfindahl–Hirschman index of competitive balance

Perhaps the most widely used means to assess competitive balance within a season in a sporting league is the Herfindahl–Hirschman index of competitive balance (Owen, Ryan, and Weatherston 2007). This index is based on assessing a measure of the spread of points share in a given season. Suppose that team i scored s_i points over the course of a season in a league involving n teams. Then one defines the HHICB index of competitive balance (HHICB) in terms of $p_i := s_i / \sum_1^n s_i$, the proportion of points achieved by team i, as

$$\text{HHICB} = n \sum_{i=1}^{n} p_i^2.$$

It is therefore simply a measure of the variability of the vector $(p_1, \ldots, p_n)$. When each team has an identical proportion of points so that $p_i = 1/n$, then HHICB=1. In Figure 1 (a) the HHICB statistic is plotted for each season from $1978/79$ to $2018/19$, the duration of the study in this paper. We remark that there is a general trend that the HHICB statistic is increasing over time lending some evidence to the hypothesis that the league has become more imbalanced over time. We also remark that this is simply a single number summary of each season and it is therefore difficult to draw any qualitative or quantitative conclusions from this.

2.2 Relative entropy as a measure of competitive balance

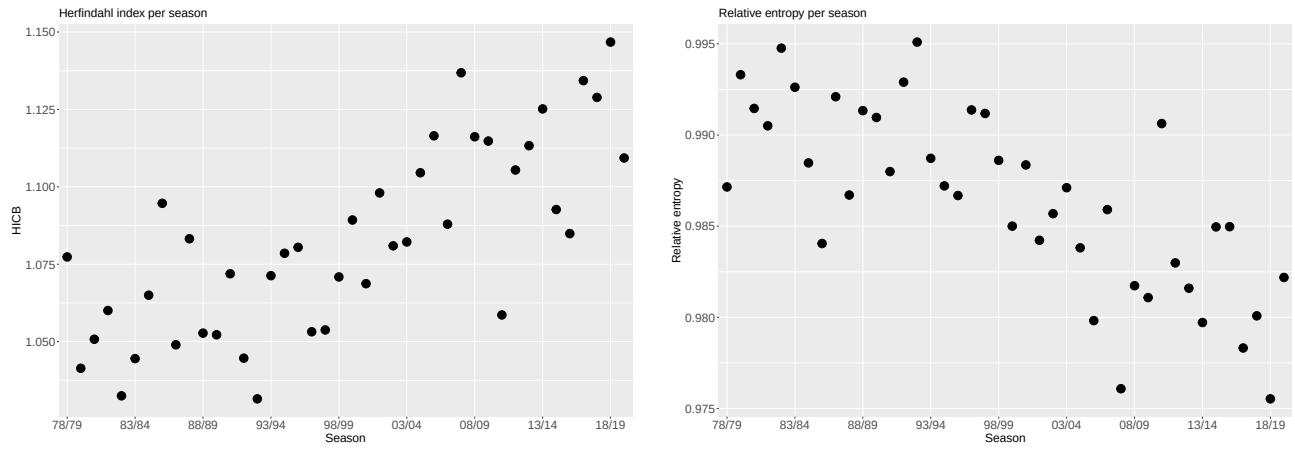
A natural approach to summarise the vector of proportion of points share among all n teams in a league, $(p_1, \ldots, p_n)$, is to use the concept of entropy (Horowitz 1997). Here one may define the relative entropy for a given season as:

$$\frac{\sum_{i=1}^{n} p_i \log(p_i)}{\log(1/n)}.$$

This statistic takes a maximum value of 1 in the case where p_i, the proportion of points share for team i is $1/n$ for all n teams and this corresponds to the case of a perfectly balanced league. While lower values of relative entropy correspond to a more imbalanced league. Figure 1 (b) displays the relative entropy statistic for each season over the course of this study. Here there is a general trend towards lower relative entropy over time, once again coherent with the hypothesis that premier league has generally become more imbalanced over time.

2.3 An alternative statistical modelling approach

Each of the two measures of competitive balance outlined in the previous sections are somewhat limited as they are simply univariate statistics and therefore it is not straightforward to make any qualitative conclusions such as whether a season is balanced. Moreover, if there is evidence to suggest that a season is imbalanced, it would be useful to give an indication as to the nature of this imbalance, for example, which teams, if any, are stronger than the rest. To date, to the best of our knowledge, there is no literature which provides such a statistical framework. We aim to address this shortcoming here and develop a fully probabilistic model of competitive balance using a stochastic block modelling framework. The proposed network model offers multiple outcomes with information about the quality of the leagues and the relative competitiveness. Generally, the method we propose is more computationally demanding but richer in terms of results and inference we obtain from the final output it offers. Essentially the idea, which we now develop, is based on a statistical model which probabilistically partitions teams into blocks (or clusters) so that, broadly speaking, the multinomial outcome (a win, draw or loss for the home team) when any two teams in the same block play one



(a) (b)

Figure 1: (a) The Herfindahl–Hirschman index of competitive balance (HHICB) is plotted for each season. This illustrates a general trend that HHICB is increasing over time and so consistent with the notion that the premier league has become more imbalanced over time. (b) Relative entropy is plotted for each season. Here high values of relative entropy correspond to more balanced league and this plot suggests that the premier league is gradually becoming more imbalanced over time.

another is the same. That is to say that if we estimate that a single cluster (consisting of all teams) is most probable, then this provides evidence that the league is balanced. Moreover, in the case where the model estimates support for a league with more than one cluster, we assign probabilities of membership (or allocation) for each team to any of the blocks. This in turn would allow one to assess the presence of blocks or clusters of teams which are broadly competitive with each other. Additionally, by integrating over the uncertainty in the number of blocks we can estimate the number of teams in the strongest block and this provides another means to assess competitive balance. This perspective is particularly important for seasons in which there is broadly equal support for a one or a two block model, hence accounting for this uncertainty is an important factor to accommodate. Assessing the constituent teams in the strongest block of teams over time is of interest in the context of the English premier league as there is anecdotal evidence of the emergence of a *big-six* block of teams over the past decade or so, namely, a collection of teams which are stronger than the remaining teams leading to a competitively imbalanced league. Our work aims to provide Bayesian framework to tackle these types of questions.

3 Representing the outcome for a season as a results matrix

To begin, we introduce the format of the data which we analyse each season. In particular, we represent the entire collection of results for every game in a single season in the form of a matrix. Here we consider the typical league scenario where each team plays every other team twice, once at home and once away from home. Therefore in a league with N teams, there are $N \times (N-1)$ fixtures. The results of all these fixtures can be summarised in a $N \times N$ matrix, R. Each cell r_{ij} contains the result of team i playing against team j when team i plays at *home*. Clearly, the diagonal entries are missing as no team plays against itself. The matrix for the season $2018/19$ is displayed in Figure 2 (a).

In the modelling framework considered in this paper we take as outcome variable for every game, the three categorical variables, "win", "draw" or "loss", denoted by $1, 2$ or 3, respectively. Therefore this leads us to summarise the matrix R with a transformed matrix $\boldsymbol{y}$ which we term a results matrix,

$$\boldsymbol{y} = \begin{pmatrix} - & y_{12} & \dots & y_{1j} & \dots & y_{1N} \\ y_{21} & - & \dots & y_{2j} & \dots & y_{2N} \\ \dots & \dots & - & \dots & \dots & \\ y_{i1} & y_{i1} & \dots & - & \dots & y_{iN} \\ \dots & \dots & \dots & \dots & - & \dots \\ y_{N1} & y_{N2} & \dots & y_{Nj} & \dots & - \end{pmatrix}, \tag{1}$$

where $y_{ij} \in \{1, 2, 3\}$, for $i = 1, \dots, N; \quad j = 1, \dots, N; j \neq i$. The categorical outcomes for season $2018/19$ are displayed in Figure 2 (b), where the categorical variables "win", "draw" or "loss" are displayed using the colours green, yellow and red, respectively.

Home\Away	ARS	BOU	BHA	BUR	CAR	CHE	CRY	EVE	FUL	HUD	LEI	LIV	MCI	MUN	NEW	SOU	TOT	WAT	WHU	WOL
Arsenal	~	5~1	1~1	3~1	2~1	2~0	2~3	2~0	4~1	1~0	3~1	1~1	0~2	2~0	2~0	2~0	4~2	2~0	3~1	1~1
Bournemouth	1~2	~	2~0	1~3	2~0	4~0	2~1	2~2	0~1	2~1	4~2	0~4	0~1	1~2	2~2	0~0	1~0	3~3	2~0	1~1
Brighton & Hove Albion	1~1	0~5	~	1~3	0~2	1~2	3~1	1~0	2~2	1~0	1~1	0~1	1~4	3~2	1~1	0~1	1~2	0~0	1~0	1~0
Burnley	1~3	4~0	1~0	~	2~0	0~4	1~3	1~5	2~1	1~1	1~2	1~3	0~1	0~2	1~2	1~1	2~1	1~3	2~0	2~0
Cardiff City	2~3	2~0	2~1	1~2	~	1~2	2~3	0~3	4~2	0~0	0~1	0~2	0~5	1~5	0~0	1~0	0~3	1~5	2~0	2~1
Chelsea	3~2	2~0	3~0	2~2	4~1	~	3~1	0~0	2~0	5~0	0~1	1~1	2~0	2~2	2~1	0~0	2~0	3~0	2~0	1~1
Crystal Palace	2~2	5~3	1~2	2~0	0~0	0~1	~	0~0	2~0	2~0	1~0	0~2	1~3	1~3	0~0	0~2	0~1	1~2	1~1	0~1
Everton	1~0	2~0	3~1	2~0	1~0	2~0	2~0	~	3~0	1~1	0~1	0~0	0~2	4~0	1~1	2~1	2~6	2~2	1~3	1~3
Fulham	1~5	0~3	4~2	4~2	1~0	1~2	0~2	2~0	~	1~0	1~1	1~2	0~2	0~3	0~4	3~2	1~2	1~1	0~2	1~1
Huddersfield Town	1~2	0~2	1~2	1~2	0~0	0~3	0~1	0~1	1~0	~	1~4	0~1	0~3	1~1	0~1	1~3	0~2	1~2	1~1	1~0
Leicester City	3~0	2~0	2~1	0~0	0~1	0~0	1~4	1~2	3~1	3~1	~	1~2	2~1	0~1	0~1	1~2	0~2	2~0	1~1	2~0
Liverpool	5~1	3~0	1~0	4~2	4~1	2~0	4~3	1~0	2~0	5~0	1~1	~	0~0	3~1	4~0	3~0	2~1	5~0	4~0	2~0
Manchester City	3~1	3~1	2~0	5~0	2~0	6~0	2~3	3~1	3~0	6~1	1~0	2~1	~	3~1	2~1	6~1	1~0	3~1	1~0	3~0
Manchester United	2~2	4~1	2~1	2~2	0~2	1~1	0~0	2~1	4~1	3~1	2~1	0~0	0~2	~	3~2	3~2	0~3	2~1	2~1	1~1
Newcastle United	1~2	2~1	0~1	2~0	3~0	1~2	0~1	3~2	0~0	2~0	0~2	2~3	2~1	0~2	~	3~1	1~2	1~0	0~3	1~2
Southampton	3~2	3~3	2~2	0~0	1~2	0~3	1~1	2~1	2~0	1~1	1~2	1~3	1~3	2~2	0~0	~	2~1	1~1	1~2	3~1
Tottenham Hotspur	1~1	5~0	1~0	1~0	1~0	3~1	2~0	2~2	3~1	4~0	3~1	1~2	0~1	0~1	1~0	3~1	~	2~1	0~1	1~3
Watford	0~1	0~4	2~0	0~0	3~2	1~2	2~1	1~0	4~1	3~0	2~1	0~3	1~2	1~2	1~1	1~1	2~1	~	1~4	1~2
West Ham United	1~0	1~2	2~2	4~2	3~1	0~0	3~2	0~2	3~1	4~3	2~2	1~1	0~4	3~1	2~0	3~0	0~1	0~2	~	0~1
Wolverhampton Wanderers	3~1	2~0	0~0	1~0	2~0	2~1	0~2	2~2	1~0	0~2	4~3	0~2	1~1	2~1	1~1	2~0	2~3	0~2	3~0	~

(a)

(b)

Figure 2: The matrix of results of the 2018/19 Premier League season. (a) Cell entries correspond to the result when a home team (row) plays an away team (column). For example, when Burnley played at home against Fulham, they won $2-1$, while in the return fixture, when Fulham were at home against Burnley, Fulham won $4-2$. (b) The results in (a) summarised in a results matrix by categorising each result as a win, draw or loss corresponding to the colour green, yellow or red, respectively.

We remark that the results matrix $\boldsymbol{y}$ can be considered as an adjacency matrix of a directed network where each node represents a team and an edge from node i to node j represents the result of the match when team i plays at home against team j, where the edge takes a value in the set $\{1, 2, 3\}$. In many networks, we observe sparsity in the relational variable. For example, in a social network where the dyadic relational variable is binary, for instance a friendship relationship recorded as 1 if the two nodes are friends, and 0 otherwise, typically most dyads will take the value 0. However this is not the case here as we always record a categorical outcome between each pair of nodes in the network. Hence, the network under study is dense and complete but does not allow self-loops.

