Instituto Tecnológico Autónomo de México

Aplicación del Modelo de Estimación de Espacios Latentes para Fútbol

Tesis
que para obtener el título de
Licenciado en Matemáticas Aplicadas
Presenta
Sebastián Serra Peña

Asesor: Doctor Simón Lunagómez Coria

CIUDAD DE MÉXICO 2024

Autorización para difusión

"Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada *Aplicación del Modelo de Estimación de Espacios Latentes para Fútbol*, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación."

Sebastián Serra Peña

_____

Firma

Laura Peña (Mamá): *Por tu esfuerzo y sacrificio que has dedicado para proporcionarme la mejor educación posible. Tu apoyo incondicional, amor y paciencia infinita han sido pilares en mi vida. Reconozco que en ocasiones he fallado en corresponder a tu dedicación como debería, por lo que te pido disculpas de todo corazón.*

Emilio Serra (Papá): *Te agradezco por enseñarme el valor de la pasión en todo lo que hago y por mostrarme cómo enfrentar la adversidad sin temor. Reconozco que en ciertas ocasiones no he sido el mejor ejemplo y que mis juicios y falta de confianza pueden haber afectado nuestra relación. Me disculpo sinceramente por ello.*

Laura Rivera (Abuela): *Mi más profundo agradecimiento por el amor y apoyo incondicional que siempre me has brindado. Tu presencia ha sido un faro de luz en mi vida, y siempre estaré agradecido por tus consejos sabios y tu cariño inquebrantable.*

Juan Serra Buxade (Abuelo) (†): *Agradezco profundamente las lecciones de vida que me impartiste, especialmente la importancia de seguir mis pasiones sin buscar la aprobación de los demás. Aunque a veces tu personalidad fue difícil de comprender, admiro tu lucha y determinación hasta el final.*

Amigos: *Le debo un gracias a esos animales que un día se convirtieron en amigos. Mi gratitud hacia ustedes, mis queridos amigos, es inmensurable. Su amistad ha sido clave en mi vida, y deseo sinceramente que encuentren la felicidad en todo lo que hagan. Que si están contentos, yo estaré contento. Aunque eso comporte que otros sean infelices.*

Simón Lunagómez (Asesor): *Agradezco tu paciencia y orientación durante los momentos de desesperación, y por mostrarme el camino correcto cuando más lo necesitaba.*

Manuel Mendoza (Profesor): *Mi más sincero agradecimiento por infundirme entusiasmo por la estadística Bayesiana cuando pensaba que debía abandonarla. Su pasión por el tema ha sido inspiradora y ha marcado una diferencia significativa en mi camino académico.*

Índice

Resumen

La presente tesis se centra en el análisis táctico de fútbol utilizando un modelo de estimación gráfica mediante espacios latentes y un enfoque Bayesiano. El análisis táctico pre-partido, que busca estimar las características del juego del rival, plantea desafíos debido a la falta de información completa sobre el equipo contrario. Los métodos estadísticos tradicionales carecen de precisión al considerar una amplia gama de variables en el fútbol cuando se dispone de una cantidad limitada de datos sobre el equipo contrario. Por lo tanto, se propone utilizar la Estadística Bayesiana, permitiendo incorporar información a priori y actualizar el modelo posterior basado en los partidos más recientes.

La metodología comienza con un análisis descriptivo de los datos para construir una red social que describa el evento específico a analizar. Luego, se formula un modelo de espacios latentes basado en distancias, siguiendo el enfoque propuesto por Hoff, Raftery y Handcock en su artículo *Latent Space Approaches to Social Network Analysis*. Este modelo permite estimar posiciones latentes dados los datos observados, considerando las características de los jugadores y las interacciones entre ellos.

Además, se llevará a cabo un análisis detallado de las propiedades de los grafos, como la centralidad de intermediación para identificar jugadores clave y analizar su influencia en la red social futbolística. Esto proporcionará información valiosa sobre la estructura de las interacciones y las dinámicas de juego dentro del equipo.

El resultado final de la tesis incluye una evaluación de los resultados del modelo y una discusión sobre posibles mejoras, limitaciones y aplicaciones del análisis táctico a nivel de reporte para un hipotético club de fútbol que se enfrenta al rival. Se espera que esta investigación proporcione herramientas más precisas y fidedignas para el análisis táctico en el fútbol, mejorando la toma de decisiones estratégicas y el rendimiento del equipo en el campo.

1. Introducción

1.1. Motivación

Dos pasiones han sido los pilares en la creación de esta tesis: las matemáticas y el fútbol. Criado en un hogar donde las ecuaciones y los goles eran igualmente apreciados, estas dos influencias se entrelazaron formando parte de mi ser. Con padres actuarios y un arraigo familiar al fútbol catalán, el mundo numérico y el terreno de juego se convirtieron en mi doble naturaleza.

Ingresar al ámbito de las Matemáticas Aplicadas avivó una fascinación latente. El pensamiento de fusionar la abstracción matemática con la dinámica del fútbol me cautivaba. En geometría analítica, delineé las zonas de influencia de los jugadores, pintando el campo con las líneas invisibles que definen su alcance mediante las regiones de Voronoi. El cálculo diferencial e integral desveló la estrategia financiera de un club, donde la optimización actuaba como el pegamento financiero. En el mundo probabilístico, los goles esperados emergieron como destellos de certeza en un océano de incertidumbre.

La teoría de gráficos abrió nuevas perspectivas, presentando al equipo como una red interconectada, cada jugador un nodo que tejía estrategias con los demás. Fue en el vasto mar de la estadística, con su enfoque Bayesiano, que encontré el punto de partida de mi exploración intelectual.

En el corazón de mi motivación late una pregunta audaz: ¿es posible desentrañar la incertidumbre inherente al fútbol, predecir las estrategias de un rival con datos limitados, todo mediante la estadística Bayesiana? Mi convicción inquebrantable me impulsa a responder afirmativamente. Esta tesis es el producto de esa creencia.

Mi último semestre me regaló la oportunidad de participar en el curso de Analista Táctico de Fútbol, impartido por el prestigioso Fútbol Club Barçelona y su Barça Innovation Hub. Los resultados obtenidos abrieron puertas insospechadas. Los datos, gentilmente cedidos, se transformaron en la paleta con la que dar vida a este escrito.

Esta tesis es un mosaico de saberes: una fusión de análisis táctico, teoría de gráficos y estadística Bayesiana que convergen en un todo coherente. Mi

pasión por el fútbol y las matemáticas se entremezclan en estas páginas, como una coreografía precisa donde cada paso tiene su significado y cada dato es un eslabón en la estrategia que late en el corazón del juego.

Más que una tarea académica, esta tesis es un homenaje a la pasión que ha guiado mi vida. Cada línea que escribo rinde tributo al amor por el juego y al asombro por la belleza de las matemáticas. Que estas palabras reflejen mi determinación, el eco de mi entusiasmo y la evidencia de mi compromiso.

1.2. Antecedentes

El análisis táctico de fútbol desempeña un papel fundamental en la comprensión de los patrones de juego y en la toma de decisiones estratégicas durante los partidos. Para llevar a cabo un análisis efectivo, se requiere de herramientas y metodologías que permitan modelar y representar de manera adecuada la complejidad inherente a la interacción entre los jugadores y las dinámicas del juego. En este contexto, la Teoría de Gráficas ha surgido como una herramienta útil para el modelado y análisis táctico de fútbol.

Sin embargo, la estimación de los parámetros de un modelo basado en grafos en el contexto del análisis táctico de fútbol plantea desafíos significativos debido a la alta dimensionalidad de los datos y la incertidumbre del juego en sí. Además, los métodos tradicionales de estimación de modelos de grafos pueden resultar insuficientes para capturar las complejidades del juego, lo que limita su capacidad para proporcionar una representación precisa de las interacciones entre los jugadores.

Uno de los desafíos más comunes en este análisis táctico es la estimación de las características del juego del equipo rival antes de un partido. En este contexto, surge un problema fundamental relacionado con la abundancia de datos altamente variables. Los enfoques estadísticos clásicos carecen de una estimación precisa debido a que el uso de una muestra de datos grande puede implicar información desactualizada y poco relevante sobre el rival, como su plantilla, el entrenador, sus entrenamientos, los ciclos de preparación y las lesiones, que son factores que pueden cambiar rápidamente y usualmente lo hacen.

1.2.1. Enfoque Bayesiano

Para abordar esta problemática, se propone un enfoque basado en la Estadística Bayesiana. Este enfoque permite tratar este problema como uno de decisión en ambiente de incertidumbre en el cual se puede aprovechar el conocimiento a priori obtenido al observar al rival y actualizar el modelo de acuerdo con la información proveniente de los últimos tres partidos disputados por el equipo rival en la misma competición que el nuestro.

El uso de muestras resumidas se convierte en un problema más fácil de aterrizar utilizando este enfoque, ya que el análisis estadístico clásico no logra proporcionar estimaciones precisas bajo estas circunstancias. La Estadística Bayesiana nos brinda la capacidad de integrar tanto el conocimiento previo como la información más reciente, permitiendo obtener estimaciones mejor informadas sobre las características del juego del rival.

1.2.2. Preguntas Específicas del Estudio

En este contexto, se plantean las siguientes preguntas específicas que se abordarán a lo largo del estudio:

1. ¿Cómo se determina la relevancia de un jugador en un equipo cuando este tiene posesión del balón?

2. ¿Cuál es la metodología para identificar el perfil táctico de un equipo cuando se encuentra en posesión del balón?

3. ¿Dada una estructura en posesión, cómo se identifica el camino más probable para que el balón llegue desde la línea defensiva hasta la línea ofensiva?

1.3. Objetivo

Se anticipa que la incorporación de espacios latentes junto con un enfoque de estimación Bayesiano tenga el potencial de mejorar la precisión y la capacidad de representación de los modelos gráficos empleados en el análisis táctico de fútbol. Es importante aclarar que la mejora de la precisión no implica simplemente una reducción del error, como se aborda en enfoques frecuentistas, sino que en el contexto Bayesiano, se refiere a una mejor captura y

cuantificación de la incertidumbre epistémica. Esta característica permite una interpretación más robusta de las interacciones entre los jugadores y, por ende, ofrece perspectivas más profundas para la toma de decisiones estratégicas durante los partidos.

1.4. Hipótesis

Se espera que la utilización de espacios latentes en combinación con un enfoque de estimación Bayesiano mejore la precisión y la capacidad de representación de los modelos gráficos utilizados en el análisis táctico de fútbol. Además, se espera que este enfoque permita una mejor comprensión de las interacciones entre los jugadores, lo que a su vez puede proporcionar nuevas perspectivas para la toma de decisiones estratégicas previo y durante los partidos.

1.5. Metodología

Análisis descriptivo de los datos: En primer lugar, se llevará a cabo un análisis descriptivo de los datos disponibles con el objetivo de comprender la estructura subyacente y evaluar la viabilidad de construir un grafo o una red a partir de estos datos. Esto implicará explorar las características de los datos, como la naturaleza de las variables, la distribución de los valores y posibles relaciones entre ellas.

Construcción del grafo: Utilizando la teoría de gráficas, se procederá a estructurar un grafo que represente las relaciones relevantes para el suceso específico que se desea analizar en el contexto del análisis táctico de fútbol. Esto implicará identificar los jugadores y definir las conexiones ellos.

Formulación de un modelo: Siguiendo el enfoque propuesto por Hoff, Raftery y Handcock en su artículo *Latent Space Approaches to Social Network Analysis*, se formulará un modelo que permita estimar los espacios latentes con base a un modelo de distancias. Este modelo capturará las características subyacentes no observables que influyen en las interacciones dentro de la red social.

Estimación del parámetro de posición: Una vez formulado el modelo, se llevará a cabo el proceso de estimación del parámetro de posición. Esto implica utilizar técnicas de inferencia estadística, como métodos de máxima

verosimilitud o métodos Bayesianos, para estimar los valores de los parámetros del modelo que mejor se ajusten a los datos observados.

Evaluación de los resultados y discusión: Se evaluarán los resultados obtenidos a partir del modelo estimado y se llevará a cabo un análisis detallado de los mismos. Se discutirán posibles mejoras o ajustes necesarios en el modelo, así como las limitaciones identificadas durante el proceso de estimación. Además, se presentará un análisis táctico específico a nivel de reporte que sería útil para un hipotético club de fútbol que se enfrenta al rival en cuestión, destacando las principales conclusiones y recomendaciones estratégicas derivadas del análisis de la red social estimada.

Es importante destacar que esta metodología requerirá de la utilización de herramientas y software especializados para el análisis de redes (R [*igraph*, *network*]) y la estimación de modelos de espacios latentes (R [*latentnet*]). Asimismo, se recomienda realizar validaciones cruzadas y pruebas de sensibilidad (R [*tidyverse*, *dplyr*, *MASS*]) para garantizar la robustez de los resultados obtenidos.

2. Marco Teórico

2.1. Fundamentos Teóricos

2.1.1. Teoría de Gráficas (grafos)

La aplicación de la teoría de grafos en el presente trabajo de tesis reviste una trascendental importancia, no únicamente en relación al resultado final obtenido, sino también en virtud de las propiedades que un grafo ostenta. Dichas propiedades no solo influyen en el desenlace, sino que también desempeñan un papel crucial en las etapas subsiguientes al proceso de estimación de las posiciones latentes, pues permitirá visualizar un grafo a ser analizado.

Motivado por lo anterior, se emprende la tarea de establecer una serie de definiciones fundamentales. Estas definiciones adquieren un valor sustancial en el contexto del análisis táctico en el ámbito del fútbol.

Definición 1 (Grafo)**.** *Un* ***grafo*** *es un objeto que consiste en dos conjuntos llamados conjunto de vértices y conjunto de aristas. El conjunto de vértices V es finito y no-vacío. El conjunto de aristas E puede ser vacío, pero si no lo es tiene al menos dos subconjuntos de elementos que forman parte de V. Así se defina a un grafo como una función G tal que recibe ambos conjuntos y da una regla de asociación entre ellos.*

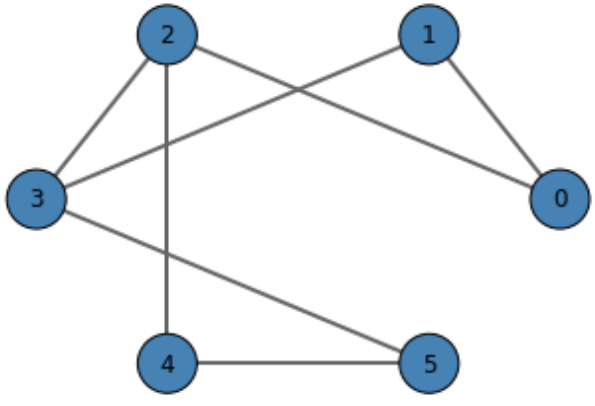

Figura 1: $G(V, E)$ con $V = \{0, 1, ..., 5\}$ y $E = \{(0, 1), (0, 2), ..., (4, 5)\}$

Un grafo se configura como una representación gráfica de las interrelaciones existentes entre diversos elementos. En esta analogía, los vértices son concebidos como puntos discretos, mientras que las aristas materializan las conexiones que los enlazan. En el contexto que nos compete, los vértices denotan a los jugadores, al tiempo que las aristas representan las vinculaciones que emergen de sus interacciones tácticas.

Cuando aludimos a las interacciones tácticas, nos estamos refiriendo a una medida específica, abarcando la utilización de toques de balón, la posesión, las intercepciones, entre otros aspectos relevantes. La elección de un enfoque gráfico, como el grafo, se fundamenta en su visualización, ya que facilita la creación de un gráfico que retrata a los jugadores en el terreno de juego, así como las posibles conexiones tácticas que se manifiestan durante el desarrollo de un partido.

Definición 2 (Vértices)**.** *Los* ***vértices*** *son los elementos individuales del grafo. Cada vértice puede representar un objeto, entidad o elemento del sistema que se está modelando.*

En nuestra narrativa futbolística, los vértices desempeñan el papel central. Cada uno de estos vértices encarna a un jugador, y es precisamente esta colección de entidades la que adquiere vida en la red interconectada que nos encontramos edificando.

Definición 3 (Aristas)**.** *Las* ***aristas*** *son los enlaces que conectan los vértices del grafo. Cada arista representa una relación o una conexión entre dos vértices.*

Las aristas, a su vez, constituyen los vínculos que entrelazan a nuestros protagonistas. Cada arista ejemplifica una relación táctica, tal como una asistencia entre jugadores.

Definición 4 (Peso)**.** *El peso de una arista es el valor que se le puede o no asignar a una arista dentro de un grafo. Este puede ser una medida de distancia, importancia o incluso probabilidad.*

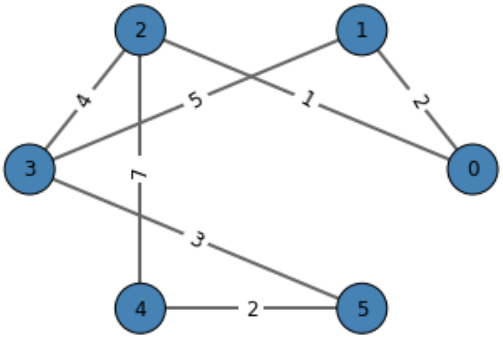

Figura 2: $G(V, E)$ con pesos representados con números en las aristas

El peso de una arista permite conferir una importancia adicional a ciertas acciones tácticas. Un ejemplo de esto sería considerar un pase largo como más significativo que un pase corto en el contexto táctico. En esencia, asignar pesos a las aristas permite establecer una jerarquía de relevancia entre diferentes movimientos y decisiones en el juego. Este enfoque proporciona una forma de cuantificar y expresar la ponderación que se concede a distintas acciones, lo que a su vez contribuye a una evaluación más matizada y detallada de las estrategias tácticas. Los pesos en las aristas añaden una dimensión adicional al análisis del grafo, permitiendo capturar de manera más precisa las sutilezas y las prioridades en la ejecución de tácticas específicas al permitir discernir entre pases que recorren mayor distancia, lejanía entre jugadores o probabilidad de interacción entre los jugadores.

