Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Modeling of group differences in passing networks of an NBA team

Relatore:	*Studente:*
Prof. Bruno Scarpa	Alvise Zanardo
Correlatore:	*Matricola N.*
Ph.D. Daniele Durante	1106821

Anno Accademico 2015/2016

Contents

List of Figures

Introduction

The passing fundamental is one of the most important key aspects of the game of basketball. Throughout the course of a single game between 250 and 350 passes occur on average, under several different circumstances. Professional coaches embrace their philosophy through plays that involve specific patterns and adaptations, in which passing is a core and crucial element. Great examples are represented by the triangle offense by Coach Phil Jackson, or the Princeton offense, which has its roots in college basketball. With the recent development in optical tracking systems, nearly every single instant of an entire game is traced and turned into huge multidimensional and complex raw data. Adapting statistical models, data mining and machine learning techniques it is possible to pre–process such data to extract several types of information, including passes, team performance, and others.

Since a pass is a connection between two teammates or, under a different perspective, between two areas of the court, this collection of ties can be naturally translated into a *network*–valued observation. The term *network analysis* refers to the analysis of the relationships structures among a set of interacting units, called *nodes*, and their underlying patterns. In the recent years, network analysis has garnered a substantial interest in different applied fields covering biostatistics, neuroscience, social science, and also sports. This type of data representation fits perfectly with the concept of passing: nodes can be identified as players (or possibly positions on the court), while a *tie* is represented by the pass itself.

When studying this structured type of data, usual statistical approaches often fail to capture the important traits of networks. It is therefore necessary to provide specific methods and models that embrace the complexity and the properties typical of a network. This is particularly true when the focus is on assessing group differences in the probabilistic generative mechanisms associated with groups of multiple network observations. In basketball, as in other sports, games are only won or lost. Coaches, players themselves, but also fans and television broadcasters, are always interested in understanding the reasons why a team performs better in certain situations than in others. Through this work, we want to inspect differences between wins and losses for an NBA team using passing networks data.

The structure of the thesis is defined as follows. Chapter 1 features a quick review of how statistics and basketball are connected, with a focus on research in the field of network analysis. In Chapter 2 we detail the adopted pre-processing procedure to retrieve the passing networks from the raw tracking data. Some descriptive analyses and classical statistical models for networks are considered in Chapter 3. Lastly, Chapter 4 features a recently proposed Bayesian nonparametric model for undirected data, and provides a novel generalization of it for directed networks. Refer to Table A.1 in the Appendix for all the basketball-related terms that will be used in this work.

Chapter 1

Statistics and Basketball

The relationship between statistical analysis and basketball has a long history, dating back to even before the official National Basketball Association was instituted in 1946. Slowly but steadily the amount of information gathered increased, starting from 1894/1895 with very simple score sheets for the games of the Springfield YMCA league (in which the creator of the game himself Dr. James Naismith played); in late 60s *box scores* featured the addition of *rebounds*, *field goals* made/attempted and other simple statistics, up until the 1996/1997 season when the NBA Stats division (`http://nba.com/stats`) started collecting full *play-by-play* logs. The most recent development is the processing of player tracking data provided by the company STATS LLC with the recently implemented SportVU® system (`http://www.stats.com/sportvu-basketball`, see Section 2.1 for details).

Researchers and professional statisticians have answered to the evolution in the amount and richness of data with increasingly complicated analyses, starting from early examples in literature studying team and individual performance (Elbel and Allen, 1941). This early paper started questioning the utility of merely tracking scores and considered different measuring factors for wins/losses such as assists, violations etc, that only a few years later became part of official *box scores*. More recently, in the last 20 years the use of possession statistics started gaining popularity due to their different approach that focuses on offensive and defensive metrics per 100 possessions, adapting for any pace a team can possibly have. Following the success in baseball statistics represented by *Sabermetrics* (Grabiner, 1994) by Bill James, Professor Dean Oliver created *APBRmetrics* (`http://www.apbr.org`), named after the Association for Professional Basketball Research. After publishing various articles about the topic, his work culminated with his first book *Basketball on paper* (Oliver, 2004), that illustrates how powerful statistical tools can help in explaining game results, player and team performance etc., when evaluated per possession or per minute (instead of doing it per game). Other notable and more recent works on basketball statistics include *Basketball Analytics: Spatial Tracking* (Shea, 2014), that uses optical tracking data provided by SportVU and others to investigate game strategy, player evaluation, player types, and prospect potential. The author introduces new measures of a player's scoring and playmaking efficiency, quantifies the offense spacing and defense stretching, and demonstrates several ways in which the NBA game has changed over the years.

Lately, research has also tried to answer specific questions such as "Who is the most productive player for his team?" or "How do we measure an expected value of an on-going play?" making use of the powerful optical tracking data. Notably, researchers at Harvard University have been working profusely on different complex topics. For example, Miller et al. (2014) proposed methods to extract spatial patterns from NBA shooting data using Gaussian, Poisson, Log-Gaussian Cox Processes and Non-Negative Matrix Factorization; they also provided interesting insights on how differently shooting frequency and shooting efficiency are characterized (Miller et al., 2014). Franks et al. (2015) additionally proposed new defensive metrics that shed light on this important aspect of the game. Their work included a Hidden Markov model to recognize defensive–matchups, estimates of percentage of contested shots, and many other new metrics.

All these works were developed in order to satisfy the growing interest in basketball analytics requested by teams themselves and a widespread constantly growing fanbase. For coaching staffs, it is an extremely powerful tool that helps them to better understand their team capabilities and flaws, providing a bigger picture that is not observable from a single possession or even a single game. For players, it can be used to understand their productivity and consequently upgrade their game. This is the reason why STATS LLC has been providing teams and players, as well as betting companies, a wide variety of statistical insights they could work with.

1.1 Network analysis in basketball

The topic of this thesis is to study basketball data from a network perspective. In particular the overarching goal is to understand how team performance relates to team passing structures. This type of statistical approach to basketball data is still in its infancy, although few examples can be found in the literature.

Gupta et al. (2015) review how network analysis concepts can be used to analyze and characterize team, and individual behaviors in basketball, making use of SportVU data and *play-by-play* logs. The authors build the networks around three different types of nodes: *start of play*, that identifies *inbounds, steals, rebounds* as in a start of new possession; *end of play*, indicating events like *shots, turnovers, offensive fouls*; *players*, as players' ids. The article focuses on the characterization of games in terms of descriptive indices such as *entropy* and *degree centrality*, also comparing college team Ohio State University games with NBA counterparts. Fewell et al. (2012) provide similar descriptive analyses this time representing nodes as players' positions instead of individual players, inspecting teams' differences. The data were taken from the first round of the National Basketball Association *playoffs* of the 2009/2010 season. These two works both focused on single game networks individually.

As far as we know, at the time we started working on this thesis project no study had been done on analyzing passing networks' data in professional basketball other than descriptive analyses. It is therefore of interest to explore this new field in order to provide insights on this key aspect of basketball mechanics through statistical inference.

Chapter 2

The data and the pre–processing procedure

The aim of this thesis is to study the passing behaviour of a basketball team, its underlying characteristics and how it is related to team performance. To accomplish this goal the focus is on the data from the first half of the 2015/2016 season of the United States of America basketball pro league, comprising information on more than 600 games. Refer also to the National Basketball Association official website for more details. The team under analysis is the Golden State Warriors, for which data on 26 games are available: 23 consecutive wins and 3 losses. A list of the games with opponents, date and outcome is available in Table 2.1. Because of the incredible amount of information contained in these files (around 100 MB per game, for a total of 635 files) the decision of considering only one team was made, although the whole procedure can be repeated for any team in the data. The choice of Golden State lies on the personal interest in finding differences between *quarters* where the team performed well or badly during and after the record setting streak of 24 consecutive wins [1], according to differences in score (plus–minus). As only 3 games — out of 26 — were losses, the passing behaviour is studied on a *quarter* basis. This allows an increased amount of information on bad performances, as there were several games whose result ended up in favor of Golden State where some quarters were lost.

[1] The second game of the season vs Houston Rockets was not available in the data.

Table 2.1: The schedule of the 26 analyzed games.

date	home	visitor	result
2015-10-27	*Golden State Warriors*	New Orleans Pelicans	W
2015-10-31	New Orleans Pelicans	*Golden State Warriors*	W
2015-11-02	*Golden State Warriors*	Memphis Grizzlies	W
2015-11-04	*Golden State Warriors*	Los Angeles Clippers	W
2015-11-06	*Golden State Warriors*	Denver Nuggets	W
2015-11-07	Sacramento Kings	*Golden State Warriors*	W
2015-11-09	*Golden State Warriors*	Detroit Pistons	W
2015-11-11	Memphis Grizzlies	*Golden State Warriors*	W
2015-11-12	Minnesota Timberwolves	*Golden State Warriors*	W
2015-11-14	*Golden State Warriors*	Brooklyn Nets	W
2015-11-17	*Golden State Warriors*	Toronto Raptors	W
2015-11-19	Los Angeles Clippers	*Golden State Warriors*	W
2015-11-20	*Golden State Warriors*	Chicago Bulls	W
2015-11-22	Denver Nuggets	*Golden State Warriors*	W
2015-11-24	*Golden State Warriors*	Los Angeles Lakers	W
2015-11-27	Phoenix Suns	*Golden State Warriors*	W
2015-11-28	*Golden State Warriors*	Sacramento Kings	W
2015-11-30	Utah Jazz	*Golden State Warriors*	W
2015-12-02	Charlotte Hornets	*Golden State Warriors*	W
2015-12-05	Toronto Raptors	*Golden State Warriors*	W
2015-12-06	Brooklyn Nets	*Golden State Warriors*	W
2015-12-08	Indiana Pacers	*Golden State Warriors*	W
2015-12-11	Boston Celtics	*Golden State Warriors*	W
2015-12-12	Milwaukee Bucks	*Golden State Warriors*	L
2015-12-30	Dallas Mavericks	*Golden State Warriors*	L
2016-01-13	Denver Nuggets	*Golden State Warriors*	L

2.1 The SportVU data

As mentioned in Chapter 1, the data processed in this work are the result of the highly sophisticated SportVU optical tracking system used by STAT LLC. Starting November 2015 the NBA Stats division began hosting the files regularly on their website (http://nba.com/stats). These open data were available until January 23rd, 2016 when the NBA decided to stop the service due to technical reasons.

The system, originally designed for military use, was developed in 2005 by an Israeli engineer named Miky Tamir whose background is in missile tracking and advanced optical recognition. After being used for soccer matches in Israel, the technology was purchased by STATS LLC in 2008 to provide a similar service for basketball. With its constant movement and only 11 elements to track (5 home players, 5 away players and the ball) basketball would make full use of this overflowing stream of data that could possibly provide much richer insights than the current statistics. In the 2009–2010 season the first tests were carried out with a few teams [2] willing to explore this new field in basketball statistics; these were followed by a couple team in the following year, to eventually reach all 30 teams in 2013. The tracking system works through the use of six computer vision cameras installed in the rafters of the arenas, equally divided per half court; they collect two–dimensional coordinates for all players and a third dimension (height) for the ball [3], 25 times per second. With a rough calculation, given that NBA games last at least 48 minutes without counting important moments when the clock is not running that are usually tracked, this means that for each game the technology collects at least $48*60*25 = 72000$ different "pictures". Although still not perfect, either in the tracking or in the collecting step, SportVU is now able to correctly gather data for more than the 99.9% of the total *moments* (as they are called in the files) for the whole game; unfortunately from time to time weird and sometimes very funny behaviors are registered, such as a player running 50 feet in less than half a second. Therefore it is fundamental to carefully pre–process the raw data in order to avoid corrupted information substantially affecting the final analyses. Due to the large amount of information, this pre–processing step should be automatic and use state-of-the-art data mining and filtering procedures.

2.2 Structuring and filtering the data

We acquired the tracking data available from the official website of the NBA [4] in JSON (JavaScript Object Notation, `http://json.org/`) format, and parsed it using the software `R` (R Core Team, 2014). The structure of the data is represented by several nested lists with various info about the game, stored as *double* or *character* class. The detailed list is provided in Figure 2.1. The list named *events* contains a number e_m of items that correspond to *play-by-play* elements such as a shot, a rebound, a turnover etc, and comprise all the time units (in 25ths of seconds) that are relevant for that specific event. This means

[2] Namely Dallas Mavericks, Houston Rockets, Oklahoma City Thunder and San Antonio Spurs.

[3] Height will be probably provided for players in the near future as well.

[4] As of January 23rd, 2016 the data are no longer available.

- **json object**: List of **3** items:
  - *gameid*: the unique id of the match
  - *game date*: complete date of the match in YYYY-MM-DD format
  - *events*: List of $\boldsymbol{E}$ items:
    - e^{th} element: List of **4** items:
      - *eventid*: e, event progressive identifier (numeric)
      - *visitor/home*: List of **4** items:
        - *name*: complete name of the team, as in "Golden State Warriors"
        - *teamid*: the numeric id of the team
        - *abbreviation*: as in "GSW"
        - *players*: List of $\boldsymbol{P_t}$ (t is the team index) items:
          - p_t^{th} element: List of **5** items:
            - *lastname, firstname, playerid, jersey, position*
      - *moments*: List of $\boldsymbol{M}$ items:
        - m^{th} element: List of **6**:
          1. - a numeric var. for the *period*
          2. - a numeric var. for the *time* in Unix format[1]
          3. - a numeric var. for the *game clock* (from 720.00 to 0)
          4. - a numeric var. for the *shot clock* (from 24.00 to 0)
          5. - **always NULL**
          6. List of $\boldsymbol{J_m}$ (usually $\boldsymbol{J_m} = 11$):
             (a) - *teamid*
             (b) - *playerid*
             (c) - $\boldsymbol{x}$: the coordinate relative to the longer side of the court (0,94) in feet
             (d) - $\boldsymbol{y}$: the coordinate relative to the shorter side of the court (0,50) in feet
             (e) - $\boldsymbol{z}$ (only for the ball): distance from the ground (in feet)

Figure 2.1: The complete indexing of the JSON object for a single example game. (1) The Unix format counts the amount of milliseconds from January 1st, 1970.

that many of these elements, like a missed shot and the relative rebound that follows, have overlapping time units. Hence, the variable containing the time in *Unix* format is being used to only select all the unique single moments so that there are no repeated datapoints. While doing so, we are also interested in storing relevant moments where the ball is not necessarily alive (the ball is alive if the game clock is running), like the case of *inbounds*. To do so a variable checking for inbound status is created for each moment, so that it is 1 if the ball is being inbounded, that is when it's coming from outside the court and the game clock is not running (otherwise we would count as inbound also a dribble or a pass that is temporarily out of the imaginary plane that cuts the air vertically rising from the boundary lines, e.g. when a player is diving out of bounds to maintain possession), and 0 otherwise. Unfortunately, a couple of unmanageable inconveniences might still occur and make inbound data irretrievable: a player could be inbounding the ball while being out of bounds with his feet, but close to the line so that the ball is inside the court rectangle; alternatively, the tracking system could possibly be off until the ball is alive.

After these checks, if the moment is not in the list of unique *Unix* times and the clock is running, or the ball is being inbounded, we update the list containing all moments information. This features a list itself with: a dataframe with players and ball coordinates and ids, plus all the time–related variables. This way, we have a flexibly manageable set of data to be processed later. It is also important to note that in order to treat all the games the same, all the points have been "mirrored" if necessary so that every match would have Golden State attacking from the left (corresponding to $x = 0$) to the right side (corresponding to $x = 94$), for each of the four quarters. At this stage, we have clean data regarding the positions of players and ball and several other features that allow us to visualize interesting traits such as the heatmaps presented in Figure 2.2; these plots display the distribution of the movements by two different players in the 4 different periods of a sample game, both in the defensive end (left) and offensive end (right). Warmer colors correspond to a higher concentration.

At the same time, we need other types of data that are not stored in the JSON files to obtain additional important information. This group comprises the features that will serve as the differencing variable for the networks later in the analyses, and some other details that can help in the human pass–recognizing step that will be explained later in this section. These infos are stored in what is called the *play-by-play* of the game, which can be directly grabbed or scraped from the NBA official website via a quick function in `R`, simply knowing the official *gameid* of the game. Out of all the data stored in the *play-by-play*, we

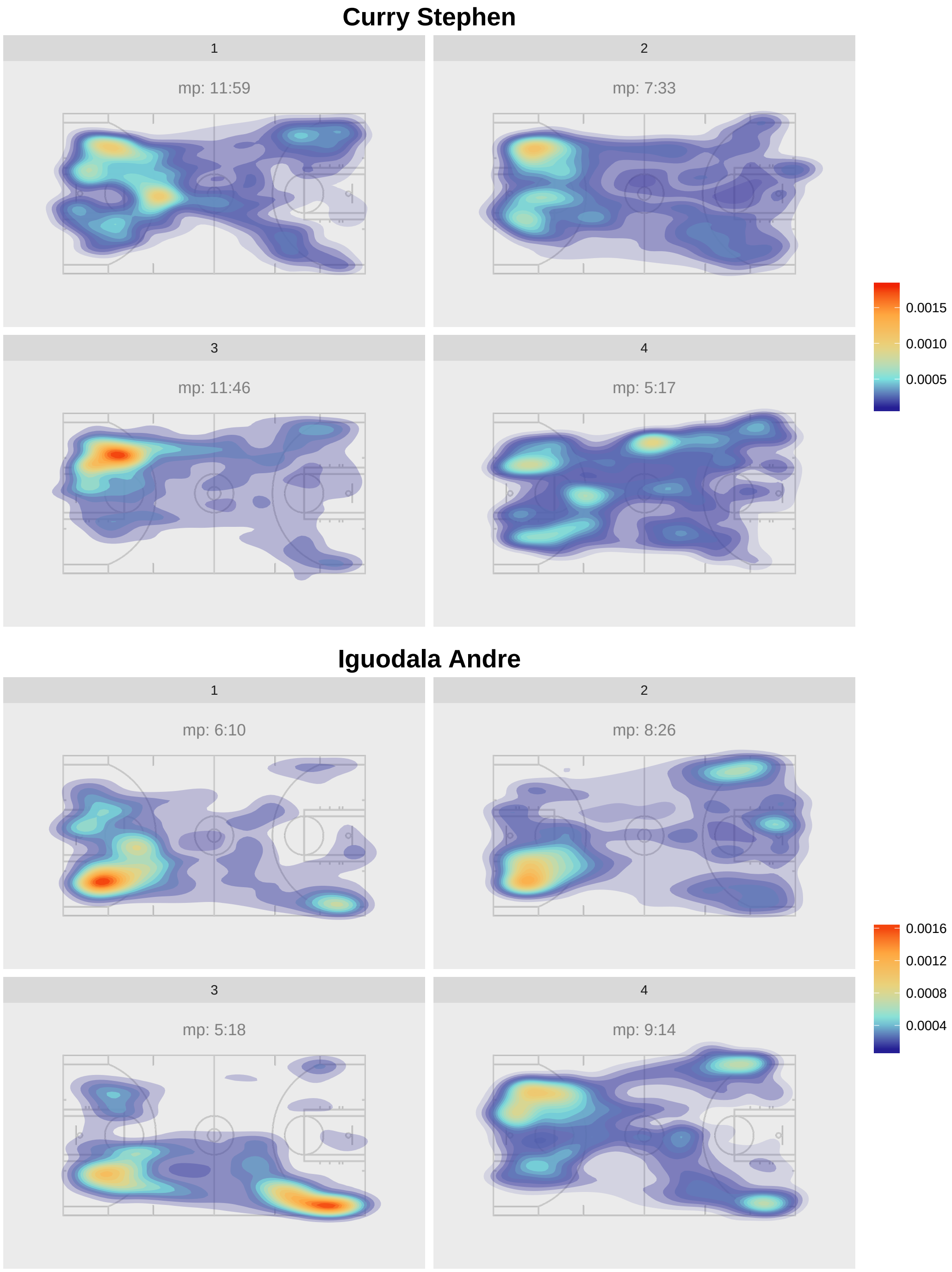

Figure 2.2: An example movements heatmap off the cleaned data from one game. The four rectangles refers to game quarters; "mp" stands for minutes played in that quarter in this case by point guard Stephen Curry and guard/forward Andre Iguodala respectively.

are interested in the partial scores for each period (to characterize the periods in terms of plus–minus) and the description of events. The latter contain numerical *ids* that specify the type of event happening, such as *left side three–point* shot or *alley–oop pass*[5]; they will be used in the next step to help recognizing possession and passes in the human–regulated classification stage that will provide training and test sets to obtain the data later used in the analyses.