4 The stochastic block model

The stochastic block model (SBM) of (Nowicki and Snijders 2001) has proven to be incredibly useful in the analysis of relational network data. The core idea of an SBM, which we make precise in the subsequent section, in the context of a binary network is to partition the nodes of the network into blocks such that the probability of an edge between any two nodes depends on an unobserved (or latent) variable assigning each node to one of K blocks or clusters in the network. Typically, if two nodes are assigned to the same block, then the probability that they are connected by an edge is greater than if they belong to different blocks. One of the contributions of this paper is to explain how this framework can be extended to the context of football results data by developing a novel Bayesian stochastic block model which can be applied to the results matrix illustrated in the previous section.

4.1 Specification of the block model

As outlined before, the aim of our stochastic block model is to partition the N teams in a league, into K blocks in such a way that the probability of a win, draw or loss for the home team is, broadly speaking, similar when any two teams in the same block play against one and other. But equally, that the probability outcome between blocks, that is, where any team from one block plays at home against another team from a different block also has a similar probability of a win, draw or a loss. Essentially we wish to model the following scenarios. Matches involving teams within a block tend to have similar outcomes, while matches between teams from different blocks tend also to have similar outcomes. Crucially, the probability of a given outcome depends on the blocks to which each team is assigned. One of the key objective is to infer the most likely value of K. In particular, if we deem that $K = 1$ has most support, then we have some evidence that the league is balanced.

Here we develop the model and introduce some notation. We interchangeably use the terms nodes and teams, as from before there is an analogy between $\boldsymbol{y}$, the results matrix (1) and an adjacency matrix of a network. We make two main assumptions:

- For a K block (or cluster) model, each node (or team), i for $i = 1, \ldots, N$, belongs to one of the blocks with membership or allocation label, $z_i \in \{1, \ldots, K\}$.
- The distribution of the relational structure $\boldsymbol{y} = (\boldsymbol{y_{ij}})_{1 \leq i \neq j \leq N}$ is assumed to be conditionally independent given the latent variable of cluster memberships, $\boldsymbol{z} := (z_1, \ldots, z_N)$.

We now describe the various elements of our model.

Distribution for the allocation vector $\boldsymbol{z}$: We assume that the entries of $\boldsymbol{z}$ are independent and identically distributed following a multinomial distribution:

$$z_i|\boldsymbol{\theta}, K \overset{iid}{\sim} Multi(1, \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)), \text{ for } i = 1, \ldots, N, \tag{2}$$

where $P(z_i = k|\boldsymbol{\theta}, K) = \theta_k$ is the probability that node i belongs to cluster k, $\theta_k > 0,\ k = 1, \ldots, K$ and $\sum_{k=1}^{K} \theta_k = 1$. We can write this probability mass function compactly as:

$$p(z_i|\boldsymbol{\theta}, K) = \prod_{k=1}^{K} \theta_k^{I(z_i=k)},$$

where $I(A)$ is the indicator function I defined as 1 if condition A is satisfied and 0 otherwise. Thus, the distribution of the partition of the N nodes into K clusters conditional on $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is:

$$\pi(\boldsymbol{z}|\boldsymbol{\theta}, K) = \prod_{i=1}^{N} Multi(z_i; 1, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{I(z_i=k)}. \tag{3}$$

Prior for $\boldsymbol{\theta}$: We assume a vague conjugate prior for the vector $\boldsymbol{\theta}$ following a Dirichlet distribution of dimension K with vector of concentration parameters $\boldsymbol{\gamma}$:

$$\boldsymbol{\theta}|K \sim Dir(\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k, \ldots, \gamma_K)).$$

We choose to set all concentration parameters to the same value, thus resulting in a symmetric prior. In particular, we set all γ_k's equal to $\gamma_0 = 1$ yielding a uniform prior. Another possibility would be to set $\gamma_0 = \frac{1}{2}$ corresponding to a non informative Jeffrey's prior.

Blocks interaction probabilities: Following the stochastic block model framework, the relational pattern of the nodes depends on the block probabilities of each game outcome. The $K \times K \times 3$ array, $\boldsymbol{p}$, encoding the block interaction probabilities is therefore of the form:

$$\boldsymbol{p} = \begin{pmatrix} \underline{p}^{11} & \underline{p}^{12} & \ldots & \underline{p}^{1l} & \ldots & \underline{p}^{1K} \\ \underline{p}^{21} & \underline{p}^{22} & \ldots & \underline{p}^{2l} & \ldots & \underline{p}^{2K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \underline{p}^{k1} & \underline{p}^{k2} & \ldots & \underline{p}^{kl} & \ldots & \underline{p}^{kK} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \underline{p}^{K1} & \underline{p}^{K2} & \ldots & \underline{p}^{Kl} & \ldots & \underline{p}^{KK} \end{pmatrix}$$

where,

$$\underline{p}^{kl} = (p_1^{kl}, p_2^{kl}, p_3^{kl}) \text{ and } \sum_{\omega=1}^{3} p_\omega^{kl} = 1,$$

for all $k = 1, \ldots, K$ and $l = 1, \ldots, K$. In other words, p_1^{kl}, p_2^{kl} or p_3^{kl} is the probability that a team allocated to block k playing at home against a team allocated to block l, wins, draws or loses, respectively. Therefore, the random variable denoting the result (win, draw or loss) for any given team in block k playing at home against any team in block l, playing away, has the same probability mass function $\underline{p}^{kl}$.

Distribution of the relational pattern of $\boldsymbol{y}$: As in (Nowicki and Snijders 2001), we model the distribution of edges between nodes conditionally on the block memberships. Additionally, we model that edges between nodes are identically distributed with parameters given by the interaction matrix $\boldsymbol{p}$. In contrast to (Côme and Latouche 2015),

where a Bernoulli distribution is chosen, the dyadic relation in our scenario takes categorical values. Here we model the observation y_{ij} conditional on the latent allocations z_i, z_j as a multinomial distribution,

$$\begin{aligned} f(y_{ij}|z_i, z_j, \boldsymbol{p}, K) &= \prod_{k=1}^{K}\prod_{l=1}^{K} Multi(y_{ij}; 1, \underline{p}^{kl})^{I(z_i=k)I(z_j=l)} \\ &= \prod_{k=1}^{K}\prod_{l=1}^{K}\left\{\prod_{\omega=1}^{3}\left(p_{\omega}^{kl}\right)^{I(y_{ij}=\omega)}\right\}^{I(z_i=k)I(z_j=l)} \end{aligned}$$

for $i, j = 1, \ldots, N, i \neq j$.

Thus, given the latent structure $\boldsymbol{z}$, the block interactions array $\boldsymbol{p}$ and K, we can write the likelihood of the relational pattern $\boldsymbol{y}$ as

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{p}, K) &= \prod_{i=1}^{N-1}\prod_{\substack{j=1 \\ j\neq i}}^{N} f(y_{ij}|z_i, z_j, \boldsymbol{p}, K) \\ &= \prod_{i=1}^{N-1}\prod_{\substack{j=1 \\ j\neq i}}^{N}\prod_{k=1}^{K}\prod_{l=1}^{K}\left\{\prod_{\omega=1}^{3}\left(p_{\omega}^{kl}\right)^{I(y_{ij}=\omega)}\right\}^{I(z_i=k)I(z_j=l)} . \end{aligned} \tag{4}$$

Prior for the block interaction probabilities: We assume that the entries of $\boldsymbol{p}$ are mutually independent and that each $\underline{p}^{kl}$ follows a conjugate prior from a 3-dimensional Dirichlet distribution:

$$\underline{p}^{kl} \sim Dir(\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)), \quad \text{for } k = 1, \ldots, K, \quad \text{and } l = 1, \ldots, K.$$

We set all the hyperparameters $\beta_1, \beta_2, \beta_3$ to 1 leading to a uniform distribution.

Prior for K: We treat the number of blocks or clusters as a random variable and choose a probability mass function for K which is distributed as a zero-truncated Poisson random variable with $\lambda = 1$ restricted to $1 \leq k \leq K_{max}$, where K_{max} is an user specified upper limit on the plausible number of blocks. We note that the justification for this prior was developed in (Nobile 2005) in the context of mixture models and used in several papers including (Nobile and Fearnside 2007) and (Wyse and Friel 2012). It consists of a Poisson distribution with rate parameter equal to 1 conditioned on $K > 0$:

$$K \sim Poi(1|K > 0),$$

that is,

$$\pi(K|K > 0) = \frac{Poi(1)}{1 - Poi(K = 0)} = \frac{1}{K!(e-1)}. \tag{5}$$

Therefore this prior probability mass function is proportional to $\frac{1}{K!}$. As such, this prior reflects the fact that an SBM with K block contains $K!$ permutations of the block labels, due to the label switching phenomena which is well understood for finite mixture models. This prior therefore assigns equal prior mass to each of these $K!$ possible relabellings. We return to the issue of label switching in Section 5.4.

4.2 Collapsing the stochastic block model

Following the specification of stochastic block model, we can write out the full joint posterior distribution of all model parameters as

$$\pi(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}, K|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{z}, K)\pi(\boldsymbol{p}|K)\pi(\boldsymbol{z}|\boldsymbol{\theta}, K)\pi(\boldsymbol{\theta}|K)\pi(K). \tag{6}$$

Our interest here is primarily the joint posterior distribution of the latent allocation vector and number of blocks, $\pi(\boldsymbol{z}, K|\boldsymbol{y})$. By virtue of the fact that we have chosen a conjugate prior for $\boldsymbol{\theta}$ and $\boldsymbol{p}$ we can integrate out both of these vectors from the posterior distribution yielding a collapsed posterior distribution,

$$\begin{aligned} \pi(\boldsymbol{z}, K|\boldsymbol{y}) &= \int_{\Theta}\int_{\boldsymbol{P}} f(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{z}, K)\pi(\boldsymbol{p}|K)\pi(\boldsymbol{z}|\boldsymbol{\theta}, K)\pi(\boldsymbol{\theta}|K)\pi(K)d\boldsymbol{\theta}d\boldsymbol{p} \\ &= \int_{\boldsymbol{P}} f(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{z}, K)\pi(\boldsymbol{p}|K)d\boldsymbol{p} \times \int_{\Theta}\pi(\boldsymbol{z}|\boldsymbol{\theta}, K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\theta} \times \pi(K) \\ &= f(\boldsymbol{y}|\boldsymbol{z}, K)\pi(\boldsymbol{z}|K)\pi(K). \end{aligned} \tag{7}$$

Details of the integration carried out in (7) are explained in Appendix A. Following (21) and (24) the collapsed posterior of the model has the form:

$$\pi(\boldsymbol{z}, K|\boldsymbol{y}) \propto \prod_{k=1}^{K}\prod_{l=1}^{K}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}\cdot\prod_{k=1}^{K}\Gamma(n_k+1)\frac{\Gamma(K)}{\Gamma(N+K)}\times\frac{1}{K!}, \tag{8}$$

where we define

$$N_{kl}^{\omega}=\sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} I(y_{ij}=\omega)I\left(z_i=k\right)I\left(z_j=l\right),$$

for $\omega=1,2,3$ and for $k,=1,\ldots,K$ and $l,=1,\ldots,K$, therefore allowing for the possibility that $k=l$. Also,

$$n_k=\sum_{i=1}^{N} I\left(z_i=k\right),\ k=1,\ldots,K.$$

Therefore N_{kl}^{ω} counts the number of times that the outcome ω was observed for all games involving a team allocated to block k playing at home against a team allocated to block l. While n_k accounts for the number of nodes/teams allocated to block k.

5 Bayesian estimation of the stochastic block model

Over time, many inferential strategies have been developed for the stochastic block model and its variants. For example, variational Bayes EM algorithm or integrated likelihood variational Bayes (Latouche, Birmele, and Ambroise 2012). Moreover, model selection criteria based on the integrated complete likelihood (Biernacki, Celeux, and Govaert 2000) have been developed for bipartite and binary networks (Rastelli, Latouche, and Friel 2018). Additionally, (McDaid et al. 2013) developed a novel MCMC algorithm for an SBM for a network with binary edges and this is the framework which we develop here for our model.

The overall objective is to develop an algorithm to sample from the posterior distribution $\pi(\boldsymbol{z}, K|\boldsymbol{y})$ (8). In so doing, this will allow us to estimate the posterior distribution of the number of blocks, $\pi(K, \boldsymbol{y})$, but also to estimate the posterior distribution $\pi(\boldsymbol{z}|K, \boldsymbol{y})$, so that we can assign probabilities to the allocation of teams to each of K blocks. We also highlight, similar to (Nobile and Fearnside 2007) that although our objective is to estimate the posterior distribution of K, we do not require a dimension changing MCMC algorithm such as reversible jump MCMC. This is because the dimension of the model is encoded in the vector $\boldsymbol{z}$ which is of fixed dimension N. We note that this strategy has also been employed in (Wyse and Friel 2012) in the context of the latent block model.

We now describe the MCMC algorithm used to sample from (8), the posterior distribution $\pi(\boldsymbol{z}, K|\boldsymbol{y})$. The algorithm is based on three move types:

Move type MK: Metropolis move to insert or remove an empty cluster. This move changes the current state of K but not the allocation vector $\boldsymbol{z}$.

Move type M-GS: Metropolis-within-Gibbs move that updates all components of the allocation vector $\boldsymbol{z}$ but does not change the number of clusters.

Move type AE: Metropolis-Hastings move to absorb or eject a cluster. This move affects both $\boldsymbol{z}$ and K.

We now present an overview of the pseudocode in Algorithm 1 below before providing more details on each move type in turn.