Definición 5 (Grafo no Dirigido)**.** *Un grafo no dirigido, denotado como* $G = (V, E)$*, consiste en un conjunto de vértices* V *y un conjunto de aristas* E*, donde cada arista es una conexión no direccionada entre dos vértices distintos. Formalmente, un grafo no dirigido se define como:*

- V*: Conjunto de vértices, donde* $V = \{v_1, v_2, ..., v_n\}$*, siendo* n *el número de vértices en el grafo.*
- E*: Conjunto de aristas, donde* $E = \{e_1, e_2, ..., e_m\}$*, siendo* m *el número de aristas en el grafo.*
- *Cada arista* e_i *en* E *se define como un par de vértices* (v_j, v_k)*, donde* v_j *y* v_k *son vértices distintos en* V*.*

En un grafo no dirigido, la relación entre dos vértices a través de una arista es simétrica, es decir, si hay un defensa que se relaciona con un centrocampista, entonces el recíproco es también cierto, pues el cenrtocampista también se relaciona con ese mismo defensa. En palabras más sencillas, no representan una dirección.

Definición 6 (Camino)**.** *Un* ***camino*** *en un grafo se refiere a una secuencia de aristas que conectan una secuencia de vértices. El camino puede ser de longitud variable y puede pasar por varios vértices y aristas.*

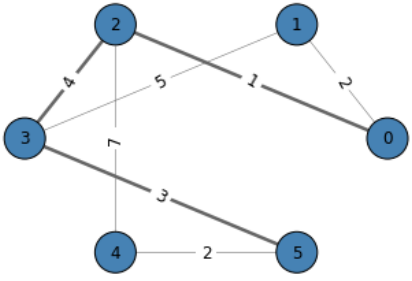

Figura 3: $G(V, E)$ con el camino 0 - 2 - 3 - 5 en línea gruesa

Un camino se puede interpretar en nuestro contexto como un itinerario táctico; es semejante a trazar una ruta de juego entre distintos elementos dentro de la red, pasando por circunstancias específicas que potencialmente puedan influir en el transcurso del partido. Esta analogía se asemeja a la acción de concretar un recorrido entre diversos vértices en el grafo que podrían determinar la trayectoria de un partido.

Si analizamos los pases más factibles de ser ejecutados entre un equipo inicial de once jugadores, nuestra atención se dirigirá a evaluar los caminos más cortos desde la línea defensiva hacia la ofensiva. ¿La razón? Descubriremos la vía con la probabilidad más alta de que el balón alcance la zona de tres cuartos de cancha. Más aún, si nuestra prioridad es el camino de longitud mínima, esta estrategia podría resultar óptima en el contexto de un contraataque.

En este sentido, el análisis de caminos en el grafo puede arrojar valiosa información acerca de las posibles tácticas a emplear en función de la disposición

de los jugadores en el campo, optimizando así las posibilidades de avance y éxito durante el juego.

Definición 7 (Ciclo)**.** *Un* ***ciclo*** *en un grafo se produce cuando existe un camino cerrado que comienza y termina en el mismo vértice, pasando por diferentes vértices y aristas en el proceso.*

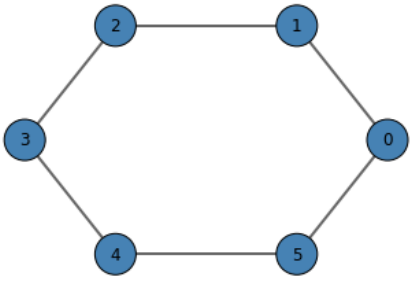

Figura 4: Ejemplificación de un ciclo en un grafo

Un ciclo, en este contexto, se configura como un bucle que culmina en el punto de origen. Este ciclo representa una secuencia de acciones que se reiteran en el transcurso del juego, quizás reflejando una táctica recurrente que un equipo emplea con el fin de mantener el control de la pelota. Siguiendo la secuencia de ejemplos previamente mencionados, es posible interpretar a los ciclos como estrategias destinadas a conservar la posesión del balón.

La lógica detrás de esta noción radica principalmente en la preferencia de ciertos equipos por mantener el control del balón. En este sentido, la aplicación de ciclos tácticos puede ser más pertinente para equipos que favorecen la posesión continua. Esta estrategia se fundamenta en una premisa lógica: el equipo adversario se ve limitado en su capacidad de anotar si no está en posesión del balón.

Así, los ciclos tácticos pueden ser interpretados como herramientas destinadas a asegurar el dominio del juego, ofreciendo al equipo la oportunidad de dictar el ritmo y el flujo del partido, al tiempo que dificulta las opciones de ataque del oponente al mantener la pelota en su poder.

Propiedad. *El **grado** de un vértice en un grafo se refiere al número de aristas que están conectadas a ese vértice. En un grafo no dirigido, el grado de un vértice es igual al número de aristas incidentes en ese vértice.*

El grado de un vértice denota su prominencia táctica. Cuanto mayor sea el número de aristas que se conecten a un jugador, mayor será su influencia en las interacciones del equipo. Esta característica adquiere un valor sustancial en un contexto descriptivo, ya que proporciona una explicación cuantitativa, precisa y visualizable sobre la cantidad de conexiones que un jugador es capaz de establecer con otros en el equipo.

Un jugador cuyo grado sea elevado refleja mayor habilidad para interactuar con otros compañeros, según la métrica específica en consideración. Algunas de las métricas más comunes son el número de pases entre dos jugadores, número de toques en una misma zona y otras de las cuales se hablará después. Sin embargo, es importante resaltar que la interpretación del grado está intrínsecamente vinculada a la métrica empleada. En otras palabras, la comprensión derivada del grado varía en función de la visión táctica particular que se esté evaluando. Esta visión táctica resalta el cómo quiere jugar un equipo en particular.

En resumen, el concepto de grado dentro de un grafo táctico brinda información esencial acerca de la participación y la influencia de un jugador en el esquema de juego del equipo, proporcionando una herramienta valiosa para analizar y comprender las dinámicas tácticas en el terreno de juego.

Propiedad. *La **densidad** de un grafo se refiere a la proporción de aristas presentes en el grafo en relación con el número total de aristas posibles. Se calcula dividiendo el número de aristas existentes entre el número total de aristas posibles. Sea n el número de vértices y m el número de aristas, en un grafo no dirigido la densidad Δ se define formalmente como:*

$$\Delta = \frac{2m}{n(n-1)}$$

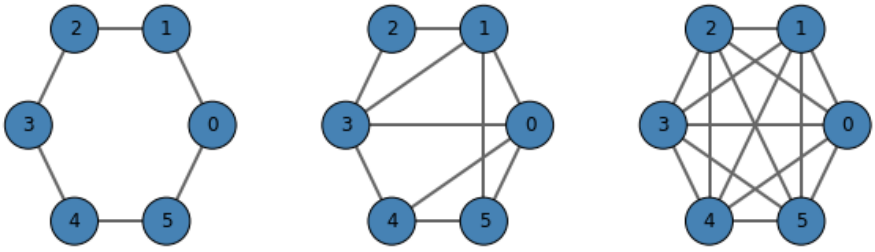

Figura 5: Grafos con densidades 0.4, 0.66, 1.0 respectivamente

En la Figura 5 tenemos tres grafos con seis vértices. El segundo tiene diez aristas y por tanto su densidad sería $2(10)/6(5) = 0.66$. Lo mismo ocurre para el resto de grafos.

Hasta este punto, hemos abordado propiedades individuales a nivel de vértice, es decir, a nivel de jugador. Sin embargo, ahora nos adentramos en una descripción global del grafo, del equipo en su conjunto. La densidad de un grafo encarna la intensidad de la táctica en juego. Si la red contiene numerosas aristas, significa que la táctica se entrelaza de manera densa entre los elementos. Los equipos con una mayor densidad en términos de pases reflejan una tendencia a mantener un control considerable del balón, no solo entre compañeros cercanos, sino también al realizar transiciones entre diferentes líneas del equipo. Esta noción evoca al famoso *fútbol total* promovido por Johan Cruyff.

En contraposición, densidades más bajas corresponden a equipos que no priorizan la interconexión entre todos los jugadores. Estos equipos pueden orientarse hacia un enfoque más vertical en el ataque, o, como se describe en inglés, pueden optar por *estacionar el autobús*, haciendo énfasis en una estrategia defensiva y restringiendo las posibilidades de interacción en la red táctica.

En última instancia, la densidad de un grafo resalta la naturaleza y la extensión de las tácticas utilizadas por un equipo, proporcionando una perspectiva valiosa sobre su estilo de juego y su enfoque tanto en la posesión del balón como en la organización defensiva.

La centralidad de intermediación es una métrica empleada en el análisis de grafos con el propósito de discernir la relevancia relativa de los vértices en función de su ubicación y contribución a las rutas más cortas presentes en la

red.

La **centralidad de intermediación** de un vértice destaca la posición del mismo en relación con los caminos más cortos existentes entre otros vértices. Esencialmente, mide en qué medida un vértice funge como intermediario en la comunicación entre otros vértices dentro de la red. Un alto valor de centralidad de intermediación sugiere que el vértice es crucial para la conectividad y la eficiencia en la transmisión de información o influencia entre distintos componentes de la red.

La centralidad de intermediación es una herramienta esencial en el análisis táctico, ya que permite destacar la relevancia estratégica de ciertos jugadores en función de su capacidad para actuar como intermediarios clave en las comunicaciones o su papel destacado en las estructuras modulares del equipo.

Propiedad. *La **centralidad de intermediación** mide el grado en que un vértice actúa como puente o intermediario en la comunicación entre otros vértices dentro de un grafo. Formalmente, la centralidad de intermediación de un vértice se calcula como la proporción de los caminos (sin ciclos) entre todos los pares de vértices que pasan por ese vértice. Un vértice con alta centralidad de intermediación tendrá un mayor control sobre la comunicación y la transferencia de información dentro del grafo.*

La fórmula matemática para calcular la centralidad de intermediación de un vértice v en un grafo $G = (V, E)$ se puede expresar como:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{C_{st}(v)}{C_{st}}$$

donde C_{st} es el número total de caminos desde el vértice s al vértice t, y $C_{st}(v)$ es el número de esos caminos que pasan por el vértice v.

La centralidad de intermediación se asemeja al papel de un jugador que funge como el *puente* en una táctica. Cuando un jugador presenta una alta centralidad de intermediación, desempeña el papel crucial de ser el enlace principal en el flujo de información táctica. En otras palabras, este jugador ejerce la función de jugador clave en la transmisión eficiente de la estrategia y la coordinación entre diferentes componentes del equipo. En pocas palabras a mayor centralidad de un jugador, hay más caminos en donde el balón pasa

por ese jugador y no otros. Su posición estratégica lo convierte en un elemento fundamental para mantener la cohesión y la eficacia en las interacciones tácticas dentro del equipo.

Esta métrica de centralidad resulta invaluable para la identificación de vértices clave en un grafo, permitiendo una comprensión profunda de su función en la transmisión de comunicación, la difusión de información o la transferencia de recursos estratégicos. La centralidad de intermediación se enfoca en la capacidad de un vértice para ejercer control sobre el flujo de información en la red y evalúa la conexión de un vértice con otros vértices influyentes presentes en el grafo.

En cierto sentido, los vértices con alta centralidad son semejantes a los *amigos influyentes* dentro de la red social táctica. Estos jugadores ejercen un papel determinante en la configuración y el éxito de las estrategias predominantes en el equipo, al estar conectados con otros jugadores altamente influyentes en el esquema de juego. Esta medida resalta la importancia de estos jugadores en la dinámica táctica, ya que su posición central favorece la propagación efectiva de tácticas y la cohesión en la ejecución de las estrategias del equipo.

Definición 8 (Subgrafo)**.** *Un* ***subgrafo*** G' *de un grafo* G *es un grafo cuyos conjuntos de vértices y aristas son subconjuntos de los de* G*. Se denota de la siguiente forma:*

$$G' \subset G$$

Un subgrafo se configura como un subconjunto táctico en el marco de un juego más amplio. Se trata de focalizarse en un grupo específico de jugadores y examinar cómo interactúan en situaciones particulares dentro del contexto general del partido. En esencia, es como analizar una sección específica del terreno de juego y estudiar las relaciones tácticas entre un subconjunto determinado de jugadores, lo que brinda una visión detallada de sus movimientos, estrategias y colaboraciones en momentos concretos del juego. Este análisis a nivel de subgrafo puede revelar patrones tácticos específicos y contribuir a una comprensión más profunda de cómo opera un equipo en diferentes contextos dentro del partido.

Definición 9 (Isomorfismo)**.** *Dos grafos se consideran isomorfos si pueden ser reorganizados de tal manera que sean estructuralmente idénticos. En otras pa-*

labras, los grafos isomorfos tienen la misma estructura de conexiones, aunque los nombres o etiquetas de los vértices pueden ser diferentes.

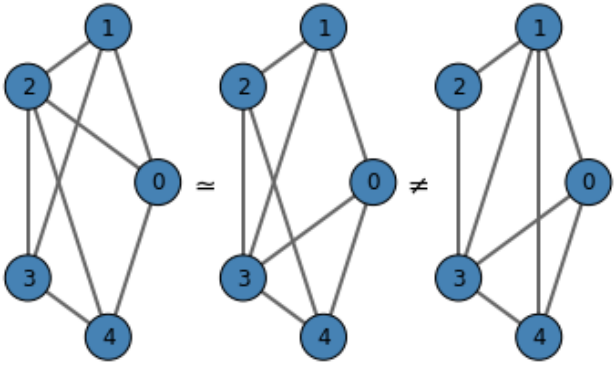

Figura 6: El primer grafo es isomorfo al segundo, pero el tercero no es isomorfo a ninguno de los otros.

En la Figura 6 podemos observar que el primer grafo G_1 y el segundo G_2 son isomorfos pues si reestructuramos a G_2 de tal forma que el vértice 3 ocupe la posición del vértice 2, entonces tenemos la misma estructura de conexiones aunque con etiquetas diferentes. El tercer grafo G_3 en cambio no es ismorfo en ningún sentido, pues no hay forma de reordenar los vértices para obtener la esrtuctura de G_1 o G_2.

Dos grafos isomorfos se refieren a dos equipos que comparten estrategias similares. A pesar de que los jugadores en ambos equipos pueden tener nombres distintos, los patrones tácticos son idénticos en términos de su estructura y relaciones. Esta noción implica que las configuraciones tácticas y las interacciones entre los jugadores en ambos equipos son esencialmente iguales, aunque los individuos involucrados puedan variar. El concepto de isomorfismo en este contexto subraya la similitud fundamental en las decisiones tácticas y las dinámicas de juego adoptadas por ambos equipos, lo que puede sugerir influencias mutuas o la adopción de enfoques tácticos compartidos provenientes de entrenadores, estilos de juego o estrategias comunes.

Definición 10 (Matriz de Adyacencia)**.** *La matriz de adyacencia es una representación matricial de un grafo, donde las filas y columnas representan los*

vértices del grafo. El valor en la intersección de la fila i y la columna j indica si existe una arista que conecta los vértices i y j.

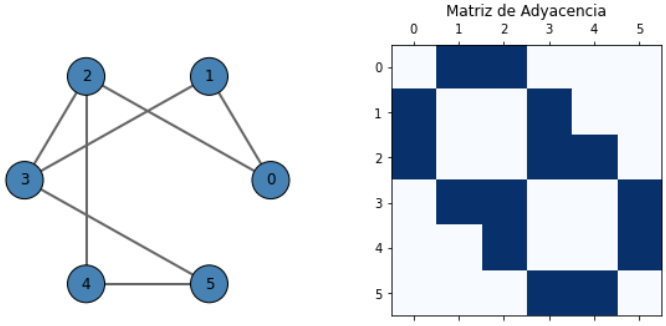

Figura 7: Construcción de un grafo a partir de una martriz de adyacencia

La matriz de adyacencia es la representación tabular de las conexiones entre jugadores en un grafo. Esta matriz proporciona una visualización rápida y concisa de las interacciones entre los jugadores y cómo contribuyen a la táctica global del equipo. En esencia, cada celda de la matriz muestra si existe una conexión (arista) entre dos jugadores particulares en la red táctica. Este enfoque visual permite una comprensión inmediata de quiénes están vinculados entre sí en términos tácticos y cómo se integran en la estrategia general del equipo. La matriz de adyacencia es una herramienta valiosa para analizar la estructura de las relaciones en un equipo y evaluar la colaboración y la cohesión entre los jugadores en diversas situaciones de juego.

En algunos casos, las aristas de un grafo pueden tener asociados valores numéricos o **pesos** que representan alguna medida o atributo. Estos pesos pueden representar distancias, costos, tiempos, etc. Los grafos con peso son utilizados para modelar situaciones donde se requiere considerar la magnitud de las relaciones entre los vértices.