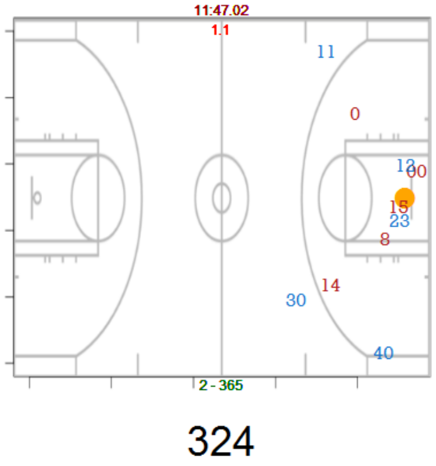

Figure 2.3: An example image used for human classification.

In order to be able to classify the data into possession (1 if the Golden State Warriors have the ball, 0 otherwise) and recognize passes, including the moments where a change of possession or a pass is happening, a set of around 21000 images (one image per 25th of a second, meaning this roughly corresponds to a little more than one period of one game) was produced so that *manual human classification* could be performed. These images featured the positions of all players identified by their jersey number and colored differently according to the team. The aforementioned time data, moment id, and additional event labels were also provided in the images to help distinguish critical plays such as *alley–oops* and shots, or made and missed shots etc. Specifically, with the term *manual human classification* we mean the procedure with which we manually classified all moments by looking at sequences of images like the example in Figure 2.3. The data were stored in *.csv* files, whose single lines contained: the jersey number of the player involved in the action; an event label, e.g. "RE" for received the ball, "H" for having the ball (as in possession) etc.; the moment when the new event started happening. All events concerning the opponent team were marked with an "X" so that they could be easily distinguished from the ones when Golden State had the ball, and filtered out. The black number at the bottom is the unique *moment* of the game based on *Unix* time; the

[5] To understand which numbers corresponded to what event, a sample game was inspected while checking the explicit description available in the *play-by-play* data.

two green numbers directly above are respectively the event and the moment number as in the JSON files (see Figure 2.1); Golden State Warriors are always displayed with a blue color while opponents are colored in red; the ball is identified as orange and changes its size accordingly to its height; the bright red number on top shows the *shot clock* while crimson red indicates the *game clock*.

2.3 From clean data to passes

The aim of this project is to analyze passing networks. Therefore, all the passes happening in every game need to be retrieved out of the whole stream of data for all 26 games. In order to do so, after collecting the human–classified data, a 2-step procedure is carried out. Firstly, a statistical model is fit to select only the moments in which Golden State has the ball, to consequently decrease the number of possessions to process in the next stage. This step is fundamental since by doing so we avoid processing data that we are not interested in (we are only focusing on one team), saving a considerable amount of time and memory space while fitting the models. Secondly, a pass recognizing model is designed to ultimately get data on when and where the act of passing started and ended, and who were the players involved [6]. To fit the models, some useful additional variables are created specifically to deal with the two different problems (see details below).

In accomplishing the above goals, we estimate several different types of models with different sets of variables to perform a majority vote classifier; this improved significantly the performance of single models. The models we used are: Random Forests (Breiman, 2001) with varying tuning parameters such as number of variables to possibly split at each node; Generalized Linear Models (Nelder and Baker, 1972), with forward stepwise variable selection; adaBoost (Freund and Schapire, 1996), extreme gradient Boosting (Friedman, 2001) and Bagging (Breiman, 1996), with tuning parameters on the growth of trees; k–nearest neighbours (Altman, 1992), with tuning parameter k. In the two steps the response variables are going to be, respectively: *possession*, that equals 1 when Golden State has possession in that specific moment (as in 25ths of a second), and 0 otherwise; *pass*, that is 1 if any Golden State player is in the act of passing the ball or has just received it, and 0 otherwise.

[6] This last information is not being used for the analyses in the next chapters, but throughout the data pre–processing step we gathered as many details as possible for possible future applications that would still not impact too much on the processing time.

2.3.1 Detecting possession

In this first step we focused on selecting all moments in which Golden State appears to be in possession of the ball, filtering away everything else (that is both when opponent team has possession and when the ball is loose). This way we will only deal with the time when any player of the Warriors' team is potentially able to be in the act of passing. To do this, we design the majority vote classifier previously mentioned for a total of 11 different models that feature, in addition to information already obtainable from the clean data (ball's x and z coordinates, and the game clock):

- the average of the x coordinates of Golden State players on the court and the opponents respectively (two different variables), since usually defenders are closer to the basket and therefore to the baselines;
- the distance from the ball by the closest Warrior, and the closest opponent, as usually who is closer to the ball also has possession in that moment. In some models the difference of these two quantities was also considered;
- convex hull area: a dichotomous variable having value 1 when the area of the polygon formed by Golden State players is bigger than the one formed by the opponents. Usually when the area is bigger, this means most players are outside the three–point line, suggesting that they are on offense;
- a differenced shot clock variable with $lag = 5$, since the shot clock is mainly a possession clock (when there are no offensive rebounds). A small lag is applied because usually shot clock operators take some time before actually assessing the possession by one of the two teams.

Since these models treat each moment as independent from all other moments, they sometimes predict consecutive moments to have different values. We then decided to use a minimum cutoff of 25 moments (corresponding to a whole second), that means that at least a second has to pass between a double change of possession. This operation is done on both the single predictions and the majority vote prediction, to provide robustness. This also turned out to improve performance significantly. In addition to test set performance, the predictions were visually tested with the help of images identical to the ones provided for human classification. After these operations, the final fit turned out to be close to perfect with just a few situations where possession to Golden State was assigned 1–2 seconds later than the actual moment when it happened.

We then proceed to estimating possession for all 26 games, to ultimately filter away all datapoints for which the majority vote adjusted prediction is equal to 0. These moments are the only ones we will feed the passes' recognition model with.

2.3.2 Detecting passes

In this step, we extract some other variables that might help in recognizing the moments when passes and receptions happen. Among these are:

- ball's *speed* and *diffspeed*: using consequent moments we track the space covered by the ball in a 25th of second, without considering changes in height. Since the optical tracking system is not 100% reliable, extremely high values are capped, together with values that have an unlikely big difference with respect to the previous moment. *diffspeed* is the differentiated version of this variable, since we expect a big variation in speed to be a possible warning for a pass happening;

- difference in ball's height (*diffz*): since passes can be performed in many different ways (*chest pass*, *lob*, *bounce pass*, *alley–oop* etc.), they also feature different variation in distance from the ground;

- *angle* and *diffangle*: we measure the angle that the trajectory of ball is tracing with respect to the lower sideline. Although similar to the speed variable in terms of information provided, this also accounts for quick passes that do not change speed considerably but change direction. This value is also smoothed (with a moving average) to regularize shaky optical mistracking. *diffangle* features a lagged version of *angle*;

- other variables used for the possession model.

Our goal is therefore to predict if a single moment has the "passing status". Similarly to what has been done for possession, we wanted to combine the classifiers to improve the total accuracy; although passes are harder to predict because they happen quite fast and sometimes comprehend challenging situations like an *alley–oop* that might be intended a shot (even to the human eye), the models surprisingly almost never presented critical false positives. This means that when they were predicting a pass was happening, it was indeed happening. Moreover, we applied a "continuity" correction as for possession predictions, but this time for a much shorter period since passes can happen quite quickly, so for each single classifier the threshold value was 3/25ths of a second. Because of the models' particular trait of very high specificity ($TN/(TN+FP)$[7]), we did not opt for a majority vote classifier. Instead, the "passing moment" (that is a moment when we predict a pass is happening) response variable is set to 1 whenever even just one model predicts $pass_m = 1$, where m indicates the m-th moment considered. Although initially this might not make sense, it is reasonable from a combining classifier perspective: imagine we have 11 people looking at a video trying to spot the instant in which a particular event happens. Since they might be focusing on different parts of the screen or might not consider that event to be happening, it is reasonable to think that at some point even just one of them will be pointing a difference out while the others will not. As a matter of fact, this approach granted a high performance in passes recognition close to 95%. The only situation in which all models seemed to hobble was in the case of *hand–off* passes. However, these very delicate situations almost always result in a pass that does not heavily impact on the play [8], and most importantly they do not imply a substantial movement of the ball in the court. Hence, we decided not to focus on this flaw. Also, as we will explain later, we will not consider passes happening in the same court area.

A *reality* correction

Unfortunately, the aforementioned models still had some critical deficiency. In fact, after fitting a solid base set for all quarters in every game, there were some situations that the models could not handle well. Among these we found that, in few occasions:

- a model was predicting $pass = 1$ when the same player was responsible for both the passing and the receiving act. For example this might happen in the case of a near turnover where the ball is later recollected. Clearly this should not be considered, so we used the tracked player id variable to filter all these passes away;
- a model had a prediction that lasted for too much, e.g. the pass was received after 5 seconds. We did not want it to consider these as passes even though there might be

[7] Here TN denotes "true negatives", as the number of observations for which the model correctly predicts a non–passing moment, while FP ("false positives") indicates the number of true non–passing moments that the model actually labels as passing.

[8] If a hand–off pass opens up a higher chance at a shot it is usually not due to the pass but possibly a screen or a quick reaction by the receiver.

situations in which this could possibly be the case, so these were deleted;

- a pass was predicted to be happening while no player was in the range of 5 feet: this situation was mainly due to mishandled optical tracking, so we decided to filter these away as well.

Finally, we had a sufficiently reliable passes dataset that we could use to create the passing networks. A first look at the players' positions at the moment of initiating or receiving a pass is available in Figures 2.4 and 2.5 respectively. These heatmaps measure the activity concentration in the court divided for quarters for which the resulting plus–minus was favorable (at least +0) or non–favorable (negative, at least −1). As far as we can judge, we cannot make any assumptions on the difference between wins a losses and therefore a more structured and focused analysis is needed. As it provides a valid option for this project, we decided to use statistical network analysis as a tool to answer the question "Is there a difference in passes networks between wins and losses?".

2.4 Building the networks

As a final step of the data pre–processing procedure, we want to actually turn the passes data into networks. Previous works available in the literature consider the interactions between single players (Clemente et al., 2015) or positions such as *point guard, shooting guard, forward, center* etc (Fewell et al., 2012), but they both have flaws in a bigger picture perspective, especially in aligning the nodes when multiple passes networks are considered. In fact, not everyone happens to play every quarter due to injuries or coaches decisions, and moreover some of the players might be considered as filling the same positions while having often very different personal traits. To avoid these issues and facilitate alignment of nodes in multiple networks, the players should be replaced with other types of nodes. This is also motivated by the fact that in the NBA coaches are famous for establishing "systems" that are not entirely built around single players, but rather follow a philosophy that takes advantage of different aspects of the game. A few examples could be Lakers' triangle offense by Coach Phil Jackson, of Spurs' continuous ball movement by Coach Gregg Popovich. The final court division used is available in Figure 2.6.

To make the problem more tractable, we also decided to treat the networks as binary: this means that a tie is registered if at least one pass was observed between two different areas of the court. As stated earlier, we do not include *self–loops* in the analysis, because

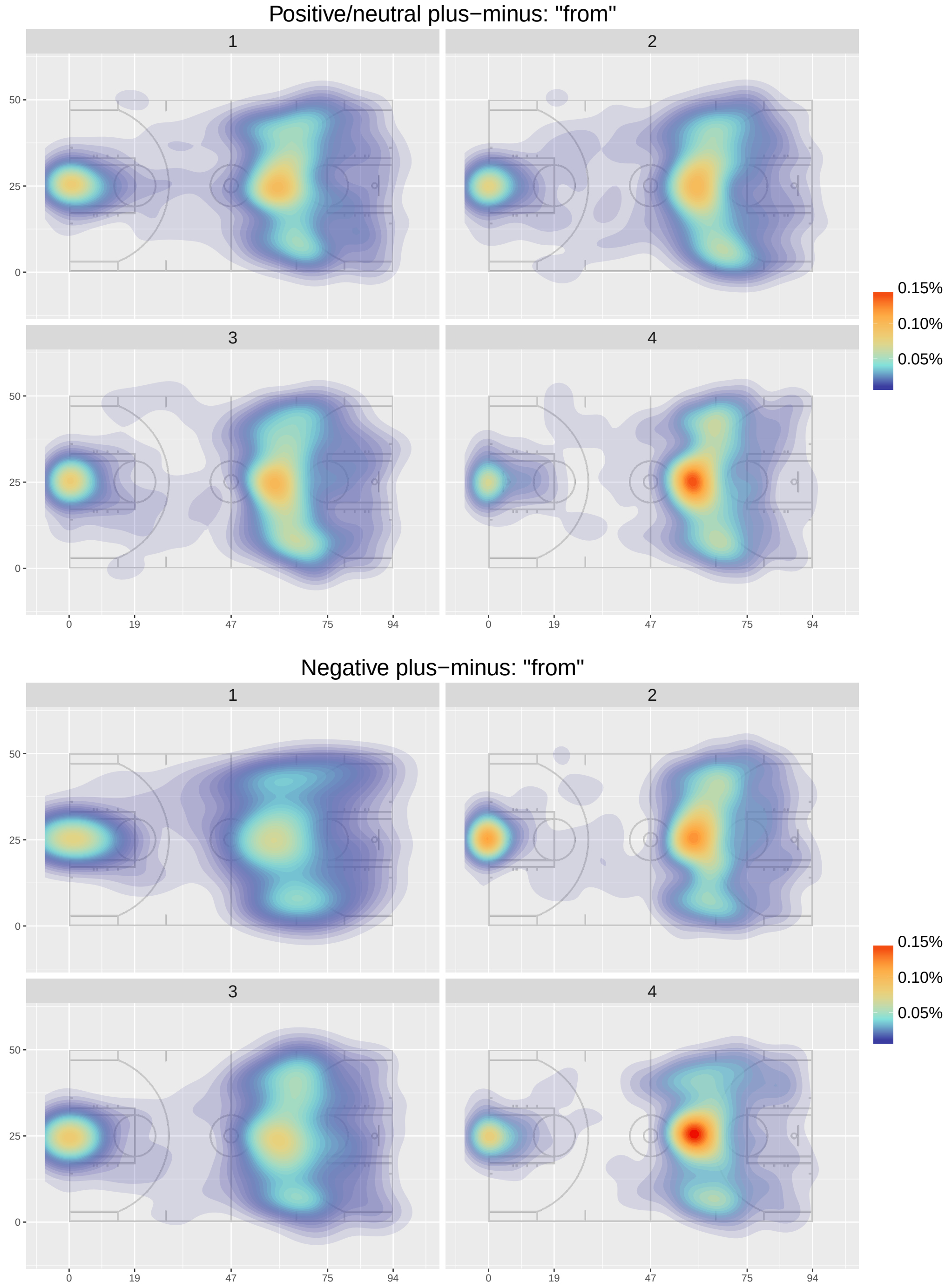

Figure 2.4: Heatmap displaying the kernel estimated density of the positions of players initiating the act of passing the ball for quarters with at least +0 plus–minus (TOP) and a negative plus–minus (BOTTOM). These plots refer to all the 26 analyzed games.

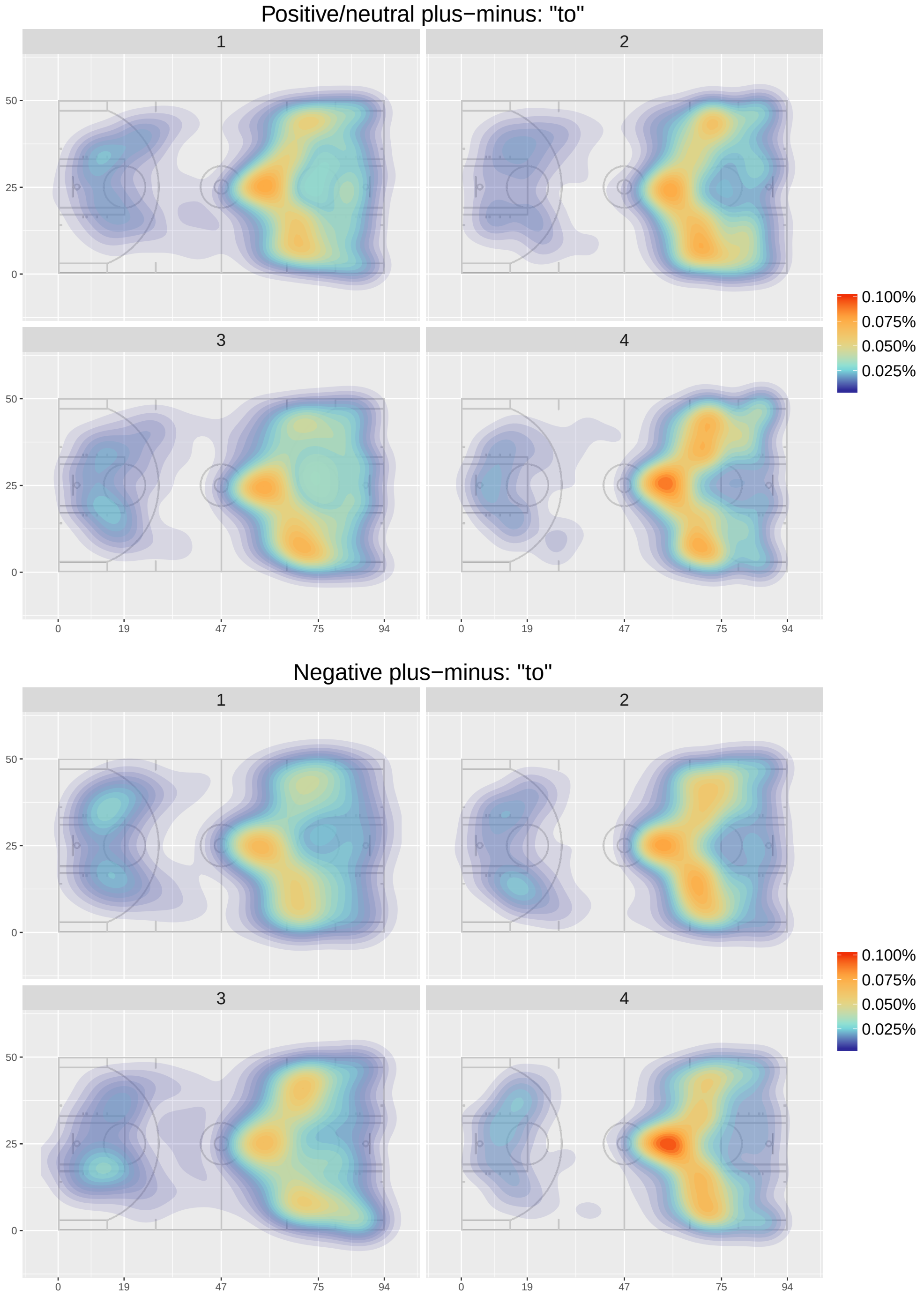

Figure 2.5: Heatmap displaying the kernel estimated density of the positions of players receiving the pass for quarters with at least +0 plus–minus (TOP) and a negative plus–minus (BOTTOM). These plots refer to all the 26 analyzed games.

usually passes happening in the same area do not have an impact on the play. At first, we will considered this connection as undirected, while in Section 4.2 networks will feature the information of pass direction. As for the time units, since dividing in single plays would result in extremely sparse networks, and an entire match would collapse a lot of different nuances in passing dynamics, games are divided into *quarters* (or equivalently *periods*). This way, our final dataset comprises 4 networks for each one of the 26 games, resulting in a total number of 104. To avoid complications, overtime periods are omitted in the creation of the final dataset, since their duration is of a reduced time of 5 minutes compared to a normal period lasting 12 minutes. In the next chapters, we will take a look at these networks with descriptive measures and consider different statistical models to answer our research question.