5.1 Move-type MK

When this move is chosen, the algorithm selects with probability 0.5 to increase or decrease the number of clusters by inserting or deleting an empty cluster, respectively. In the case of an *insert* proposal, an empty cluster is added and the label with smallest value available is set to this cluster. This move is accepted with probability (9). When the current number of clusters K is equal to the maximum allowed, K_{max}, an *insert* proposal is rejected with probability 1 and the number of clusters remains at $K=K_{max}$. When proposing to *delete* a cluster, the algorithm first checks if some clusters are empty and the cluster with highest label value is selected. This move is accepted with probability (10). Otherwise, if $K=1$, the proposal to *delete* a cluster is rejected and the number of clusters remains at $K=1$. Notice that the move-type MK affects K but leaves the allocation vector $\boldsymbol{z}$ unchanged. If the proposal to *insert* an empty

Algorithm 1 MCMC algorithm to sample from $\pi(\boldsymbol{z}, K|\boldsymbol{y})$.

We begin with an initial state $(\boldsymbol{z}^1, K^1)$.
while iteration $s < S$ **do**
  With equal probability select a move type MK, M-GS or AE.
  if *MK is selected* **then**:
    if an ***insert*** attempt is selected and accepted **then**
      K is updated and increased by 1 and an empty cluster is added
      $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$ and $K^{(s+1)} = K^{(s)} + 1$
    if a ***remove*** attempt is selected and accepted **then**
      K is updated and decreased by 1 and an empty cluster (if there is one) is removed.
      $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$ and $K^{(s+1)} = K^{(s)} - 1$
    otherwise
      The new state is set equal to the current state: $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$ and $K^{(s+1)} = K^{(s)}$
  if *M-GS is selected* **then**:
    The value of K is unchanged: $K^{(s+1)} = K^{(s)}$
    The allocation vector is updated to $\boldsymbol{z}^{(s+1)}$
  if *AE is selected* **then**:
    With probability p_K^e or $1 - p_K^e$ select an *absorption* attempt or *ejection* attempt, respectively.
    if an ***absorption*** attempt is selected and accepted **then**
      $K^{(s+1)} = K^{(s)} - 1$ and $\boldsymbol{z}$ is updated to $\boldsymbol{z}^{(s+1)}$
    if an ***ejection*** attempt is selected and accepted **then**
      $K^{(s+1)} = K^{(s)} + 1$ and $\boldsymbol{z}$ is updated to $\boldsymbol{z}^{(s+1)}$.
    otherwise
      $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$ and $K^{(s+1)} = K^{(s)}$

cluster is accepted, there will be an additional label but no nodes will be assigned to this new cluster. In case of a *remove* proposal, the number of clusters decreases, but again the allocation vector $\boldsymbol{z}$ will remain unchanged. Below we provide an algorithmic description of this move type. For more details of the derivation of the the acceptance probabilities below, see Appendix B.1.

At iteration s with current state $(\boldsymbol{z}^{(s)}, K^{(s)}) = (\boldsymbol{z}, K)$:

1. An *insertion* or a *removal* of a cluster is proposed with probability 0.5.
2. If the ***insertion attempt*** is selected:

   - If $K = K_{max}$, the new state is set equal to the current state: $K^{(s+1)} = K_{max}$.
   - If $K < K_{max}$, we accept $K^{(s+1)} = K + 1$ as the new state with acceptance probability: $min\,[1, \alpha]$, where:

$$\alpha = \frac{K}{(N + K)(K + 1)} \tag{9}$$

3. If the ***delete attempt*** is selected:

   - If $K = 1$, the new state is set equal to the current state: $K^{(s+1)} = 1$.
   - If $K > 1$, we accept $K^{(s+1)} = K - 1$ as the new state with acceptance probability: $min\,[1, \alpha]$, where:

$$\alpha = \frac{K(N + K - 1)}{K - 1} \tag{10}$$

     Note that the above quantity is always greater than 1 as we always accept the proposal of deleting an empty cluster if there is one.

Notice that in all cases the new state for $\boldsymbol{z}^{(s+1)}$ will be equal to the current state $\boldsymbol{z}^{(s)}$.

5.2 Move-type M-GS

This move-type involves a standard Metropolis-within-Gibbs update of each element of the allocation vector $\boldsymbol{z}$. This sweep consists of updating the allocation of each node sequentially from z_1 through to z_N. Recall that this move-type does not change the current value of K. We describe this move type below and provide details of the calculation involved in (11) in Appendix B.2 .

At iteration s, we carry out a sweep of $\boldsymbol{z}^{(s)}$ yielding an updated allocation vector $\boldsymbol{z}^{(s+1)} = (z_1^{(s+1)}, z_2^{(s+1)}, \ldots, z_N^{(s+1)})$. To do this, we update each element z_i of $\boldsymbol{z}^{(s)}$ using a Metropolis kernel, for $i = 1, \ldots, n$, as follows:

1. Set $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$.
2. For $\{i = 1, \ldots, n\}$
   (a) Suppose the current state of $z_i^{(s)} = k_0$. Propose a new state z_i' for $z_i^{(s)}$ by sampling a new cluster label uniformly from the set $\{1, \ldots, K\} \setminus k_0$.
   (b) Denote the proposed new allocation vector as $\mathbf{z}'$, which is identical to the current state of the allocation vector, $\boldsymbol{z}^{(s+1)}$, except for its ith element, which is z_i'.
   (c) The proposed new allocation vector $\mathbf{z}'$ is accepted with probability

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{z}', K|\boldsymbol{y})}{\pi(\boldsymbol{z}^{(s+1)}, K|\boldsymbol{y})}\right) \tag{11}$$

   in which case, $z_i^{(s+1)} = z_i'$. Otherwise, $z_i^{(s+1)} = z_i^{(s)}$.

5.3 Move-type AE

This move was originally proposed in (Nobile and Fearnside 2007) and is designed to allow a change to the number of blocks, K as well as to the allocation vector $\boldsymbol{z}$. It consists of a Metropolis-Hastings pair of absorption/ejection moves. Here we choose to attempt an *ejection* with probability p_K^e, depending on the current number of blocks. In particular we set,

$$p_K^e = \begin{cases} 1, \text{ if } K = 1, \\ 1/2, \text{ if } 1 < K < K_{max}, \\ 0 \text{ , if} K < K_{max}. \end{cases}$$

Otherwise, an *absorption* move is proposed.

Ejection attempt A block j_1 is randomly selected as the ejecting block among the K available blocks. The ejected block is assigned the label $j_2 = K + 1$. Each of the nodes in the block j_1 are then randomly assigned to a new block j_2 or remain in the original block j_1. The probability for each node to be allocated to the new block j_2, that is the probability of a node to be ejected, is constant and indicated with p_E, where $p_E \sim U(0, 1)$. Note, that instead of specifying an ejection probability p_E, we integrate over the p_E in much the same manner as collapsing. The details of this are outlined in Appendix B.3.

Absorption attempt Here both the absorbing and absorbed block, j_1 and j_2, respectively, are randomly chosen from the K available labels. All nodes allocated in block j_2 are reallocated to block j_1.

This move type, as pointed out in Nobile and Fearnside (2007), leads to improved mixing. Reversibility requires that the ejected block is a random draw from the resulting $K + 1$ clusters. To achieve this, at the end of the *ejection* move we perform a *swap* between the label $j_2 = K + 1$ and a randomly selected label from the set of all available ones, $\mathcal{K}' = \{1, 2, \ldots, K + 1\}$. The candidate state $(\boldsymbol{z}', K')$ is accepted with probability $\min(1, \alpha)$, where α is the product of the joint probability ratio and the proposal ratio as detailed below.

We detail the proposal ratios and probabilities for the *ejection* and *absorption* moves below. We refer the reader to Appendix B.3 for precise details of the derivation of these quantities.

The ratio of posterior densities for *ejection attempt* is:

$$\frac{P_{prop}((\boldsymbol{z}', K') \to (\boldsymbol{z}, K))}{P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}', K'))} = \frac{(1 - p_K^e)}{p_K^e} \cdot (n_{j_1} + n_{j_2} + 1), \tag{12}$$

where,

$$n_{j_1} = \sum_{i=1}^{N} I(z_i' = j_1) \text{ is the number of nodes in the ejecting cluster after reallocation,}$$

$$n_{j_2} = \sum_{i=1}^{N} I(z_i' = j_2) \text{ is the number of nodes in the ejected cluster.}$$

To obtain α, we multiply (12) by:

$$\frac{\pi(\boldsymbol{z}', K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} = \frac{\prod_{k=1}^{K+1}\prod_{l=1}^{K+1}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{'\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{'\omega}+1))} \times \prod_{k=1}^{K+1}\Gamma(n_k'+1)}{\prod_{k=1}^{K}\prod_{l=1}^{K}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))} \times \prod_{k=1}^{K}\Gamma(n_k+1)} \cdot \frac{K}{(N+K)(K+1)} \tag{13}$$

and accept this move with probability $\min(1, \alpha)$.

For the *absorption attempt* the proposal ratio takes the form:

$$\frac{P_{prop}((\boldsymbol{z}', K') \to (\boldsymbol{z}, K))}{P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}', K'))} = \frac{p_K^e}{(1-p_K^e)} \cdot \frac{1}{(n_{j_1}+n_{j_2}+1)}. \tag{14}$$

The corresponding posterior density ratio appears as:

$$\frac{\pi(\boldsymbol{z}', K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} = \frac{\prod_{k=1}^{K-1}\prod_{l=1}^{K-1}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{'\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{'\omega}+1))} \times \prod_{k=1}^{K-1}\Gamma(n_k'+1)}{\prod_{k=1}^{K}\prod_{l=1}^{K}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))} \times \prod_{k=1}^{K}\Gamma(n_k+1)} \cdot \frac{K(N+K-1)}{K-1}. \tag{15}$$

Once again, α is then the product of (14) and (15) and the move is accepted with probability $\min(1, \alpha)$.

5.4 Label switching phenomenon and correction

In the framework outlined above, the discrete labels in the allocation vector $\boldsymbol{z}$ are not identifiable by the model. This is because the likelihood is invariant to permutations of the labels of $\boldsymbol{z}$, leading to the well-known label switching phenomenon. This scenario arises naturally when using MCMC methods and, as a result, it is necessary to employ an algorithm to correct for this. In particular we use the online relabelling algorithm proposed in (Nobile and Fearnside 2007) which relies on the algorithm of (Carpaneto and Toth 1980) to correct the output of a sample from the posterior distribution of $\boldsymbol{z}$ from the MCMC algorithm described above. The relabeling algorithm which we use first orders this sample by the number of non-empty blocks in increasing order to yield an ordered sample $\boldsymbol{Z} := \{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(S)}\}$. The algorithm then iterates over $\boldsymbol{Z}$ comparing the current allocation vector to all previously relabeled states. The current state is then relabeled by permuting its labels so that a distance to all the preceding relabeled states is minimised. Here we define a distance between any two allocation vectors, $\boldsymbol{z}$ and $\boldsymbol{z}'$ as

$$D(\boldsymbol{z}, \boldsymbol{z}') = \sum_{i=1}^{N} I(z_i \neq z_i'), \tag{16}$$

where I is the usual indicator function used previously, so that this distance records the number of locations where the entries of $\boldsymbol{z}$ and $\boldsymbol{z}'$ differ. Further, let σ denote a permutation of the integers $1, \ldots, K$, that is, a bijection from the set $\{1, \ldots, K\}$ onto itself, and let $\sigma\boldsymbol{z}$ denote the relabeled allocation vector by applying the permutation to the labels of the vector $\boldsymbol{z}$. This algorithm was implemented using the `R` package `collpcm` and can be briefly described as follows:

1. The sample of S allocation vectors from the MCMC algorithm are ordered by the number of non-empty blocks and in decreasing order with respect to the number of labels used yielding an sequence of allocation vectors $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(S)}\}$.
2. Set $\sigma^{(1)}$ to be the identity permutation.

3. For $\{s = 2, \ldots, S\}$
   Find the permutation $\sigma^{(s)}$ which minimises the sum of distances from the permuted allocation vector $\boldsymbol{z}^{(s)}$ to all preceding permuted allocation vectors,

$$\sigma^{(s)} = \arg\min_{\sigma} \sum_{t=1}^{s-1} D(\sigma \boldsymbol{z}^{(s)}, \sigma^{(t)} \boldsymbol{z}^{(t)}),$$

   where $D(\sigma \boldsymbol{z}^{(s)}, \sigma^{(t)} \boldsymbol{z}^{(t)})$ is the distance function (16).
4. The relabeled sequence of allocation vectors

$$\{\sigma^{(1)} \boldsymbol{z}^{(1)}, \ldots, \sigma^{(S)} \boldsymbol{z}^{(S)}\}$$

   is returned.