Visualizemos los pesos en las aristas como valores que asignas a las acciones tácticas. Estos pesos pueden representar la significancia de un pase preciso o la relevancia de un movimiento específico en el juego. En otras palabras, los pesos reflejan la ponderación que se otorga a ciertas interacciones tácticas en función de su impacto en la estrategia general. Cada peso se convierte en una medida cuantitativa que subraya la importancia relativa de una acción táctica

en comparación con otras. A través de la asignación de pesos, se logra una evaluación más precisa y detallada de las decisiones tácticas y su influencia en la consecución de los objetivos del equipo.

2.1.2. Redes Sociales

Las redes sociales son una forma particular de redes en las cuales los nodos (vértices) representan individuos o actores sociales, y las conexiones (aristas) entre ellos representan las interacciones o relaciones sociales. Estas interacciones pueden ser de diversos tipos, como amistad, colaboración, intercambio de información, influencia, entre otros. El estudio de las redes sociales se enfoca en comprender cómo se estructuran y cómo influyen en diversos fenómenos sociales.

En una red social, el conjunto de nodos puede representar a personas, organizaciones, clústers o cualquier otro tipo de entidad social. La conexión entre dos nodos se establece cuando existe una relación o interacción entre ellos, y esta conexión puede ser simbolizada por una arista en el grafo que representa la red.

La representación de una red social se realiza mediante una matriz denominada sociomatriz, que es una matriz cuadrada de dimensiones $n \times n$, donde n es el número de nodos en la red. En esta matriz, los valores $Y_{(i,j)}$ indican la existencia o ausencia de una conexión entre los nodos i y j. Si $Y_{(i,j)} = 1$, significa que existe una conexión entre los nodos i y j, mientras que si $Y_{(i,j)} = 0$, indica que no hay conexión entre ellos. En pocas palabras la sociomatriz de la teoría de redes sociales y la matriz de adyacencia de teoría de grafos son lo mismo con diferentes nombres. Por lo anterior de ahora en adelante se usará sólo el nombre de **matriz de adyacencia**.

Es importante mencionar que, en una red social, las conexiones suelen ser bidireccionales, es decir, si existe una conexión de i a j, es decir; $i \to j$, también existe una conexión de j a i, es decir; $j \to i$. Esto se refleja en la simetría de la matriz de adyacencia, donde $Y_{(i,j)} = Y_{(j,i)}$ si la red es no dirigida. Esta bidireccionalidad implica que la relación entre dos nodos puede ser recíproca y que la información o influencia puede fluir en ambas direcciones.

Además, se ha observado que las relaciones en las redes sociales tienden a ser transitivas. Esto significa que, si hay una conexión entre los nodos i y j, y otra conexión entre los nodos j y k, es más probable que exista también una conexión directa o indirecta entre los nodos i y k. Esta propiedad de transitividad indica que los nodos que están conectados a través de intermediarios comparten similitudes o están más cercanos en el espacio social de la red.

2.1.3. Análisis Táctico de Fútbol

Las redes sociales también son aplicables al ámbito del fútbol, donde los nodos representan jugadores y las conexiones entre ellos reflejan las interacciones y relaciones dentro del vestuario o del campo. Imaginemos una red social en la que los nodos son jugadores y las conexiones representan la interacción en forma de pases entre ellos durante un partido.

En esta red social futbolística, la sociomatriz se construiría de manera que los valores $Y_{(i,j)}$ indiquen si existe una conexión entre los jugadores i y j, es decir, si se han realizado pases entre ellos durante el partido. Si $Y_{(i,j)} = 1$, significa que ha habido un pase exitoso de i a j, mientras que si $Y_{(i,j)} = 0$, indica que no se ha establecido una conexión directa entre ellos.

Al profundizar en el análisis táctico del fútbol, la distribución del juego se convierte en un aspecto crucial para el cuerpo técnico. Este término se refiere a cómo el equipo distribuye y controla la posesión del balón en el campo.

Para los entrenadores, entender la distribución del juego significa comprender cómo fluye la pelota entre los diferentes jugadores y áreas del campo. Esto incluye observar cómo se mueve el balón desde la defensa hacia el centro del campo y hacia el ataque, así como identificar los patrones de juego que el equipo utiliza para crear oportunidades de gol.

La distribución efectiva del juego permite al equipo mantener la posesión del balón, controlar el ritmo del partido y crear oportunidades de ataque. Un equipo que distribuye bien el juego puede mantener la presión sobre el oponente y desestabilizar su defensa, lo que aumenta las posibilidades de marcar goles.

Además, la distribución del juego también está estrechamente relacionada

con la comunicación y la coordinación entre los jugadores. Un equipo que se comunica bien en el campo puede mover el balón de manera eficiente y anticiparse a las jugadas del oponente.

Para el cuerpo técnico, comprender la distribución del juego les permite identificar áreas de mejora en la táctica del equipo. Pueden analizar qué jugadores están involucrados en la construcción de juego y cómo se relacionan entre sí en el campo. Esto les ayuda a diseñar estrategias específicas para optimizar la distribución del balón y mejorar el rendimiento general del equipo.

Es importante destacar que este es solo un ejemplo de cómo las redes sociales pueden aplicarse al análisis táctico del fútbol, pero en este trabajo se mostrará un análisis más profundo de este tipo de redes sociales al trabajar con la incertidumbre de las acciones que pueden ejecutar los jugadores y estimarlas por medio de elementos de Estadística Bayesiana, volviendo a estas redes sociales, en la estimación de espacios latentes.

2.1.4. Espacios Latentes

Imaginemos una biblioteca que tiene a su disposición todos los libros que se han escrito a lo largo de la historia. Esta colección de libros es tan grande que es imposible organizarla a manera de orden alfabético, pero da la casualidad que tenemos un mapa mágico de la biblioteca. Este mapa en lugar de enseñarte la ubicación de cada libro en cada estante agrupa cada libro por su contenido, temática y estilo. La biblioteca representa el espacio de alta dimensión en donde se guardan los datos y el mapa mágico que agrupa los datos es a lo que llamamos **espacio latente**. Cuando se trabaja con modelos como el expuesto en esta tesis, lo que se busca es intentar obtener este mapa, como una versión comprimida de la información que explica de mejor forma la biblioteca.

Este espacio que se creará es todo lo que necesitamos. Al navegar alrededor de él podemos generar variaciones de lo aprendido a manera de patrones o relaciones que no las podríamos ver sin el espacio latente porque habría demasiada información al adentrarnos solos a la biblioteca sin el mapa.

Cada libro en la biblioteca puede considerarse un punto en el espacio de alta dimensión, pero dentro del espacio latente, este libro es representado no

por su ubicación física en un estante, sino por una **posición latente** que refleja su esencia: el género, la trama, el estilo de escritura, las emociones que evoca o cualquier otra característica.

La posición latente es como las coordenadas de cada libro en el mapa mágico: no te dice dónde está el libro físicamente, sino dónde se sitúa en un paisaje conceptual formado por las relaciones y similitudes con otros libros. Por ejemplo, en este mapa mágico, una novela de misterio de Agatha Christie podría estar cercana a otras obras de ficción detectivesca, mientras que un libro de texto de física cuántica estaría en una región completamente diferente, agrupado con otros textos científicos.

Del mismo modo, en este modelo, las posiciones latentes de los jugadores no indican su posición en el campo, sino su rol y estilo de juego dentro de la compleja dinámica del equipo, sus tendencias de posicionamiento, y su influencia en el juego. Al aprender y analizar estas posiciones latentes, somos capaces de entender y predecir comportamientos y patrones que serían demasiado complejos si intentáramos analizar todos los datos brutos a la vez.

Los datos del problema se pueden copnstruir como una matriz de adyacencia $Y \in \mathbb{R}^{n \times n}$ con entradas $Y_{(i,j)}$ que denotan la relación entre un actor (vértice) i con otro j, e información covariada $X_{(i,j)}$. Cuando se habla de información covariada nos referimos a información que se tiene sobre la relación de i con j. En nuestro caso esto puede ser un vector de métricas sobre la interacción de los jugadores como el porcentaje de pases completados entre los dos, zona promedio del campo ocupada o incluso métricas más avanzadas como *amenaza esperada* (*Expected Threat*) o la cadena de gol esperado (*Expected Goal Chain*). También se añade un vector parametral β a estimar que contendrá el intercepto y los coeficientes que acompañan a las covariables. El parámetro más importante en este modelo es Z que representa las posiciones latentes.

El enfoque es binario y parte de una regresión logística. Se toma independencia condicional $(Y_{(i,j)}|Z_i, Z_j, X_{(i,j)}, \beta \overset{\text{iid}}{\sim} Ber(p))$ como base para el modelo asumiendo que la presencia o ausencia de una relación entre actores es independiente al resto del sistema.

En un partido cualquiera si un 8 (centrocampista / interior) se relaciona con un 11 (extremo) en el sentido de posesión, es decir; los dos tienden a ocu-

par el mismo espacio del campo y aparte este mismo 8 se relaciona en el mismo sentido con el 4 (defensa central) ocurre lo siguiente. La relación $8 \rightarrow 11$ es completamente independiente respecto a la relación $4 \rightarrow 8$, pues el hecho de que exista esta segunda relación no implica la existencia o no existencia de la relación $4 \rightarrow 11$. A lo largo de la construcción del modelo observaremos que existe una reciprocidad y transitividad de estas relaciones.

En este caso particular tenemos aún vestuario completo de jugadores, tanto el once inicial como los suplentes. Con esto tenemos un total de 17 jugadores y por tanto una cantidad total de $17^2 = 289$ relaciones cada una de ellas será estimada acorde a características observable de los jugadores y sus relaciones entre ellos. Estas características son las que definirán la dimensionalidad de nuestra biblioteca. En este caso tenemos la posesión del jugador, el número de toques por zona, el porcentaje de pases completados y las interacciones entre jugadores (¿quién le pasó el balón a quién?). Esto es en total cuatro dimensiones de características, cada una con 289 datos que usaremos para crear nuestro espacio latente.

Nuestro interés está en conocer con que probabilidad se relacionan los jugadores. Esto quiere decir que necesitamos la probabilidad de la matriz de adyacencia Y. Gracias a que suponemos independencia condicional para cada relación ahora podemos plantearlo como el producto de cada entrada de la propia matriz de adyacencia. A estas probabilidades se les conoce como **probabilidades de díadas**

$$\mathbb{P}(Y \mid Z,\ X,\ \theta) = \prod_{i,j} \mathbb{P}(Y_{(i,j)} \mid Z_i,\ Z_j,\ X_{(i,j)},\ \beta), \tag{2.1}$$

donde:

- $X_{(i,j)}$ es información covariada
- β es un vector parametral a estimar
- Z son las posiciones latentes a estimar

El modelo que usaremos es uno de distancia y una parametrización conveniente de $\mathbb{P}(Y_{(i,j)} \mid Z_i,\ Z_j,\ X_{(i,j)},\ \beta)$ es el modelo de regresión logística (*logit*), en donde la probabilidad de una relación depende de la distancia *euclidiana*

entre $Z = (Z_i,\ Z_j)$, así como las covariables $X = X_{(i,j)}$ y $\beta = (\beta_0,\ \beta_1^\top)$. Dicha parametrización es la siguiente:

$$\mathbb{P}(Y_{(i,j)} \mid \eta_{(i,j)}) = \frac{e^{\eta_{(i,j)} Y_{(i,j)}}}{1 + e^{\eta_{(i,j)}}} \tag{2.2}$$

Si se divide la expresión 2.2 por su complementario, es decir, si se construyen sus *odds* (la probabilidad de estar relacionado entre la probabilidad de **no** estar relacionado). El uso de *odds* en la parametrización de un modelo logístico facilita las matemáticas y la interpretación estadística debido a que los *log-odds* se extienden sobre toda la línea real, permitiendo una relación lineal entre los predictores y la respuesta en escala *log-odds*. Además, las propiedades derivativas de la función logística simplifican los cálculos de optimización necesarios para la estimación de parámetros, como aquellos en la maximización de la verosimilitud, donde la transformación logarítmica convierte productos en sumas, haciéndola computacionalmente más eficiente y matemáticamente más manejable.

$$\begin{aligned} \text{odds}(Y_{(i,j)} = 1 \mid Z,\ X,\ \beta) &= \frac{\mathbb{P}(Y_{(i,j)} = 1 \mid Z,\ X,\ \beta)}{1 - \mathbb{P}(Y_{(i,j)} = 1 \mid Z,\ X,\ \beta)} \\ &= exp\{\beta_0 + \beta_1^\top X_{(i,j)} - |Z_i - Z_j|\} \end{aligned}$$

Si ahora realizamos su transformación con el logaritmo natural, se obtiene el siguiente resultado.

$$\begin{aligned} \eta_{(i,j)} &= \log \text{odds}(Y_{(i,j)} = 1 \mid Z_i,\ Z_j,\ X_{(i,j)},\ \beta_0,\ \beta_1) \\ &= \beta_0 + \beta_1^\top X_{(i,j)} - |Z_i - Z_j| \end{aligned} \tag{2.3}$$

Por tan sólo un momento supongamos el modelo sin covariables:

$$\beta_0 - |Z_i - Z_j|$$

Al analizar esta función parametral (sin covariables) es intuitivo ver que esa norma final de la diferencia entre Z_i y Z_j modela las distancias entre los jugadores i y j, pero si se reemplaza esta norma por un conjunto de distancias $\{d_{(i,j)}\}$, tales que satisfagan la desigualdad del triángulo $d_{(i,j)} \leq d_{(i,k)} + d_{(k,j)}\ \forall\{i,j,k\}$ ahora podemos interpretar mejor el modelo. A mayor distancia menor será la probabilidad de una relación y en su caso contrario a menor distancia, mayor la

probabilidad de relación. En este caso particular modelaremos estas distancias en una dimensión baja ($\mathbb{R}^2$) por razones de **parsimonia** y poder contar con un modelo interpretable.

Este modelo es inherentemente recíproco: Si $i \to j$, entonces la distancia $d_{(i,j)}$ probablemente sea tan corta, al ser un hecho que estos vértices se encuentran relacionados, volviendo así el evento $j \to i$ igual de probable, permitiendo así la reciprocidad del modelo, pues $d_{(i,j)} = d_{(j,i)}$. La transitividad no está asegurada, sin embargo si los tres nodos i, j, k se encuentran cerca en el espacio latente tal que $i \to j$ y $j \to k$, entonces existirá una alta probabilidad que $i \to k$.

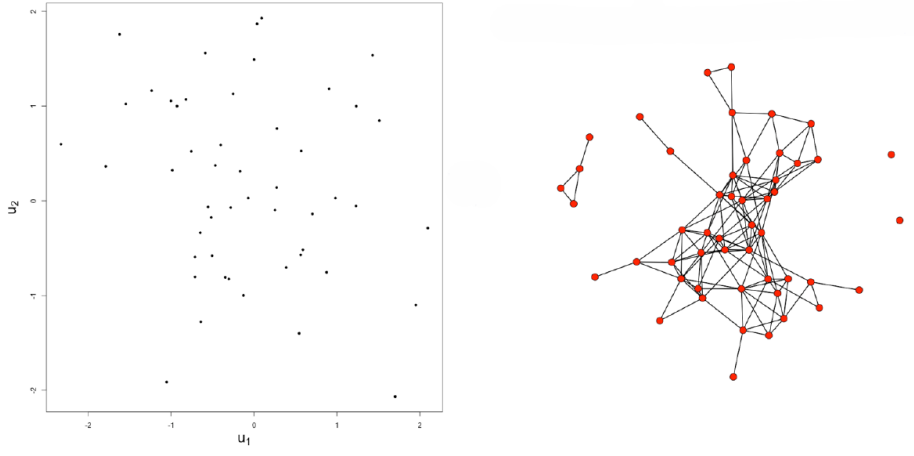

Figura 8: Jugadores cercanos en el espacio latente, son proclives a tener una probabilidad de estar relacionados

2.1.5. Estadística Bayesiana

En la búsqueda por desentrañar los secretos detrás de las tácticas del fútbol, se ha vuelto esencial recurrir a enfoques analíticos más avanzados. Una herramienta poderosa en esta búsqueda es la aplicación de estadística Bayesiana.

Se tomará una posición subjetivista en cuanto a la interpretación de la probabilidad. La explicación y defensa de esta postura se encuentran en el libro de *Theory of Probability* escrito por de Finetti y el siguiente fragmento del

prefacio resume de manera precisa el núcleo del argumento.

> *Lo único relevante es la incertidumbre: el alcance de nuestro propio conocimiento e ignorancia. El hecho real de si los eventos considerados están en algún sentido determinados, o conocidos por otras personas, y así sucesivamente, no tiene ninguna consecuencia.*
>
> (de Finetti, 1974, pp. Preface, xi-xii)

La estadística Bayesiana es una teoría racionalista de las creencias en contextos de incertidumbre, con el objetivo de caracterizar cómo debe actuar un individuo para evitar ciertos tipos de inconsistencias conductuales no deseadas. La teoría establece que la maximización de la utilidad esperada proporciona la base para la toma de decisiones racionales y que el teorema de Bayes proporciona la clave para la forma en que las creencias deben encajar entre sí a la luz de la evidencia. El objetivo es establecer reglas y procedimientos para personas preocupadas por la incertidumbre. La teoría no es descriptiva, en el sentido de pretender modelar el comportamiento real. Más bien, es prescriptiva, en el sentido de decir, "si deseas evitar la posibilidad de consecuencias no deseadas, debes actuar de la siguiente manera" (Bernardo & Smith, 2009, p. 4).