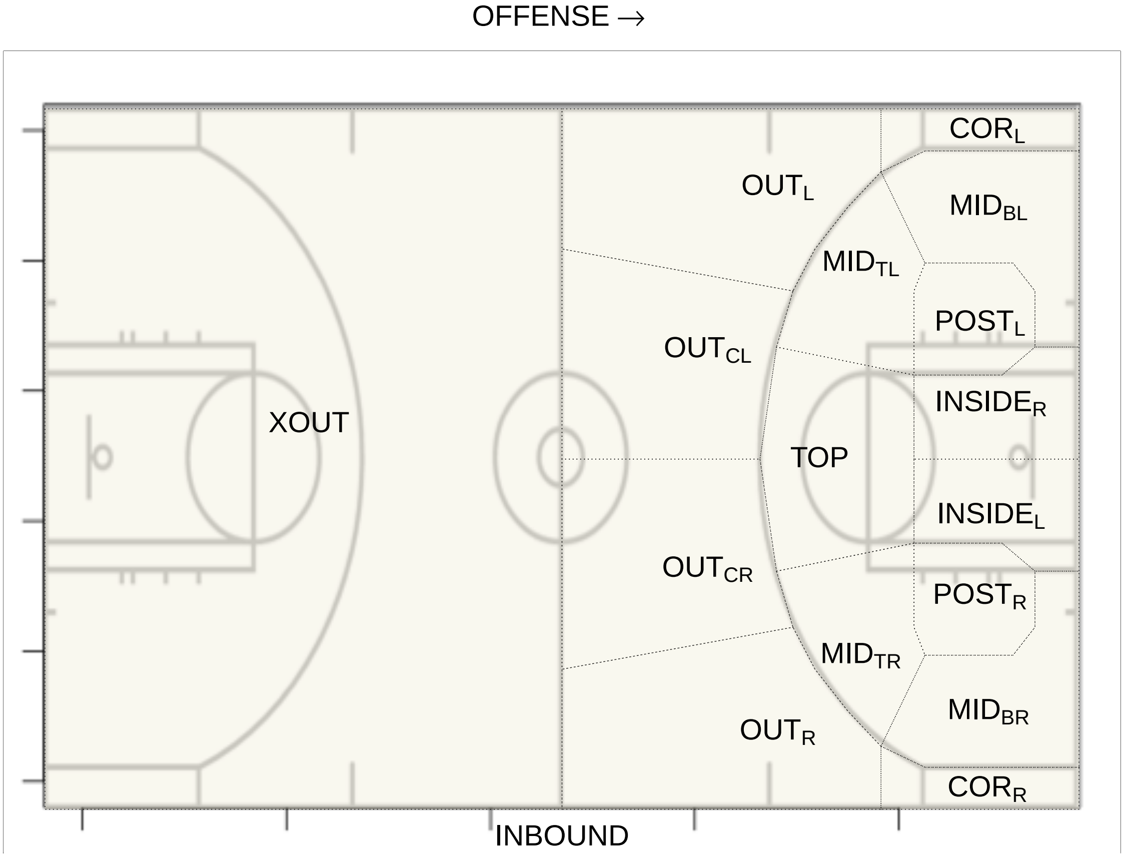

Figure 2.6: The sections of the court used to create the zones' network.

Chapter 3

Passing networks: a first look and classical models

The focus of this chapter is on providing a first study of the passing networks described in Chapter 2. This is accomplished via descriptive analyses and inference under classical statistical models for networks. Commonly used descriptive statistics are computed for each network in the two groups, positive/neutral plus–minus quarters and negative ones. At this stage, the networks are considered binary and undirected, meaning that a tie is formed when in that particular quarter at least one pass was made between two different areas of the court, no matter what the direction. This means that the edges' value is either 0 (no pass between two areas) or 1 (one or more passes).

A first insight is given by the distribution of passes that happened in each quarter, computed for the two groups of networks; the resulting plot is presented in Figure 3.1[1]. Apart from the third quarter differences in the passes distributions are more evident, no sensible group variations are displayed in general. In both cases the fourth quarter is also the one characterized by highest variability with an almost flat distribution from 50 to over 90. Table 3.1 presents the number of networks observed for the two groups per quarter. Out this simple contingency table, we are able to see that usually in the first quarter Golden State prevails on the opponent, while the fourth quarter is generally more balanced. This last insight should not surprise since Oakland's team has won a great number of games by a sensible margin that was already established at the end of the third quarter. This situation allows the coach to take superstars out and let them rest, while sending in bench players to get some minutes that they probably wouldn't have had in a close game. This

[1] An important note: when checking the distribution of passes every pass is being counted, even if repeated between the same pair of nodes; from Section 3.1 onwards a binary structure is considered.

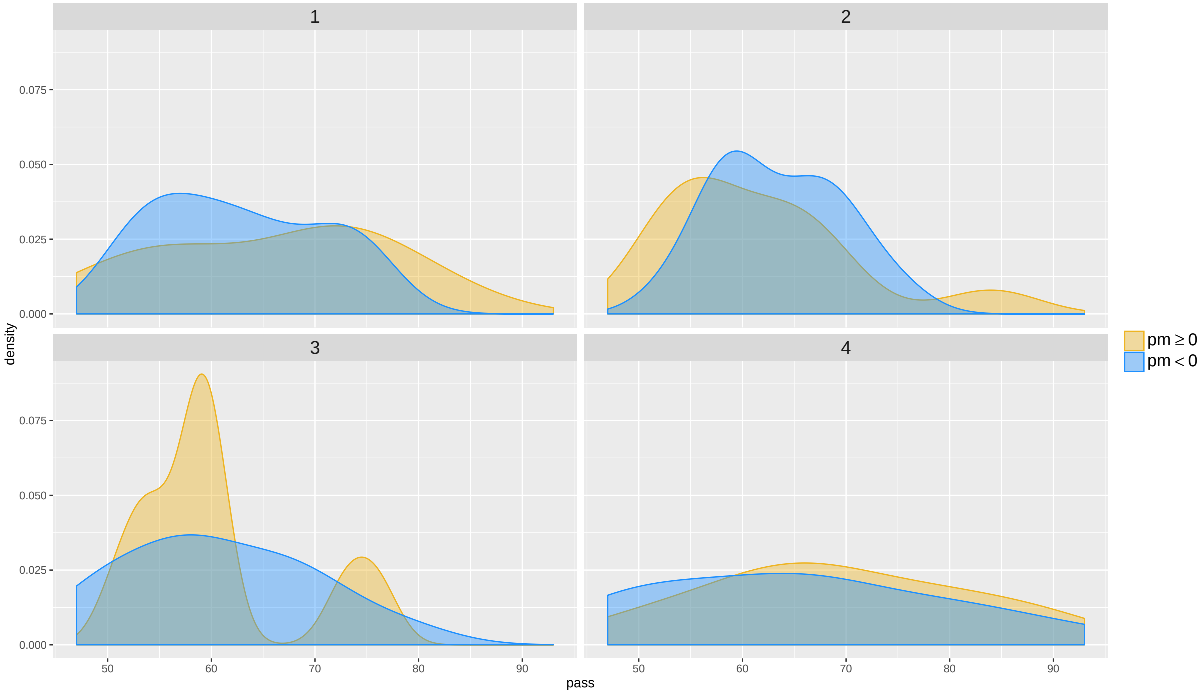

Figure 3.1: Distribution of the number of passes performed per quarter; color indicates positive/neutral plus–minus networks (dark yellow) and negative (blue).

is often referred to as *garbage time* since it does not impact on the win/loss result.

Table 3.1: Positive/negative plus–minus partitioning of networks for each quarter.

	q_1	q_2	q_3	q_4	*total*
Negative	7	11	10	12	40
Positive/Neutral	19	15	16	14	64

Before moving to statistical modeling and inference, we study the undirected binary networks from a descriptive perspective through the use of some of the many statistics available to characterize the properties of this nonstandard type of data.

3.1 Networks' descriptive analyses

In order to provide a first assessment of potential differences in the passing networks across won and lost periods, the initial focus is put on networks' descriptive statistics. These measures are computed for each network and their empirical distribution is shown separately for the won and lost quarters to highlight potential group differences. The type of descriptive statistics considered a described below. These measures are divided into *global*, i.e. considering the networks as a whole, and *local* measures, that consider the features of the single nodes.

3.1.1 Global measures

Global measures are quantities that consider each network as a whole and therefore result in one single index per network.

- *Density*: the relative frequency of the total number of ties in the network. Its range is $(0, 1)$, with 0 corresponding to no ties and 1 to "every node is connected with every other node".
- *Transitivity*: the percentage of observed closed triangles out of all the possible triangles (both open and closed). A triangle or triplet consists of three connected nodes. This quantity is also known as *clustering coefficient.*
- *Average path length*: the average of all the shortest path lengths, where the shortest path is a local measure at the edge level that indicates the minimum number of observed ties that need to be traversed to connect node u to node v.
- *Diameter and radius*: respectively the maximum and the minimum distance between all the nodes. Distance is defined as *geodesic distance*, the length of the shortest path between two nodes u and v. These two measures can also be defined as maximum and minimum *eccentricity.*
- *Degree variance*: the variance of the local measure *degree* defined below. Gives a rough idea on how differently nodes interact in terms of number of ties.

3.1.2 Local measures

Local measures focus on single nodes and are sometimes averaged to be transformed into the global scale.

- *Degree of node v*: the sum of nodes $u \neq v$ that are connected to node v, $v = 1, \ldots, V$. For undirected network, it can be interpreted as an "activity" measure.

- *Betweenness centrality of node v*: defined in Freeman (1977) by the following expression:

  $$C_b(v) = \sum_{u \neq v \neq w} \frac{n_{uw}(v)}{n_{uw}}$$

  where n_{uw} is the total amount of shortest paths between nodes u and w, and $n_{uw}(v)$ is the number of shortest paths between u and w that pass through v. In words, it corresponds to the importance of a node in efficiently connecting other nodes that are not directly tied together.

- *Closeness centrality of node v*: defined in Sabidussi (1966) by the following expression:

  $$C_c(v) = \frac{1}{\sum_{w \neq u} d(u, w)}$$

  where $d(\cdot, \cdot)$ is the aforementioned *geodesic distance*. It roughly corresponds to how close a node is to the others in terms of path.

Figure 3.2 shows no clear differences in the distribution of the selected descriptive statistics with the only exception of diameter. These plots are not surprising, since we are considering two groups that are not clearly distinct. In general we can observe these traits:

- the density distribution is concentrated around the values 0.20 and 0.25, stating that roughly only $20 - 25\%$ of all the possible 136 ties are observed in each network.

- given the spatial structure of the networks, a relatively high level of transitivity is observed on average compared to the relatively low level of density. This makes sense

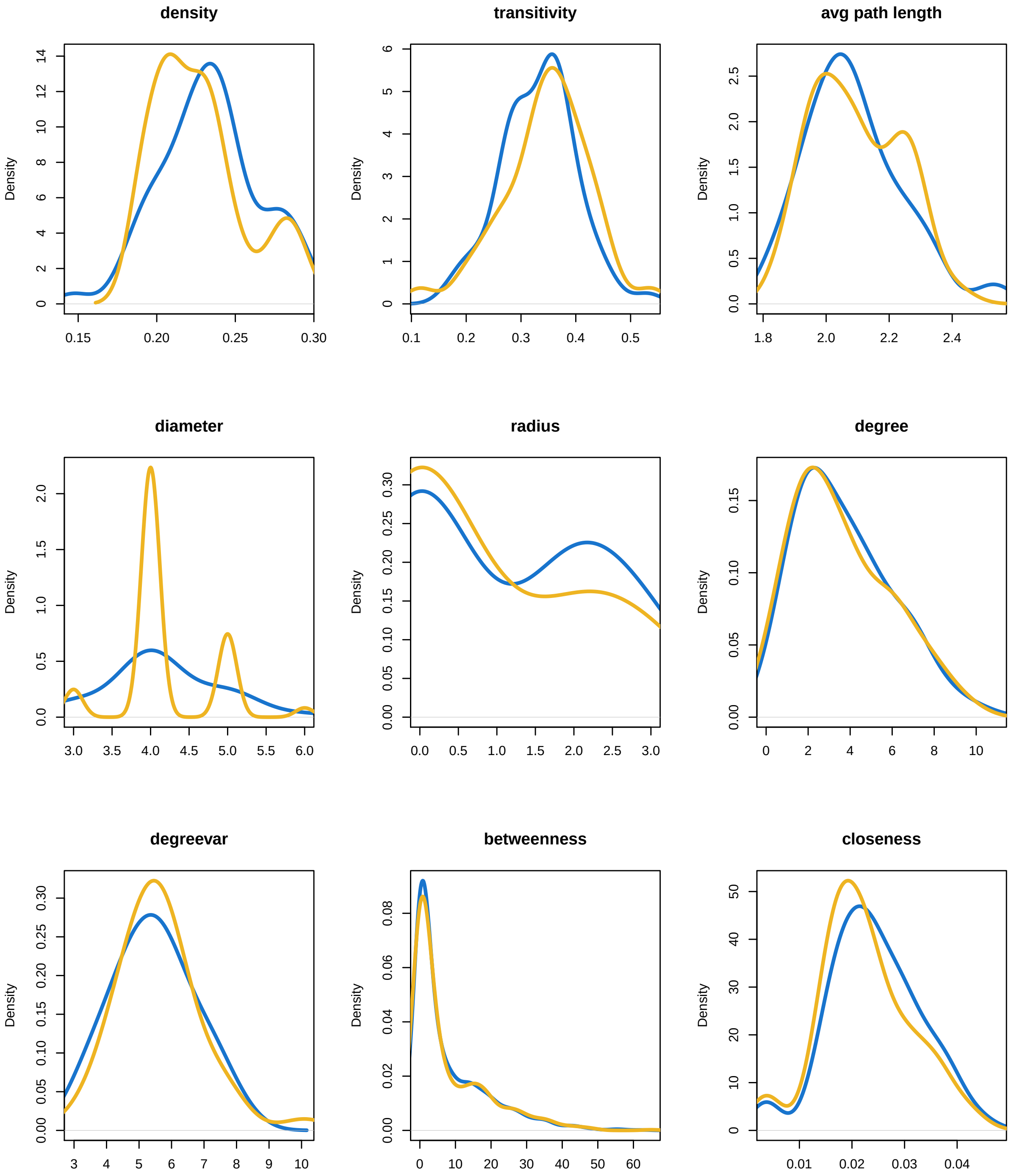

Figure 3.2: The selected descriptive statistics for the two groups. Positive plus–minus quarters are identified by the dark yellow line while negative quarters by the blue line.

when considering the fact that most passes happen between adjacent areas, and a third area is generally close to both.

- average path length's distribution is set at around 1.9 and 2.4, meaning that on average it takes around 2 passes to get from any zone to any other one. For positive quarters this seems to be a little bit higher, suggesting that sometimes extrapasses are more effective to reach more distant areas.
- diameter shows the biggest difference although it might not be as significant as it looks; for positive quarters the longest path is mostly 4, while for positive ones there is more variability. However, the range is between 3 and 6 for both groups.
- degree's most frequent value is set between 2 and 3, with a variability mostly between 5 and 6.
- betweenness plots presents a very skewed distribution, which is coherent with the heatmaps shown in Chapter 2, where a sensible concentration of passes was shown to be going through the central outer areas while many others only had a few.
- similarly to the other plots, closeness is no exception as far as differences are concerned.

3.2 Exponential Random Graph Models

A first simple approach to network analysis is represented by Exponential Random Graph Models (ERGM, Erdös and Rényi (1959), Holland and Leinhardt (1981) and more recently Snijders et al. (2006) and Robins et al. (2007b)). This class of models characterizes the probability of a given network as function of its summary measures, under an exponential family representation.

Many different models fall under the wide class of ERGMs. Among these stand Markov graphs (Frank and Strauss, 1986), based on the Markov assumption that in terms of networks translates into the following statement: two or more edges are considered independent if they do not share any node, conditioned to the rest of the network. Wasserman and Pattison (1996) generalized this concept with p^* (p-star) models, whose specification is shown in (3.1).

$$Pr(\mathcal{A} = A; \theta) = \exp\{\theta^T g(A) - k(\theta)\} \tag{3.1}$$

This equation indicates the general characterization for the probability distribution of these models, where $\mathcal{A}$ is the network random variable of which the network A is a realization, with total number of nodes V and edges denominated A_{uv} (in the undirected case $|A| = V \cdot (V-1)/2$); θ is a set of p parameters; $g(A)$ is a vector of arbitrary statistics; $k(\theta)$ is a normalizing constant.

In order to estimate the parameters we would have to know $k(\theta)$. Since generally this value is hard to compute, other methods are used to get an approximation the likelihood. Among them are *Markov Chain Monte Carlo* Bayesian methods, *simulated maximum likelihood* and lastly *pseudo-likelihood*, who has several different specifications. We can maximize the *pseudo-likelihood* function (3.2) to obtain $\hat{\theta}$.

$$pseudo - L(\theta) = \prod_{u<v} P(\mathcal{A}_{uv} = A_{uv} | \mathcal{A}_{-uv} = A_{-uv}; \theta) \tag{3.2}$$

Since each element of this product is a Bernoulli variable, and its conditional probability can be reformulated as a logistic regression problem, 3.2 is equivalent to fitting a simple GLM with response $A = \{A_{21}, \dots, A_{V1}, \dots, A_{uv}, \dots, A_{V,V-1}\}$ and matrix of covariates $\Delta = \{g(1, A_{(-uv)}) - g(0, A_{(-uv)})\}_{u>v}$. The obtained estimate holds asymptotic consistency when $V \to \infty$, even though the standard errors are only an approximation. As the name suggests, the *pseudo-likelihood* function is not exactly a likelihood function. However, it holds similar properties such as consistency, asymptotic distributions etc., so we use it as we would do with usual likelihoods, hopefully getting the same results.

Since ERGM models can only deal with single networks, a first glimpse at the difference between quarters in which the Golden State Warriors had a positive/neutral and negative plus–minus is provided by the comparison of the most extreme results. These are represented by the third quarter of the November, 2nd game vs. Memphis (Golden State ended up winning by an astonishing 50 points differential) which had a positive +25 differential, and the fourth quarter of the December 8th game at Indiana where the plus–minus was −20, even though this game still resulted in a win for Oakland's team on the road.