6 Analysis of over 40 seasons of the English first division/ Premier League

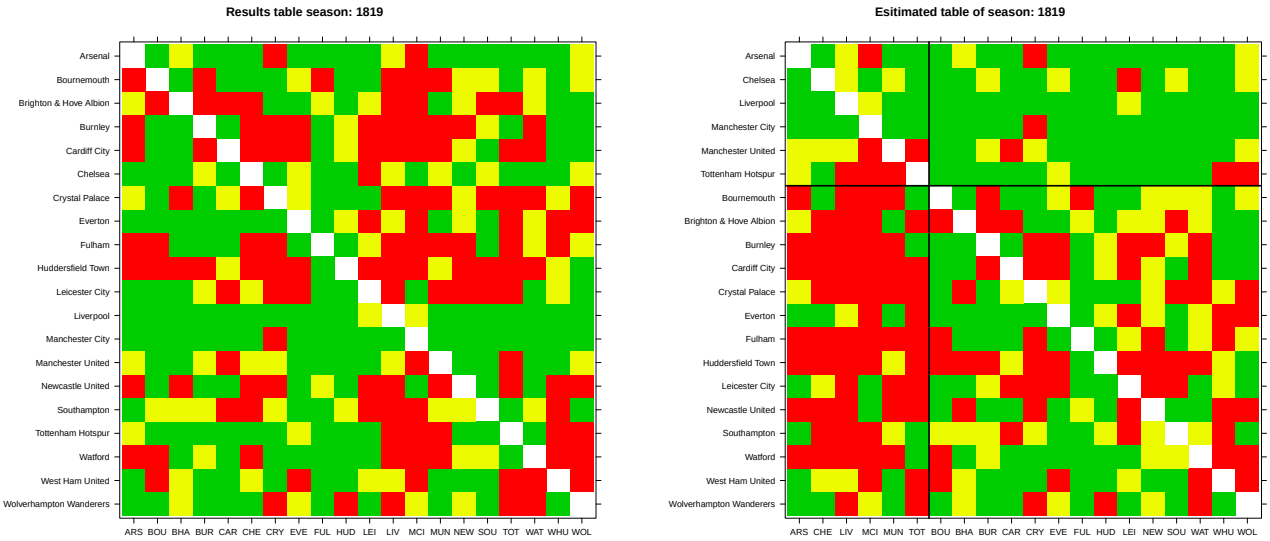
In this section, we present a detailed analysis for the 2018/19 Premier League season, highlighting some of the important aspects of the output from the model. We then extend our analysis to each season over the past 42 years from $1978/79$ to $2018/19$ encompassing the end of the old English first division through to the inception of the Premier League. Analysing each season over this length of time will allow us to provide insights into changes in competitive balance. For each season, the MCMC algorithm described in Section 5 was run for $200,000$ iterations, discarding the first $50,000$ as burn-in iterations and took approximately 3 minutes to complete. This was followed by the label switching correction algorithm which took only a few seconds to run. The MCMC algorithm was coded in `R` and a link to this code and the data used in this paper can be found here `https://github.com/basins95/Football_SBM`.

6.1 Background to the English football league and premier league

The English football league is one of the oldest football leagues in the world dating back to 1888 when the league consisted of only 12 clubs. It grew rapidly with the introduction of a second division in 1892 and today consists of 4 divisions. Interestingly, no club has been ever present in top division. Everton holds the record of most seasons in the top division, missing only 4 seasons in total, while Arsenal and Aston Villa have both amassed over 100 seasons. A persistent feature of the English football leagues is that of relegation from the first division and promotion from the second division. Similarly, for the other lower divisions. In the more recent past, encompassing the study period of this paper, the English first division has undergone several changes in its structure and format. We outline these briefly here. In $1978/79$ the first division consisted of 22 teams. This was reduced to 21 teams for a single season in $1987/88$ before reducing further to 20 teams until $1990/91$. It then reverted to 22 teams from $1991/92$ to $1994/95$, before changing once again in $1995/96$ to 20 where it has remained to date. Another important change which occurred over the period of study in this paper was the adoption of *3-points-for-a-win* in $1981/82$. The motivation for this change was to encourage more attacking playing tactics than the previous *2-points-for-a-win* as it was widely accepted that teams would often settle for a draw since the difference was only one point compared to a win. We note that this change occurred towards the beginning of our study, so the effect of this might have minimal effect on the conclusions which we make. Although on the other hand, our model does not account for the number of points won, so may be agnostic to this change in any case. However, perhaps the most significant change has been the introduction of the English Premier League in $1992/93$. This heralded massive increases in revenue primarily through satellite television payments to clubs. The first TV deal between the Premier League and the television companies generated revenue of around 40 million pounds. This has increased dramatically over time to 5.14 billion between 2016 and 2019 (Ernst & Young 2017). As such it is not an understatement to say that the introduction of the Premier League has had a transformative effect on football.

6.2 Analysis of the 2018/19 premier league season

Here we present a detailed analysis of the $2018/19$ season as this provides us with the opportunity to discuss some of the salient features of our model. To begin we present in Figure 3 a summary of the outcome for each team in terms of the percentage of wins, draws and losses over the course of the season. As expected, this illustrates that teams that are placed higher tend to have more wins. It also illustrates the disparity in terms of percentage of wins between teams with a high final position in the league compared to those whose final position was towards the bottom of the league.



(a) (b)

Figure 4: Match grid for the 2018/19 season. Each entry of this grid corresponds to a match where a home team (row entry) plays against an away team (column entry). The outcome of each match is represented using green, yellow and red colours corresponding to a win, draw and loss, respectively, for the home team. (a) Teams listed in alphabetical order. (b) Teams listed by most likely block membership, a posteriori. The solid horizontal and vertical lines (which also coincides with final league position) separates each team in their most likely block.

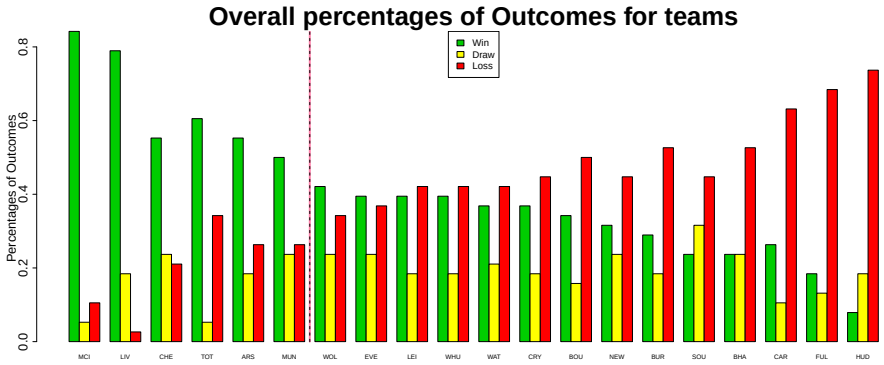


Figure 3: Barplot of the percentage of overall outcomes (win, draw, loss) for each team for Season 2018/19. Teams are listed in decreasing order from left to right according to their position in the final league table.

As before, the data for this season is presented in Figure 4 (a) in the form of a results grid, where we again use the convention that cell colours represent the outcome for a home team (corresponding to the rows of the grid) playing against an away team (the columns of the grid) where the colours green, yellow and red correspond to a win, draw and loss, respectively for the home team.

The MCMC algorithm 1 was implemented using each of the three moves types, MK, GS and AE following the description in Section 5. The posterior distribution of K is presented in Table 6 and indicates that there is overwhelmingly most support, a posteriori, for a model with two blocks. While the posterior distribution of the allocation variable $\boldsymbol{z}$ indicates that the strongest block is comprised of Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham. In fact this is also illustrated in Table 2 which presents the estimated posterior probability that each team belongs to the strongest block, conditional on $K = 2$. Here we have omitted $\pi(K = 4|\boldsymbol{y})$

Table 2: Final league table for the Premiere League Season 2018/19 showing the number of points for each team. For comparison, the estimated posterior probability of allocation of each team to the top block is also presented. The solid horizontal line separates the 6 teams most likely to be in the top block, a posteriori.

	Points	$P(\text{top block}\|\boldsymbol{y}, K=2)$
Manchester City	98	1.00
Liverpool	97	1.00
Chelsea	72	0.90
Tottenham Hotspur	71	0.80
Arsenal	70	0.89
Manchester United	66	0.84
Wolverhampton Wanderers	57	0.22
Everton	54	0.06
Leicester City	52	0.01
West Ham United	52	0.01
Watford	50	0.00
Crystal Palace	49	0.00
Bournemouth	45	0.00
Newcastle United	45	0.00
Burnley	40	0.00
Southampton	39	0.00
Brighton & Hove Albion	36	0.00
Cardiff City	34	0.00
Fulham	26	0.00
Huddersfield Town	16	0.00

as this had an estimated probability of 2×10^{-4}. Note that further output of the model for this season is presented in Appendix C. We also note that permuting the rows and columns of the match grid matrix in Figure 4 (b) by final league position reveals the block structure of the results. Here the solid vertical and horizontal lines separate the two blocks of teams. We make the following remarks to illustrate some important features of the estimated block structure in Figure 4 (b).

1. The top right quadrant corresponds to when a team in the strongest block plays at home against a team in the weakest block. Here the colours are predominately green, indicating that the home team generally won.
2. While the bottom left quadrant corresponds to the opposite situation, when a team in the weakest block plays at home against a team in the strongest block. Here the colours are predominately red, indicating the away team (in the strongest block) generally won.
3. Finally, note that the top right and bottom left quadrants, corresponds to the within-block matches for the strong block and weak blocks, respectively. Each blocks contain a mix of all three colours, indicating that these teams were generally quite balanced.

Table 1: Season 2018/19: Posterior probabilities of K.

K	1	2	3
$\pi(K\|\boldsymbol{y})$	0.0	0.98	0.02

It is instructive to further explore the six teams in the strongest block. Interestingly, the final league positions of these top 6 teams is not quite in agreement with the posterior probability of membership to this block, as presented in Table 2. In particular, our results suggest that Tottenham Hotspur may have slightly overachieved in the sense that there is less posterior support that they belong to the top block (0.80) than the two team immediately below them in the final league table, Arsenal and Manchester United, whose posterior probability of membership to the top block of team is 0.89 and 0.84, respectively. It is interesting therefore to interrogate the results of these three teams to compare how each fared against each other. This subset of the data is presented in Table 3. Essentially, it reveals that Arsenal's head-to-head record against Tottenham and Manchester United of 2 wins and 2 draws is superior than the

Table 3: Season 2018/19: Head-to-head records of Tottenham, Arsenal and Manchester United. This indicates that Tottenham's record against both teams (1 Win, 1 Draw, 2 Losses) was identical to Manchester United's but much worse that Arsenal's record against both teams (2 Wins, 2 Draws). This is also consistent with Arsenal being estimated a higher posterior membership to the top block than Tottenham despite having a lower overall league position.

	Tottenham	Arsenal	Manchester United
Tottenham	–	$1-1$	$0-1$
Arsenal	$4-2$	–	$2-0$
Manchester United	$0-3$	$2-2$	–

Table 4: Posterior probability of K, expressed as a percentage, over the last 42 seasons. The block model with highest posterior probability is coloured accordingly.

	Number of clusters			
Season	1	2	3	4
78/79	1.17	96.74	2.08	0.01
79/80	97.57	2.39	0.04	0.00
80/81	30.55	69.10	0.35	0.00
81/82	97.25	2.72	0.02	0.00
82/83	99.80	0.20	0.00	0.00
83/84	99.15	0.85	0.00	0.00
84/85	42.34	57.22	0.44	0.00
85/86	0.00	99.81	0.19	0.00
86/87	99.49	0.51	0.00	0.00
87/88	12.41	87.26	0.33	0.00
88/89	99.20	0.79	0.01	0.00
89/90	98.66	1.32	0.02	0.00
90/91	49.31	50.07	0.61	0.00
91/92	94.10	5.89	0.01	0.00
92/93	98.85	1.15	0.00	0.00
93/94	27.98	71.08	0.94	0.00
94/95	22.19	74.59	3.21	0.01
95/96	48.05	51.68	0.27	0.00
96/97	99.63	0.37	0.00	0.00
97/98	98.14	1.85	0.01	0.00
98/99	0.07	99.78	0.16	0.00
99/00	64.34	35.27	0.39	0.00
00/01	88.26	11.06	0.68	0.00
01/02	0.13	99.37	0.49	0.01
02/03	59.29	40.19	0.52	0.01
03/04	6.45	88.39	5.12	0.04
04/05	0.00	99.79	0.21	0.00
05/06	0.15	97.08	2.73	0.04
06/07	5.45	92.37	2.17	0.01
07/08	0.00	93.81	5.92	0.27
08/09	0.00	99.24	0.76	0.00
09/10	0.00	95.30	4.67	0.03
10/11	79.87	20.09	0.05	0.00
11/12	1.68	96.91	1.40	0.02
12/13	0.00	99.64	0.36	0.00
13/14	0.00	98.96	1.04	0.00
14/15	7.30	88.70	3.98	0.03
15/16	75.37	24.38	0.25	0.00
16/17	0.00	98.99	1.01	0.00
17/18	0.00	97.95	2.03	0.02
18/19	0.00	97.94	2.04	0.02
19/20	2.10	96.47	1.41	0.01

head-to-head record of either of Tottenham (against Arsenal of Manchester United) or Manchester United (against Arsenal of Tottenham) both of which is 1 win, 1 draw and 2 losses. This therefore is consistent with the fact that the posterior probability of membership to the top block is higher for Arsenal than for Tottenham, despite the latter having a higher league position than the former.

6.3 Analysis of four decades of the English Premier League

Here we extend our analysis to each season over the past 42 years from 1978/79 to 2019/20. This period of time encompasses the end of the old English first division and the inception of the Premier League in 1992/93 and allows an examination of any changes in the competitive balance of the league. We begin by providing a summary of the output of our model for each season. This is then followed by a more detailed analysis of the block structure resulting from the most recent 10 seasons. In particular, we explore whether there is evidence to suggest that a *big-six* group of teams emerged over this time period. A summary of the output of our analysis on a season-by-season basis is presented in Table 4 where we displayed the estimated posterior probability for a one block model through to a four block model.