Desde el principio, el desarrollo de la teoría presume necesariamente un marco de discurso bastante formal, dentro del cual se pueden describir eventos inciertos y acciones disponibles, y se enuncian axiomas de comportamiento racional, los cuales no profundizaré aquí, pero cualquier persona interesada en consultarlos puede encontrarlos en el libro *Bayesian Theory* de José M. Bernardo y Adrian F. M. Smith.

Definición 11. *Si* $\{\eta_{(i,j)},\ i,j \in V\}$ *son hipótesis únicas e independientes, entonces para cada evento* Y *(datos),*

- $I\!P(\eta_{(i,j)}),\ i,j \in V$ *es la* ***distribución a priori*** *de* $\eta_{(i,j)}$*;*
- $I\!P(Y \mid \eta_{(i,j)}),\ i,j \in V$ *son* ***probabilidades*** *de* $\eta_{(i,j)}$ *dado* Y*;*
- $I\!P(\eta_{(i,j)} \mid Y),\ \forall i,j \in V$ *es la* ***distribución posterior*** *de* $\eta_{(i,j)},\ i,j \in V$*;*
- $I\!P(Y)$ *es llamada la* ***distribución predictiva*** *de* Y*.*

En este contexto, la distribución a priori ($\mathbb{P}(\eta_{(i,j)})$) representa la información previa disponible sobre la función parametral $\eta_{(i,j)}$. Puede ser informativa, reflejando conocimiento sólido previo, o no informativa, expresando falta de conocimiento específico. En particular describe la incertidumbre que se tiene sobre los parámetros.

La simulación, en particular el Método de Montecarlo por Cadenas de Markov (MCMC), es un método muy utilizado en estadística Bayesiana debido a la complejidad inherente de muchos modelos probabilísticos y la dificultad para calcular analíticamente la distribucion posterior.

Nuestro objetivo es actualizar nuestras creencias subjetivas sobre nuestra función parametral $\eta_{(i,j)}$ a medida que se observan nuevos datos. Esto se logra mediante la regla de Bayes:

$$\mathbb{P}(\eta_{(i,j)}|Y) \propto \mathbb{P}(Y|\eta_{(i,j)})\ \mathbb{P}(\eta_{(i,j)})$$

Así se actualiza la información previa (distribución a priori) con la probabilidad de $\eta_{(i,j)}$ dado Y. Sin embargo, en muchos casos, la forma exacta de la distribución posterior no tiene una solución analítica simple.

Aquí es donde entra en juego la simulación, especialmente el MCMC. MCMC permite aproximar la distribución posterior generando muestras de forma iterativa. En lugar de depender de cálculos analíticos, MCMC utiliza métodos numéricos para explorar el espacio de parámetros de manera eficiente y muestrear de la distribución posterior.

La simulación MCMC es especialmente valiosa cuando nos enfrentamos a modelos complejos o a situaciones con múltiples parámetros interdependientes. Su utilidad se destaca en la efectiva gestión de modelos jerárquicos y en escenarios donde la estructura del modelo no se ajusta fácilmente a métodos analíticos convencionales. Para llevar a cabo la simulación, se utilizará un algoritmo ampliamente conocido: el algoritmo de Metropolis-Hastings (véase Apéndice I).

Por otro lado, el concepto de *burn-in* es importante mencionarlo, pues es el periodo inicial de una cadena de Markov en la que se permiten transiciones antes de considerar que la cadena ha alcanzado la distribución estacionaria. Durante este período, la cadena se estabiliza y se aleja de la dependencia de

las condiciones iniciales. El *burn-in* se puede entender como una fase de entrenamiento de la simulación. Este proceso es fundamental para garantizar la validez de las inferencias realizadas posteriormente.

En el marco de la estadística Bayesiana, la selección de parámetros óptimos, como el parámetro $\eta_{(i,j)}$ en modelos estadísticos, se conceptualiza no solo como un problema de estimación, sino también como un problema de decisión. Esto se debe a que, más allá de calcular probabilidades, la inferencia bayesiana involucra tomar decisiones basadas en estas probabilidades, considerando las consecuencias de tales decisiones.

2.1.6. Función de Pérdida en Inferencia Bayesiana

Una vez obtenida la distribución posterior del parámetro $\eta_{(i,j)}$, es esencial decidir cuál valor del mismo deberíamos usar. Esta decisión se guía por una **función de pérdida** que cuantifica el costo asociado con la elección de un valor estimado cuando el valor verdadero es otro. En este modelo en particular se usará la función de pérdida cuadrática que se define como:

$$L(\eta_{(i,j)}, \hat{\eta}_{(i,j)}) = (\eta_{(i,j)} - \hat{\eta}_{(i,j)})^2, \tag{2.4}$$

donde $\hat{\eta}_{(i,j)}$ es una estimación puntual de $\eta_{(i,j)}$. El objetivo es seleccionar $\hat{\eta}_{(i,j)}$ para minimizar el valor esperado de esta pérdida dado la distribución posterior de $\eta_{(i,j)}$, es decir:

$$\hat{\eta}_{(i,j)} = \text{argmin}_{\hat{\eta}_{(i,j)}} \mathbb{E}[(\eta_{(i,j)} - \hat{\eta}_{(i,j)})^2 | Y].$$

Este procedimiento no solo implica estimar $\eta_{(i,j)}$ sino tomar una decisión sobre cuál valor de $\eta_{(i,j)}$ minimiza el error cuadrático medio esperado bajo la distribución posterior, lo cual constituye un problema de decisión.

2.1.7. Decisión Bajo Incertidumbre

La gestión de la incertidumbre es un principio fundamental de la inferencia Bayesiana. Elegir un estimador en este contexto implica ponderar diferentes valores posibles del parámetro, cada uno con su respectiva probabilidad, evaluados según el costo de la función de pérdida. Esto subraya que la decisión final en inferencia Bayesiana depende de cómo se define y evalúa el costo de

estar equivocados en nuestra estimación.

2.1.8. Utilidad Práctica de los P-values en Contextos Aplicados

Aunque la estadística Bayesiana típicamente no emplea **p-values** para determinar la credibilidad de los resultados, su uso sigue siendo prevalente en muchas aplicaciones prácticas, incluyendo el análisis del ámbito deportivo. Andrew Gelman, en su artículo *P Values and Statistical Practice* (Gelman, 2013), reconoce que, aunque los p-values no se interpretan adecuadamente como probabilidades posteriores del modelo nulo, pueden tener valor práctico bajo ciertas condiciones.

Gelman discute cómo los p-values pueden aproximarse a las probabilidades posteriores bajo distribuciones a priori que contienen poca información en comparación con los datos. En contextos aplicados, especialmente en aquellos donde la comunicación de resultados estadísticos es esencial y debe ser accesible, los p-values ofrecen un marco familiar para muchos investigadores y usuarios finales.

> *Bajo ciertas condiciones, un p-value unilateral para una mediana a priori proporciona un límite inferior aproximado sobre la probabilidad posterior de que la estimación puntual esté en el lado equivocado de esa mediana.*
>
> (Gelman, 2013)

Esta interpretación permite utilizar p-values como una herramienta complementaria en la inferencia estadística, proporcionando un punto de referencia útil sobre el impacto de la información a priori en la probabilidad posterior. Además, en muestras de tamaño pequeño, común en análisis de fútbol, las probabilidades posteriores dependen significativamente de la distribución a priori, haciendo que los p-values ofrezcan una perspectiva adicional útil para interpretar los efectos observados.

En esta tesis, empleo los p-values no como pruebas definitivas de hipótesis, sino como indicadores de áreas donde los modelos muestran discrepancias notables que merecen más investigación o confirmación a través de métodos

Bayesianos más rigurosos. Esta aproximación asegura que se mantenga una interpretación prudente y contextualizada de los resultados, en línea con las mejores prácticas estadísticas y las necesidades específicas del análisis en el deporte.

> *La interpretación directa de los p-values como probabilidades posteriores puede ser un punto de partida útil, si reconocemos que tales resúmenes estiman sistemáticamente la fuerza de las afirmaciones de cualquier conjunto de datos en particular.*
>
> (Gelman, 2013)

2.1.9. Paqueterías en *R*

Dentro del ámbito del análisis de redes, la biblioteca *igraph* ha sido fundamental en la investigación. Esta herramienta de código abierto y eficiente se ha convertido en el recurso principal para comprender las relaciones complejas presentes en las interacciones tácticas. Su núcleo en C garantiza operaciones rápidas y efectivas, y su capacidad para conectarse con lenguajes de alto nivel como *R* ha facilitado su integración en esta investigación.

A lo largo de este trabajo de investigación, se han aprovechado las capacidades de *igraph* en *R* para explorar la estructura y las conexiones de las redes tácticas en el fútbol. Desde crear y modificar nodos y aristas hasta calcular métricas de centralidad y detectar patrones estratégicos, igraph ha sido un recurso valioso en cada etapa del análisis.

Sin embargo, donde *igraph* ha brillado especialmente es en la visualización. La capacidad de traducir las complejas relaciones en representaciones visuales ha sido esencial para comprender las interacciones tácticas. Esta librería ha permitido transformar patrones y conexiones en visualizaciones claras y significativas.

En el contexto de esta investigación, la librería *network* ha demostrado ser una herramienta esencial para el manejo de datos relacionales. Esta librería ofrece una gama de funciones para crear y modificar objetos de red, lo que ha facilitado el trabajo en el análisis de las interacciones complejas presentes en el estudio.

La librería es especialmente valiosa debido a su capacidad para representar diversos tipos de datos representados como matrices de adyacencia. Desde relaciones simples hasta conexiones más complejas, esta librería ha brindado la flexibilidad necesaria para abordar una amplia gama de situaciones. Además, su soporte para atributos de vértices, aristas y grafos ha aportado mucho al análisis, pues permitió incorporar información adicional en el modelo.

En la práctica, *network* ha agilizado la creación y modificación de objetos de red, lo que ha sido fundamental para desarrollar los análisis. La capacidad de agregar atributos a vértices, aristas y grafos ha enriquecido la representación de las relaciones y ha proporcionado una visión más completa de los datos. Esto ha sido particularmente útil al explorar las complejas interacciones presentes en el análisis táctico del fútbol.

La librería *latentnet* se utiliza para ajustar modelos de efectos aleatorios de espacio latente, donde la probabilidad de una red g en un conjunto de nodos es el producto de probabilidades de díadas 2.1. Cada una de estas probabilidades de díadas es un modelo lineal generalizado con un componente lineal que se explicará más a detalle en el siguiente capítulo.

El resultado de un ajuste de modelo de variables latentes es un objeto. Por lo tanto, las funciones summary, print y plot se aplican a los ajustes. La función plot tiene muchas opciones específicas para modelos de variables latentes.

En el enfoque de análisis de redes tácticas en el fútbol, se ha aprovechado la potencia de la librería *latentnet* para ajustar modelos de efectos aleatorios de espacio latente. Al incorporar covariables de díadas, posiciones en el espacio latente y efectos de emisor y receptor, se ha podido desglosar las complejidades de las interacciones tácticas.

En la especificación de los modelos se ha utilizado la sintaxis proporcionada por la librería para definir las relaciones y características relevantes. La flexibilidad de esta librería ha permitido crear un modelo preciso y personalizado, lo que ha sido esencial para abordar la naturaleza única de las interacciones tácticas.

Además, han brindado herramientas para evaluar y visualizar el modelo. Gracias a las funciones de resumen, impresión y gráficos específicos para los

ajustes de variables latentes, se puede comunicar de manera efectiva los hallazgos y representar visualmente las relaciones identificadas en el análisis.

3. Diseño de Investigación

3.1. Metodología de investigación

Para esta investigación se requiere inicialmente de la elaboración de un modelo gráfico, es decir; de un grafo inicial construido a partir de conocimiento sobre el equipo rival a estudiar y dada la enorme cantidad de datos se utilizarán tan sólo características particulares de los jugadores, generando así un grafo que representará al equipo rival en dichas características. Iniciamos el proceso con el análisis de posesión relevante del equipo rival.

Al hablar de posesión relevante nos referimos a aquellas variables que muestran una tenencia del balón de manera puntal por jugador. Dichas variables serían los toques de balón (tanto de forma absoluta como porcentual referente al número de toques totales del equipo), porcentaje de posesión por zona del campo, pases completados (tanto de forma absoluta como porcentual). La zona en la que un jugador tiene posesión del balón es sumamente importante, pues es la característica que diferencia a una posesión cualquiera a una posesión relevante. A manera de ejemplo digamos que hay dos equipos A y B, el equipo A tiene un número total de toques del balón de 345 en un partido con posesión del 60 % y el equipo B tiene 210 toques y 40 % de la posesión del balón, ambos en un mismo partido. Estos son datos clásicos que salen en la televisión y gran parte de la gente diría que el equipo A tuvo mayor posesión relevante. Sin embargo, esto no es necesariamente cierto sólo con esos datos. Digamos que el equipo A tuvo dicha posesión en su propia mitad del campo y que la mayoría de pases fueron pases de control al portero o sobre la línea defensiva. Mientras que el equipo B tuvo la mayoría de su posesión en el último tercio del campo rival. El equipo con mayor posesión relevante bajo esta nueva información fue el equipo B aún con menos toques de balón y menor porciento de posesión, pues tuvieron más relevancia en el juego de posesión.

Es por esto que la posesión relevante incluye una clasificación de los toques y de la posesión acorde a las zonas del campo en las que estas fueron registradas. La división de las zonas del campo es la siguiente:

- Defensa Área Penal (1 a 3)
- Primer Tercio (Defensa) (4 a 6)

- Segundo Tercio (Mediocampo) (7 a 12)
- Última Tercio (Ofensiva) (13 a 15)
- Ataque Área Penal (16 a 18)

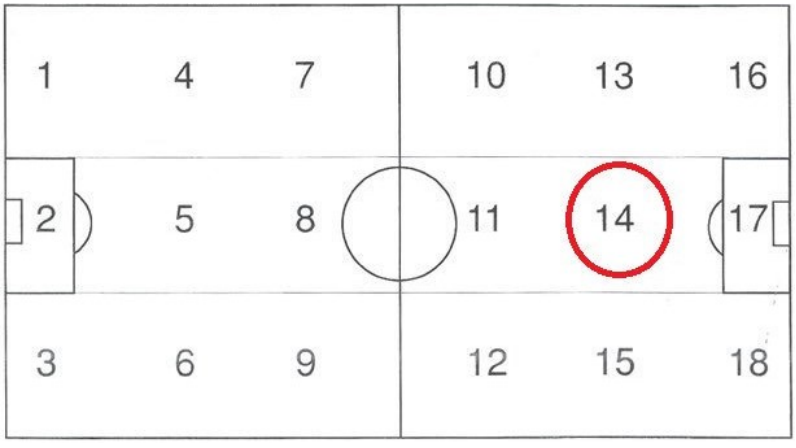

Figura 9: Representación del campo en 18 zonas

Existen muchas forma de definir estas zonas, pues estas zonas tienden a ser definidas por el entrenador en turno. En ocasiones estas divisiones no pueden ser representadas a manera matricial o son demasiado específicas. Por esta razón se opta por una división más clásica en la que simplemente se parte el campo en 3×6. Esta es una división clásica que muchos equipos adoptan para evitar confusiones tácticas a la hora de alinear a los jugadores. En este caso particular se puede notar el énfasis en la zona 14. Esta es una zona muy importante en el equipo, pues es justamente el espacio entre la línea defensiva del rival y su mediocampo. El tener posesión relevante ahí es una de las mejores señales de dominancia en el juego y esta será muy importante en el análisis táctico final.

Tras definir estas zonas del campo se hará la agrupación de variables por zona y por jugador y particularmente haremos una limpieza de los datos. La limpieza se hará eliminando a los jugadores que tuvieron una cantidad de minutos promedio por partido menor a veinte y al portero. La racionalización detrás de este criterio de tiempo jugado proviene de la creencia en que no podemos juzgar el rendimiento de un jugador con menos de veinte minutos

de juego en tres partidos. En este caso eliminamos al portero principalmente por el número de toques que tiene y dado que nos consta que su posesión del balón siempre será en su propia área penal y no existe una distribución de su presencia en el campo en otro lado.

Iniciaremos con la construcción inicial de un grafo que funcionará como entrada de datos inicial para el modelo. La construcción parte del grafo $G = (V, E)$, donde el conjunto de vértices es el conjunto de jugadores que tiene el rival. La elección de las aristas parte de una idea más creativa. Tenemos jugadores como vértices, pero ¿cómo decidimos si un vértice se relaciona o no con otro en términos de posesión? Para ello generamos una matriz de adyacencia dadas ciertas reglas.