The analysis carried out here is provided by the some interesting aspects of the networks' relationships features such as: density, node type differences in terms of side of the court (left, right, central, or in-out of the three point line), homophily, reciprocity and particular structures such as alternating k-stars or triangles. See Robins et al. (2007a) for a detailed review on such effects for p^* models. The selected approach is *forward stepwise*, starting from the simple standard density effect up until court area homophily, other node attributes

Table 3.2: Coefficients for the maximum positive margin (+25) network in the chosen ERGM.

parameter	Estimate	Std. Error	p-value
edges	-2.4973	1.0758	0.0218 *
kstar2	5.4482	1.8899	0.0046 **
kstar3	-1.4537	0.5223	0.0062 **
altkstar	-6.5025	2.3862	0.0073 **
nodefactor.in3.out	1.0131	0.4583	0.0288 *
triangle	-0.3311	0.4612	*0.4741*

Table 3.3: Coefficients for the maximum negative margin (−20) network in the chosen ERGM.

parameter	Estimate	Std. Error	p-value
edges	-1.5592	1.5322	*0.311*
kstar2	1.0133	0.62080	*0.105*
kstar3	-0.17025	0.10425	*0.105*
altkstar	-1.76391	1.08801	*0.107*
nodefactor.in3.out	-0.01237	0.41462	*0.976*
triangle	0.65214	0.32519	0.047 *

and more complicated networks structures' effects whose addition entails an easier and more accurate estimation, and therefore interpretation of the other simpler parameters (e.g. inside/outside three-point line).

Tables 3.2 and 3.3 present the results for the two networks regarding effects for: density (**edges**), which was the only common significant effect, classical k-star structures, i.e. number of ties from the same node (**kstar**), alternating k-star, that consider all k-stars in one take but with a decay factor $\lambda = 2$ discounting the effect as k grows (**altkstar**), inside/outside the three-point line node attribute (**nodefactor.in3.out**) and a triangles' effect; these were estimated for both models in order to better compare them and correspond in order to the quantities displayed in (3.3).

$$Pr(\mathcal{A} = A; \theta) = \exp\left\{ \sum_{k=1}^{3} \theta_k S_k(A) + \theta_4 \sum_{k=2}^{V-1} (-1)^k \frac{S_k(A)}{\lambda^{k-2}} + \theta_5 x_{in3} + \theta_6 T(A) - k(\theta) \right\} \tag{3.3}$$

.

These results are showing quite a clear difference between the networks, both in the distinction between inside and outside the three-point line passing and in the tested network structures. Interpretability of the parameters is given in terms of conditional *odds ratio* similarly to Generalized Linear Models; the **edges** term corresponds to a GLM's intercept and consequently acts as a reference point for further terms (and is equivalent to a **kstar1** parameter). Ergo, considering the **triangle** coefficient, relative to the amount of triangular structures in the network, implies a positive effect of around $+20\%$ chance of resulting in a tie if two nodes have in common one or more connected areas of the court they are tied to in the $+25$ network, while it's non significant for the -20 one. Moreover, for the **in3** coefficient, the probability of having a tie happening outside the arc is 14% higher than a pass inside for the first "better" network, while reduced to non significant difference for the second "negative" one. The remarkably positive value that refers to the 2-star composition together with the fact that the triangle effect is not significant for the positive margin network, implies that two areas that share a common node are not more likely to be communicating with each other. Lastly, regarding the **altkstar** parameter, a negative value is observed: this means that, given that the weights for consecutive k-stars decreases when k increases because of the λ set to the value 2, networks with high degree nodes are improbable, so that nodes tend not to be hubs, with a smaller variance between the degrees (Robins et al., 2007a).

What does this mean in terms of passing dynamics? For example, even though this is an ambitious interpretation, it might suggests that the highly negative margin quarter case lightly presents relevant triangular structures between different areas of the court, even though nothing can be said about which zones this holds for, while for positive margin quarters passes interacting with the outside the three point line area are less likely to happen. An important note is to be made about the density of the two analyzed cases: the "negative" one seems to have a higher density implying a higher number of passes. This could possibly imply that in big wins, passing is more efficient. Other covariates effects such as left-right side, or other typical networks' structures and characteristics ended up not being significant in either cases.

Although capable of giving powerful insights for small networks, especially when characterized by categories and multiple node covariates, ERGMs fail in this case under many aspects: only one quarter can be considered at a time unless multiple networks are collapsed into one, leading to a loss of information and infeasible interpretation of the results in view of the purpose of the thesis. Therefore nothing but a coefficients' comparison is possible to examine the win/loss contrast; the overwhelming number of different effects that can be included makes the choice of the model a muddled process that can wind up in a continuous trial-and-error game. Most importantly, it does not take advantage of the fact that several observations of passing networks for wins and losses are available, at least not in a convenient way.

3.3 Latent space models

One possibility to account for multiple network observations is to consider latent space models (Hoff et al., 2002). This widely used and studied class of models (Handcock et al. (2007), Hoff (2003) and Krivitsky et al. (2009) to cite a few) relies on the idea that each node $v \in \mathcal{N}$ can be represented as a point z_v in a low-dimensional latent space $Z \in \Re^k$, with k adequately small. The probability of a tie between two nodes is higher the closer these two points are in the Z space, given the covariates. A popular choice for the distance measure is the Euclidean one, although different measures are possible. In latent space models for networks, each potential tie has a value modeled by a GLM, with a distribution whose density is f. This density is parameterized by its expected value, which is a function of the linear predictor $\eta_{u,v}$, as shown in (3.4). Estimating the quantities of interest is conveniently achieved using a Bayesian approach, choosing non–informative diffuse priors and MCMC methods to sample from the posterior distribution.

$$\begin{aligned} Pr(\mathcal{A} = A|\beta, x, Z) &= \prod_{(u,v)} Pr(\mathcal{A}_{u,v} = A_{u,v}|\beta, x, Z) \\ Pr(\mathcal{A}_{u,v,} = A_{u,v}|\beta, x_{\cdot,u,v}, |Z_u - Z_v|) &= f(A_{u,v}|\mu) = \binom{t}{A_{u,v}} \mu^{A_{u,v}} (1-\mu)^{t-A_{u,v}} \\ \mu &= g^{-1}\big(\eta_{u,v}(\beta, x_{\cdot,u,v}, |Z_u - Z_v|)\big) \\ \eta_{u,v}(\beta, x_{\cdot,u,v}, |Z_u - Z_v|) &= \sum_{k=1}^{p} x_{\cdot,u,v}\beta_k - |Z_u - Z_v| \end{aligned} \tag{3.4}$$

According to the purpose stated at the beginning of this section, two different sets were created through the sum of all the positive plus–minus and all the negative plus–minus networks, respectively represented by 64 and 40 single quarters. In order to model win–loss group differences using the above formulation, the networks associated to wins are modeled

Table 3.4: Summary for parameters' posteriors, positive margin networks' Latent Space model.

$parameter$	Estimate	2.5%	97.5%	$Pr(outsideCI)$
(Intercept)	-0.11084	-0.18461	-0.0450	$<$ 2.2e-16 ***
nodefactor.in3.out	1.48620	1.41408	1.5666	$<$ 2.2e-16 ***

separately from those relative to losses, using a different latent space model for each group. Within each group, the multiple observed networks are assumed to be independent and identically distributed from the corresponding latent space model. As a result, leveraging the conditional independence of the ties in (3.4), inference for each of the two latent space models can be accomplished under a binomial specification for $f(\cdot)$, letting y_{uv} be the sum of the ties from u to v observed for the networks associated with the group under analysis. This also means that, probably unrealistically, we assume that all positive/neutral and all negative plus–minus periods respectively come from the same f_1 and f_2 distribution.

The performance of different models was tested, with varying latent space dimension $d \in \{1, 2, 3\}$ using the classical latent space model proposed in Hoff et al. (2002); the inclusion of the in/out the three point line covariate (whose parameter is β_1) was also tested, considerably improving in terms of BIC performance reducing it by about 200 (from 1070 to 880) and 700 (from 1760 to 1010) for the "wins" and "losses" models respectively.

Models were estimated via the `latentnet` package (Krivitsky and Handcock, 2009) run through the software R (R Core Team, 2014), via MCMC. Posterior inference under each model, relies on two chains of 20000 MCMC samples after a burn–in of 15000. In order to improve mixing a thinning of 15 was additionally considered. These settings granted satisfactory convergence for each of the parameters, with a good mixing and negligible autocorrelation. Note that the coordinates of the latent space Z are invariant under rotations and reflections. Following this property, the plots have been adapted so that they could be compared side by side with the same reference points.

The first thing that leaps out from Figure 3.3 is the ability of the latent space Z to resemble the spatial disposition of the areas presented in Chapter 2 (Figure 2.6). The left/right disposition is arguably clear, in both representations. Being the closeness of the points in the graph, and hence the Z coordinates, directly proportional to the estimated probability of two nodes being tied together, a first remark is that in both cases, the further

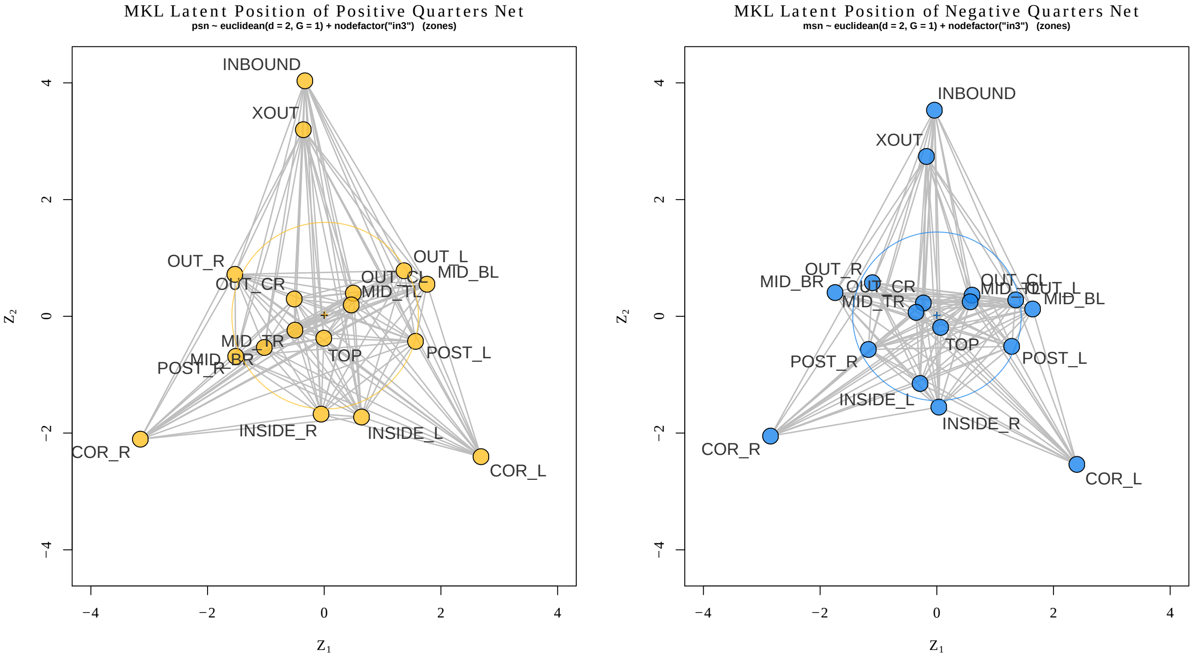

Figure 3.3: Minimum Kullback-Leibler Latent Positions of the two separate models built on positive and negative plus/minus networks with the node covariate effect "inside/outside three-point line".

Table 3.5: Summary for parameters' posteriors, negative margin networks' Latent Space model.

$parameter$	Estimate	2.5%	97.5%	$Pr(outsideCI)$
(Intercept)	-0.21657	-0.27459	-0.1539	$<$ 2.2e-16 ***
nodefactor.in3.out	1.43327	1.28664	1.5727	$<$ 2.2e-16 ***

the areas are in the court, the lower the probability of a pass being made between these zones, as common sense would suggest. However, this rule has some slight exceptions. For example in the "positive" network the two mid-bottom areas(MID_{BL} and MID_{BR}) seem to have a different role between left and right: the first one is closely tight to the OUT_L section, implying an important role in the connection of these two, while this doesn't hold for the right side. On the other hand the "negative" network does not present this behaviour. This remark is confirmed by the "yellow cross" displayed in Figure 3.4, that shows the distances' differences in the latent spaces between the two models. First the distance matrix is built from the latent space coordinates of the "positive" and "negative" networks, separately; then the difference between these two matrices is computed (*losses distances – wins distances*). If a tile is colored in blue it means that there is a smaller distance between that particular pair of nodes for the losses model compared to the win model, while yellow denotes the opposite. Therefore the general "blue-ish" color of the plot states that areas are usually closer (more likely to be tied) in the negative model than in the positive one; this also points out a higher density. However, unsurprisingly, the differences in the two models are quite small, as it is also stated by the coefficients shown in Tables 3.4 and 3.5.

After analyzing single networks with ERGMs and groups of networks with Latent Space models, we are still uncertain about whether or not there is a difference between quarters when the Golden State Warriors prevailed and the ones where they lost. Motivated by this need, we move on to the next chapter to consider a joint model for testing this difference.

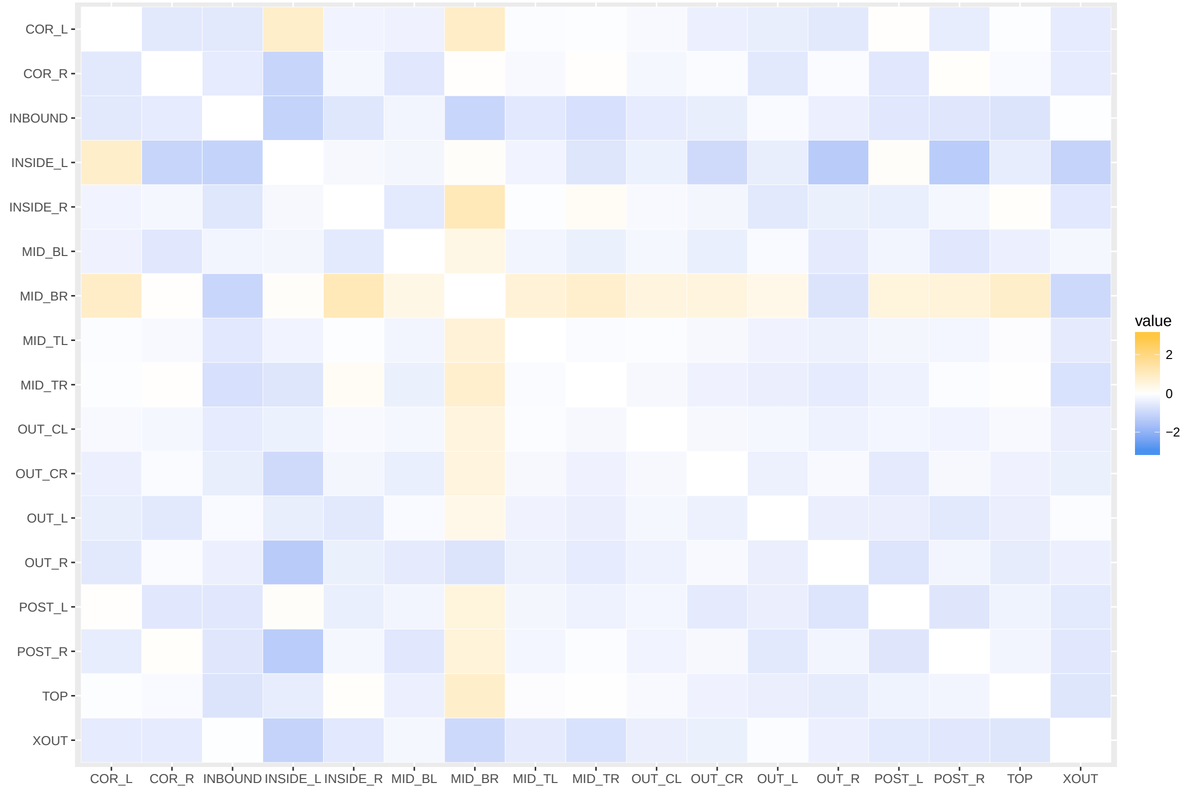

Figure 3.4: Heatmap displaying the differences between the euclidean distances among the nodes arising from the latent spaces in the two groups.

Chapter 4

A joint Bayesian nonparametric model

In Chapter 3 we firstly presented simpler models that took into consideration only single networks at a time, providing insufficient validity and inefficiency for testing group differences (ERGMs). Secondly, latent space models allowed to account for multiple observations of passing networks associated with lost and won quarters. However, the assumption of a unique latent space model underlying the multiple networks associated with each group, may be unrealistic, collapsing network variability around an averaged structure. Therefore we are looking for a model that is more flexible in characterizing the joint distribution of the random variable generating the multiple passing networks and allows for formal testing of differences between lost and won quarters.

4.1 Undirected networks model

4.1.1 The general idea

A recent development that fulfills these characteristics is represented by the Bayesian nonparametric model proposed in Durante et al. (2016). In their work, the authors wanted to provide a valid tool for modeling replicated binary undirected network data via the use of a mixture of low–rank factorizations. Durante and Dunson (2016) generalized the previous model to include global and local testing to assess evidence of group differences, adjusting for multiplicity.

Maintaining the notation previously used in Chapter 3, we define/recall the following quantities:

- y_i is the membership variable that indicates to which group the i–th network belongs to, with 1 representing positive margin quarters and 2 the negative ones. It is the realization of the random variable $\mathcal{Y}$

- A_i indicates the adjacency matrix of i–th network, generated from random variable $\mathcal{A}$

- operator $\mathcal{L}$ extracts the lower triangle vector of matrix A_i so that:

  $$\mathcal{L}(A_i) = (A_{i[21]}, A_{i[31]}, \dots, A_{i[V1]}, A_{i[V2]}, \dots, A_{V(V-1)})^T$$

  with each edge $A_{i[uv]}$ taking values in $\{0, 1\}$, for $v = 2, \dots, V$, $u = 1, \dots, V - 1$ and $A_{i[uv]} = A_{i[vu]}$

- indicator l maps the pair of nodes $v = 2, \dots, V$, and $u = 1, \dots, V-1$ to $1, \dots, V(V-1)/2$

- $p_{\mathcal{Y},\mathcal{L}(A)}(y, \boldsymbol{a}) = pr(\mathcal{Y} = y, \mathcal{L}(A) = \boldsymbol{a})$ is the joint probability mass function for the random variable $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$

As the number of nodes grows, even for small values of V any parametric model seems unsuitable unless a comparable number of subjects, in this case quarters, is available; this scenario is fairly impossible. A nonparametric approach is chosen in order to maintain flexibility when defining the network–valued random variable density. However, a fully nonparametric model is not viable due to dimension of the sample space being $2^{V(V-1)/2}$ for random variable $\mathcal{L}(A)$. Hence, to decrease the dimensionality, a dependent mixture of low–rank factorizations is considered while latent space models are used as kernels. This specification also allows to exploit the fact that network configurations share a common underlying structure with respect to edge probabilities; this mixture thus envelops information for the whole network, efficiently borrowing information across networks and within each network (Durante et al., 2016).