Overall, Table 4 indicates there was most support each season, a posteriori, for either a one block or a two block model, however the number of seasons where a two block model has most support, a posteriori, increases considerably over the past two decades. In particular, over the first half of this study period there is no strong support for either a one or a two block model. In fact for some seasons there is broadly equal posterior support for either model, for example, seasons 84/85, 90/91, 95/96. In fact, of note is season 84/85 where the second block consisted of a single team, Stoke

City, who were bottom of the league in that season recording only 3 wins from a possible 42 matches. This situation changes considerably when we analyse the final two decades of this study. The right hand side of Table 4 indicates that there is typically most support from a two block model, but further that the posterior probability for a two block model is over 0.85 for almost every season since $2003/04$, providing strong evidence that the league has become more competitively imbalanced since then. There are a few exceptions to this, most notably, season $15/16$ when Leicester City famously won their first league title. However, in general, this analysis suggests a structural change in the nature of the competitiveness of the league since the turn of the millennium.

Additionally, there was quite limited support for a three block model, for example, from the beginning of the study until $02/03$ almost all seasons presented very little, if any, posterior support for this model, with the exception of season $94/95$, for which a three block model was attributed 0.03 probability. Conversely, from $2003/04$ to the end of the study, there were some seasons which gave some modest support to a three block model. For example, in seasons $03/04$, $07/08$, $09/10$, there was approximately 0.05 posterior probability for this model. Again, this is consistent with the idea that the league was more imbalanced over the past two decades.

6.3.1 Marginal posterior allocation of a team to the strongest block

It is informative to estimate the marginal probability that a team belongs to the strongest block by integrating over the uncertainty in the number of blocks. This is particularly important for seasons where there is broadly equal support for a one or a two block model. For each value of k, post label switching we associate the block label 1 with the strongest block of teams. We estimate the marginal posterior probability that team i is in the strongest block averaging over all possible number of blocks as follows,

$$\begin{aligned} \pi(z_i = 1|\boldsymbol{y}) &= \sum_{k=1}^{K_{max}} \pi(z_i = 1, K = k|\boldsymbol{y}) \\ &= \sum_{k=1}^{K_{max}} \pi(z_i = 1|\boldsymbol{y}, K = k)\pi(K = k|\boldsymbol{y}). \end{aligned} \tag{17}$$

In Figure 5 we display the estimated marginal posterior allocation of each team to the strongest block following (17) for each season. Moreover, we again use the convention to colour each season according to the block model with most support (sand and blue corresponding to seasons in which a one block model or a two block model, respectively, has most support). Notice in Figure 5 that for seasons from $1978/79$ to around $2002/03$ the spread of posterior probability of allocation to the top block for each team is quite different to that of the seasons which follow. In particular, from $2003/04$ to $2019/20$ we see that there are typically a relatively large group of teams (in a two block season) which have very small posterior probability of being allocated to the strongest block, indicated in the bottom right hand corner of Figure 5. In other words, there is a strong separation of teams in terms of posterior allocation to the strongest block, compared to the first half of seasons in this study. In addition, we make the important remark that the number of teams allocated the strongest block for each of the seasons from $2003/04$ to $19/20$, that is to say, those with high posterior probability of allocation to the strongest block, typically consists of a small number of teams. This analysis leads us in Section 6.4 to explore the composition of teams in the strongest block over the past two decades.

Following from the analysis presented above and in Figure 5 we can explore the number of teams of teams estimated to belong to the strongest block, by again integrating over the uncertainty in k. To do this, we simply examine (17) for each team and investigate if this posterior probability is greater than 0.5. If so, we assign this team to the strongest block of teams. For each season, this allows one to estimate the size of the strongest block of teams for each season in turn. This number serves as a useful alternative index of competitive balance for each season. This information is presented in Figure 6. We highlight for certain seasons that although a two block model has most posterior support, the estimated number of teams in the strongest block can be close to or equal to the total number of teams in the league that season. This is the case for seasons $1990/91$ and $1995/96$ where there was broadly equal posterior support for a one and a two block model, integrating over the uncertainty in the number of blocks indicates that all 20 are estimated to be in the strongest block suggesting quite a balanced league for these seasons.

An overall analysis of Figure 6 indicates that the number of teams in the strongest block is quite large for the majority of seasons in the first half of the study period until around $2003/04$. From $1978/79$ to $2003/04$ the estimated size of the top block contained more than half the total number of teams in the league each season, ranging from 11 to 22 teams for 20 out of 25 season. This is in stark contrast with the seasons which followed, where the size of the top block ranged from 2 to 7 teams for 15 out of 17 season.

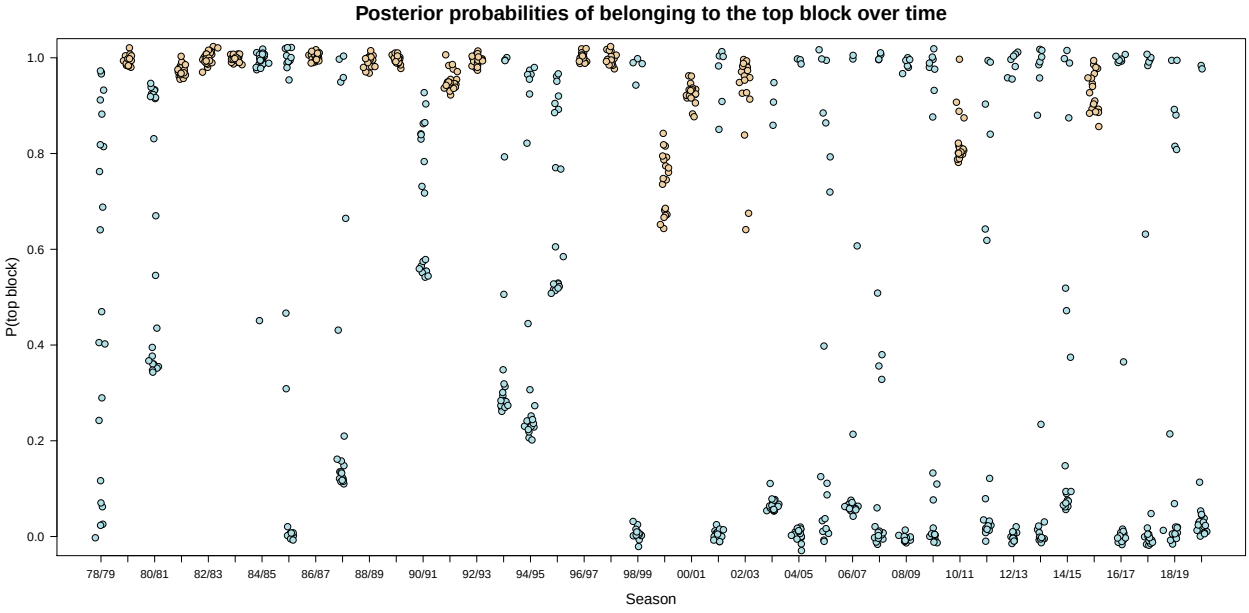


Figure 5: The posterior allocation probability of belonging to the strongest group of teams over the 42 seasons under study. For each season the colour indicates whether the league was partitioned into a single cluster (sand colour) or two (light blue). Each point represents the estimated posterior probability of allocation for a team. Note that within each season the points have been jittered.

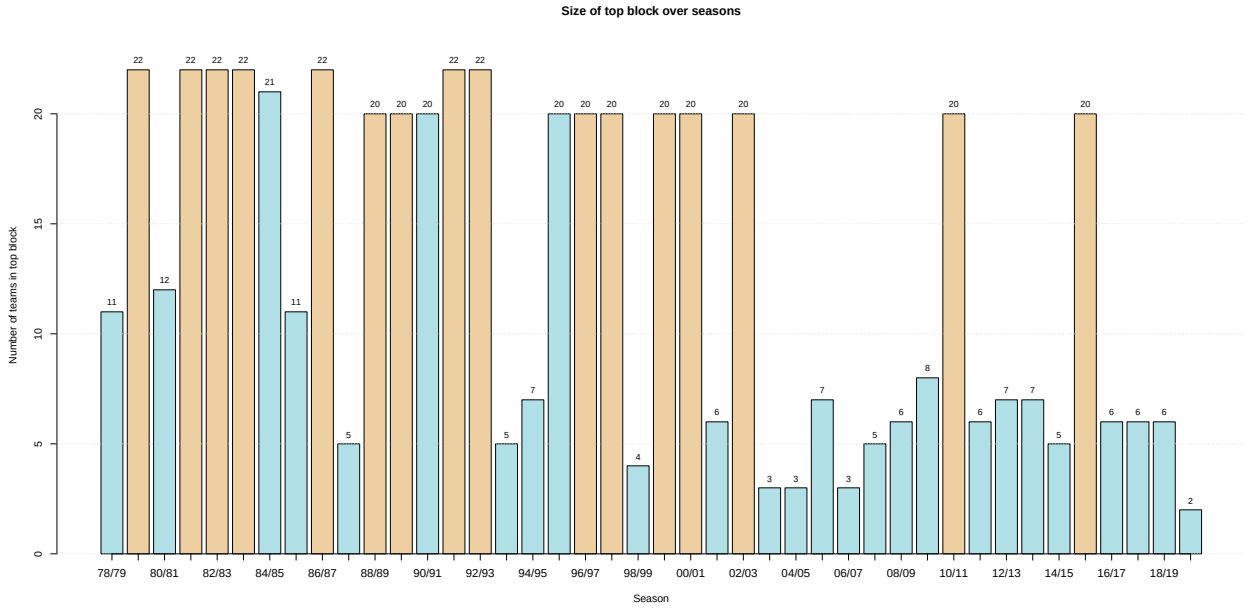
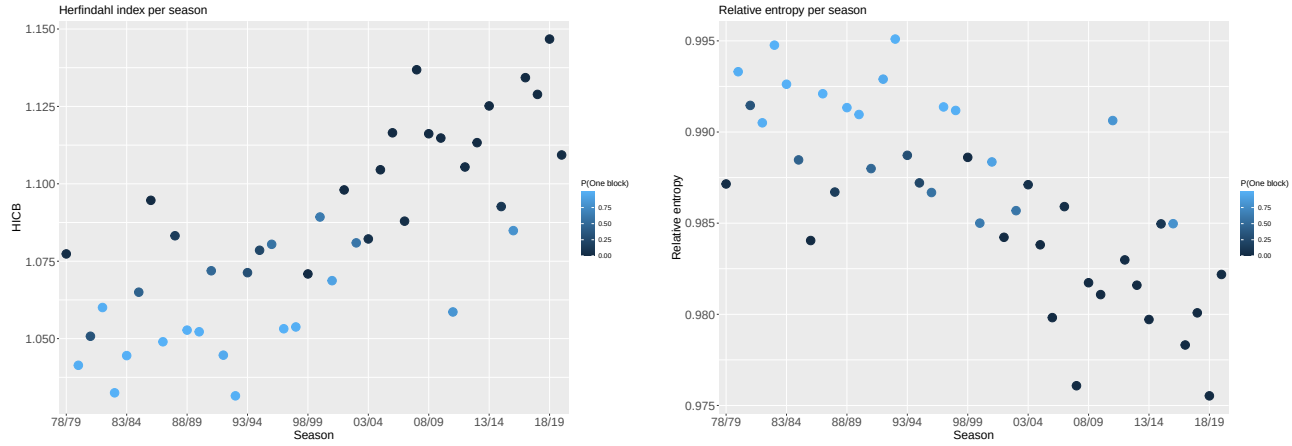


Figure 6: Barplot displaying the posterior estimate of the number of teams allocated to the strongest block each season. Single and two block seasons are coloured sand and blue, respectively. This illustrates that the size of the strongest block has generally decreased during the second half of the study period.



(a) (b)

Figure 7: (a) The Herfindahl–Hirschman index of competitive balance (HHICB) and (b) the relative entropy is plotted for each season where the posterior probability of a one block model, $\pi(K = 1|\boldsymbol{y})$, is overlaid for each season. Both plots illustrates that, in general, seasons with a lower HHICB score (higher relative entropy, respectively) tend to correspond to seasons where a single block had the most posterior support.

6.3.2 Validation of the SBM approach using two standard indices

Here we revisit the statistics outlined in Section 2. We once again display both HHICB and relative entropy for each season. But now, in Figure 7 we overlay the posterior probability of $K = 1$ for each season. Generally, we see both plots illustrates that, in general, seasons with a lower HHICB score (higher relative entropy, respectively) tend to correspond to seasons where a single block had the most posterior support. Effectively, the HHICB and relative entropy scores are broadly in agreement with the results of our stochastic block model and provide a validation of our SBM approach.

6.4 Evidence for the emergence of a *big-six* groups of teams

In this section we explore if there is some statistical evidence for two well known phenomena in popular discourse: the dominance of the *big-four* (a group of four teams, namely, Arsenal, Chelsea, Liverpool and Manchester United) during the mid-2000's and the emergence of a *big-six* from 2010 with the addition of Manchester City and Tottenham Hotspur. To investigate this, consider Table 5 where we present the posterior allocation of teams to the strongest block of teams from $2000/01$ to $2019/20$ following Section 6.3.1. This indicates that, apart from $2019/20$, Arsenal, Chelsea and Manchester United have always been allocated to the strongest block. Additional, with the exception of five seasons, Liverpool have been also allocated to the strongest block. This gives some credence to the notion of a *big-four* block of teams since the early 2000s. On the other hand, our analysis shows that Manchester City and Tottenham were not allocated to the top-ranked block of teams until season $2009/10$ which is when the former was purchased by the Abu Dhabi United Group. However, since then, both teams have consistently been allocated to the top block of teams. We also remark that $2019/20$ is quite different to all of the seasons which have preceded it, since in this season the strongest block consisted of only Liverpool and Manchester City. In fact, for this season Liverpool amassed a record tally of 99 points. Note that the third placed team, Manchester United achieved 66 points.