- Si un jugador i tiene un porcentaje de posesión y de toques menor al 20 % en una zona en particular, entonces la posesión relevante será cero en dicha zona, esto porque al tener tan pocos toques en un zona no se fue relevante en la misma. Sin embargo esto genera una duda, ¿Qué pasa si el porcentaje del jugador se distribuye de manera uniforme en todas las zonas? En ese caso entonces podemos asegurar que su posesión relevante no puede ser alta, pues es un jugador que practicamente va a donde quiere cuando quiere sin hacer caso a posición o rol. En caso de tener un porcentaje superior al 20 % se hará una ponderación de dos porcentajes dada por:

  $$pr_i = 0.60 \times posesión + 0.40 \times toques$$

  La ponderación asignada a la posesión (60 %) en comparación con los toques (40 %) refleja la importancia táctica de controlar el juego: poseer el balón implica no solo recibirlo, sino retenerlo, facilitando la construcción de jugadas y dictando el ritmo de juego. Así, esta fórmula ponderada no solo cuantifica la actividad, sino que valora más la posesión sostenida, para imponer presencia táctica en una zona.

- Si dos jugadores tienen una posesión relevante diferente de cero, entonces se calculará:

  $$d = \frac{pr_i + pr_j}{max\{pr_i,\ pr_j\}}$$

- Sea $M \in \mathbb{R}^{17\times 17}$ la matriz de adyacencia, entonces $M_{i,j} = d$

De esta forma relacionaremos a un jugador con otro si y sólo sí tienen una posesión relevante en una zona particular del campo con un valor d que define el peso de la arista. Los datos ocupados para esta construcción se basan en los datos recabados desde inicio de temporada hasta tres partidos antes del enfrentamiento. Ahora dada la matriz de adyacencia construida podemos entender a este grafo como el preprocesamiento de los datos. Estos datos se codifican como un grafo con la necesidad de explicar reglas de asociación para crear la visualización del propio modelo gráfico y además cada uno de los nodos en este grafo le corresponderá otro en el espacio latente. Estos nodos en el espacio latente no contarán por el momento con ninguna arista entre ellos, pues nuestro objetivo es estimar la probabilidad con la que existirán estas relaciones.

Retomando el enfoque Bayesiano tenemos ahora que llevar a cabo la inferencia Bayesiana, para ello primero tenemos que recordar la regla de Bayes y como está nos ayudará al proceso de inferencia y la propia exploración de la distribución posterior.

$$\mathbb{P}(\eta_{(i,j)}|Y_{(i,j)}) \propto \mathbb{P}(Y_{(i,j)}|\eta_{(i,j)})\ \mathbb{P}(\eta_{(i,j)})$$

La primer probabilidad ya la conocemos y la desarrollamos en el Capítulo 2 ecuación 2.2. Esta se encuentra multiplicando a la probabilidad de nuestra función parametral $\eta_{(i,j)}$ que conocemos como nuestra distribución a priori. Esta distribución ha de estar condicionada a un conocimiento previo, permitiendo generar una distribución a priori a ser después actualizada con el resto de información para conocer la distribución posterior y en su defecto la predictiva para poder contestar nuestras preguntas planteadas desde la introducción de este trabajo.

La pregunta es ahora ¿cómo se define esta distribución a priori? Afortunadamente *Handcock* nos da la solución en la paquetería de R al definir una distribución a priori a cada parametro dentro de $\eta_{(i,j)}$ de la siguiente forma:

$$\eta_{(i,j)} = \beta_0 + \beta_1 X_{(,j)} - |Z_i - Z_j|,$$

donde:

- $\beta_0 \sim N(0, \psi^2)$, para el intercepto.
- $\beta_1 \sim N(0, \psi^2)$, para el coeficiente de la covariable.
- $Z_k \overset{\text{iid}}{\sim} N_2(0, \sigma^2 I)$, para las posiciones latentes.

Una vez planteadas estas surge una duda respecto a ψ^2 y σ^2, pues estas aún no están definidas, pero para la gran mayoría de casos en el análisis de redes se ha llegado a una heurística que las define. Al no ser muy relevante en el enfoque de esta tesis, la heurística estará definida en el Apéndice I.

Este es el modelo de la estructura en posesión, la información de pases completos también se encuentra dentro del modelo a manera de covariable $(X_{(i,j)})$, es decir; información única entre dos jugadores. Lo anterior implica que el grafo estimado tiene sus conexiones gracias a datos de posesión y su clasificación se hará por medio de los datos de relación entre jugadores. Antes de enfocarnos en el desarrollo del modelo hemos de mostrar como obtuvimos los datos y como los trabajamos.

Ahora presento el grafo dado como preprocesamiento de los datos para tener una idea inicial del equipo rival. Donde las conexiones como ya se mencionó provienen de las métricas de posesión y en este caso por razones de visualización, el tamaño de los nodos es dependiente de la información de pases completos de cada jugador.

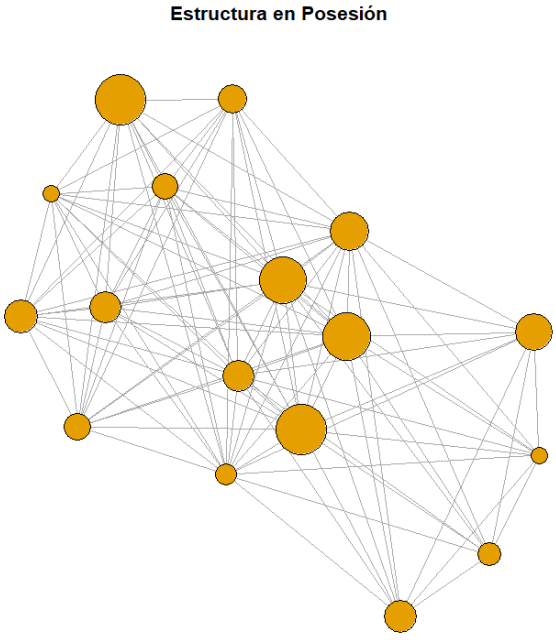

Figura 10: Grafo preprocesado del equipo rival

3.2. Fuentes y recolección de datos

La fuente de nuestros datos proviene de StatsBomb, una empresa dedicada a la recolección y análisis de datos de fútbol. En particular el quipo estudiado con este modelo como si fuese nuestro rival es el Fútbol Club Barcelona (FCB). La descarga se puede hacer directamente con ellos o por medio de un concurso que se abrió en el 2022 Proveniente de Statsbomb que cuenta ya con una enorme lista de datos con muchas de características separadas por variables y cada una organizada en un archivo *.csv*.

3.3. Análisis y Modelado

Como se habló en el Capítulo 2 el modelo con el que evaluaremos estos datos es un modelo de distancia euclidiana y un detalle importante es que nuestros datos de posesión generan un grafo no dirigido y por tanto la reciprocidad ($Y_{(i,j)} = Y_{(j,i)}$) existe inherentemente.

En contraste con otros modelos, la estimación de los parámetros en la *log-verosimilitud* del modelo de independencia condicional es relativamente sencillo.

$$\begin{aligned}
\log \mathbb{P}(Y \mid \eta) &= \log \prod_{i \neq j} \mathbb{P}(Y_{(i,j)} \mid \eta_{(i,j)}) \\
&= \sum_{i \neq j} \log \mathbb{P}(Y_{(i,j)} \mid \eta_{(i,j)}) \\
&= \sum_{i \neq j} \log \frac{e^{\eta_{(i,j)} Y_{(i,j)}}}{1 + e^{\eta_{(i,j)}}} \\
&= \sum_{i \neq j} [\log(e^{\eta_{(i,j)} Y_{(i,j)}}) - \log(1 + e^{\eta_{(i,j)}})] \\
&= \sum_{i \neq j} [\eta_{(i,j)} Y_{(i,j)} - \log(1 + \exp^{\eta_{(i,j)}})],
\end{aligned}$$

donde η es la función parametral con intercepto β_0, posiciones latentes Z_i y Z_j y covariables $X_{(i,j)}$ (pases completos) acompañado por el coeficiente β_1. Así, métodos basados en verosimilitud como **máxima verosimilitud** e **inferencia Bayesiana** son factibles.

Una complejidad añadida al modelo es el problema de terminar haciendo inferencia sobre una infinidad de posiciones latentes Z. Las rotaciones en el espacio euclidiano mantienen las distancias entre puntos porque conservan la longitud de los vectores. Las reflexiones no alteran las distancias relativas entre puntos, ya que conservan la distancia entre ellos. Las traslaciones desplazan todos los puntos por la misma cantidad en la misma dirección, manteniendo así la diferencia de distancia entre cualquier par de puntos. De esta forma las distancias entre puntos en el espacio euclidiano son invariantes bajo rotaciones, reflexiones y traslaciones. Entonces, para cada matriz $n \times 2$ de posiciones latentes Z existe una infinidad de otras posiciones que dan la misma *log-verosimilitud*, es decir;

$$\log \mathbb{P}(Y|Z) = \log \mathbb{P}(Y|Z^*)$$

Para cualquier Z^* que sea equivalente a Z bajo las operaciones de reflexión, rotación o traslación. Sea $[Z]$ la clase de posiciones equivalentes a Z bajo las

operaciones anteriores, entonces para cada $[Z]$, existe un conjunto de distancias entre vértices al cual llamaremos una **configuración**.

Nuestro interés está en evitar la infinidad de otras posiciones que dan la misma *log-verosimilitud* y por ello aplicamos a la matriz Z la transformación de **Procrustes**. El método Procrustes busca encontrar una transformación que minimice la discrepancia entre dos conjuntos de puntos, representados por las matrices Z y Z_0. Estas matrices contienen las posiciones de los puntos en las configuraciones que se están comparando.

La definición de la transformación Procrustes se expresa como:

$$Z^* = \text{argmin}_{TZ}\ \text{tr}(Z_0 - TZ)^\top (Z_0 - TZ), \tag{3.1}$$

donde T es una matriz de transformación que representa rotaciones, reflexiones y traslaciones. El objetivo es encontrar la matriz Z^* que minimiza la suma de las diferencias cuadráticas entre los puntos de Z y Z_0 bajo estas transformaciones.

La solución Z^* se conoce como la transformación de Procrustes de Z, y representa la configuración de puntos más cercana a Z_0 en términos de la distancia euclidiana. Es importante destacar que al asumir que Z y Z_0 están centradas en el origen, entonces podemos dar una formulación alternativa para calcular Z^*:

$$Z^* = Z_0 Z^\top (Z Z_0^\top Z_0 Z)^{1/2} Z,$$

Dada información a priori sobre los parámetros $\beta_0,\ \beta_1$ y Z, nuestro procedimiento para muestrear la distribución posterior es el siguiente:

1. Se identifica un estimador de máxima verosimilitud $\hat{Z}$ de las posiciones latentes Z, centrado en el origen ($Z_0 = \hat{Z}$) que será nuestro valor inicial y con él construimos una Cadena de Markov como se muestra a continuación:

   a) Muestrea una propuesta $\check{Z} \sim N(Z_k, \sigma^2)$

   b) Calcular la tasa de aceptación:

$$\alpha = \frac{\mathbb{P}(Y|\check{Z},\ \beta_{0,k},\ \beta_{1,k},\ X)\pi(\check{Z})}{\mathbb{P}(Y|Z_k,\ \beta_{0,k},\ \beta_{1,k},\ X)\pi(Z_k)},$$

1) Generar un número aleatorio uniforme $u \in [0, 1]$
2) Si $u \leq \alpha$, entonces se acepta $Z_{k+1} = \check{Z}$
3) Si $u > \alpha$, entonces se rechaza $Z_{k+1} = Z_k$

c) Calcula y guarda $\tilde{Z}_{k+1} = \text{argmin}_{TZ_{k+1}} \text{tr}(\hat{Z} - TZ_{k+1})^\top (\hat{Z} - TZ_{k+1})$

2. Actualiza β_0 y β_1 con el algoritmo de **Metrópolis-Hastings**.

Gracias a que cada configuración puede ser representada por su Procrusteana, la distribución posterior de la configuración alrededor de $\hat{Z}$ es representada por muestras de $\tilde{Z}$ provenientes de la Cadena de Markov.

En este enfoque, las entradas de la matriz de adyacencia, denotada como $Y = [Y_{(i,j)}]$ y asociada con un grafo $G = (V, E)$, son condicionalmente independientes y siguen una distribución Bernoulli:

$$Y_{(i,j)}|\beta_0,\ \beta_1,\ X_{(i,j)},\ Z_i,\ Z_j\ \sim\ \textit{Bernoulli}(g(\beta_0 + \beta_1 X_{(i,j)} + \mu(Z_i,\ Z_j))),\ (3.2)$$

donde:

- $g(\cdot)$ es una función de liga, por ejemplo, la función logit inversa *expit(x)* o la función de distribución de una Normal estándar $\Phi(x)$.
- β_0 es el intercepto.
- β_1 es el coeficiente asociados con la covariable (pases completos $X_{(i,j)}$).
- $Z_i = (Z_{i,1},\ Z_{i,2})$ es un vector de variables latentes definidas en un espacio euclidiano 2-dimensional conocido como Espacio Latente.
- $\mu(\cdot,\ \cdot)$ es una función simétrica, que puede tomar diferentes formas:
  - **Modelo de distancia:** $\mu(Z_i,\ Z_j) = -|Z_i - Z_j|$. Esta es la que se usó.
  - **Modelo bilineal:** $\mu(Z_i,\ Z_j) = Z_i^\top Z_j$.
  - **Modelo factorial:** $\mu(Z_i,\ Z_j) = Z_i^\top \Lambda Z_j$, donde $\Lambda = diag(\lambda_1,\ \lambda_2)$ es una matriz diagonal de 2×2.

3.4. Especificaciones del Modelo

Para aplicar este enfoque, es necesario especificar:

1. La dimensión latente 2.
2. La distribución de las variables latentes.
3. La forma funcional de $\mu(\cdot,\ \cdot)$.
4. La función liga.

4. Análisis de Resultados

En este capítulo, se profundizará en el análisis de los resultados derivados de la implementación del modelo previamente planteado. La metodología Bayesiana adoptada se apoya en la aplicación del método *Monte Carlo Markov Chain* (**MCMC**), una técnica eficaz para explorar la distribución posterior de los parámetros del modelo.

La elección del MCMC se fundamenta en la necesidad de abordar la distribución compleja de los parámetros. Este enfoque numérico facilita la generación de muestras aleatorias. La justificación radica en su capacidad para gestionar la dimensionalidad del espacio de parámetros. El método es capaz de proporcionar estimaciones de parámetros y predicciones que son confiables y estables incluso cuando se enfrenta a diferentes condiciones y supuestos. Esto implica que el MCMC puede adaptarse a la complejidad de la distribución de los parámetros y proporcionar resultados consistentes, independientemente de la naturaleza de la incertidumbre presente en los datos.

En cuanto a los diagnósticos MCMC, se lleva a cabo un análisis exhaustivo que garantiza la convergencia del algoritmo y la calidad de las muestras generadas. Se emplea el trazado de autocorrelación, asegurando la fiabilidad de los resultados obtenidos.

La distribución posterior, al ser representativa de la incertidumbre en los parámetros y su impacto en la formación de los espacios latentes, se explora minuciosamente. Este análisis implica examinar la media posterior de las posiciones latentes. Al tratarse de un modelo gráfico, esta media condensa la información de múltiples iteraciones del MCMC, ofreciendo una perspectiva consolidada de las posiciones latentes.

Dicha representación gráfica, derivada de la media posterior, proporcionará una visión detallada de cómo la incertidumbre en los parámetros se traduce específicamente en la distribución del espacio latente. Cada punto en esta distribución refleja una posible configuración de las posiciones latentes, brindando así una perspectiva holística de la complejidad del sistema modelado. Este enfoque permite capturar de manera efectiva la diversidad de escenarios potenciales y la influencia de la incertidumbre en la estructura subyacente del modelo. En resumen, la visualización de la distribución posterior, centrada en

la media de las posiciones latentes, se convierte en un recurso para comprender la incertidumbre y la complejidad en el contexto del análisis Bayesiano de espacios latentes.

En la sección dedicada a las caracterísicas del grafo con incertidumbre, se examina cómo el modelo Bayesiano de espacios latentes resume características como la densidad, transitividad, asortatividad, distancia promedio, grado promedio y desviación estándar del grado. Se analiza la influencia de la incertidumbre en estos parámetros.

Este capítulo se presenta como un pilar para interpretar y contextualizar los resultados, ofreciendo una visión detallada del modelo y su impacto en las propiedades de los grafos bajo estudio.

Antes de iniciar los diagnósticos, es necesario abordar un aspecto delicado. En el diseño del modelo, se propuso una metodología general para aplicar a los datos del problema. Sin embargo, obtener estos datos resultó ser particularmente desafiante. En nuestro caso, utilizamos datos sobre toques de balón y posesión por zona para construir nuestra matriz de adyacencia en el Capítulo 3, donde discutimos la covariable denotada como X, que representa los pases completos. Al evaluar el modelo general 2.3, obtuvimos resultados de inferencia sobre el intercepto β_0, el coeficiente de los pases completos β_1 además de nuestras posiciones latentes Z_i y Z_j. También realizamos pruebas sin considerar los pases completos, lo que nos llevó al modelo más simple:

$$\eta_{(i,j)} = \beta - |Z_i - Z_j| \tag{4.1}$$

Este modelo implica inferencia únicamente en el intercepto y las posiciones latentes, prescindiendo de las covariables.

Al comparar los resultados de ambos modelos, observamos que no hubo diferencias significativas en la distribución posterior ni en los resultados finales en términos de grafos y análisis táctico. Por consiguiente, concluimos que seleccionar el modelo más simple (4.1) es apropiado para realizar todos los diagnósticos.