In evaluating evidence of a global dependence between the group membership and the networks' related generating random variable $\mathcal{L}(A)$, we are formally testing the null hypothesis:

$$H_0 : p_{\mathcal{Y},\mathcal{L}(A)}(y, \boldsymbol{a}) = p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(\boldsymbol{a}) \tag{4.1}$$

for all $y \in \{1,2\}$ and $\boldsymbol{a} \in \{0,1\}^{V(V-1)/2}$, versus

$$H_1 : p_{\mathcal{Y},\mathcal{L}(A)}(y, \boldsymbol{a}) \neq p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(\boldsymbol{a}) \tag{4.2}$$

for at least some y and $\boldsymbol{a}$; $p_{\mathcal{Y}}(y)$ identifies the marginal probability mass function of the membership variable and $p_{\mathcal{L}(A)}(\boldsymbol{a})$ the unconditional pmf for the network random variable $\mathcal{L}(A)$. This way, the tests are not performed on networks' structural properties or summary statistics such as density, transitivity etc, but directly on their probability mass function. As far as our example of basketball networks is concerned, H_0 corresponds to no differences in passing dynamics between periods in which Golden State won/tied the score and when they lost; however, this only tests for global differences and does not accommodate for specific edges diversities. To provide this, after assessing a global dependence, areas connections denoted as $\mathcal{L}(A)_l \in \{0,1\}, l = 1, \ldots, V(V-1)/2$, are inspected via multiple local tests that lead to the null hypothesis

$$H_{0_l} : p_{\mathcal{Y},\mathcal{L}(A)}(y, a_l) = p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(a_l) \tag{4.3}$$

for all $y \in \{1,2\}$ and $a_l \in \{0,1\}$, versus

$$H_{1_l} : p_{\mathcal{Y},\mathcal{L}(A)}(y, a_l) \neq p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(a_l) \tag{4.4}$$

for at least some y and a_l.

These tests are made possible by a flexible specification that is able to maintain the important traits of the networks while reducing dimensionality and simplifying the derivation of the probabilities needed for (4.1) – (4.2) and (4.3) – (4.4).

4.1.2 Model specification

To perform the aforementioned test, a convenient expression for $p_{\mathcal{Y},\mathcal{L}(A)}$ is needed. It is derived from the following factorization:

$$p_{\mathcal{Y},\mathcal{L}(A)}(y, \boldsymbol{a}) = p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)|y}(\boldsymbol{a}) = pr(\mathcal{Y} = y) pr(\mathcal{L}(\mathcal{A}) = \boldsymbol{a} | \mathcal{Y} = y) \tag{4.5}$$

as it is always possible to derive the joint pmf as product of the marginal for the grouping variable $\mathcal{Y}$ and the conditional pmfs $p_{\mathcal{L}(A)|y}$. This way, hypotheses (4.1) – (4.2) can be reformulated as

$$H_0 : p_{\mathcal{L}(A)|1}(\boldsymbol{a}) = p_{\mathcal{L}(A)|2}(\boldsymbol{a}) \tag{4.6}$$

for all network configurations $\boldsymbol{a}$, versus

$$H_1 : p_{\mathcal{L}(A)|1}(\boldsymbol{a}) \neq p_{\mathcal{L}(A)|2}(\boldsymbol{a}) \tag{4.7}$$

for some $\boldsymbol{a}$.

To provide a flexible representation of the conditional pmf for the networks given the group, while reducing dimensionality and allowing simple testing, Durante and Dunson (2016) define

$$p_{\mathcal{L}(A)|y}(\boldsymbol{a}) = pr(\mathcal{L}(\mathcal{A}) = \boldsymbol{a}|\mathcal{Y} = y) = \sum_{h=1}^{H} \nu_{hy} \prod_{l=1}^{V(V-1)/2} (\pi_l^{(h)})^{a_l}(1 - \pi_l^{(h)})^{1-a_l} \tag{4.8}$$

where ν_{hy} denotes the group specific mixture probabilities for the h–th component, with $\sum_{h=1}^{H} \nu_{hy} = 1$, $\nu_{hy} \in (0, 1)$ for all $y \in \{1, 2\}$ and $h \in 1, \ldots, H$; H is the total number of mixture components and $\boldsymbol{\pi}^{(h)} = (\pi_1^{(h)}, \ldots, \pi_{V(V-1)/2}^{(h)})$ is the edge probability vector relative to the h–th component. Its formal definition is

$$\boldsymbol{\pi}^{(h)} = \{1 + \exp(-\mathbf{Z} - \mathbf{D}^{(h)})\}^{-1}, \quad \mathbf{D}^{(h)} = \mathcal{L}(\mathbf{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\mathbf{X}^{(h)T}) \tag{4.9}$$

where $\boldsymbol{Z} \in \Re^{V(V-1)/2}$ is a vector that indicates a shared similarity effect that conveys easier centering of different mixture components and improve computational performance (Durante et al., 2016); $X^{(h)} \in \Re^{V \times R}$ is a matrix whose rows are node–specific latent coordinate vectors, weighted for $\Lambda^{(h)}$, a diagonal matrix with R non–negative elements $\lambda_r^{(h)}$. Typically, $R \ll V$. After (4.9) it is simple to note that the probability of an edge l for the pair of nodes u and v in the h–th component increases with Z_l and $\mathcal{L}(\mathbf{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\mathbf{X}^{(h)T})_l = \sum_{r=1}^{R} \lambda_r^{(h)} X_{vr}^{(h)} X_{ur}^{(h)}$. This characterization is an adaptation of existing concepts in literature with regards to latent variable modeling for single networks (Nowicki and Snijders (2001), Airoldi et al. (2008) and Hoff et al. (2002) as seen in Chapter 3).

The generating process for $\{y_i, \mathcal{L}(\mathbf{A}_i)\}$ is outlined in the following steps:

1. the grouping variable y_i is sampled from $p_{\mathcal{Y}}$
2. given $y_i = y$, the latent indicator $G_i \in \{1, \ldots, H\}$ is obtained

3. following (4.9) the edges denoted as $\mathcal{L}(\mathbf{A}_i)_l$ for network $\mathcal{L}(\mathbf{A}_i)$ are sampled from conditionally independent Bernoulli variables given y_i, h and consequently $\boldsymbol{\pi}^{(h)}$.

This way, networks that are in the same mixture component h share the same probability vector, with the probability assigned to each component being specific to each group. In case of global group differences these mixing probabilities are different across groups, whereas these mixing probabilities are equal when no group differences are found.

4.1.3 Global and local testing procedures

It is then almost straightforward to uniquely define global testing. Formally, the null hypothesis will be

$$H_0 : (\nu_{11}, \dots, \nu_{H1}) = (\nu_{12}, \dots, \nu_{H2}) \qquad \text{versus} \qquad H_1 : (\nu_{11}, \dots, \nu_{H1}) \neq (\nu_{12}, \dots, \nu_{H2}) \tag{4.10}$$

This leads to a unique characterization of the global hypotheses displayed in (4.1) – (4.2).

To provide a local testing procedure, the authors make use of the model–based version of Cramer's V proposed in Dunson and Xing (2009), that measures the association between two variables similarly to Pearson's χ^2. This results into the quantity ρ to be computed for each pair of nodes l:

$$\rho_l^2 = \sum_{y=1}^{2} \sum_{a_l=0}^{1} \frac{\{p_{\mathcal{Y},\mathcal{L}(A)}(y, a_l) - p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(a_l)\}^2}{p_{\mathcal{Y}}(y) p_{\mathcal{L}(A)}(a_l)} = \sum_{y=1}^{2} p_{\mathcal{Y}}(y) \sum_{a_l=0}^{1} \frac{\{p_{\mathcal{L}(A)|y}(a_l) - p_{\mathcal{L}(A)}(a_l)\}^2}{p_{\mathcal{L}(A)}(a_l)} \tag{4.11}$$

Being $\rho_l \in (0, 1)$, the local association is absent when $\rho_l = 0$, denoting no difference across the groups in terms of edge l probabilities, and it is therefore stronger when closer to 1. Computation of ρ is available from posteriors' derivation of the quantities of interest in the following way:

- $p_{\mathcal{L}(A)|y}(1) = 1 - p_{\mathcal{L}(A)|y}(0) = \sum_{h=1}^{H} \nu_{hy} \pi_l^{(h)}$
- $p_{\mathcal{L}(A)}(1) = 1 - p_{\mathcal{L}(A)}(0) = \sum_{y=1}^{2} p_{\mathcal{Y}}(y) \sum_{h=1}^{H} \nu_{hy} \pi_l^{(h)}$

4.1.4 Priors specification and posterior derivation

Since working in a Bayesian setting, priors' distributions need to be set. The authors specify independent priors for $p_{\mathcal{Y}}$, $\mathbf{Z}$, $\mathbf{X}^{(h)}$, $\boldsymbol{\lambda}^{(h)}$, with $h = 1, \dots, H$; plus, the mixture components $\boldsymbol{\nu}_y = (\nu_{1y}, \ldots, \nu_{Hy})$, $y \in \{1,2\}$, in order to induce a prior Π on the joint pmf $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, \boldsymbol{a})$ that leads to easy posterior derivation, allows for testing and has profitable asymptotic behaviour. The selected priors therefore maintain the flexibility that characterizes the dependent mixture model. As for $p_{\mathcal{Y}}$ being the pmf of a categorical variable with two levels, let $p_{\mathcal{Y}}(2) = 1 - p_{\mathcal{Y}}(1) \sim \text{Beta}(a, b)$, with a and b properly chosen hyperparameters. Following Durante et al. (2016), Gaussian priors are chosen for $\mathbf{Z}$, as well as $\mathbf{X}^{(h)}$, and multiplicative inverse gammas for $\boldsymbol{\lambda}^{(h)} \sim \text{MIG}(a_1, a_2)$ for each component with a_1, a_2 being hyperparameters. This choice provides a convenient adaptive shrinkage. Priors for mixtures' probabilities are induced as follows, with $\boldsymbol{v} = (v_1, \ldots, v_H)$ and $\boldsymbol{v}_y = (v_{1y}, \ldots, v_{Hy})$:

$$\nu_y = (1-T)\boldsymbol{v} + T\boldsymbol{v}_y, y \in \{1,2\} \tag{4.12a}$$

$$\boldsymbol{v} \sim \text{Dir}(1/H, \ldots, 1/H), \quad \boldsymbol{v}_y \sim \text{Dir}(1/H, \ldots, 1/H) \quad y \in \{1,2\} \tag{4.12b}$$

$$T \sim \text{Bern}\{\text{pr}(H_1)\} \tag{4.12c}$$

Here T denotes the test result being $T = 0$ in the case of H_0 and $T = 1$ for H_1; hence, under H_1 different mixture probabilities are independently generated, whereas they are equal across groups in the H_0 setting. As for the Dirichlet priors, small values are chosen in order to allow for automatic deletion of redundant components (Rousseau and Mengersen, 2011). To get $\text{pr}(H_1)$, the full conditional for $\text{pr}(T = 1|-) = \text{pr}(H_1|-) = 1 - \text{pr}(H_0|-)$ is retrieved:

$$\begin{aligned}\text{pr}(H_1|-) &= \frac{\text{pr}(H_1)\prod_{y=1}^{2}\int(\prod_{h=1}^{H} v_{hy}^{n_{hy}})d\Pi_{v_y}}{\text{pr}(H_0)\int(\prod_{h=1}^{H} v_h^{n_h})d\Pi_v + \text{pr}(H_1)\prod_{y=1}^{2}\int(\prod_{h=1}^{H} v_{hy}^{n_{hy}})d\Pi_{v_y}} \\ &= \frac{\text{pr}(H_1)\prod_{y=1}^{2}\{\text{B}(\boldsymbol{\alpha} + \bar{n}_y)/\text{B}(\boldsymbol{\alpha})\}}{\text{pr}(H_0)\text{B}(\boldsymbol{\alpha} + \bar{n})/\text{B}(\boldsymbol{\alpha}) + \text{pr}(H_1)\prod_{y=1}^{2}\{\text{B}(\boldsymbol{\alpha} + \bar{n}_y)/\text{B}(\boldsymbol{\alpha})\}}\end{aligned} \tag{4.13}$$

with $n_{h_y} = \sum_{i:y_i=y} I(G_i = h), n_h = \sum_{i=1}^{n} I(G_i = h), \bar{\boldsymbol{n}}_y = (n_{1y}, \ldots, n_1Hy), \bar{\boldsymbol{n}} = (n_1, \ldots, n_H)$, $\boldsymbol{\alpha} = (1/H, \ldots, 1/H)$ and $\text{B}(\cdot)$ being the multivariate Beta function. The second part of (4.13) is obtained exploiting the Dirichlet–multinomial conjugacy.

Despite being an excellent setting for global testing, (4.12) is impractical to characterize local null hypotheses $H_{0_l} : \rho_l = 0$ versus $H_{1_l} : \rho_l \neq 0$ for each $l \in \{1, \ldots, V(V-1)/2\}$; it is then necessary to reformulate the hypotheses as $H_{0_l} : \rho_l \leq \epsilon$ versus $H_{1_l} : \rho_l > \epsilon$ with ϵ

usually chosen to be around 0.1. This allows for simple estimation of $\hat{\text{pr}}\big(H_{1_l}|\{y, \mathcal{L}(\mathcal{A})\}\big)$ as the proportion of Gibbs samples for which $\rho_l > \epsilon$.

Lastly, to compute the posterior a Gibbs sampler is designed as follows:

1. $p_{\mathcal{Y}}(1)$ is sampled from the full conditional $p_{\mathcal{Y}}(1)|- \sim \text{Beta}(a + n_1, b + n_2)$, being $n_y = \sum_{i=1}^{n} I(y_i = y)$.

2. update the mixture grouping variable G_i for each $i = 1, \dots, n$ from the probabilities:

$$\text{pr}(G_i = h) = \frac{\nu_{hy_i} \prod_{l=1}^{V(V-1)/2} (\pi_l^{(h)})^{\mathcal{L}(A_i)_l} (1 - \pi_l^{(h)})^{1-\mathcal{L}(A_i)_l}}{\sum_{q=1}^{H} \nu_{qy_i} \prod_{l=1}^{V(V-1)/2} (\pi_l^{(q)})^{\mathcal{L}(A_i)_l} (1 - \pi_l^{(q)})^{1-\mathcal{L}(A_i)_l}}$$

   for $h = 1, \dots, H$ and each $\boldsymbol{\pi}^{(h)}$ characterized as in (4.9).

3. conditioning on G_i, $\boldsymbol{Z}, \boldsymbol{X}^{(h)}$ and $\boldsymbol{\lambda}^{(h)}$ for each h are updated through the use of Pólya–gamma data augmentation scheme for Bayesian logistic regression (see note below for details) developed in Polson et al. (2013), as detailed in Durante et al. (2016).

4. sample the testing indicator T from a Bernoulli distribution with parameter p equal to probability (4.13).

5. based on the result of T:

   - if $T = 0$, let $\nu_y = \boldsymbol{v}$ for both groups, with $\boldsymbol{v}$ updated from the full conditional Dirichlet $(v_1, \dots, v_H)|- \sim \text{Dir}(1/H + n_1, \dots, 1/H + n_H)$
   - else if $T = 1$, update each ν_y independently from $(v_{1y}, \dots, v_{Hy})|- \sim \text{Dir}(1/H + n_{1y}, \dots, 1/H + n_{Hy})$.

Since we do not know in advance how many components H or latent space dimensions R will be needed, these dimensions are set in the algorithm at conservative upper bounds allowing the shrinkage priors on these quantities to adapt to the dimensions required to characterize the observed data.

A note on Pólya–gamma data augmentation scheme

A key tool to derive the Gibbs sampler described above, is the Pólya–Gamma data augmentation. This method was developed by Polson et al. (2013) and provides a reliable way to perform posterior inference in the case of Bayesian logistic regression using a data augmentation step.

Given a binomial likelihood on y_i (like in our case), with a p–dimensional input X_i and a vector of weights β with a Gaussian prior as in (4.14)

$$\begin{aligned} \text{Likelihood} : y_i|X_i, \beta &\sim \text{Binom}\big(n_i, (1 + \exp(-x_i^T \beta)^{-1}\big) \\ \text{Prior} : \beta &\sim \mathcal{N}(b, B) \end{aligned} \tag{4.14}$$

we want to sample the posterior for β. This can be done via the use of a Pólya–gamma distributed latent variable, through these two steps:

$$\begin{aligned} \text{Pólya–gamma} : \omega_i|\beta &\sim \text{PG}(n_i, x_i^T \beta) \\ \text{Posterior} : \beta|y, \omega &\sim \mathcal{N}(\mu_\omega, \Sigma_\omega) \end{aligned} \tag{4.15}$$

where $\Sigma_\omega = (X^T \text{diag}(\omega) X + B^{-1})^{-1}$ and $\mu_\omega = \Sigma_\omega(y - \mathbf{1}_N n/2 - B^{-1} b)$, being $\mathbf{1}_N = 1_1, \ldots, 1_N$ a N–length vector of 1s.

This is made possible by the representations of logodds–parameterized binomial likelihoods in terms of mixtures of Gaussians with respect to a Pólya–gamma distribution $p(\omega)$. With $X \sim \text{PG}(b, c)$, $b > 0, c \in \Re$:

$$p(\omega) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

with $g_k \sim \Gamma(b, 1)$ being independent gamma random variables. This leads to the following integral identity:

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega$$

where $\kappa = a - b/2$, $\omega \sim \text{PG}(b, 0)$ with $b > 0$, and $\psi = x_i^T \beta$ is a linear function of predictors. Given these conditions, the integrand is the kernel of a Gaussian likelihood in β. Moreover, the implied conditional distribution for ω given ψ, also follows the Pólya–gamma distribution. This way, a Gibbs sampler is able to get these quantities with a Gaussian distribution that draws for the main parameters, and the Pólya–gamma draws for a single layer of latent variables (Polson et al., 2013).

4.1.5 Application to passing networks

In this section the model just presented is being fit on the Golden State Warriors passing networks and testing is performed to evaluate the difference between positive and negative margin quarters, i.e. periods with a plus–minus of at least $+0$ in the first case, and at least -1 in the second as we did in the previous chapter.

As mentioned, the case of positive/neutral and negative margin quarter networks comprises 104 units, with all four quarters for each game (overtimes are omitted). Also, the networks are considered as binary and undirected with no self–loops, as the model requires. To evaluate global dependence we run the Gibbs sampler as provided in Section 4 in Durante et al. (2016) for 10000 iterations and reject the first 2000 as burn–in to grant satisfactory convergence. Note that because of the range of ρ being $(0, 1)$, in case of zero evidence of local differences, $\rho = 0$; it may look like the posterior distribution is stuck and therefore implies high autocorrelation, but it actually is a stable indication that there is practically no evidence of diversity across the groups. Relative convergence plots are available in Appendix B as we assessed convergence via Potential Scale Reduction Factors (Gelman and Rubin, 1992) and mixing via traceplots and effective sample sizes. The parameters H and R are both set at 10 to allow for sufficient flexibility and dimensionality reduction in the model.

As a result, we accept the null hypothesis since the estimated posterior probability of the alternative is $\hat{\text{pr}}\big(H_1|\{y, \mathcal{L}(\mathcal{A})\}\big) = 0.5134$; although not exactly being close to 0, this value expresses that there is no strong evidence of dependence between the groups and the networks generating variable. Such a statement is confirmed by the analyses on the quartiles of ρ_l shown in Figure 4.1: the overall pure white color in all three frames implies that most of the whole distribution of each ρ_l lies below the suggested threshold 0.1 denoting no change in edge probability. Accordingly, Figure 4.2 does not display any stable pattern for the computation of the difference between the estimated edge probabilities in the two groups. These quantities correspond to the vectors $\bar{\boldsymbol{\pi}}_y$ whose elements are $\bar{\boldsymbol{\pi}}_{yl} = p_{\mathcal{L}(\mathcal{A})_l|y}(1) = \text{pr}\{\mathcal{L}(\mathcal{A})_l = 1|\mathcal{Y} = y\}$, for $y \in \{1, 2\}$, re–arranged in matrix form. In case of complete independence, these plots would show an overall pure white coloration as the difference is almost always 0; since we obtain very lightly colored representations, we confirm the initial conjecture of no dependence.