Overall our analysis strongly supports the idea that a *big-six* groups of teams is a persistent feature of the Premier League for the past decade. It is also interesting to note that the composition of the strongest block of teams has been remarkably stable over time. Notably, there were exceptions where some teams appear for consecutive seasons in the strongest block, including Everton from $2007/08$ to $2009/10$ and again from $2012/13$ to $2103/14$. While Aston Villa appeared in the strongest block for the consecutive seasons $2008/09$ and $2009/10$. Additionally, Newcastle United appear in seasons $2001/02$, $2005/06$ and $2011/12$.

We supplement more detail to the analysis above in Figure 8 by displaying the posterior probability of allocation to the strongest block for each season for Arsenal, Chelsea, Liverpool, Manchester city, Manchester United, Tottenham Hotspur over the entire course of the study. This provides additional information by illustrating that the posterior probability of allocation to the strongest team can vary considerably for each team and over each season. For example,

Table 5: Posterior allocation of teams to the strongest block for each season since $2001/02$. Here the symbols ✓and ✗ denote allocation or not, respectively, to the strongest block. Seasons with the highest probability of a single block model are highlighted in colour. Additional teams which were allocated to the strongest block are also listed per season. Since $2009/10$ Manchester City have each been ever present in this block. Similarly for Tottenham Hotspur apart from $2019/20$. Prior to $2009/10$, Manchester City were not allocated to the strong block. Also, note that each of the other four teams (with the exception of Liverpool for some seasons) have been in the strongest block over this time period.

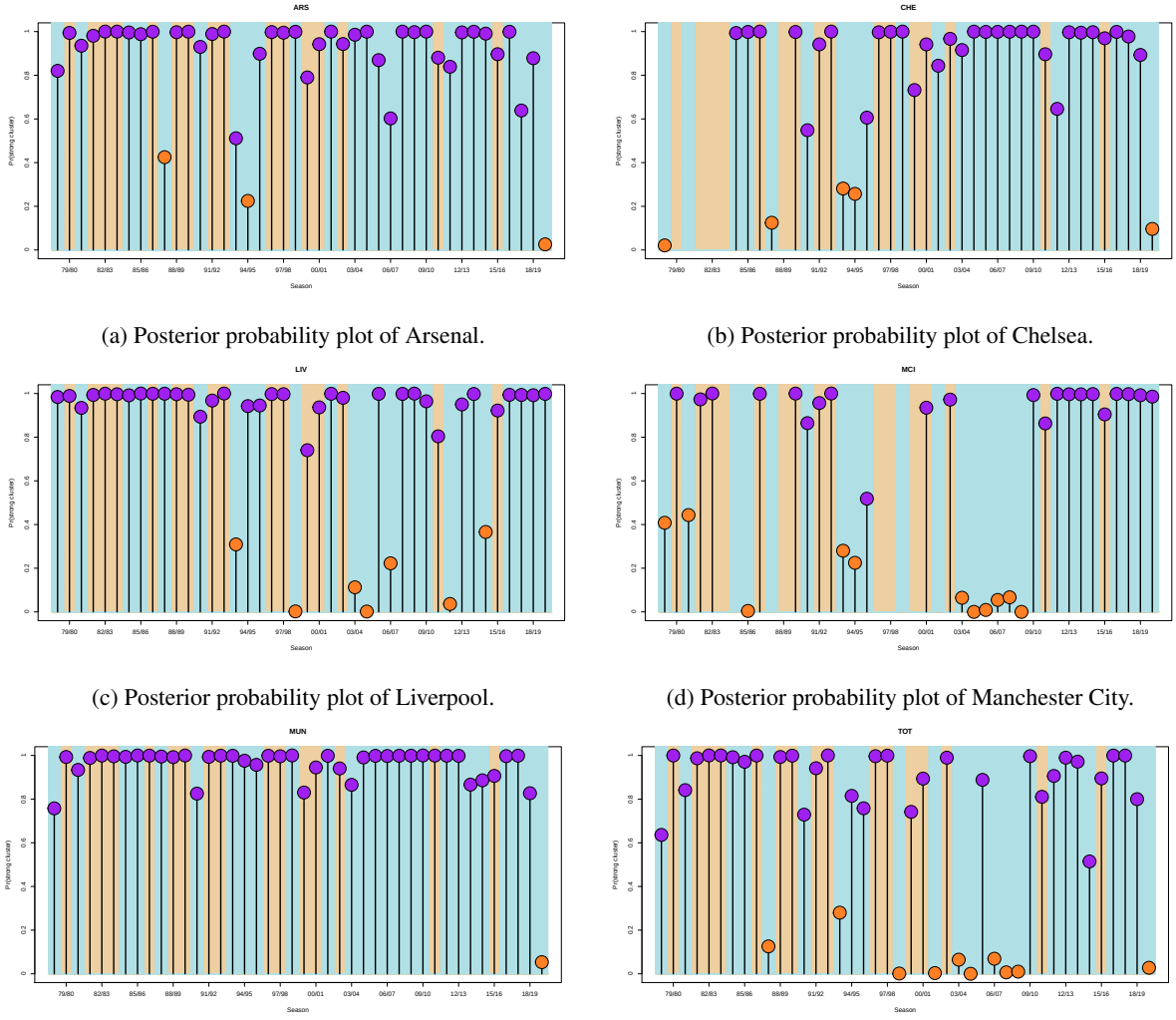
	Arsenal	Chelsea	Liverpool	Man City	Man Utd	Tottenham	Additional teams
19/20	✗	✗	✓	✓	✗	✗	
18/19	✓	✓	✓	✓	✓	✓	
17/18	✓	✓	✓	✓	✓	✓	
16/17	✓	✓	✓	✓	✓	✓	
15/16	✓	✓	✓	✓	✓	✓	
14/15	✓	✓	✗	✓	✓	✓	
13/14	✓	✓	✓	✓	✓	✓	Everton
12/13	✓	✓	✓	✓	✓	✓	Everton
11/12	✓	✓	✗	✓	✓	✓	Newcastle
10/11	✓	✓	✓	✓	✓	✓	
09/10	✓	✓	✓	✓	✓	✓	Everton, Aston Villa
08/09	✓	✓	✓	✗	✓	✗	Everton, Aston Villa
07/08	✓	✓	✓	✗	✓	✗	Everton
06/07	✓	✓	✗	✗	✓	✗	
05/06	✓	✓	✓	✗	✓	✓	Blackburn, Newcastle
04/05	✓	✓	✗	✗	✓	✗	
03/04	✓	✓	✗	✗	✓	✗	
02/03	✓	✓	✓		✓	✓	
01/02	✓	✓	✓		✓	✗	Leeds, Newcastle
00/01	✓	✓	✓		✓	✓	
Total no. of Seasons in the top block (out of 20)	19/20	19/20	15/20	11/20	19/20	13/20	

in $2017/18$ there is much less support for Arsenal's inclusion in the strongest block (the estimated probability in this case turns out to be 0.63).

7 Conclusions

In this paper we have developed a stochastic block model to allow a probabilistic assessment of competitive balance in the English football league. The important aspect of this model is that it allows one to assess the likely number of blocks of teams in a league, where a block consists of a number of teams such that the probability mass function of the outcome (a win, draw, or loss) is estimated to be the same for a match involving any two teams in that block. Similarly, a match involving a team from one block playing against any team from another block has its own between-block probability mass function for the outcome of that match. In contrast to previous appraches, our modelling approach yields a richer understanding of the nature of the competitiveness of a league through estimation of the posterior probability of the number of blocks, together with the most likely allocation of teams to each block. But also it allows one to estimate the number of teams allocated to the strongest block of teams, by integrating over the posterior uncertainty in the number of blocks.

In terms of our analysis of the English Premier League, our analysis has uncovered evidence that the league was quite balanced from around 1980 to 2003. However, subsequent to that, there is strong evidence that the league has become more imbalanced since from $2003/04$ we see an emergence of seasons where two blocks are most probable, a posteriori. This is further supported by our analysis which shows that the estimated number of teams in the strongest block of teams is typically quite large from 1980 to around 2003, in constrast to the second half of the study period where the number of teams estimated to be in strongest block is, in general, much smaller. In addition, our analysis suggests the emergence of a so-called *big-six* teams (Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hospur) around 2010 as during this time period all six teams, with only a few exceptions have always been present in the strongest block of teams.



(a) Posterior probability plot of Arsenal.

(b) Posterior probability plot of Chelsea.

(c) Posterior probability plot of Liverpool.

(d) Posterior probability plot of Manchester City.

(e) Posterior probability plot of Manchester United.

(f) Posterior probability plot of Tottenham Hotspur.

Figure 8: Plot of posterior probability of allocation to the strongest block for some selected teams over the entire study period. The x-axis indicates each season in chronological order, while the y-axis displays the posterior probability of inclusion in the strongest block where each team is indicated with a purple dot if its posterior allocation probability is greater than 0.5 or with a red one otherwise. If no dot is present for any time period it means that team was not playing in the First division/Premier League for that season. The background indicates how many blocks have been chosen with highest posterior probability. In particular, a sand coloured background is used for single-block model and a light blue colour is for a two block model.

This paper could be extended in several directions. For example, the stochastic block model does not directly model for the number of goals scored by either team. In fact, there is a steady literature on statistical models for football match data beginning with (Dixon and Coles 1997), where a Poisson GLM framework is used to model the number of goals scored by either team. This has been extended by several authors, including (Karlis and Ntzoufras 2003) to the bivariate Poisson setting. While (Rue and Salvesen 2000) extend the Dixon and Coles framework to a Bayesian dynamic GLM setting. There would therefore be interest in extending our stochastic block model to accommodate models for the number of goals scored by both teams using some of these frameworks.

Acknowledgements

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number 12/RC/2289_P2.

Datasets and code

All of the datasets and `R` code used to reproduce the results and figures in this paper can be found at https://github.com/basins95/Football_SBM

References

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*(7), 719–725.

Brandes, L. and E. Franck (2007). Who made who? an empirical analysis of competitive balance in european soccer leagues. *Eastern Economic Journal 33*(3), 379–403.

Buzzacchi, L., S. Szymanski, and T. Valletti (2003). Equality of opportunity and equality of outcome: Open leagues, closed leagues and competitive balance. *Journal of Industry, Competition and Trade 3*(3), 167–186.

Cairns, J. (1987). Evaluating changes in league structure: the reorganization of the Scottish football league. *Applied Economics 19*(2), 259–275.

Carpaneto, G. and P. Toth (1980). Algorithm 548: Solution of the assignment problem [h]. *ACM Transactions on Mathematical Software (TOMS) 6*(1), 104–111.

Côme, E. and P. Latouche (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling 15*(6), 564–589.

Dixon, Mark, J. and G. Coles, Stuart (1997). Modelling assoiation football scores and inefficiences in the football betting market. *Applied Statistics 46*(2), 265–280.

Eckard, E. W. (1998). The ncaa cartel and competitive balance in college football. *Review of Industrial Organization 13*(3), 347–369.

Ernst & Young (2017). Stoke city football club: Economic and social impact assessment.

Evans, R. (2014). A review of measures of competitive balancein the 'analysis of competitive balance' literature. *Birkbeck Sport Business Centre Research Paper Series 7*(2).

Fort, R. and J. Maxcy (2003). Competitive balance in sports leagues: An introduction. *Journal of Sports Economics 4*(2), 154–160.

Haan, M., R. H. Koning, and A. van Witteloostuijn (2008). Competitive balance in national european soccer competitions. In J. Albert and R. Koning (Eds.), *Statistical Thinking in Sports*, pp. 63–76. Taylor & Francis Group.

Herfindahl, O. C. (1950). Concentration in the u.s. steel industry.

Hirschman, A. O. (1945). *National Power and the Structure of Foreign Trade*. Berkeley.

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review 54*(5), 761–762.

Horowitz, I. (1997). The increasing competitive balance in major league baseball. *Review of Industrial Organization 12*, 373–387.

Humphreys, B. (2002). Alternative measures of competitive balance. *Journal of Sports Economics 3*(2), 133–148.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393.

Koning, R. H. (2000). Balance in competition in Dutch soccer. *Statistician 49*(3), 419–431.

Koop, G. (2004). Modelling the evolution of distributions: An application to major league baseball. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 167*(4), 639–655.

Latouche, P., E. Birmele, and C. Ambroise (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling 12*(1), 93–115.

Manasis, V., V. Avgerinou, I. Ntzoufras, and J. Reade (2013). Quantification of competitive balance in european football: development of specially designed indices. *IMA Journal of Management mathematics 24*(3), 363–375.

Manasis, V., I. Ntzoufras, and J. Reade (2021). Competitive balance measures and the uncertainty of outcome hypothesis in european football. *arXiv:1507.00634 [stat.AP]*.

McDaid, A. F., T. B. Murphy, N. Friel, and N. J. Hurley (2013). Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis 60*, 12–31.

Nobile, A. (2005). Bayesian finite mixtures: a note on prior specification and posterior computation. Technical report, University of Glasgow.

Nobile, A. and A. T. Fearnside (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing 17*(2), 147–162.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association 96*(455), 1077–1087.

Owen, P. D., M. Ryan, and C. R. Weatherston (2007). Measuring Competitive Balance in Professional Team Sports Using the Herfindahl-Hirschman Index. *Review of Industrial Organization 31*(4), 289–302.