En circunstancias diferentes, habríamos realizado pruebas adicionales considerando otras covariables o incluso un vector de covariables. El modelo 2.3 habría mantenido la misma estructura y resultados teóricos, siendo especial-

mente útil para lidiar con la alta dimensionalidad que implicaría el agregar más covariables. Aunque nuestra investigación mantiene su rigor, la disponibilidad limitada de datos nos impidió incorporar covariables en este estudio.

4.1. Diagnóstico del MCMC

Estos diagnósticos se utilizan para evaluar la convergencia y la autocorrelación de las Cadenas de Markov de Monte Carlo (MCMC) utilizadas en el análisis del modelo. Los siguientes gráficos muestran la autocorrelación de las muestras tomadas a diferentes intervalos de tiempo (*lags*) para las variables. La autocorrelación puede indicar si existen o no problemas de convergencia o autocorrelación serial en las muestras de MCMC, por ello se busca que la autocorrelación disminuya rápidamente a medida que aumenta el *lag*.

El concepto de *effective sample size (ESS)*, o tamaño efectivo de la muestra, es una medida importante en el diagnóstico del MCMC. El ESS muestra la cantidad de muestras independientes e informativas que se obtendrían de la cadena en lugar de la muestra autocorrelacionada real. En otras palabras, indica cuántas muestras verdaderamente útiles se tienen después de tener en cuenta la autocorrelación entre las muestras. Se califica al ESS no por el número puntual de estas muestras, sino pro el porciento respecto al total de muestras, es decir $\frac{\text{ESS}}{\text{Total de muestras}} \times 100$. La definición puntual del ESS y el cómo calcularlo se encuentra en el Apéndice I, Definición 13.

Un porcentaje bajo de muestras independientes e informativas sugiere que las muestras están altamente correlacionadas, lo que significa que se está aprovechando poco la información disponible. Por el contrario, un porcentaje alto indica que se tiene una buena cantidad de muestras independientes e informativas no correlacionadas entre ellas.

El cálculo del ESS evalúa la precisión de las estimaciones y para determinar cuántas muestras son necesarias para obtener una estimación confiable de los parámetros del modelo. En el diagnóstico del MCMC, se busca que el ESS sea lo más alto posible, lo que indica que se están aprovechando eficientemente las muestras generadas por la cadena MCMC. Por lo tanto, al examinar los gráficos de autocorrelación frente a los lags, uno de los objetivos es identificar si la autocorrelación disminuye rápidamente a medida que aumenta el lag, lo

que sugeriría que se está obteniendo un ESS adecuado y que la cadena MCMC está convergiendo de manera efectiva.

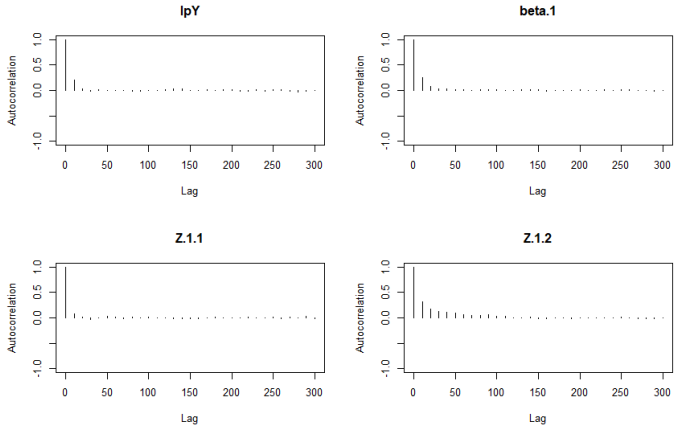

Figura 11: Autocorrelación para los datos Y, el intercepto β y las posiciones latentes Z

Como podemos notar, la rápida caída de la autocorrelación en nuestras variables indican que no parece haber problemas con las muestras de MCMC y nos indican que no hay problemas de convergencia. Un ESS como el obtenido en este caso (2706.422) con porciento respecto al total de 67.66 % indica que más de la mitad de las muestras son independientes y proporcionan una cantidad significativa de información para la estimación de parámetros y la realización de inferencias en el modelo. Este valor sugiere que la cadena MCMC ha generado muestras efectivas, lo que resulta en estimaciones confiables y precisas de los parámetros del modelo.

4.2. Cadena de MCMC e Histograma de la β

La **cadena de MCMC** es una herramienta esencial en nuestro análisis, pues modela la evolución dinámica de las muestras de los parámetros del modelo a lo largo de las iteraciones del proceso de MCMC. Esta cadena facilita la exploración de la distribución posterior de los parámetros, lo cual es fundamental para comprender la incertidumbre asociada a nuestras estimaciones.

A medida que la cadena avanza y se estabiliza, alcanza una distribución estacionaria, que idealmente refleja esta distribución posterior. En este contexto, la función de verosimilitud desempeña un papel crucial, ya que cuantifica la probabilidad de observar nuestros datos dadas las distintas configuraciones de parámetros en el modelo propuesto, influenciando así la forma y la convergencia de la cadena de MCMC hacia la distribución estacionaria.

La cadena de MCMC actúa como un registro detallado del recorrido de nuestro algoritmo, permitiéndonos observar cómo evoluciona la adaptación entre el modelo y los datos a través de diferentes combinaciones de parámetros. Este seguimiento es crucial para evaluar la convergencia de la cadena; específicamente, nos ayuda a verificar que, con suficientes iteraciones, la cadena se estabiliza en torno a una distribución estacionaria. Este punto de estabilidad indica que hemos alcanzado una solución consistente y que las iteraciones adicionales no divergen hacia valores atípicos o incoherentes.

Por otro lado, el **histograma de** β es una herramienta visual para comprender la distribución posterior del intercepto de nuestro modelo. Este histograma ofrece un resumen visual de los posibles valores de β derivados de nuestras 4000 iteraciones MCMC, lo cual nos permite apreciar la incertidumbre y la variabilidad del intercepto en nuestro análisis.

Al observar la distribución, podemos obtener información sobre la incertidumbre en torno a nuestro intercepto. ¿Dónde se concentra la mayor probabilidad en la distribución posterior? ¿Cuál es la dispersión alrededor de ese valor central? ¿Hay valores extremos que podrían tener un impacto significativo en nuestras conclusiones? Estas son las preguntas que podemos responder al examinar esta distribución.

En resumen, la cadena de MCMC nos muestra cómo nuestro modelo se ajusta a los datos a lo largo del tiempo, mientras que la distribución posterior de la β nos permite explorar la incertidumbre de uno de nuestros parámetros más importantes. Ambos son componentes esenciales para evaluar y comprender la calidad de nuestro análisis y la solidez de nuestras conclusiones.

En la Figura 12, se presenta tanto la cadena de Markov como la distribución posterior de β. La media de la distribución está indicada por una línea punteada azul y por líneas rojas el intervalo de credibilidad del 95 % está tam-

bién destacado. Esto implica que existe un 95 % de probabilidad de que el valor verdadero de β se encuentre dentro de este intervalo.

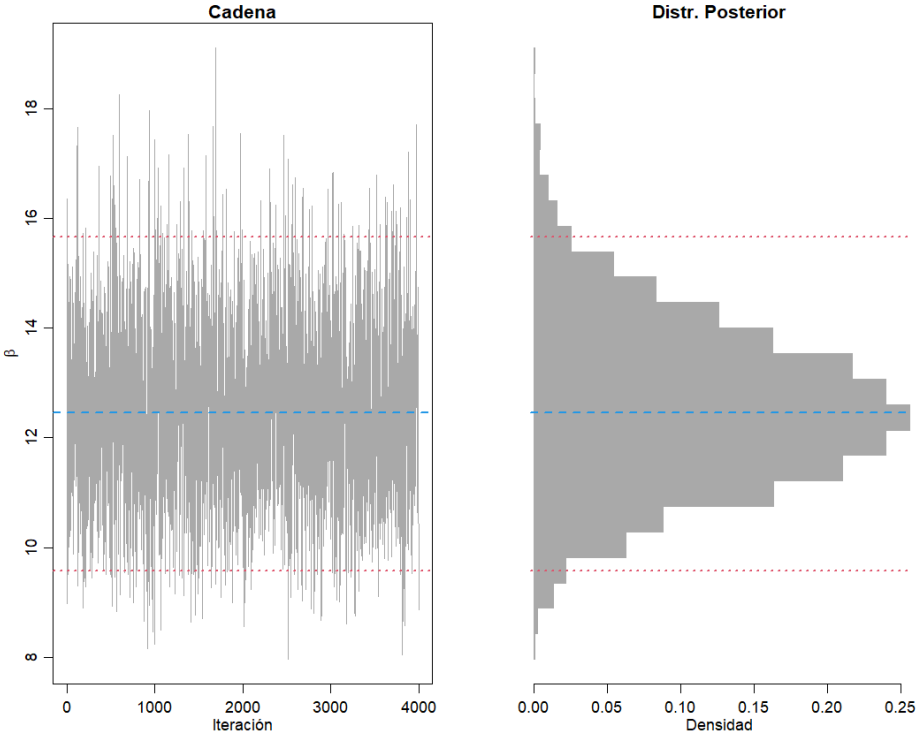

Figura 12: Cadena de MCMC y distribución posterior de β

4.3. Resúmen del Ajuste del Modelo

Este resúmen proporciona información crucial sobre la calidad del ajuste del modelo. Aqui están los puntos clave:

- **Fórmula:** Indica la fórmula utilizada para el modelo. En este caso está modelando la objeto llamado *fcb* (el grafo preprocesado) en función de una distancia euclidiana en un espacio latente bidimensional ($d = 2$).
- **Attribute:** Se refiere a la característica que se está modelando. En este caso, se trata de las aristas en el grafo.
- **Model:** Indica el tipo de modelo utilizado, que es un modelo Bernoulli como se indica en 3.2.

- **MCMC Sample:** Muestra el tamaño de la muestra MCMC (4,000 iteraciones) y cómo se tomaron las muestras (cada 10 iteraciones después de un *burn-in* de 10,000 iteraciones).
- **Intercept:** Proporciona la estimacion del intercepto en el modelo. En este caso, hubo un intercepto con una estimación de 12.4576 con pérdida definida en 2.4.
- **Significance Codes:** Estos códigos indican la significanca estadística del intercepto. *** significa altamente significativo (esto quiere decir que tiene un p-value menor a 0.001). Si bien el concepto de significancia estadística no tiene cabida en el marco Bayesiano se pueden ocupar de la forma en la que se estableció en la sección 2.1.8.

Aqui los resultados puntuales:

```
==========================
Summary of model fit
==========================

Formula:   fcb ~ euclidean(d = 2)
Attribute: edges
Model:     Bernoulli
MCMC sample of size 4000,
draws are 10 iterations apart,
after burnin of 10000 iterations.

Intercept:
             Estimate    2.5%  97.5% 2*min(Pr(>0),Pr(<0))
(Intercept)  12.4576  9.5451 15.677       < 2.2e-16 ***
---
```

4.4. Resultados de cuantiles, precisión y probabilidades

El resumen proporcionado a continuación ofrece una visión del comportamiento y la eficacia de la cadena de Markov Monte Carlo (MCMC) utilizada para la obtención de la distribución posterior β.

```
-----------------------------------------------------
Cuantil (q) = 0.025
Precision (r) = +/- 0.0125
Probabilidad (s) = 0.95
-----------------------------------------------------
```

- **Cuantil** $(q) = 0.025$: Este valor indica el cuantil inferior del intervalo de credibilidad al 95 %. Este cuantil demuestra que el 2.5 % de la distribución acumulada de los valores de β está por debajo de este umbral, ofreciendo un punto de referencia para entender la distribución de valores inferiores del parámetro.
- **Precisión** $(r) = +/- 0.0125$: La precisión reportada es la mitad del ancho del intervalo de credibilidad al 95 %, muestra que el intervalo es suficientemente estrecho como para garantizar una estimación precisa de β. Esto es esencial para asegurar la confiabilidad de las inferencias hechas a partir del modelo.
- **Probabilidad** $(s) = 0.95$: Hay una probabilidad del 95 % de que el valor verdadero de β se encuentra dentro de este intervalo, basado en la información proporcionada por los datos.

4.5. Media Posterior del Intercepto

La media posterior del intercepto ($\hat{\beta} = 12.45765$) es la media de la distribución posterior del intercepto obtenido a partir de las muestras posteriores generadas por las simulaciones MCMC. El valor $\hat{\beta}$ es una estimación puntual del intercepto del modelo. La elección de la media como estimador puntual proviene de la optimización planteada en 2.4 que es la función de pérdida y se busca minimizar la pérdida media.

$$\text{argmin}_{\hat{\beta}} \, \mathbb{E}[(\beta - \hat{\beta})^2 \mid Y]$$

Desarrollando el término cuadrático obtenemos lo siguiente:

$$\begin{aligned}
\mathbb{E}[(\beta - \hat{\beta})^2 \mid Y] &= \mathbb{E}[\beta^2 - 2\beta\hat{\beta} + \hat{\beta}^2 \mid Y] \\
&= \mathbb{E}[\beta^2 \mid Y] - \mathbb{E}[2\beta\hat{\beta} \mid Y] + \mathbb{E}[\hat{\beta}^2 \mid Y] \\
&= \mathbb{E}[\beta^2 \mid Y] - 2\hat{\beta}\mathbb{E}[\beta^2 \mid Y] + \hat{\beta}^2
\end{aligned}$$

Se encuentra el mínimo mediante el cálculo de la derivada respecto a $\hat{\beta}$ igualando a cero.

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\beta}}\ \mathbb{E}[\beta^2 \mid Y] - 2\hat{\beta}\mathbb{E}[\beta^2 \mid Y] + \hat{\beta}^2 = -2\mathbb{E}[\beta \mid Y] + 2\hat{\beta} = 0$$
$$\implies\ \hat{\beta} = \mathbb{E}[\beta \mid Y]$$

Y sólo como confirmación se calcula la segunda derivada para revisar que en efecto sea mínimo.

$$\frac{\mathrm{d}^2}{\mathrm{d}\hat{\beta}^2}\ \mathbb{E}[(\beta - \hat{\beta})^2 \mid Y] = 2 > 0$$

Al ser la segunda derivada positiva, entonces se asegura que el estimador encontrado es en efecto mínimo.

4.6. Inferencia sobre las Posiciones Latentes

En este punto, cabe destacar que se han generado muestras de las posiciones latentes y posteriormente se ha aplicado la transformación de Procrustes. Para recordar como se explicó en el cálculo de esta transformación 3.1, esta transformación se emplea específicamente para alinear dos conjuntos de puntos en un espacio euclidiano. En otras palabras, nos permite comparar y ajustar las posiciones latentes de nuestros datos de una manera más coherente y comprensible.

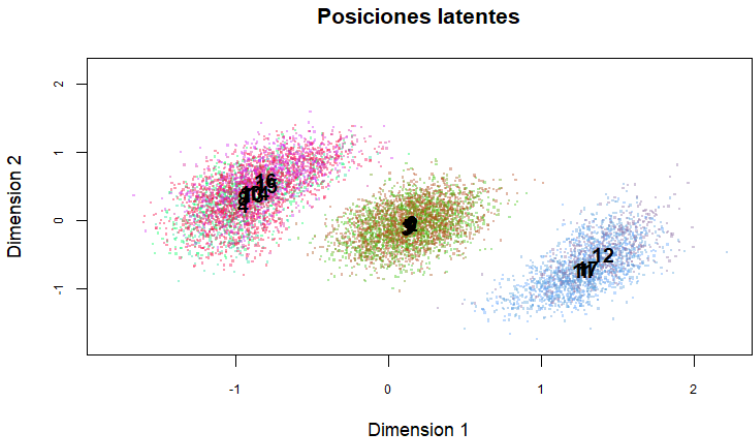

Figura 13: Representación de las posiciones latentes

La Figura 13 muestra una representación visual de las posiciones latentes de los jugadres en el espacio latente, derivadas de muestras de la distribución posterior. Los colores varían de acuerdo con su posición en el espacio latente, con una transparencia establecida para permitir que se vea la densidad de las muestras.

La interpretación de las posiciones latentes en el espacio bidimensional revela patrones informativos sobre las interacciones entre jugadores en el contexto del juego. Las agrupaciones visibles en el gráfico no son aleatorias; más bien, sugieren la formación de clústers dentro del grafo que reflejan dinámicas subyacentes del juego.

Estos clústers, representados por las concentraciones de puntos, estan delineando grupos de jugadores que a menudo se mueven de manera coordinada en el terreno de juego por zonas similares. La disposición de los jugadores en estos grupos no solo resalta sus posiciones físicas en el campo, sino que también proporciona perspectivas sobre sus roles tácticos. Por ejemplo, los defensores centrales y los mediocampistas defensivos pueden agruparse debido a su tendencia a intercambiar pases frecuentemente como parte de la construcción de juego desde la defensa.

La tendencia observada en los colores dentro de cada grupo refuerza esta interpretación, reflejando diferencias en la posesión y toques de balón. Los co-

lores representan cambios en las variables subyacentes que son la frecuencia de toques y la naturaleza de las posesiones. Por ejemplo, un gradiente de color dentro de un grupo de jugadores indica una jerarquía en la cantidad de interacciones o una tendencia a participar en jugadas clave que avanzan el juego hacia la zona de ataque.