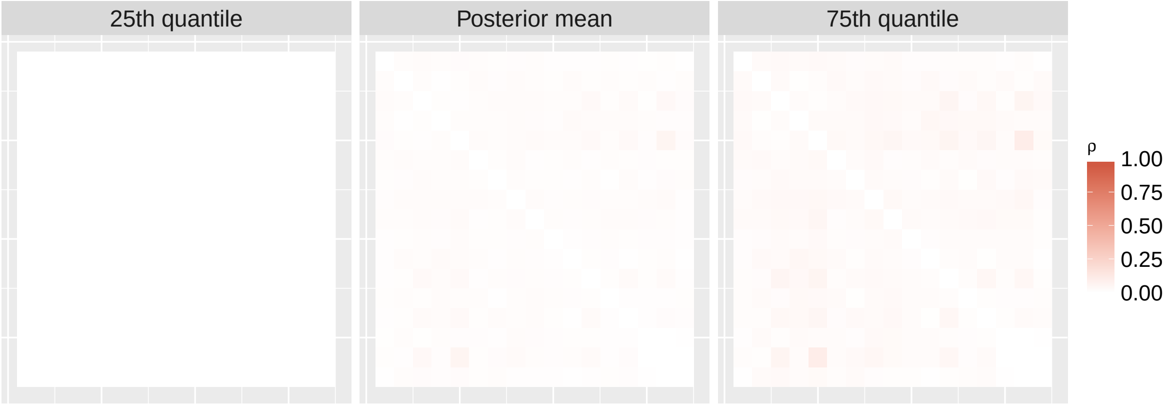

Figure 4.1: Mean and quartiles of the posterior distribution of ρ_l for the $+0/-1$ dataset. Here pairs of court areas identified by l are arranged in matrix form.

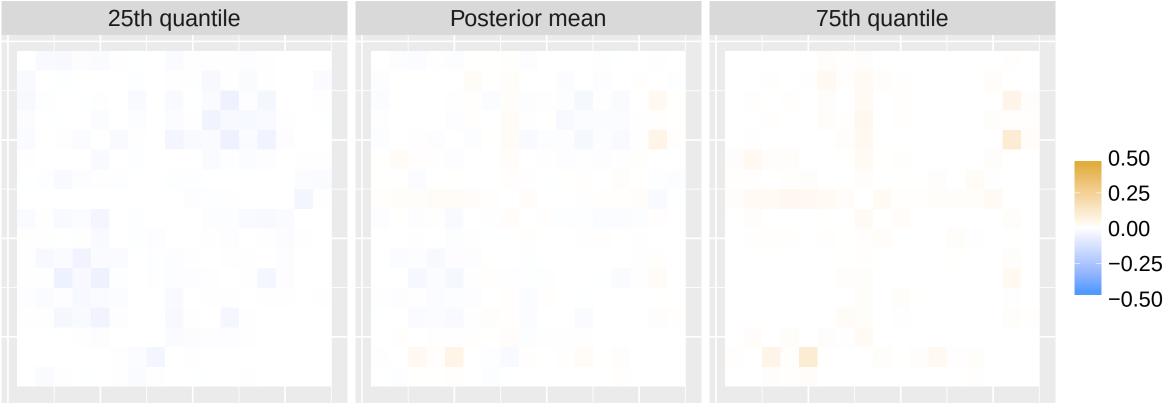

Figure 4.2: Mean and quartiles of the posterior distribution of $\bar{\pi}_2 - \bar{\pi}_1$ for the $+0/-1$ dataset. Here pairs of court areas identified by l are arranged in matrix form.

4.2 Generalizing the model for directed networks

Until now, all networks ties have been considered to be undirected, so that a connection is made if between two areas at least one pass occurred, no matter what the direction. This is of course a simplification of reality where we know exactly from which area of the court a certain player started the act of passing and where his teammate received the ball, implying a direction.

In order to do this, a new operator that vectorizes matrices into the joining of lower and upper triangles is defined as $\mathcal{V}(A) = (\mathcal{L}(A),\mathcal{U}(A))$, with $\mathcal{L}(\cdot)$ being the function selecting the lower triangle of a matrix and $\mathcal{U}(\cdot)$ the upper triangle. This corresponds to the $vec(\cdot)$ operator without diagonal elements. The length of this vector will be $V(V-1)$, being V the number of nodes in the network, so that for a general matrix A_i:

$$\mathcal{V}(A_i) = (A_{i[21]}, A_{i[31]}, \dots, A_{i[V1]}, A_{i[V2]}, \dots, A_{V(V-1)}, A_{i[12]}, A_{i[13]}, \dots, A_{i[1V]}, A_{i[2V]}, \dots, A_{(V-1)V})^T$$

with $A_{i[uv]}$ not necessarily equal to $A_{i[vu]}$.

By replacing $\mathcal{L}(A)$ with $\mathcal{V}(A)$ we can generalize the method displayed in the previous section, with some additional adaptations and observations noted below. The indicator l now maps each pair (u, v) for which $u \neq v$. The joint pmf is now denoted by $p_{\mathcal{Y},\mathcal{V}(A)}(y, \boldsymbol{a}) = \text{pr}(\mathcal{Y} = y, \mathcal{V}(A) = \boldsymbol{a})$. The data now lie in a much bigger space, since the network configurations $\boldsymbol{a} \in \mathbb{A}_V^{dir}$ have $2^{V(V-1)}$ possible representations instead of $2^{V(V-1)/2}$; this is an important note as the sample space increases considerably. In our passing networks example with just 17 nodes, this means going from $|\mathbb{A}_{17}^{undir}| = 8.7 \cdot 10^{40}$ to $|\mathbb{A}_{17}^{dir}| = 7.6 \cdot 10^{81}$. Although being both gigantic numbers, there is a sensible increased sparsity that the model has to deal with. Similarly to (4.8) and (4.9), the model is hence defined via the following equations:

$$\begin{gathered} p_{\mathcal{V}(A)|y}(\boldsymbol{a}) = \text{pr}\big(\mathcal{V}(\mathcal{A}) = \boldsymbol{a}|\mathcal{Y} = y\big) = \sum_{h=1}^{H} \nu_{hy} \prod_{l=1}^{V(V-1)} (\pi_l^{(h)})^{a_l}(1-\pi_l^{(h)})^{1-a_l} \\ \boldsymbol{\pi}^{(h)} = \{1+\exp(-\mathbf{Z}-\mathbf{W}^{(h)})\}^{-1}, \quad \mathbf{W}^{(h)} = \mathcal{V}(\mathbf{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\mathbf{Q}^{(h)T}) \end{gathered} \tag{4.16}$$

where $\mathbf{Q}^{(h)} \in \Re^{V\text{x}R}$ is a matrix that allows for the differentiation in direction to be included in the latent space coordinates for each node and $\boldsymbol{Z}$ has now length equal to $V(V-1)$. The single l–th element of vector $\mathbf{W}^{(h)}$, $\mathbf{W}_l^{(h)}$ is hence implied by:

$$\mathcal{V}(\mathbf{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\mathbf{Q}^{(h)T})_l = \sum_{r=1}^{R} Q_{vr}^{(h)}\lambda_r^{(h)}X_{ur}^{(h)}$$

With respect to priors, $\mathbf{Q}^{(h)}$ follows the same specification of $\boldsymbol{X}^{(h)}$ in Section 4.1.4, with multivariate Gaussian priors for each row $v \in \{1, \dots, V\}$, for each $h \in \{1, \dots, H\}$. However, the Gibbs sampler needs some adjustments. Before detailing the procedure, we define two matrices that will be needed in the sampler to maintain conjugacy, $\bar{\boldsymbol{X}}^{(h)} = \boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)1/2}$ and $\bar{\boldsymbol{Q}}^{(h)} = \boldsymbol{Q}^{(h)}\boldsymbol{\Lambda}^{(h)1/2}$, so that $\boldsymbol{W}^{(h)} = \mathcal{V}(\bar{\boldsymbol{X}}^{(h)}\bar{\boldsymbol{Q}}^{(h)T})$. It is hence delineated below, adapting from Section 4 in Durante et al. (2016) and 3.2 in Durante and Dunson (2016) to the directed networks case.

1. **Sample $p_{\mathcal{Y}}(1)$ from the full conditional:**
   $p_{\mathcal{Y}}(1)|- \sim \text{Beta}(a + n_1, b + n_2)$, being $n_y = \sum_{i=1}^{n} I(y_i = y)$

2. **Allocate vectorized networks $\mathcal{V}(A_i), i \in \{1, \ldots, n\}$ to one out of the H mixture components:**
   Sample the group indicator variable G_i:

   $$\text{pr}(G_i = h|-) = \frac{\nu_{hy_i}\prod_{l=1}^{V(V-1)}(\pi_l^{(h)})^{\mathcal{V}(A_i)_l}(1-\pi_l^{(h)})^{1-\mathcal{V}(A_i)_l}}{\sum_{q=1}^{H}\nu_{qy_i}\prod_{l=1}^{V(V-1)}(\pi_l^{(q)})^{\mathcal{V}(A_i)_l}(1-\pi_l^{(q)})^{1-\mathcal{V}(A_i)_l}}, \quad \text{for each} \quad h \in \{1, \ldots, H\}$$

   Consequently create Binomial matrices:

   $$Y^{(h)} = \sum_{i:G_i=h} \mathcal{V}(A_i), \quad \text{for each component} \quad h$$

3. **If a mixture component is not empty, the Pólya–gamma augmented data is updated from the full–conditional:**

   $$\omega_l^{(h)}|- \sim \text{PG}\{n_h, Z_l + \mathcal{V}(\mathbf{X}^{(h)}\mathbf{\Lambda}^{(h)}\mathbf{Q}^{(h)T})_l\}$$

   where PG is the Pólya–gamma distribution with parameters $b > 0$ and $c \in \Re$

4. **For each component h, block–sample each row v of $\bar{\boldsymbol{X}}$ conditionally on all other parameters and $\bar{\boldsymbol{Q}}_{(-v)}$, which corresponds to $\bar{\boldsymbol{Q}}$ without the v–th row and viceversa for $\bar{\boldsymbol{Q}}$ and $\bar{\boldsymbol{X}}_{(-v)}$.**
   To do so, this step can be interpreted as a Bayesian logistic regression on $\boldsymbol{Y}_{(v)}^{(h)}$ so that, being $\boldsymbol{\Omega}_{(v)}^{(h)}$ the diagonal matrix with $v-1$ elements with the corresponding Pólya–gamma augmented data, the full conditionals are:

   $$\bar{\boldsymbol{X}}_v^{(h)}|- \sim \mathcal{N}_R\Big\{\big(\bar{\boldsymbol{Q}}_{(-v)}^{(h)T}\boldsymbol{\Omega}_{(v)}^{(h)}\bar{\boldsymbol{Q}}_{(-v)}^{(h)} + \boldsymbol{\Lambda}^{(h)^{-1}}\big)^{-1}\boldsymbol{\eta}_{v_X}^{(h)}, \big(\bar{\boldsymbol{Q}}_{(-v)}^{(h)T}\boldsymbol{\Omega}_{(v)}^{(h)}\bar{\boldsymbol{Q}}_{(-v)}^{(h)} + \boldsymbol{\Lambda}^{(h)^{-1}}\big)^{-1}\Big\}$$

where $\boldsymbol{\eta}^{(h)}_{vX} = \bar{\boldsymbol{Q}}^{(h)T}_{(-v)}(\boldsymbol{Y}^{(h)}_{(v)} - \mathbf{1}_{V-1}n_h/2 - \boldsymbol{\Omega}^{(h)}_{(v)}\boldsymbol{Z}_{(v)})$,

$$\bar{\boldsymbol{Q}}^{(h)}_{v}|- \sim \mathcal{N}_R\Big\{\big(\bar{\boldsymbol{X}}^{(h)T}_{(-v)}(\boldsymbol{\Omega}^{(h)T})_{(v)}\bar{\boldsymbol{X}}^{(h)}_{(-v)}+\boldsymbol{\Lambda}^{(h)^{-1}}\big)^{-1}\boldsymbol{\eta}^{(h)}_{vQ},\, \big(\bar{\boldsymbol{X}}^{(h)T}_{(-v)}(\boldsymbol{\Omega}^{(h)T})_{(v)}\bar{\boldsymbol{X}}^{(h)}_{(-v)}+\boldsymbol{\Lambda}^{(h)^{-1}}\big)^{-1}\Big\}$$

where $\boldsymbol{\eta}^{(h)}_{vQ} = \bar{\boldsymbol{X}}^{(h)T}_{(-v)}((\boldsymbol{Y}^{(h)T})_{(v)} - \mathbf{1}_{V-1}n_h/2 - (\boldsymbol{\Omega}^{(h)T})_{(v)}(\boldsymbol{Z}^T)_{(v)})$.
Particular attention has to be payed here since we have to feed the right quantities for $\boldsymbol{Y}, \boldsymbol{Z}$ and $\boldsymbol{\Omega}$, that differ between $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Q}}$ because of the directionality information contained in the networks.

5. **Update component–specific weight parameters for each h.**
Being $\lambda^{(h)} \sim \mathrm{MIG}(a_1, a_2)$ the multiplicative inverse gamma distributed weights, denote $\lambda^{(h)}_r = \prod_{m=1}^{r} \frac{1}{\vartheta^{(h)}_m} \quad r = 1, \ldots, R$ and sample $\boldsymbol{\vartheta}^{(h)} = (\vartheta^{(h)}_1, \ldots, \vartheta^{(h)}_R)$:

$$\vartheta^{(h)}_1|- \sim \Gamma\Bigg(a_1 + V \cdot R, 1 + \frac{1}{2}\sum_{m=1}^{R}\theta^{(-1)}_m\sum_{v=1}^{V}(\bar{\boldsymbol{X}}^{(h)}_{vm})^2 + \frac{1}{2}\sum_{m=1}^{R}\theta^{(-1)}_m\sum_{v=1}^{V}(\bar{\boldsymbol{Q}}^{(h)}_{vm})^2\Bigg)$$

$$\vartheta^{(h)}_{r\geq 2}|- \sim \Gamma\Bigg(a_1 + V \cdot (R - r + 1), 1 + \frac{1}{2}\sum_{m=1}^{R}\theta^{(-r)}_m\sum_{v=1}^{V}(\bar{\boldsymbol{X}}^{(h)}_{vm})^2 + \frac{1}{2}\sum_{m=1}^{R}\theta^{(-r)}_m\sum_{v=1}^{V}(\bar{\boldsymbol{Q}}^{(h)}_{vm})^2\Bigg)$$

where $\theta^{(-r)}_m = \prod_{t=1, t\neq r}^{m} \vartheta^{(h)}_t$ for $r = 1, \ldots, R$
Since we added a new term in the model that is directly tied to $\boldsymbol{\Lambda}$, $\bar{\boldsymbol{Q}}$ has to be included in the calculation of these multiplicative inverse gamma weights, similarly to $\bar{\boldsymbol{X}}$.

6. **Update the shared similarity vector $\boldsymbol{Z}$**

$$\boldsymbol{Z}|- \sim \mathcal{N}_{V(V-1)}(\mu_Z, \Sigma_Z)$$

where Σ_Z has diagonal elements $\sigma^2_{Z_l} = 1/(\sigma_l^{-2} + \sum_{h=1}^{H}\omega^{(h)}_l)$ and $\mu_{Z_l} = \sigma^2_{Z_l}\{\sigma_l^{-2}\mu_l + \sum_{h=1}^{H}[Y^{(h)}_l - n_h/2 - \omega^{(h)}_l \mathcal{V}(X^{(h)}\Lambda^{(h)}Q^{(h)T})_l]\}$ for each l.

7. **Update component–specific edge probabilities vectors for each h:**

$$\boldsymbol{\pi^{(h)}} = \Big(1 + \exp\{-\boldsymbol{Z} - \mathcal{V}(\bar{\boldsymbol{X}}^{(h)}\bar{\boldsymbol{Q}}^{(h)T})\}\Big)^{-1}$$

8. **Sample the testing indicator T from a Bernoulli distribution with parameter p equal to probability** (4.13)

9. **Update mixture probabilities vector $\boldsymbol{v}$:**

   - if $T = 0$, let $\nu_y = \boldsymbol{v}$ for both groups, with $\boldsymbol{v}$ updated from the full conditional Dirichlet $(v_1, \dots, v_H)|- \sim \text{Dir}(1/H + n_1, \dots, 1/H + n_H)$
   - else if $T = 1$, update each ν_y independently from $(v_{1y}, \dots, v_{Hy})|- \sim \text{Dir}(1/H + n_{1y}, \dots, 1/H + n_{Hy})$

4.2.1 Simulation studies

To check the performance of the aforementioned Gibbs sampler, a set of simulations is considered. The number of nodes V is always set at 20. For what concerns the model, mixture components and latent space dimensions are set at $H = R = 10$. The study is composed of:

1. global and local independence scenarios for populations of 50 networks equally divided into two groups.

2. a global dependence scenario where the group differences are on the joint probability mass function, but not on the edge probabilities. This simulation is based on 40 networks equally.

3. two local dependence settings where 30 edges out of 380 change across groups. In the first case posterior computation conditions on 46 networks observations, while in the second the focus is on 100 networks.

Global and local independence

For this case, we provide a simple setting where edges in the groups are generated with about the same probabilities (only randomly jittered by 1%), with a total of 50 networks equally divided in two groups.

The model correctly rejects the alternative hypothesis H_1 of dependence between the groups and the networks generating process with an estimated probability of 0.0847 out of 5000 Gibbs samples. All parameters granted highly satisfactory convergence and mixing according to PSRF, traceplots and effective sample sizes. As one would expect in case of independence, Figure 4.3 shows an overall white color even with a coloring scale restricted to $(0, 0.1)$, while $\rho_l \in (0, 1)$ for each l. In fact, the model is structured so that when the null hypothesis is accepted there is no change in the probabilities, and hence no ρ is systematically bigger than 0.1.

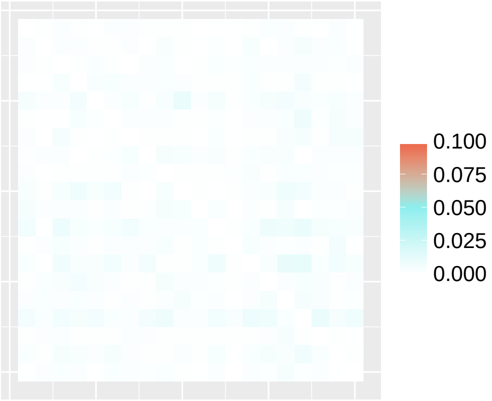

Figure 4.3: The proportion of $\rho_l > 0.1$ out of the Gibbs samples for the global and local independence simulation case, with $l \in \{1, \ldots, V(V-1)\}$.

Global dependence and local independence

For this setting we created two groups of 20 networks each that have different joint probability mass functions, but the edge probabilities — characterizing the marginals — do not change across groups. In particular, the group marked with $y = 1$ comprises a subset of 10 networks with overall tie probability $p_{1a} = 0.3$. The second subset has instead networks having edges with a tie probability $p_{1b} = 0.6$. The second group, identified by $y = 2$, contains 20 networks characterized by edges with a tie probability of $p_2 = 0.45$. This way, the generative mechanism differs in the two groups, but the edge probabilities do not

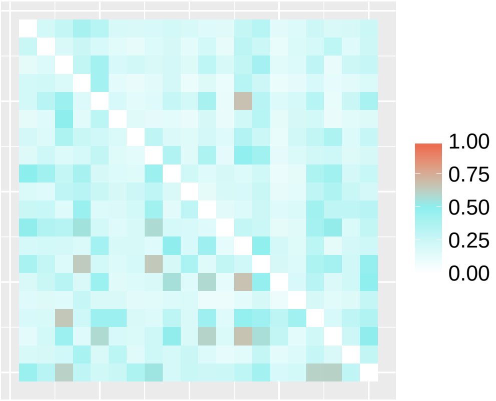

Figure 4.4: The proportion of $\rho_l > 0.1$ out of the Gibbs samples for the global dependence / local independence simulation case, with $l \in \{1, \ldots, V(V-1)\}$.

display group differences. The results show
that overall some changes are globally happening, as $\hat{\text{pr}}\big(H_1|\{y, \mathcal{L}(\mathcal{A})\}\big) = 0.996$, but these changes are due to higher–level variations and not to differences in edge probabilities across groups, as seen in Figure 4.4; although there is a diffuse sky blue coloring, no points get close to the 90% proportion. In case of great evidence, a square would be marked as bright red in case the proportion of Gibbs samples with a $\rho_l > \epsilon$ is greater than 0.9, with l again addressing all pairs of different nodes and ϵ a threshold usually chosen to be around 0.1. No particular edge is hence being indicated as evidently different across groups, consistently with the setting of the simulation.