Penn, R. and D. Berridge (2019). Competitive balance in the english premier league. *European Journal for Sport and Society 16*(1), 64–82.

Plumley, D., G. Ramchandani, and R. Wilson (2018). Mind the gap: an analysis of competitive balance in the english football league system. *International Journal of Sport Management and Marketing 18*(5), 357–375.

Rastelli, R., P. Latouche, and N. Friel (2018). Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Network Science 6*(4), 469–493.

Rue, H. and O. Salvesen (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(3), 399–418.

Vrooman, J. (1995). A general theory of professional sports leagues. *Southern Economic Journal 61*(4), 971–990.

Wyse, J. and N. Friel (2012). Block clustering with collapsed latent block models. *Statistics and Computing 22*(2), 415–428.

Appendix A Calculation of the collapsed SBM

A.1 Collapsing p

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{z},K) &= \int_{\boldsymbol{p}} f(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{p},K)\pi(\boldsymbol{p}|K)d\boldsymbol{p} \\
&= \int_{p_\omega^{kl}\in\underline{p}^{kl}\in\boldsymbol{p}} \prod_{i=1}^{N-1}\prod_{\substack{j=1\\ j\neq i}}^{N}\prod_{k=1}^{K}\prod_{l=1}^{K}\left(\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{I(y_{ij}=\omega)}\right)^{I(z_i=k)I(z_j=l)} \cdot \prod_{k=1}^{K}\prod_{l=1}^{K}\frac{1}{B(\boldsymbol{\beta})}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{\beta_\omega-1}dp_\omega^{kl} \\
&= \int_{p_\omega^{kl}\in\underline{p}^{kl}\in\boldsymbol{p}} \prod_{k=1}^{K}\prod_{l=1}^{K}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{\sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} I(y_{ij}=\omega)I(z_i=k)I(z_j=l)} \prod_{k=1}^{K}\prod_{l=1}^{K}\frac{1}{B(\boldsymbol{\beta})}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{\beta_\omega-1}dp_\omega^{kl} \qquad (18)
\end{aligned}
$$

Here we define

$$
N_{kl}^{\omega} = \sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} I(y_{ij}=\omega)I\left(z_i=k\right)I\left(z_j=l\right), \qquad (19)
$$

for $k=1,\ldots,K$ and $l=1,\ldots,K$, so that N_{kl}^{ω} counts the number of times that the outcome ω was observed for all games involving a team allocated to block k playing at home against a team allocated to block l. Thus, we can rewrite (18) as:

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{z},K) &= \int_{p_\omega^{kl}\in\underline{p}^{kl}\in\boldsymbol{p}} \prod_{k=1}^{K}\prod_{l=1}^{K}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{N_{kl}^{\omega}} \cdot \prod_{k=1}^{K}\prod_{l=1}^{K}\frac{1}{B(\boldsymbol{\beta})}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{\beta_\omega-1}dp_\omega^{kl}. \\
&= \prod_{k=1}^{K}\prod_{l=1}^{K}\frac{1}{B(\boldsymbol{\beta})}\int_{p_\omega^{kl}\in\underline{p}^{kl}\in\boldsymbol{p}}\prod_{\omega=1}^{3}\left(p_\omega^{kl}\right)^{N_{kl}^{\omega}+\beta_\omega-1}dp_\omega^{kl}, \qquad (20)
\end{aligned}
$$

Inside (20), we recognise the kernel of a 3-dimensional Dirichlet distribution of the form:

$$
(p_1^{kl},p_2^{kl},p_3^{kl}) \sim Dir(N_{kl}^{\omega=1}+\beta_1, N_{kl}^{\omega=2}+\beta_2, N_{kl}^{\omega=3}+\beta_3).
$$

Therefore, we can re-write the right hand side of equation (20) as:

$$
\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{1}{B(\boldsymbol{\beta})}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+\beta_\omega)}{\Gamma(\sum_{\omega=1}^{3}\left(N_{kl}^{\omega}+\beta_\omega\right))},
$$

where

$$
\frac{1}{B(\boldsymbol{\beta})} = \frac{\Gamma(\sum_{\omega=1}^{3}\beta_\omega)}{\prod_{\omega=1}^{3}\Gamma(\beta_\omega)}.
$$

This yields the expression

$$
f(\boldsymbol{y}|\boldsymbol{z},K) = \prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\Gamma(\sum_{\omega=1}^{3}\beta_\omega)}{\prod_{\omega=1}^{3}\Gamma(\beta_\omega)}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+\beta_\omega)}{\Gamma(\sum_{\omega=1}^{3}\left(N_{kl}^{\omega}+\beta_\omega\right))}.
$$

Recall that in our framework the concentration parameters of the prior Dirichlet distribution of $\underline{p}^{kl}$ for $k,l=1,\ldots,K$ are all set to be equal to have a uniform prior,

$$
\beta_\omega = 1, \quad \text{for} \quad \omega = 1,2,3\,,
$$

resulting in,

$$
f(\boldsymbol{y}|\boldsymbol{z},K) = \prod_{k=1}^{K}\prod_{l=1}^{K}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}\left(N_{kl}^{\omega}+1\right))}. \qquad (21)
$$

A.2 Collapsing theta

$$
\begin{aligned}
\pi(\boldsymbol{z}|K) &= \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{z}|\boldsymbol{\theta}, K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\Theta} \\
&= \int_{\boldsymbol{\Theta}} \prod_{i=1}^{N} Multi(z_i; 1, \boldsymbol{\theta}) \cdot Dir(\boldsymbol{\theta}; \boldsymbol{\gamma})d\boldsymbol{\Theta} \\
&= \int_{\boldsymbol{\Theta}} \prod_{i=1}^{N}\prod_{k=1}^{K} \theta_k^{I(z_i=k)} \cdot \frac{\Gamma(\sum_{k=1}^{K}\gamma_k)}{\prod_{k=1}^{K}\Gamma(\gamma_k)} \prod_{k=1}^{K} \theta_k^{\gamma_k-1} d\boldsymbol{\Theta} \\
&= \Gamma(\sum_{k=1}^{K}\gamma_k) \int_{\boldsymbol{\Theta}} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} I(z_i=k)} \frac{1}{\prod_{k=1}^{K}\Gamma(\gamma_k)} \prod_{k=1}^{K} \theta_k^{\gamma_k-1} d\boldsymbol{\Theta} \\
&= \Gamma(\sum_{k=1}^{K}\gamma_k) \int_{\boldsymbol{\Theta}} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} I(z_i=k)+\gamma_k-1} \frac{1}{\Gamma(\gamma_k)} d\boldsymbol{\Theta}
\end{aligned} \tag{22}
$$

We set

$$
n_k = \sum_{i=1}^{N} I\left(z_i = k\right),\ k = 1, \ldots, K, \tag{23}
$$

where each n_k accounts for the number of nodes/teams allocated to block k. Therefore we can rewrite the right hand side of (22) as:

$$
\begin{aligned}
&\Gamma(\sum_{k=1}^{K}\gamma_k) \int_{\boldsymbol{\Theta}} \prod_{k=1}^{K} \theta_k^{n_k+\gamma_k-1} \frac{1}{\Gamma(\gamma_k)} d\boldsymbol{\Theta} \\
&= \frac{\Gamma(K \cdot \gamma_0)}{\prod_{k=1}\Gamma(\gamma_0)} \int_{\boldsymbol{\Theta}} \prod_{k=1}^{K} \theta_k^{n_k+\gamma_0-1} d\boldsymbol{\Theta}\,.
\end{aligned}
$$

We recognise the integral above as the density of a $Dir(n_1 + \gamma_0, \ldots, n_K + \gamma_0)$ distribution, thus we can write:

$$
\begin{aligned}
\pi(\boldsymbol{z}|K) &= \frac{\Gamma(K \cdot \gamma_0)}{\prod_{k=1}^{K}\Gamma(\gamma_0)} \frac{\prod_{k=1}^{K}\Gamma(n_k + \gamma_0)}{\Gamma(\sum_{k=1}^{K}(n_k + \gamma_0))} \\
&= \frac{\prod_{k=1}^{K}\Gamma(n_k + \gamma_0)}{\Gamma^K(\gamma_0)} \frac{\Gamma(K \cdot \gamma_0)}{\Gamma(N + K\gamma_0)}\,.
\end{aligned}
$$

Setting the hyperparameter $\gamma_0 = 1$, leads to

$$
\pi(\boldsymbol{z}|K) = \prod_{k=1}^{K} \Gamma(n_k + 1) \frac{\Gamma(K)}{\Gamma(N + K)}. \tag{24}
$$

Collapsed posterior

Using (21) and (24) results in the expression for the collapsed posterior,

$$
\begin{aligned}
\pi(\boldsymbol{z}|\boldsymbol{y}, K) &\propto f(\boldsymbol{y}|\boldsymbol{z}, K)\pi(\boldsymbol{z}|K)\pi(k) \\
&= \prod_{k=1}^{K}\prod_{l=1}^{K} \Gamma(3) \frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega} + 1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega} + 1))} \cdot \prod_{k=1}^{K} \Gamma(n_k + 1) \frac{\Gamma(K)}{\Gamma(N + K)} \times \frac{1}{K!}\,.
\end{aligned} \tag{25}
$$

.

Appendix B Details of the MCMC move types

B.1 Acceptance probabilities of the MK move

Insert attempt

The proposal probabilities formulated in terms of an *insert* attempt are:

$$\begin{aligned} P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}', K')) &= P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}, K+1)) \\ &= Pr(\text{Insert an empty cluster}) \\ &= \begin{cases} 0.5, \text{ if } K < K_{max} \\ 0, \text{ if } K = K_{max}. \end{cases} \\ P_{prop}((\boldsymbol{z}', K') \to (\boldsymbol{z}, K)) &= P_{prop}((\boldsymbol{z}, K+1) \to (\boldsymbol{z}, K)) \\ &= Pr(\text{Delete an empty cluster}) \\ &= 0.5 \end{aligned}$$

Recall that the algorithm will reject the move to increase the number of clusters when $K = K_{max}$ and also that this move does not propose to change $\boldsymbol{z}$. Here, when $K < K_{max}$ the ratio of the posterior density at the proposed and current states can be written as:

$$\begin{aligned} \frac{\pi(\boldsymbol{z}, K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} &= \frac{f(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}|K')\pi(K')}{f(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}|K)\pi(K)} \\ &= \frac{\pi(\boldsymbol{z}|K+1)}{\pi(\boldsymbol{z}|K)} \times \frac{\pi(K+1)}{\pi(K)} \\ &= \frac{\left(\prod_{k=1}^{K+1} \Gamma(n_k+1) \frac{\Gamma((K+1))}{\Gamma(N+(K+1))}\right)}{\left(\prod_{k=1}^{K} \Gamma(n_k+1) \frac{\Gamma(K)}{\Gamma(N+K)}\right)} \times \frac{K!}{(K+1)!}. \end{aligned} \tag{26}$$

Note that since the newly inserted cluster is not allocated any nodes,

$$\prod_{k=1}^{K+1} \Gamma(n_k+1) = \prod_{k=1}^{K} \Gamma(n_k+1)$$

and so (26) reduces to

$$\begin{aligned} \frac{\pi(\boldsymbol{z}, K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} &= \frac{\left(\frac{\Gamma((K+1)}{\Gamma(N+(K+1)}\right)}{\left(\frac{\Gamma(K)}{\Gamma(N+K)}\right)} \times \frac{1}{K+1} \\ &= \frac{K}{(N+K)(K+1)}. \end{aligned} \tag{27}$$

Since we have symmetric proposal probabilities when $K < K_{max}$, the acceptance probability of the *insert attempt* is:

$$\alpha = \frac{K}{(N+K)(K+1)} \tag{28}$$

Delete attempt

For the *delete attempt*, provided that $K > 1$, the proposal probabilities are:

$$\begin{aligned} P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}', K')) &= P_{prop}((\boldsymbol{z}, K) \to (\boldsymbol{z}, K-1)) \\ &= Pr(\text{delete an empty cluster}) \\ &= \begin{cases} 0.5, \text{ if } K > 1 \\ 0, \text{ if } K = 1. \end{cases} \\ P_{prop}((\boldsymbol{z}', K') \to (\boldsymbol{z}, K)) &= P_{prop}((\boldsymbol{z}, K-1) \to (\boldsymbol{z}, K)) \\ &= Pr(\text{Insert an empty cluster}) \\ &= 0.5 \end{aligned}$$

Notice again that, due to the way the algorithm is built, we will prevent a *delete* move from taking place if $K = 1$. Here, when $K > 1$ the ratio of the posterior density at the proposed and current states can be written as:

$$\begin{aligned}\frac{\pi(\boldsymbol{z}, K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} &= \frac{f(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}|K')\pi(K')}{f(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z}|K)\pi(K)} \\ &= \frac{\pi(\boldsymbol{z}|K-1)}{\pi(\boldsymbol{z}|K)} \times \frac{\pi(K-1)}{\pi(K)} \\ &= \frac{\left(\prod_{k=1}^{K-1}\Gamma(n_k+1)\frac{\Gamma((K-1))}{\Gamma(N+(K-1))}\right)}{\left(\prod_{k=1}^{K}\Gamma(n_k+1)\frac{\Gamma(K)}{\Gamma(N+K}\right)} \times \frac{K!}{(K+1)!}\,. \end{aligned} \tag{29}$$