El análisis de las posiciones latentes va más allá de la simple proximidad física en el campo; pone de manifiesto la complejidad del fútbol como un juego tanto de espacio como de relaciones. Los jugadores que están cerca en el espacio latente comparten tanto roles similares en la estrategia del equipo, como en la formación de triángulos de pase para mantener la posesión o en la creación de oportunidades de ataque mediante movimientos sincronizados.

El modelo, al remitir estas relaciones y tendencias a través de la posesión y los toques de balón, proporciona una rica narrativa de cómo los jugadores contribuyen al flujo del juego. No se trata únicamente de las posiciones estáticas o roles asignados, sino de una dinámica interactiva que es fundamental para entender cómo las estrategias y tácticas se despliegan en el campo. Las inferencias extraídas de este modelo tendrán implicaciones significativas para el análisis táctico, la formación de equipos, y la comprensión de cómo los aspectos individuales y colectivos se entrelazan en el deporte rey.

Las etiquetas numéricas marcadas de manera más prominente representan el promedio de las posiciones latentes de los jugadores a lo largo de las muestras de MCMC, proporcionando una referencia fija a partir de la cual se puede entender la variabilidad de las muestras alrededor de estos puntos centrales. Estas medias son útiles para resumir la ubicación probable de cada nodo en el espacio latente y para interpretar la estructura del modelo.

Este enfoque visual ofrece una manera intuitiva de interpretar la distribución posterior de las posiciones latentes, facilitando la comprensión de las relaciones y dependencias modeladas dentro del espacio latente. No es meramente un conjunto de muestras de la posterior, sino una ilustración de la naturaleza probabilística y la incertidumbre en la inferencia Bayesiana.

En vista de lo anterior se observó que las muestras se dividen naturalmente en tres grupos, surge una motivación clara para abordar la agrupación de los nodos del grafo mediante *clustering* en una sección posterior 4.7. La visuali-

zación y comparación de las posiciones latentes, después de la transformación Procrustes, revelan una estructura o patrón que sugiere la existencia de agrupaciones intrínsecas en los datos.

Dado este hallazgo, la decisión de aplicar técnicas de clustering se fundamenta en la idea de que estos grupos identificados en las posiciones latentes pueden corresponder a clústers específicos. El *clustering*, en este contexto, se presenta como una estrategia lógica para organizar y comprender mejor la distribución de las entidades o elementos en el espacio de interés.

A continuación se presenta la Figura 14 de una sóla de las muestras, pero con sus correspondientes relaciones y dar una pequeña previa antes de mostrar el grafo posterior promedio.

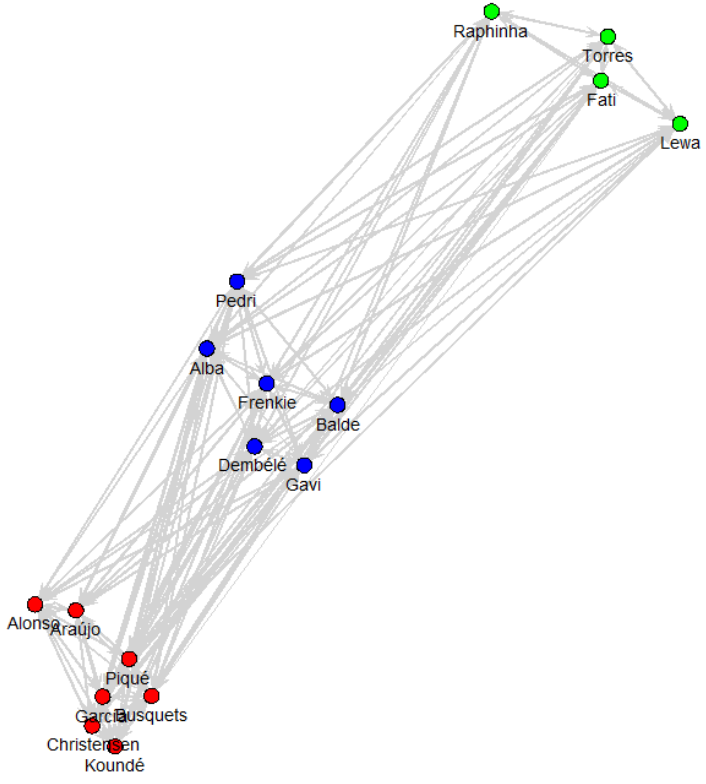

Figura 14: Posiciones latentes de una sola muestra generada por MCMC

4.7. Agrupación de Nodos

La **agrupación jerárquica**, también conocida como *clustering* jerárquico, es una técnica de análisis de datos que busca agrupar elementos similares en subconjuntos o clústers, de manera que podamos entender la estructura subyacente de nuestros datos de manera jerárquica. En nuestro estudio de la red social de los jugadores del FC Barçelona, hemos utilizado específicamente el método *average* o *enlace promedio* para realizar esta agrupación jerárquica.

4.7.1. Medición de Similitud

El proceso comienza con la medición de similitud entre todos los pares de jugadores, utilizando sus posiciones latentes para determinar cuán cercanos están en el espacio del modelo. Esta similitud se calcula mediante la distancia euclidiana, en nuestro caso, esta se ha seleccionado para reflejar las interacciones y relaciones en la red social de jugadores.

4.7.2. Creación de una Matriz de Distancia

Con las similitudes calculadas, construimos una matriz de distancia. Los elementos de la diagonal de esta matriz son ceros, reflejando la máxima similitud de un jugador consigo mismo, mientras que los elementos fuera de la diagonal representan las distancias entre diferentes jugadores, indicando así su grado de disimilitud.

4.7.3. Agrupación Inicial

El procedimiento de clustering jerárquico comienza tratando a cada jugador como un clúster individual. Luego, los jugadores o clústeres se van agrupando secuencialmente, empezando por aquellos que son más similares, según lo dictado por la matriz de distancia. El método *average* se utiliza para este propósito, fusionando clústeres basándose en la similitud promedio de sus componentes, lo que tiende a resultar en una distribución de clústeres de tamaño balanceado.

En el método de agrupación jerárquica conocido como *average*, la similitud promedio de sus componentes se refiere a cómo se calcula la proximidad entre

dos clústers. El método *average* utiliza el promedio de todas las distancias posibles entre los puntos en dos clústers diferentes.

Para dos clústers C_i y C_j, cada uno conteniendo múltiples jugadores, el método *average* considera todas las distancias entre pares de jugadores $d(x, y)$ donde x es un miembro de C_i y y es un miembro de C_j. Calcula la media de estas distancias y utiliza ese promedio como medida de disimilitud entre las dos clústeres. Matemáticamente, esto se representa como:

$$\text{Disimilitud}_{\text{promedio}}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y),$$

donde $|C_i|$ y $|C_j|$ son los tamaños de las clústers C_i y C_j, respectivamente.

Esta medida tiene la ventaja de considerar la estructura completa de las clústers, proporcionando una visión más global que puede ser menos susceptible a valores atípicos que otros métodos de enlace. Al utilizar el enlace promedio, los clústers formados tienden a representar mejor las relaciones subyacentes en el conjunto de datos. En el contexto de este análisis, este método ayuda a asegurar que los jugadores dentro de un clúster no solo estén cerca de algunos miembros de su comunidad, sino que tengan una similitud promedio con todos los miembros de la misma, reflejando así una verdadera comunidad de juego en el equipo.

4.7.4. Construcción del Dendograma

A medida que se agrupan los jugadores, se construye un dendograma, que es una representación gráfica en forma de árbol que muestra la jerarquía de clústers. En la parte superior del dendograma, encontraremos un único cluster que engloba a todos los jugadores, y en la parte inferior, tenemos clústers individuales para cada jugador. Entre estos extremos, podemos observar cómo los clústers se fusionan progresivamente en niveles más altos del dendograma.

4.7.5. Elección del Número de Clústers

Dado que el clustering jerárquico no prescribe un número específico de clústeres, podemos seleccionar el nivel de corte en el dendograma que mejor

se alinee con nuestro análisis y objetivos. Este nivel de corte determinará el número final de clústeres y la granularidad de la agrupación que consideramos adecuada para interpretar las comunidades dentro de la red. Al ver que de antemano se veían aparentemente tres comunidades, pues este fue justamente la elección que se tomó.

4.7.6. Interpretación de Resultados

Una vez que hemos seleccionado el número adecuado de clústers, podemos interpretar los resultados. Cada cluster representará un grupo de jugadores que comparten similitudes en sus posiciones latentes y relaciones en la red social. En nuestro caso, hemos identificado tres clústers que se corresponden con los roles de defensa, mediocampo y delantera en el equipo del FC Barçelona, lo que nos proporciona una valiosa comprensión de la estructura y dinámica de la red social de los jugadores.

En la siguiente figura se muestra el dendograma generado por el método, pero se cambiaron los nombres de los jugadores por números por el momento para no saltar a conclusiones tácticas por el momento.

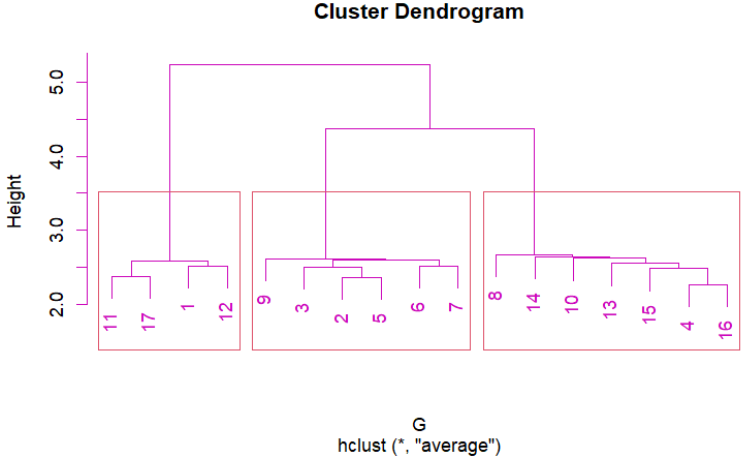

Figura 15: Dendograma de agrupación de jugadores

- **Clúster 1 *El Rol Defensivo*:** El primer agrupamiento que se ha iden-

tificado en el dendograma representa un grupo de jugadores caracterizados por desempeñar un rol defensivo en el equipo. Su función principal es la de proteger la portería y evitar que el equipo contrario atraviese la línea defensiva para anotar. Los jugadores en este cluster tienden a tener conexiones y relaciones similares en la red social, lo que sugiere que comparten estrategias y responsabilidades similares en el campo. En otras palabras, este cluster representa a los defensores cuyo objetivo es garantizar la solidez en la parte trasera y la seguridad en la zona defensiva del campo.

- **Clúster 2 *El Centro del Campo*:** El segundo cluster identificado en el dendograma agrupa a jugadores que desempeñan un rol en el centro del campo. Estos jugadores son fundamentales para la distribución del juego y la transición entre la defensa y el ataque. En este cluster, encontramos a mediocampistas que comparten características y conexiones similares en la red social. Son los encargados de crear oportunidades de gol, controlar el ritmo del partido y conectar los diferentes sectores del equipo. Este cluster representa la columna vertebral del club, cuyo papel es vital para el equilibrio y el éxito en el campo.

- **Clúster 3 *La Línea de Ataque*:** El tercer cluster del dendograma está compuesto por jugadores que ocupan posiciones adelantadas y desempeñan un rol ofensivo en e equipo. Estos jugadores son los encargados de marcar goles y crear ocasiones de peligro en el área rival. En este cluster, encontramos delanteros y jugadores de ataque que comparten similitudes en sus posiciones latentes y conexiones en la red social. Representan la vanguardia del club, cuya responsabilidad es marcar la diferencia en el último tercio del campo y convertir las oportunidades en goles.

Por último, se crean gráficos que representan las posiciones latentes promedio de los clústers en su distribución posterior. Esto nos da una idea de dónde se encuentran los clústers en promedio en el espacio latente.

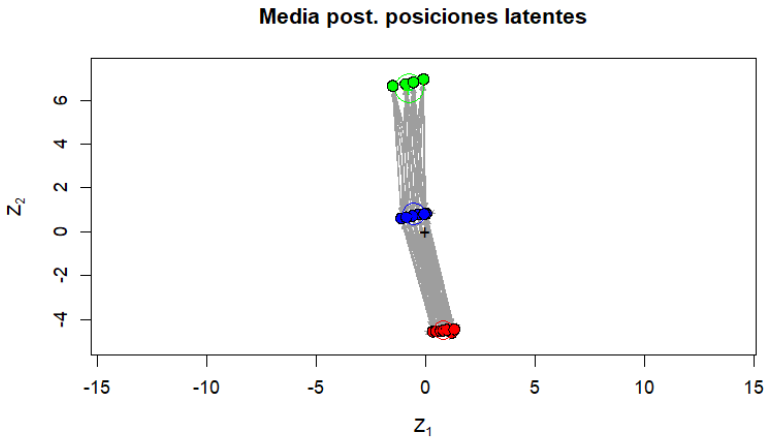

Figura 16: Media posterior de las posiciones latentes

4.8. Inferencia en las Probabilidades de Interacción

Para inferir las probabilidades de interacción entre jugadores, se parte de la premisa de que cada muestra obtenida de la distribución posterior a través de MCMC refleja una posible configuración del espacio latente y del parámetro de intercepción β. La probabilidad de interacción entre un par de jugadores i y j se infiere aplicando la función logística a la combinación de estos parámetros. La función logística, comúnmente conocida como función **expit**, es la inversa de la función **logit** que se utilizó para definir $\eta_{(i,j)}$ en la ecuación 2.3 y se define como:

$$\text{expit}(x) = \frac{1}{1 + \exp(-x)}$$

Esta función transforma los log odds x en una probabilidad en el intervalo $(0, 1)$. La distancia latente entre los jugadores i y j para cada muestra se calcula utilizando la norma euclidiana de la diferencia de sus posiciones latentes:

$$d_{i,j} = \sqrt{\sum (Z_{b,i} - Z_{b,j})^2},$$

donde $Z_{b,i}$ y $Z_{b,j}$ representan las coordenadas de los jugadores i y j para la muestra b. Con estas distancias, la probabilidad de interacción entre i y j, bajo la muestra b, se infiere como:

$$\text{expit}(\beta_b - d_{i,j}) = \frac{1}{1 + e^{-\beta_b + d_{i,j}}}$$

La probabilidad media posterior de interacción entre i y j se obtiene promediando estas probabilidades a lo largo de todas las muestras B de la distribución posterior:

$$\Pi_{i,j} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{1 + e^{-\beta_b + d_{i,j}}}$$

La matriz resultante Π es simétrica, ya que la interacción de i con j es equivalente a la de j con i, y esta matriz encapsula las probabilidades de interacción para todos los pares de jugadores en el análisis. Este enfoque garantiza una matriz de probabilidades basada en la incertidumbre que es coherente con los principios de la inferencia Bayesiana. En la Figura 17 se puede ver una visualización de esta matriz de probabilidades de interacción ordenada de tal forma que se vea en los ejes el dendograma de la Figura 15, pero ahora cambiando los números con los nombres de los jugadores.

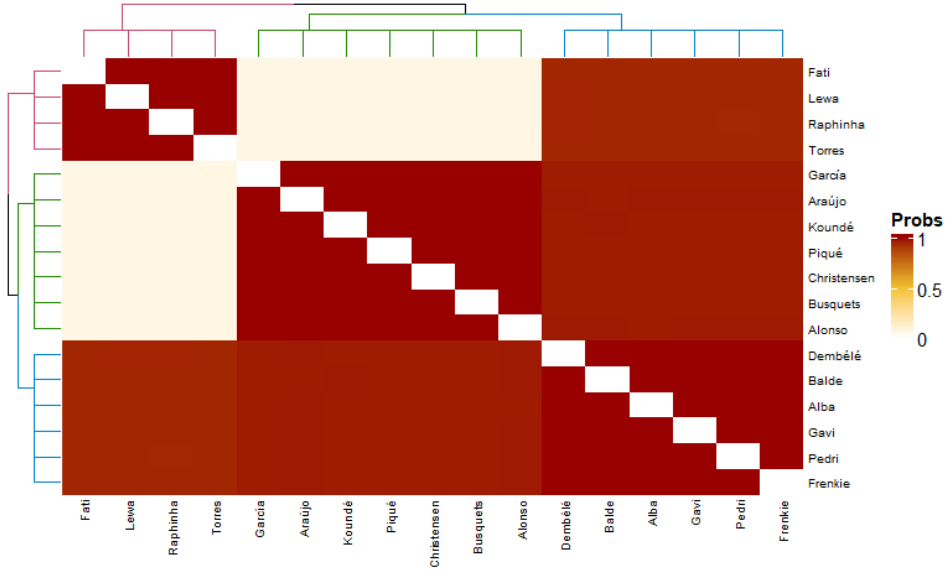

Figura 17: Mapa de calor de las probabilidades de interacción de los jugadores

4.9. Características del Grafo con Incertidumbre

En este punto de la investigación, se ha llevado a cabo un análisis para evaluar el modelo y su ajuste a los datos observados en la red social de los jugadores del FC Barçelona. Este proceso se ha basado en la generación de histogramas que nos proporcionan información sobre la relación entre las características del modelo y la realidad de nuestra red social.