Global and local dependence

To assess performance of our newly proposed method in more complex scenarios, i.e. in presence of global and local dependence, we simulate different edges probability for selected ties in the first and in the second group, for a total of 30 ties changing across groups. These are accounted to be 15 per group, with a distinctive direction: for a pair of selected nodes (u^*, v^*), $p(u^* \to v^*)$ is high and $p(v^* \to u^*)$ is small ($\to$ implies a connection). The null hypothesis is rejected following a strong evidence in favor of the alternative 0.998. The matrix of true differences is presented in Figure 4.5 along with the estimated edge probability differences in the two groups for the 46 and 100 networks settings. As it can be seen from the second and third frame, probabilities are better estimated as the number of networks increases, as common sense would suggest; this is revealed by the decreasing amount of uncertainty (i.e. more defined colors for single squares where the true probability is actually not different) in these plots.

Figure 4.6 displays the true probability and estimated proportions of Gibbs samples for which a tie shows great evidence of being different in the two groups. Consistently with what is observable for Figure 4.5, increasing the number of simulated networks also increases precision in distinguishing the ties that truly change across groups from random results. This trait is shown by the lighter coloring (i.e. proportion of significant ρ_l closer to 0) of the third frame for $n = 100$ compared with the second one for $n = 46$.

4.2.2 Application to directed passing networks

After studying the performance of the newly proposed method for directed networks, we fit the model to our passing data, that now feature the added information about the

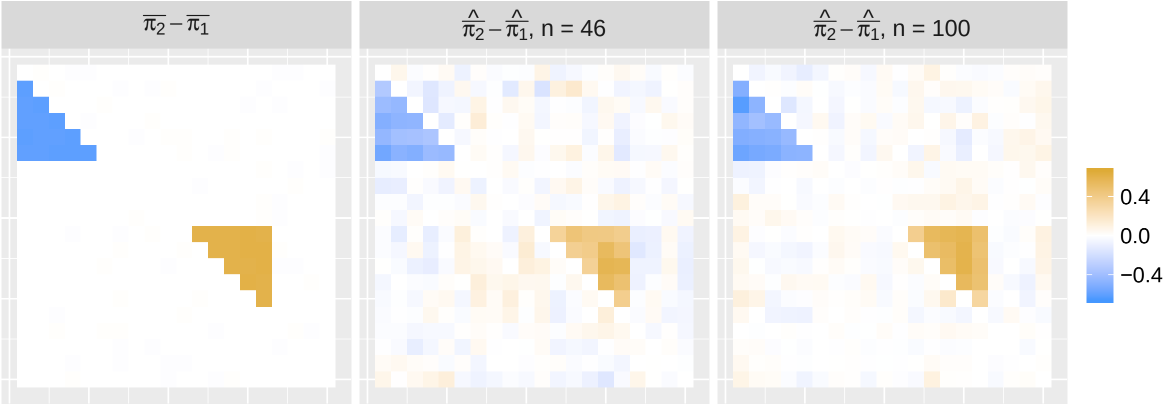

Figure 4.5: A comparison between the true difference in probabilities and the posterior means of the estimated difference of probabilities for the simulations respectively with $n = 46$ and $n = 100$, with n being the number of networks simulated.

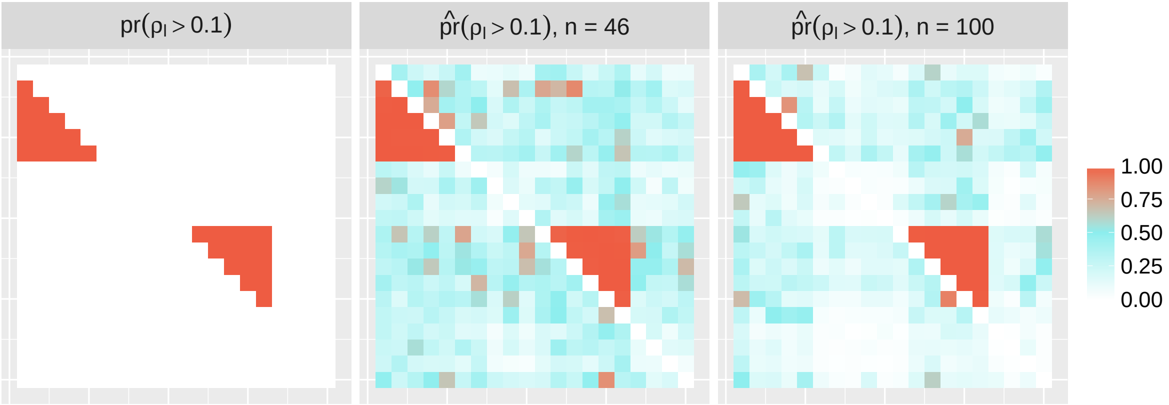

Figure 4.6: A comparison between the true probability of $\rho_l > 0.1, \quad l \in \{1, \ldots, V(V-1)\}$ and posterior means of the corresponding estimated proportion of the Gibbs samples for the simulations respectively with $n = 46$ and $n = 100$, with n being the number of networks simulated.

direction of the pass. Since the original data already had this information, the new networks are obtained by getting the area of the court from where the act of passing started and where the ball was received; we are still treating the tie as binary, hence an edge is 1 if in that particular quarter at least one pass was made between area v and u (and not from u to v).

Consistently with previous results, no evidence in favor of H_1 is found after obtaining an estimated probability of 0.0475 in the directed ties case. Such a small value implies that probabilities almost never change across groups in all Gibbs samples (after burn–in), yielding a difference of 0. This statement is confirmed by the complete white coloration of the frames in Figure 4.7 (even when displaying values in a very small range). A similar result would be displayed for the type of plot presented in Figure 4.1 that shows the distribution of the values of ρ_l, as it features an overall white color for 1^{st} and 3^{rd} posterior quartiles and posterior mean; it is here omitted due to redundancy.

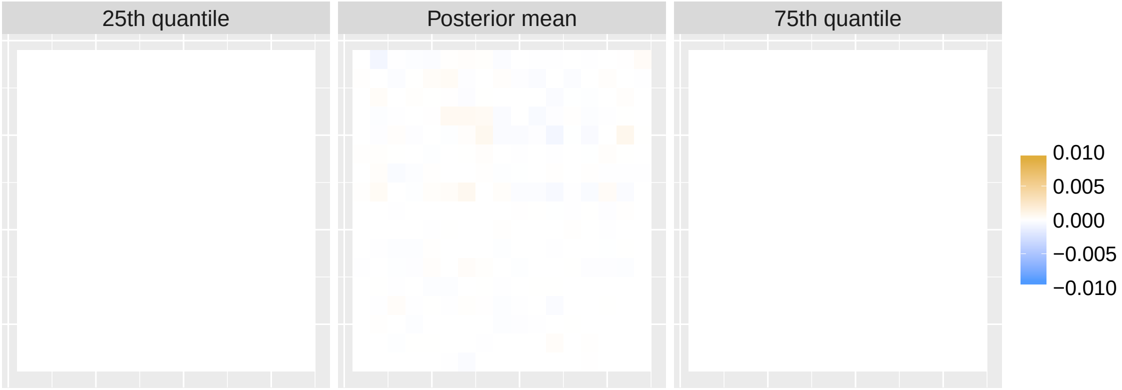

Figure 4.7: Mean and quartiles of the posterior distribution of $\bar{\pi}_2 - \bar{\pi}_1$ for the $+0/-1$ directed networks dataset. Here pairs of court areas identified by l are arranged in matrix form.

Comparing this result with what was obtained for the undirected case, we could assume that the small evidence of 0.5134 (still to be considered as no evidence) is induced by the forcing of some ties to be reciprocal while in reality they are not. While this may hold for passes right outside the three point line, that is between areas OUT_R, OUT_{CR}, OUT_{CL} and OUT_L, using the tags shown in Figure 2.6, this is probably not the case for cross–side passes (from the left side to right side with non–adjacent areas). The smaller evidence might then be induced by the fact that while using undirected networks the tie can occur for either connection, i.e. $u \rightarrow v$ or $v \rightarrow u$, contributing more to the observed ties among all quarter networks. In the directed case these are treated separately, and in case these ties happen a comparable amount of times, they do not differ too much from a possible group that has less observations for the passes between those areas. We can therefore conclude that there is no evidence of a difference between the positive/neutral and negative plus–minus quarters passing binary networks for the Golden State Warriors team.

Discussion

In this thesis project we analyzed the passing networks of an NBA team from a new perspective, considering the connections between different areas of the court. Specifically, we wanted to inspect if there were differences in these passing networks when comparing won and lost quarters by the Golden State Warriors. To do so, we considered the networks to be binary and undirected. Firstly, in Chapter 3 we used single networks models (ERGM), choosing two extreme examples for the biggest positive plus–minus and the biggest negative plus–minus quarters; secondly, we considered a latent space model that allowed to account for all the observed networks by modeling their structure via shared latent space representation. Chapter 4 provided a joint analysis of all networks via a Bayesian nonparametric model that allows joint inference on the differences in these networks, treated as undirected, between won an lost quarters. Lastly, we proposed a new method that generalized Durante and Dunson (2016) to the directed case. Accordingly to the results, there is no substantial evidence of differences in the passing networks for the two groups. This result might be caused by several factors: first of all, there is probably no actual difference in passing networks when considering edges as binary. Additionally, many factors such as the opponent team, the current lineup for the team on offense, etc., might be introducing uncertainty that we do not account for.

Although tested on win/loss group differences, the proposed method offers flexibility in terms of which grouping variable to take into consideration. For example: differences in field goal percentage, field goals made, and many others could be explored. Moreover, since we are using court areas and not single players or positions, the analysis can potentially be extended to inspect differences between two teams, or the same team in the first and second part of the season (e.g. when a new coach is hired, or after important trades are involved).

Other possible extensions involve considering batch of possessions instead of fixed time quarters, to better characterize all the nuances that a usual basketball game presents. This would imply an even more intense pre–processing procedure than the one proposed in Chapter 2, and a very precise model to recognize exactly offensive rebounds and extrapossessions that would otherwise bias the passes' counts.

An important further development of the joint Bayesian nonparametric model would feature the generalization to non–binary networks. This would allow the passing networks to retain its original characterization of counts of passes happening between two areas of the court in a quarter. Incorporating the information on weighted edges, data take the form of multivariate counts, again with network–structured dependence (Durante and Dunson, 2016). A possibility is represented by including latent variables in Poisson factor models as in Dunson and Herring (2005), among others. However, the latent variable is now responsible for both managing over–dispersion in the marginal distributions and controlling the degree of dependence, making the problem even more complicated. Canale and Dunson (2012) propose a solution via a rounded kernel method to better characterize the count variables.

Appendix A

Basketball related technical terms

Table A.1: Basketball-related terms and definitions used in this work.

term	**definition**
alley oop	a particular type of pass where the ball is thrown near the basket to a teammate jumping towards the rim that catches the ball in the air and scores, with a dunk or a lay-up
assist	a successful pass resulting in a field goal or a drawn foul leading to at least one scored free throw by the player who receives the ball
box score	an official table that contains points, minutes, rebounds, assists, steals , turnovers etc for each single player and team
bounce pass	a type of pass that bounces on the floor
chest pass	a type of pass where a player starts the pass from his chest and delivers it directly to the chest of the receiver
dunk	the act of scoring the ball directly with one or two hands usually making contact with the rim
fastbreak	on offense, the act of trying to score the ball as fast as possible before defense is able to recover
field goal	a shot with the intention of scoring from anywhere in the court. It's "attempted" if missed, and "made" if the shot is successful
free throw	an uncontested throw attempt at the basket worth one point. It is usually result of a drawn foul while attempting a shot or the result of a foul committed when the offensive team is in the bonus (i.e. the defensive team has committed at least 4 fouls in the quarter)
game clock	the main time tracking clock. It starts at 12 each quarter (5 in the case of overtimes) and ticks progressively whenever the ball is alive
inbound	the act of passing the ball from outside the boundary lines, after a deviation out of bounds, a turnover or a made basket. The game clock restarts only when the player actually receives the ball (a touch is sufficient).
lob	a slow, high-arching pass, performed in order to avoid the defense usually in the post positions
play-by-play	the collection of relevant events of the game in temporal order, with information on things such as the players involved, the game clock time, the progressive score etc
rebound	the act of collecting the ball after a missed shot in the defensive or offensive end
shot clock	a secondary clock that limits the time that teams are allowed to take shots in. It starts at 24 and is reset each time a new team gets possession, or an offensive rebound is collected after the shot has touched the rim
rim	the orange-colored metallic part of the basket, to which a net is attached
steal	the act of stealing the ball actively
three-point line	a circular line whose distance from the basket varies from 22 (in the corners) to 23.75 feet (everywhere else)
transition	the act of moving up the court after a team has just gained possession and the other squad has not established positions
turnover	an action that ends up in losing possession. The following fall under this category: bad pass, mishandle, 24-seconds violation, 8-seconds violation, 5-seconds violation, 3-seconds violation, traveling, carrying, palming, offensive foul
violation	an infraction of the rules

Appendix B

Convergence plots for the undirected networks Bayesian model

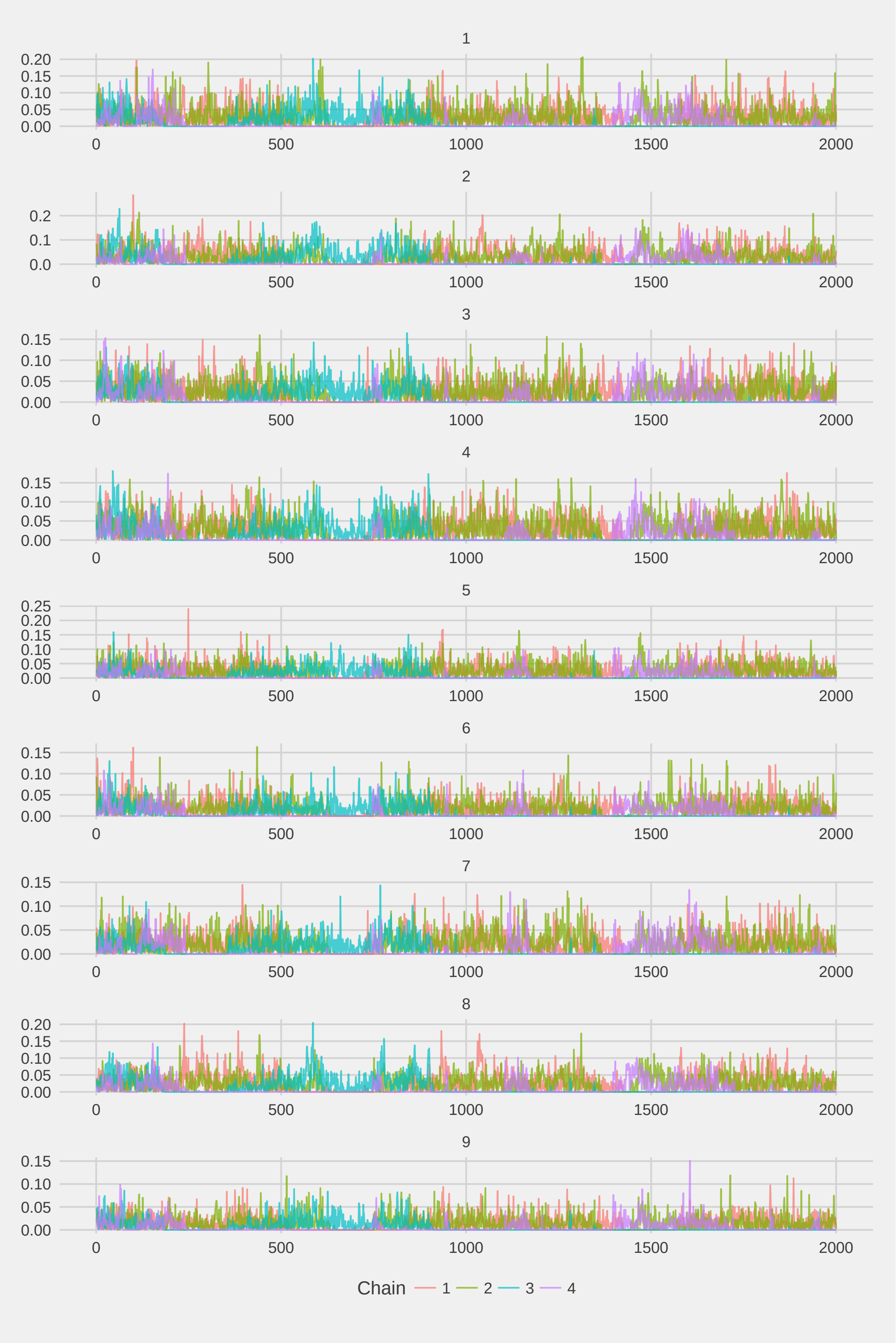

Figure B.1: Traceplots for 9 randomly selected ρ_l, $l \in \{1, \ldots, 136\}$ for the undirected data. After burn–in, the chain is divided in 4 parts which are being compared.

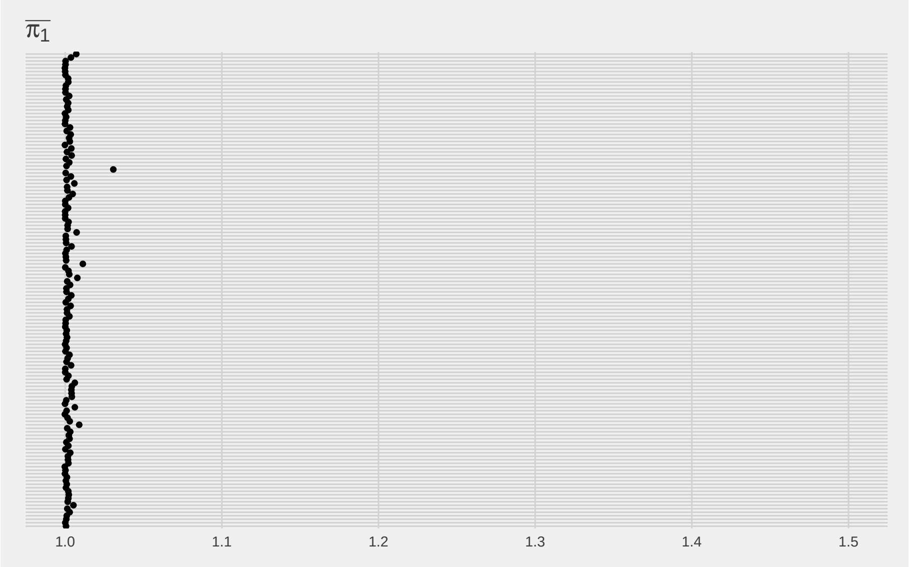

Figure B.2: Potential Scale Reduction Factors for $\bar{\pi}_{1_l}$, $l \in \{1, \ldots, 136\}$ for the undirected data. Values below 1.15 grant satisfactory convergence.