Since a *delete* move involves removing an empty cluster,

$$\prod_{k=1}^{K-1}\Gamma(n_k+1) = \prod_{k=1}^{K}\Gamma(n_k+1)$$

and so (29) reduces to

$$\frac{\pi(\boldsymbol{z}, K'|\boldsymbol{y})}{\pi(\boldsymbol{z}, K|\boldsymbol{y})} = \frac{K(N+K-1)}{K-1}\,.$$

The acceptance probability of *delete* attempt when $K > 1$, since this results in symmetric proposal probabilities, is:

$$\alpha = \frac{K(N+K-1)}{K-1}. \tag{30}$$

B.2 Acceptance probability for Metropolis-within-Gibbs move

Here we provide some details on the acceptance probability in (11) which we can express as

$$\frac{\pi(\boldsymbol{z}', K|\boldsymbol{y})}{\pi(\boldsymbol{z}^{(s+1)}, K|\boldsymbol{y})} = \frac{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega'}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega'}+1))}}{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}} \times \frac{\prod_{k=1}^{K}\Gamma(n_k^{'}+1)}{\prod_{k=1}^{K}\Gamma(n_k^{'}+1)}, \tag{31}$$

where $N_{kl}^{\omega'}$ is the equivalent quantity calculated in (19), but applied to the proposed allocation vector $\boldsymbol{z}'$,

$$N_{kl}^{\omega'} = \sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} I(y_{ij}=\omega)I\left(z_i^{'}=k\right)I\left(z_j^{'}=l\right).$$

However, recall that proposed allocation vector $\mathbf{z}'$ is identical to the current state of the allocation vector, $\boldsymbol{z}^{(s+1)}$, except for its ith element, z_i', which we suppose is allocated the label k_1. Further, suppose that the current state $z_i^{(s+1)} = k_0$. It turns out that this leads to some simplification in the first term in the right hand side of (31), since in this case,

$$N_{kl}^{\omega'} = \begin{cases} N_{kl}^{\omega'}, \text{ if } k=k_0 \text{ or } k=k_1 \text{ or } l=k_1 \text{ or } l=k_0, \\ N_{kl}^{\omega}\,, \text{ otherwise.} \end{cases} \tag{32}$$

This leads to a significant saving in the calculation of this term in (31). The second term on the right hand side of (31) can also be simplified. Since in this case, we note that $n_k^{'}$ is defined similar to (33) as

$$n_k^{'} = \sum_{i=1}^{N} I\left(z_i^{'}=k\right),\ k=1,\ldots,K, \tag{33}$$

and again we remark that $\boldsymbol{z}'$ is identical to the current state of the allocation vector, $\boldsymbol{z}^{(s+1)}$, except for its ith element, $z_i' = k_1$. This implies that

$$n_k^{'} = \begin{cases} n_{k_1}+1, \text{ if } k=k_1, \\ n_{k_0}-1, \text{ if } k=k_0, \\ n_k\,, \text{ otherwise.} \end{cases} \tag{34}$$

consequently, the second term on the right hand side of (31) can be written as

$$\begin{aligned}\frac{\prod_{k=1}^{K}\Gamma(n_k'+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} &= \frac{\Gamma(n_{k_0})\Gamma(n_{k_1}+2)}{\Gamma(n_{k_1}+1)\Gamma(n_{k_0}+1)} \\ &= \frac{n_{k_1}+1}{n_{k_0}}\,. \end{aligned} \tag{35}$$

B.3 Acceptance probabilities of the AE move

B.3.1 Ejection attempt

The proposal probabilities of an *ejection attempt* can be written as:

$$\begin{aligned} P_{prop}((Z', K') \to (Z, K)) &= P_{prop}((Z', K+1) \to (Z, K)) \\ &= Pr(\text{absorb a cluster}) \times \frac{1}{\substack{\text{\# labels for} \\ \text{absorbing cluster}}} \times \frac{1}{\substack{\text{\# labels for} \\ \text{absorbed cluster}}} \\ &= (1 - p_K^e)\frac{1}{K+1}\frac{1}{K}. \end{aligned}$$

While the reverse proposal probability is detailed as:

$$\begin{aligned} P_{prop}((Z, K) \to (Z', K')) &= P_{prop}((Z, K) \to (Z', K+1)) \\ &= Pr(\text{eject a cluster}) \times \frac{1}{\substack{\text{\# labels for} \\ \text{ejecting cluster}}} \times \frac{1}{\substack{\text{\# labels for} \\ \text{ejected cluster}}} \times Pr(\text{new allocation for Z}) \\ &= p_K^e \times \frac{1}{K+1} \times \frac{1}{K} \times \binom{n_{j_1} + n_{j_2}}{n_{j_1}} p_E^{n_{j_2}} (1 - p_E)^{n_{j_1}} \cdot \frac{p_E^{a-1}(1-p_E)^{a-1}}{B(a,a)} \end{aligned} \tag{36}$$

for $p_E \sim Beta(a, a)$ and where,

$$n_{j_1} = \sum_{i=1}^{N} I(z_i' = j_1) \text{ is the number of nodes in the ejecting cluster after reallocation,}$$
$$n_{j_2} = \sum_{i=1}^{N} I(z_i' = j_2) \text{ is the number of nodes in the ejected cluster.}$$

After integrating out with respect to the distribution of p_E, the resulting proposal probability in (36) reduces to:

$$\begin{aligned} &p_K^e \times \frac{1}{K+1} \times \frac{1}{K} \times \binom{n_{j_1} + n_{j_2}}{n_{j_1}} \frac{1}{B(a,a)} \int_0^1 p_E^{n_{j_2}+a-1}(1-p_E)^{n_{j_1}+a-1} dp_E \\ &= p_K^e \times \frac{1}{K+1} \times \frac{1}{K} \times \frac{(n_{j_1}+n_{j_2})!}{n_{j_1}! n_{j_2}!} \frac{\Gamma(2a)}{\Gamma(a)\Gamma(a)} \frac{\Gamma(n_{j_1}+a)\Gamma(n_{j_2}+a)}{\Gamma(n_{j_1}+n_{j_2}+2a)} \\ &= p_K^e \times \frac{1}{K+1} \times \frac{1}{K} \times \frac{1}{n_{j_1}+n_{j_2}+1}. \end{aligned}$$

The final equation above follows from setting $a = 1$, so that $p_E \sim U(0, 1)$, as in McDaid et al. (2013).

The posterior probability ratio for the *ejection attempt* takes the following form:

$$\begin{aligned} \frac{\pi(Z', K'|Y)}{\pi(Z, K|Y)} &= \frac{f(Y|Z')\pi(Z'|K')\pi(K')}{f(Y|Z)\pi(Z|K)\pi(K)} \\ &= \frac{f(Y|Z')\pi(Z'|K+1)\pi(K+1)}{f(Y|Z)\pi(Z|K)\pi(K)} \\ &= \frac{\prod_{k=1}^{K+1}\prod_{l=1}^{K+1}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega'}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega'}+1))}}{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}} \times \frac{\prod_{k=1}^{K+1}\Gamma(n_k'+1)\frac{\Gamma(K+1)}{\Gamma(N+K+1)}}{\prod_{k=1}^{K}\Gamma(n_k+1)\frac{\Gamma(K)}{\Gamma(N+K)}} \times \frac{K!}{(K+1)!} \\ &= \frac{\prod_{k=1}^{K+1}\prod_{l=1}^{K+1}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega'}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega'}+1))}}{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}} \times \frac{\prod_{k=1}^{K+1}\Gamma(n_k'+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \times \frac{K}{(N+K)(K+1)} \end{aligned}$$

B.3.2 Absorption attempt

The proposal probabilities of the *absorption move* are:

$$\begin{aligned} P_{prop}((Z', K') \to (Z, K)) &= P_{prop}((Z', K-1) \to (Z, K)) \\ &= Pr(\text{absorb a cluster}) \times \frac{1}{\substack{\text{\# labels for} \\ \text{absorbing cluster}}} \times \frac{1}{\substack{\text{\# labels for} \\ \text{absorbed cluster}}} \\ &= (1 - p_K^e)\frac{1}{K-1}\frac{1}{K}. \end{aligned}$$

$$\begin{aligned} P_{prop}((Z, K) \to (Z', K')) &= P_{prop}((Z, K) \to (Z', K+1)) \\ &= Pr(\text{eject a cluster}) \times \frac{1}{\substack{\text{\# labels for} \\ \text{ejecting cluster}}} \times \frac{1}{\substack{\text{\# labels for} \\ \text{ejected cluster}}} \times Pr(\text{new allocation for Z}) \\ &= p_K^e \times \frac{1}{K-1} \times \frac{1}{K} \times \binom{n_{j_1} + n_{j_2}}{n_{j_1}} p_E^{n_{j_2}} (1 - p_E)^{n_{j_1}} \cdot \frac{p_E^{a-1}(a - p_E)^{a-1}}{B(a, a)} \\ &= p_K^e \times \frac{1}{K-1} \times \frac{1}{K} \times \frac{1}{n_{j_1} + n_{j_2} + 1}. \end{aligned}$$

Where, as before, the final equation above results from setting $a = 1$. The posterior probability ratio is given by:

$$\begin{aligned} \frac{\pi(Z', K'|Y)}{\pi(Z, K|Y)} &= \frac{f(Y|Z')\pi(Z'|K')\pi(K')}{f(Y|Z)\pi(Z|K)\pi(K)} \\ &= \frac{f(Y|Z')\pi(Z'|K-1)\pi(K-1)}{f(Y|Z)\pi(Z|K)\pi(K)} \\ &= \frac{\prod_{k=1}^{K-1}\prod_{l=1}^{K-1}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega'}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega'}+1))}}{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}} \times \frac{\prod_{k=1}^{K-1}\Gamma(n_k'+1)\frac{\Gamma(K-1)}{\Gamma(N+K-1)}}{\prod_{k=1}^{K}\Gamma(n_k+1)\frac{\Gamma(K)}{\Gamma(N+K)}} \times \frac{K!}{(K+1)!} \\ &= \frac{\prod_{k=1}^{K+1}\prod_{l=1}^{K+1}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega'}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega'}+1))}}{\prod_{k=1}^{K}\prod_{l=1}^{K}\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}} \times \frac{\prod_{k=1}^{K-1}\Gamma(n_k'+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \times \frac{K}{(N+K)(K+1)} \end{aligned}$$

Appendix C Summary of output from the SBM for season $2018/19$

We reproduce the table from Section 6.2 presenting $\pi(K|\boldsymbol{y})$. Here we have omitted $\pi(K = 4|\boldsymbol{y})$ as this had an estimated probability of 2×10^{-4}.

Table 6: Posterior probabilities for values of K

K	1	2	3
$\pi(K\|\boldsymbol{y})$	0.0	0.98	0.02

Table 7: Model $K = 2$: Posterior allocation probability (as a percentage) for each team

	ARS	BOU	BHA	BUR	CAR	CHE	CRY	EVE	FUL	HUD
Cluster 1	88.94	0.06	0.00	0.00	0.00	90.34	0.39	5.98	0.00	0.00
Cluster 2	11.06	99.94	100.00	100.00	100.00	9.66	99.61	94.02	100.00	100.00
	LEI	LIV	MCI	MUN	NEW	SOU	TOT	WAT	WHU	WOL
Cluster 1	0.55	100.00	100.00	83.51	0.00	0.00	80.43	0.02	0.95	21.81
Cluster 2	99.45	0.00	0.00	16.49	100.00	100.00	19.57	99.98	99.05	78.19

Table 8: Model $K = 3$: Posterior allocation probability (as a percentage) for each team

	ARS	BOU	BHA	BUR	CAR	CHE	CRY	EVE	FUL	HUD
Cluster 1	2.88	0.23	0.80	0.96	6.41	4.61	3.53	4.87	5.29	18.32
Cluster 2	86.15	1.45	0.12	0.02	0.00	88.70	5.26	29.01	0.00	0.00
Cluster 3	10.97	98.32	99.09	99.02	93.59	6.69	91.20	66.13	94.71	81.68
	LEI	LIV	MCI	MUN	NEW	SOU	TOT	WAT	WHU	WOL
Cluster 1	2.71	53.75	55.51	7.63	6.20	0.84	11.56	2.08	4.65	8.44
Cluster 2	3.98	46.25	44.49	81.24	0.61	0.14	71.72	1.71	14.32	50.48
Cluster 3	93.31	0.00	0.00	11.13	93.19	99.02	16.73	96.21	81.03	41.08

Appendix D Team abbreviations

ARS	Arsenal
AST	Aston Villa
BIR	Birmingham City
BLP	Blackpool
BLB	Blackburn Rovers
BOL	Bolton Wanderers
BOU	Bournemouth
BHA	Brighton
BUR	Burnley
CAR	Cardiff City
CHA	Charlton Athletic
CHE	Chelsea
CRY	Crystal Palace
DER	Derby County
EVE	Everton
FUL	Fulham
HUD	Huddersfield
HUL	Hull City
IPS	Ipswich Town
LEE	Leeds United
LEI	Leicester City
LIV	Liverpool
MCI	Manchester City
MID	Middlesbrough
MUN	Manchester United
NEW	Newcastle United
NOR	Norwich City
POR	Portsmouth
QPR	Queens Park Rangers
REA	Reading
SOU	Southampton
STK	Stoke City
SWA	Swansea City
SUN	Sunderland
TOT	Tottenham Hotspur
WAT	Watford
WBA	West Bromwich Albion
WIG	Wigan Athletic
WHU	West Ham United
WOL	Wolverhampton Wanderers