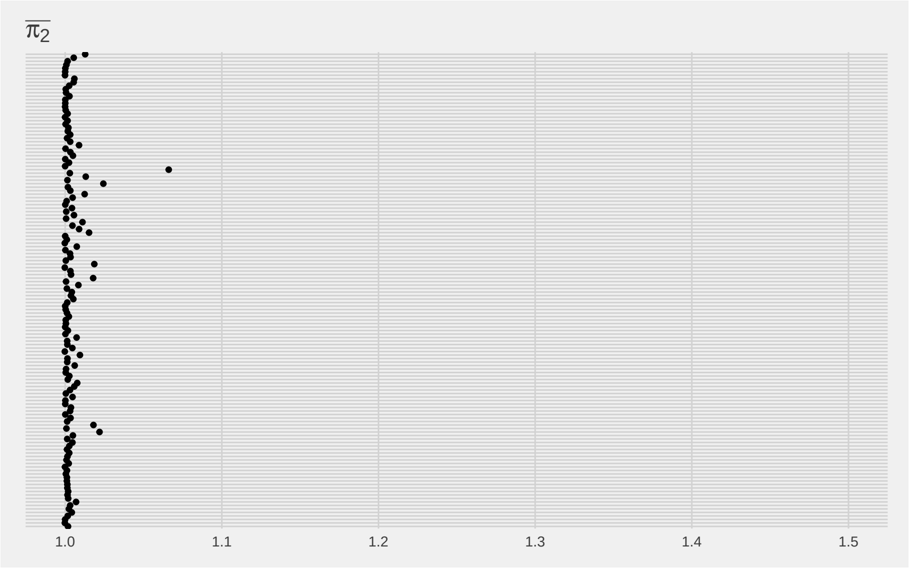

Figure B.3: Potential Scale Reduction Factors for $\bar{\pi}_{2_l}$, $l \in \{1, \ldots, 136\}$ for the undirected data. Values below 1.15 grant satisfactory convergence.

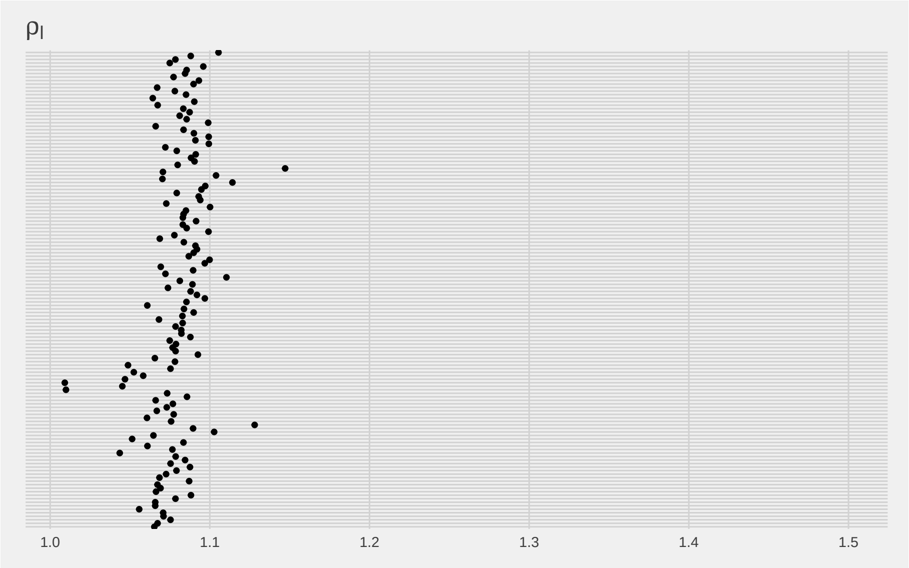

Figure B.4: Potential Scale Reduction Factors for ρ_l, $l \in \{1, \ldots, 136\}$ for the undirected data. Values below 1.15 grant satisfactory convergence.

Appendix C

Convergence plots for the directed networks Bayesian model

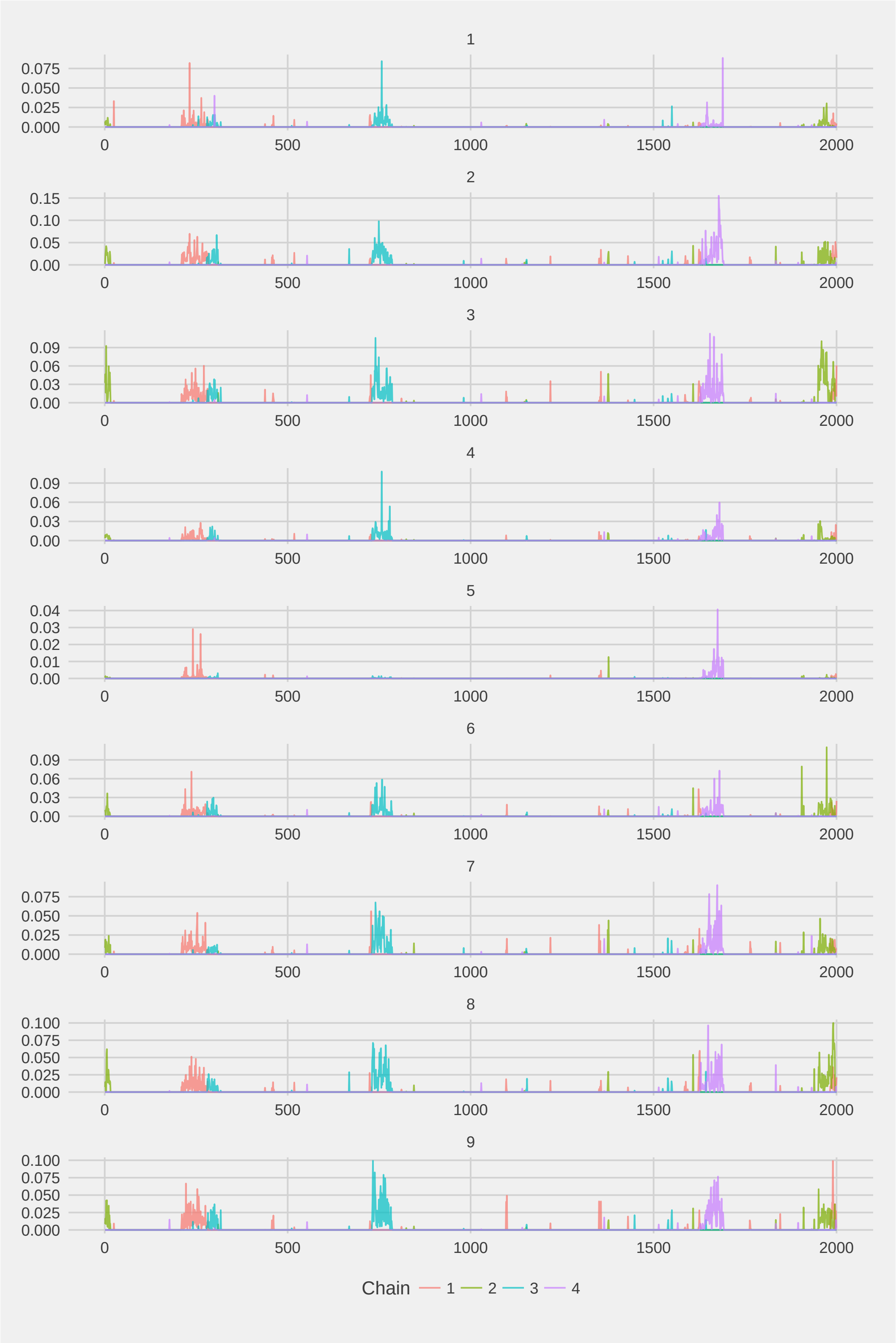

Figure C.1: Traceplots for 9 randomly selected ρ_l, $l \in \{1, \ldots, 272\}$ for the directed data. After burn–in, the chain is divided in 4 parts which are being compared.

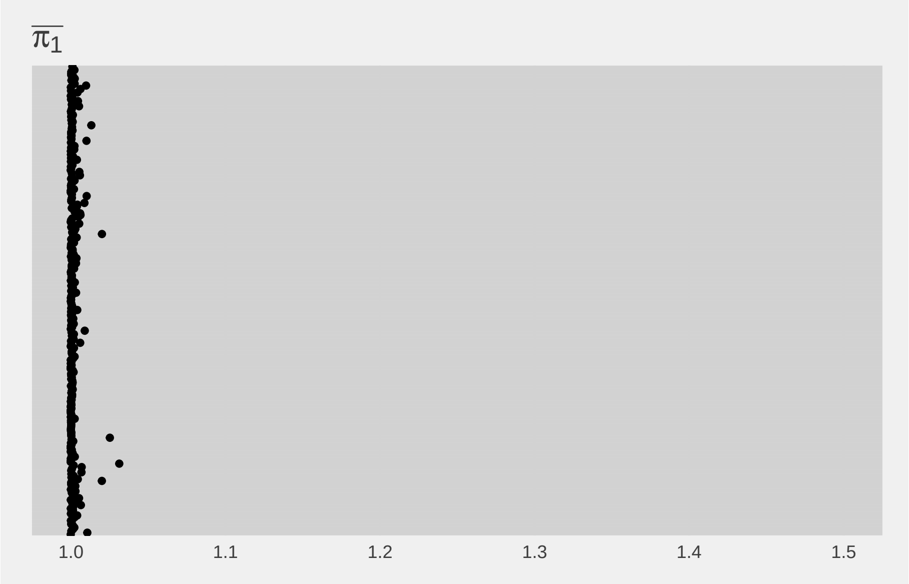

Figure C.2: Potential Scale Reduction Factors for $\bar{\pi}_{1_l}$, $l \in \{1, \ldots, 272\}$ for the directed data. Values below 1.15 grant satisfactory convergence.

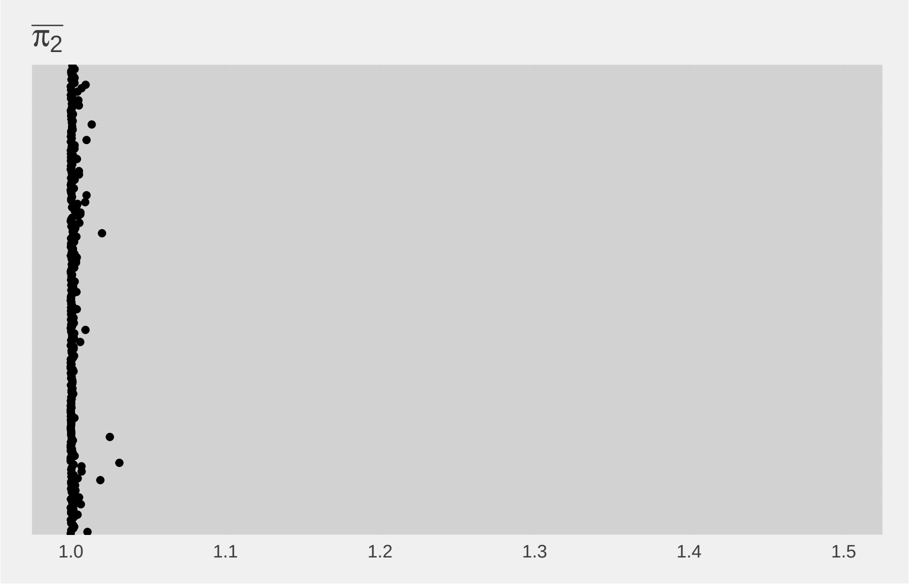

Figure C.3: Potential Scale Reduction Factors for $\bar{\pi}_{2_l}$, $l \in \{1, \ldots, 272\}$ for the directed data. Values below 1.15 grant satisfactory convergence.

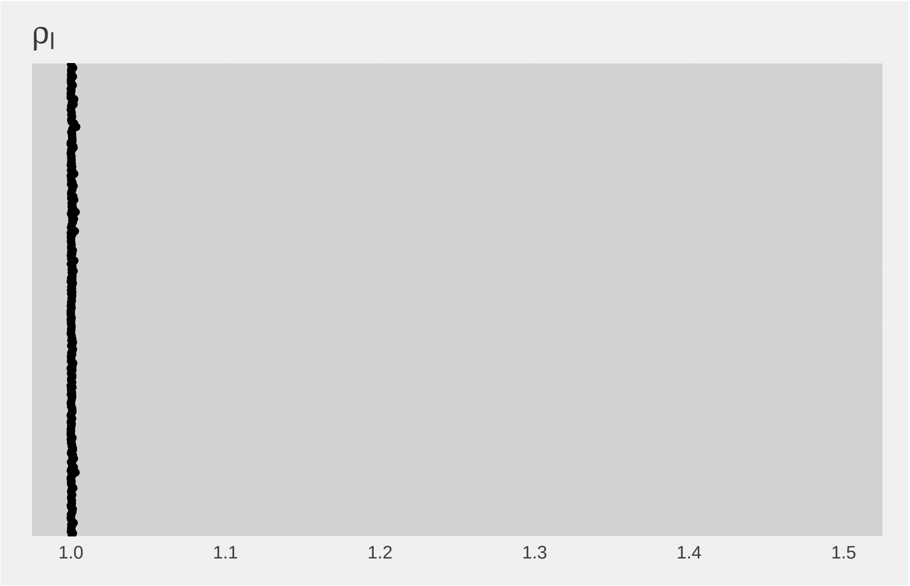

Figure C.4: Potential Scale Reduction Factors for ρ_l, $l \in \{1, \ldots, 272\}$ for the directed data. Values below 1.15 grant satisfactory convergence.

References

Airoldi, Edoardo M., Blei, David M., Fienberg, Stephen E., and Xing, Eric P. (2008). Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9 (Sep): 1981–2014.

Altman, Naomi S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46 (3): 175–185.

Breiman, Leo (1996). Bagging predictors. Machine learning, 24 (2): 123–140.

— (2001). Random forests. Machine learning, 45 (1): 5–32.

Canale, Antonio and Dunson, David B. (2012). Bayesian kernel mixtures for counts. Journal of the American Statistical Association.

Clemente, Filipe M., Martins, Fernando M.L., Kalamaras, Dimitris, and Mendes, Rui Sousa (2015). Network analysis in basketball: inspecting the prominent players using centrality metrics. Journal of Physical Education and Sport, 15 (2): 212.

Dunson, David B. and Herring, Amy H. (2005). Bayesian latent variable models for mixed discrete outcomes. Biostatistics, 6 (1): 11–25.

Dunson, David B. and Xing, Chuanhua (2009). Nonparametric Bayes modeling of multivariate categorical data. Journal of the American Statistical Association, 104 (487): 1042–1051.

Durante, Daniele and Dunson, David B. (2016). Bayesian inference and testing of group differences in brain networks. 'https://arxiv.org/pdf/1411.6506v5.pdf'.

Durante, Daniele, Dunson, David B., and Vogelstein, Joshua T. (2016). Nonparametric Bayes Modeling of Populations of Networks. Journal of the american Statistical association (to appear).

Elbel, E.R. and Allen, Forrest C. (1941). Evaluating team and individual performance in basketball. Research Quarterly. American Association for Health, Physical Education and Recreation, 12 (3): 538–555.

Erdös, Paul and Rényi, Alfréd (1959). On random graphs, I. Publicationes Mathematicae (Debrecen), 6: 290–297.

Fewell, Jennifer H., Armbruster, Dieter, Ingraham, John, Petersen, Alexander, and Waters, James .S (2012). Basketball teams as strategic networks. PloS one, 7 (11): e47445.

Frank, Ove and Strauss, David (1986). Markov graphs. Journal of the american Statistical association, 81 (395): 832–842.

Franks, Alexander, Miller, Andrew, Bornn, Luke, and Goldsberry, Kirk (2015). "Counterpoints: Advanced defensive metrics for nba basketball". In: *9th Annual MIT Sloan Sports Analytics Conference, Boston, MA*.

Freeman, Linton C. (1977). A set of measures of centrality based on betweenness. Sociometry: 35–41.

Freund, Yoav and Schapire, Robert E. (1996). "Experiments with a new boosting algorithm". In: *Icml*. Vol. 96, pp. 148–156.

Friedman, Jerome H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics: 1189–1232.

Gelman, Andrew and Rubin, Donald B. (1992). Inference from iterative simulation using multiple sequences. Statistical science: 457–472.

Grabiner, David (1994). The sabermetric manifesto. The Baseball Archive.

Gupta, Anirban, Witt, Hiroki, and Khan, Meraj (2015). Analyzing the game of basketball as networks. http://hiroki-witt.appspot.com/analyzing-basketball-games.pdf.

Handcock, Mark S., Raftery, Adrian E., and Tantrum, Jeremy M. (2007). Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170 (2): 301–354.

Hoff, Peter D. (2003). Random effects models for network data. na.

Hoff, Peter D., Raftery, Adrian E., and Handcock, Mark S. (2002). Latent space approaches to social network analysis. Journal of the american Statistical association, 97 (460): 1090–1098.

Holland, Paul W. and Leinhardt, Samuel (1981). An exponential family of probability distributions for directed graphs. Journal of the american Statistical association, 76 (373): 33–50.

Krivitsky, Pavel N. and Handcock, M. (2009). The latentnet package. Statnet project, version: 2–1.

Krivitsky, Pavel N., Handcock, Mark S., Raftery, Adrian E., and Hoff, Peter D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social networks, 31 (3): 204–213.

Miller, Andrew, Bornn, Luke, Adams, Ryan, and Goldsberry, Kirk (2014). "Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball." In: *ICML*, pp. 235–243.

Nelder, John A. and Baker, R. Jacob (1972). Generalized linear models. Encyclopedia of statistical sciences.

Nowicki, Krzysztof and Snijders, Tom A.B. (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96 (455): 1077–1087.

Oliver, Dean (2004). Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc.

Polson, Nicholas G., Scott, James G., and Windle, Jesse (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. Journal of the American statistical Association, 108 (504): 1339–1349.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Robins, Garry, Snijders, Tom A.B., Wang, Peng, Handcock, Mark, and Pattison, Philippa (2007a). Recent developments in exponential random graph (p*) models for social networks. Social networks, 29 (2): 192–215.

Robins, Garry L., Pattison, Philippa E., Kalish, Yuval, and Lusher, Dean (2007b). An introduction to exponential random graph (p*) models for social networks. Social networks, 29 (2): 173–191.

Rousseau, Judith and Mengersen, Kerrie (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (5): 689–710.

Sabidussi, Gert (1966). The centrality index of a graph. Psychometrika, 31 (4): 581–603.

Shea, Stephen M. (2014). Basketball Analytics: Spatial Tracking.

Snijders, Tom A.B., Pattison, Philippa E., Robins, Garry L., and Handcock, Mark S. (2006). New specifications for exponential random graph models. Sociological methodology, 36 (1): 99–153.

Wasserman, Stanley and Pattison, Philippa (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. Psychometrika, 61 (3): 401–425.

Acknowledgements

There are so many people around me to thank,
For this wonderful journey that I've gone through.
Although I do not like, to use a mere rank
I would like to firstly stop, and name a few.

Being blessed by two marvelous, bighearted parents
I owe them the most, the best and the rest,
As they would always grant me their felt presence
Wherever I would go, from North to West.

Long'd be the list of the friends I'd say "Hey! You!"
Even though I do not here forget their names:
Shout out to everybody who supported me, they're my crew
I hope that our friendship, gladly, remains.

Cia, Anna^2, Cittadelles and M&S,
Usuals, Milan people and Universitees;
Leuven friends and city, I just might confess:
That I miss you like with his country does a refugee.

Last but not least, as usually they say
I greatly thank Daniele for his patience and skills
'Cause along my supervisor, with passion, all the way
They drove me through this thesis, as my dream fulfills.