En la Figura 18 veremos una serie de histogramas de las características de nuestra red social. El primer histograma que se examinó representa la **densidad** de aristas en nuestra red. Cada barra refleja la densidad de aristas calculada en simulaciones del modelo. la media se representa con una línea vertical (en color azul), y las líneas punteadas rojas indican los percentiles 2.5 % y 97.5 %. La densidad de aristas se relaciona con la proporción de conexiones en la red en comparación con todas las posibles conexiones. Esto nos proporciona información sobre cuán conectada está nuestra red social.

El siguiente histograma examinado representa la **transitividad** de la red. La transitividad mide la probabilidad de que dos jugadores que están conectados a un tercer jugador también estén conectados entre sí. Al igual que en el histograma anterior, las barras representan la transitividad calculada en simulaciones del modelo, y las líneas punteadas los mismos percentiles. De forma concreta la transitividad nos ayuda a entender si existe un patrón en el que los amigos de nuestros amigos también son amigos entre sí.

En el tercer histograma, se explora la **asortatividad** de la red. La asortatividad se relaciona con la tendencia de los jugadores a conectarse con otros jugadores que comparten características similares, en este caso usando como medida de similitud el grado del jugador. Nuevamente, las barras representan la asortatividad calculada en simulaciones del modelo, con líneas punteadas en los mismos percentiles. En el contexto de nuestra red social, una asortatividad positiva sugiere que los jugadores tienden a relacionarse con otros jugadores de un nivel similar o de la misma posición.

El cuarto histograma se centra en la **distancia promedio** entre los pares jugadores en la red. Las barras representan dicha distancia calculada en las simulaciones del modelo, con líneas punteadas en los mismos percentiles. Esta métrica nos proporciona información sobre la proximidad o lejanía de los ju-

gadores en términos de conexiones en la red social.

En el quinto histograma se analiza el **grado promedio** de los jugadores en la red. El grado como vimos en el marco teórico se refiere al número de conexiones que tiene un jugador. Las barras representan el grado promedio calculado en simulaciones del modelo, y los mismos percentiles representados. El grado promedio ofrece información sobre la conectividad general de los jugadores como equipo.

Por último, se generó el histograma para la **desviación estándar del grado** en la red. La desviación estándar refleja la variabilidad en el número de conexiones de los jugadores. Las barras en este histograma representan lo dicho calculado en simulaciones del modelo y con los percentiles 2.5 % y 97.5 % remarcados con las líneas punteadas.

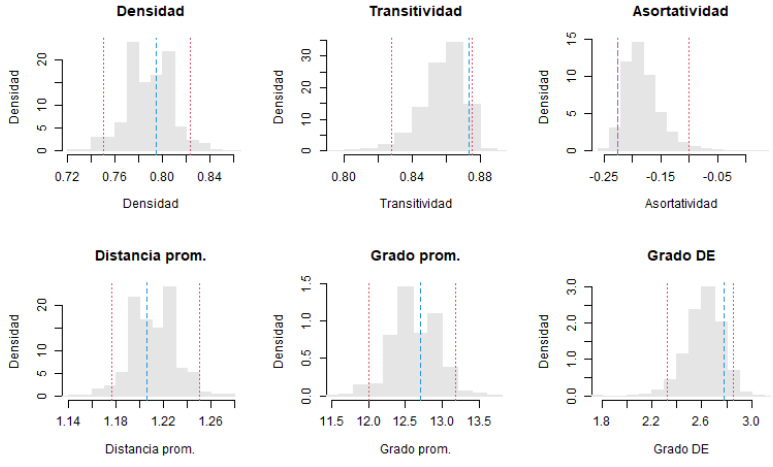

Figura 18: Histogramas de características de la red social

La densidad media del grafo es de 0.7889, lo que indica que el equipo muestra un alto nivel de conectividad en el campo. La posesión del balón es una constante y se refleja en un juego ofensivo donde los jugadores buscan constantemente oportunidades para crear jugadas y mantener el flujo de pases activo.

Con una transitividad de 0.8577, se confirma que el equipo juega con una excelente cohesión. Los jugadores tienden a moverse en triángulos dinámicos,

asegurando que la pelota circule con eficiencia y que siempre haya opciones de pase, minimizando el riesgo de intercepciones.

La asortatividad, con un valor de -0.1822, muestra que los jugadores con distintos grados de conectividad interactúan entre sí. Esto es característico de equipos que integran a todos sus miembros en el juego, desde la defensa que construye desde atrás hasta los delanteros que finalizan las jugadas.

La distancia promedio en el grafo es de 1.2111, lo que sugiere que los jugadores se posicionan a una distancia óptima para facilitar el juego combinativo. Este espaciamiento permite cubrir el campo eficientemente y estar preparados para transiciones rápidas, ya sea recuperando la pelota o cambiando el punto de ataque.

El grado promedio, que se sitúa en 12.6219, indica que cada jugador, en promedio, participa activamente en la red de pases del equipo. Este alto grado de interacción es un testimonio de la naturaleza integradora del juego del equipo, con múltiples opciones de pase y participación equitativa.

Finalmente, el grado estándar medio de 2.6179 refleja que existe cierta variabilidad en las conexiones entre los jugadores, lo cual es normal en un equipo donde diferentes roles y responsabilidades conducen a diferentes patrones de interacción. Esto podría ser indicativo de especialistas defensivos y creativos en el ataque, cada uno desempeñando su función dentro de la estrategia del equipo.

Estas características dibujan la imagen de un equipo que equilibra la posesión con propósito y la movilidad táctica, manejando el ritmo del juego y aprovechando las habilidades de sus jugadores para controlar y dominar en el terreno de juego.

4.10. Análisis del Grafo (Teoría de Grafos Aplicada)

Tras la obtención del grafo final con incertidumbre hace falta un análisis del mismo para el planteamiento de una resolución táctica del rival. Para ello se llevará a cabo un análisis preliminar con cuatro propiedades descritas en la teoría de gráficas que son el **grado** de los vértices, la **densidad** del gra-

fo, la **centralidad de intermediación** para cada uno de los vértices y los caminos más cortos de un vértice a otro. A continuación, se explicará cada uno de estos respecto a su uso en análisis táctico para fútbol, explicando a su vez su utilidad en el análisis del rival para tener una idea estructural del equipo.

Partiendo del grafo generado por el modelo de espacios latentes ahora tenemos, si bien, un grafo tremendamente similar al generado tras el preprocesamiento de los datos del Fútbol Club Barçelona en la temporada 22-23 un modelo gráfico final cuya gran diferencia es la inclusión de *incertidumbre*. Esta incertidumbre nos permite ahora analizar el grafo en términos de probabilidad de tal forma que los pesos no son una mera ponderación de pases, toques de balón o pases completos, sino que ahora los *pesos* del grafo son probabilidades de interacción entre jugadores dado su rendimiento.

Además como se vió en el agrupamiento de nodos se puede notar en el grafo una clara distinción de tres clústers, no necesariamente determinados por su posición ni su rol en el campo, sino más bien características propias (latentes) que sólo se ven reflejadas en el espacio latente. Más aún, al separar y analizar por su cuenta cada clase podemos notar que cada clúster no es más que un **subgrafo**, por lo que ahora se nombrarán así a estos clústers.

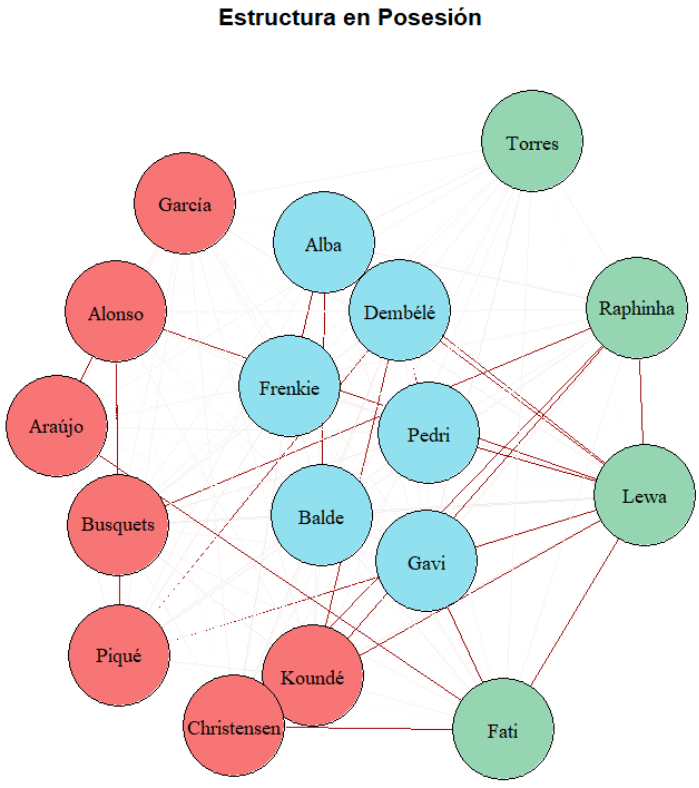

Figura 19: Estimación final del grafo en estructura de posesión del FC Barçelona 22-23

El subgrafo presentado en la Figura 20 destaca claramente la estructura defensiva del equipo, pero también subraya el rol pivotal de Sergio Busquets en el esquema táctico. Busquets, aunque catalogado como centrocampista, se muestra aquí como un eje central en la retaguardia, fungiendo como un enlace vital entre la defensa y el mediocampo. El patrón de conexiones revela que Busquets actúa como un pivote defensivo, ocupando una posición clave que facilita la transición del balón desde la defensa al ataque. Su presencia y ubicación en este subgrafo ilustran su rol de *regista*, dictando el ritmo y la dirección del juego desde una posición profunda, característica de un centrocampista defensivo.

La red también ilustra una conexión notable entre Piqué y Alonso, dos jugadores con una capacidad comprobada para iniciar jugadas desde atrás. Piqué, conocido por su habilidad para leer el juego y su comodidad con el balón

en los pies, a menudo se convierte en un punto de inicio para los ataques. Alonso, por otro lado, puede ofrecer una presencia física y técnica, así como una opción para los balones aéreos y la distribución del juego.

La baja conexión entre Araújo y García sugiere una mala asociación defensiva, probablemente cuando forman una pareja central en la línea de defensa que intenta combinar juventud con fuerza. Araújo se destaca por su fortaleza y habilidad para enfrentamientos uno a uno, mientras que García aporta buen posicionamiento, pero ya sea que no son una buenadupla defensiva o simplemente no han jugado tanto juntos.

La inclusión de Christensen y Koundé indica su importancia en la red defensiva, tal vez ofreciendo fortalezas en la cobertura lateral y la capacidad de cerrar espacios o avanzar para unirse al ataque, lo que refleja una tendencia moderna de defensas centrales con habilidades ofensivas.

Este subgrafo no solo captura las conexiones entre defensas en términos de zonas de trabajo defensivo, sino que también puede estar revelando una estrategia más amplia de construcción de juego. El grado entre estos jugadores sugieren un sistema diseñado para crear plataformas para transiciones efectivas sobre la línea. Este diseño táctico permite al equipo mantener la posesión y controlar el juego desde la parte trasera, preparándose para explotar la ofensiva con base en una estructura defensiva de mucho toque de balón.

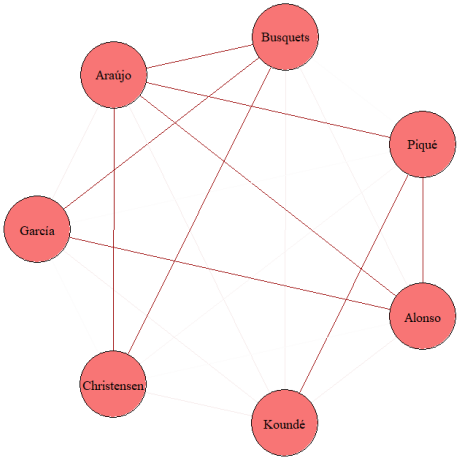

Figura 20: Subgrafo ajustado para visualizar a los defensas

El subgrafo azul resalta la intrincada red de pases y movimientos entre los mediocampistas y jugadores de posiciones ofensivas laterales del equipo, evidenciando la fluidez táctica y la inteligencia espacial que definen su estilo de juego. Frenkie de Jong aparece como un mediocampista central fundamental, y su conexión con otros mediocampistas, Pedri y Gavi, destaca una sinergia en la sala de máquinas del equipo. Las líneas más oscuras que lo unen con estos jugadores muestran la sólida base de su juego de posesión, probablemente impulsando el ataque desde el centro del campo con una combinación de pases cortos y movimiento inteligente.

La presencia de Dembélé en el subgrafo, con líneas que demuestran una baja participación significativa en la construcción del juego, es una revelación táctica. A pesar de ser un extremo derecho cuya función principal sería la de desbordar y atacar por las bandas, su rol aquí indica una mayor responsabilidad en la fase de creación, ofreciendo una dimensión adicional al ataque del equipo y confundiendo a las defensas contrarias con su imprevisibilidad. Sin embargo parece ocupar las mismas zonas que estos mediocampistas, pero sin relacionarse con ellos.

Balde y Alba, aunque nominalmente laterales izquierdos, también muestran una fuerte implicación en el juego ofensivo. No se limitan a funciones defensivas o meramente a apoyar el ataque por su banda; en cambio, las conexiones sugieren que participan activamente en la progresión del balón, posiblemente ofreciendo amplitud y apoyo constante en el ataque, creando así más opciones y espacio para sus compañeros de equipo.

Este patrón de interacciones subraya una filosofía de juego que valora la contribución de todos los jugadores en la creación y el mantener la posesión, independientemente de su posición nominal. El equipo se beneficia de la flexibilidad táctica que estos jugadores aportan, capaces de adaptar sus roles y posiciones según las necesidades del momento.

Sin embargo, este nivel de involucramiento ofensivo de los laterales y de un extremo en la fase de construcción puede tener sus desventajas. Puede dejar al equipo vulnerable a contraataques si no se realiza una transición defensiva rápida y efectiva cuando se pierde la posesión. Además, este enfoque requiere un alto nivel de resistencia y disciplina táctica de estos jugadores para mantener un equilibrio entre sus responsabilidades ofensivas y defensivas durante todo el

partido, cosa que por sus resultados se demuestra que no se hizo esa temporada.

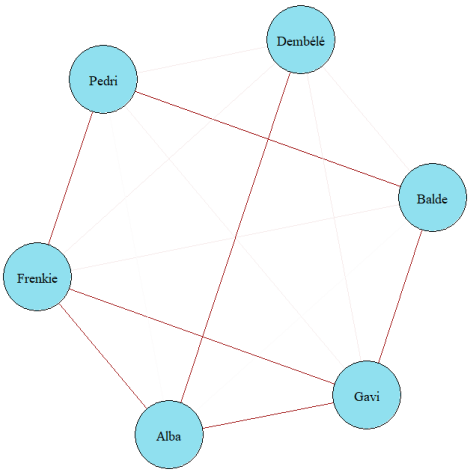

Figura 21: Subgrafo ajustado para visualizar a los centrocampistas

El subgrafo muestra un panorama alarmante respecto a la dinámica ofensiva del equipo, con Robert Lewandowski como la única figura prominente en términos de interacción significativa. La evidente desconexión entre los atacantes plantea serias dudas sobre la eficacia de la estrategia ofensiva del equipo.

Lewandowski, como eje central, parece estar en una lucha solitaria para mantener viva la delantera. Su aparente aislamiento con pocos enlaces hacia Ferran Torres, Ansu Fati y Raphinha sugiere una falta de cohesión en la línea ofensiva que podría resultar en una limitación grave de las oportunidades de gol. El equipo parece depender en exceso de las jugadas individuales de Lewandowski, quien, a pesar de su probada calidad y capacidad para influir en el juego, necesita un soporte y juego en equipo mucho más integrado para maximizar su impacto.

La ausencia de líneas más oscuras, que indicarían una mayor probabilidad de interacción, apunta a una falta de combinación entre los jugadores de ataque. Esto no solo pone en peligro la capacidad del equipo para desplegar un ataque multi-dimensional y predecible, sino que también plantea un reto para desestabilizar defensas adversarias que pueden concentrarse en neutralizar a Lewandowski como única amenaza clara.

Este análisis gráfico sugiere que el equipo necesita revisar urgentemente su aproximación táctica en el terreno ofensivo. Se requiere una mayor integración y entendimiento entre los delanteros para crear un frente de ataque dinámico y variado que pueda explotar las diferentes habilidades de cada jugador. Sin una mayor interacción y colaboración, el equipo corre el riesgo de caer en un ataque unidimensional y fácilmente contenible. Hecho que fue demostrado al ver sus resultados de esa temporada cuyos partidos terminaban con resultados de pocos goles o simplemente una anotación al primer tiempo contra equipos de mucho menor nivel.

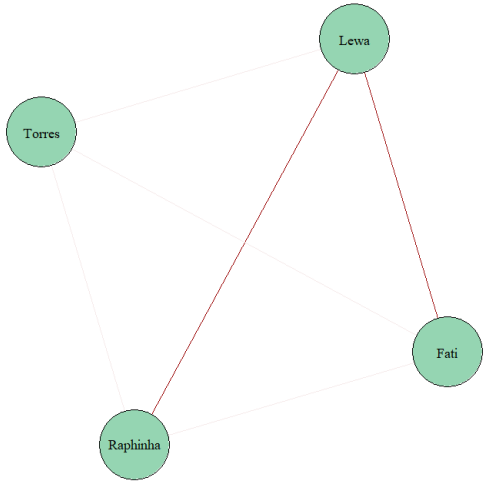

Figura 22: Subgrafo ajustado para visualizar a los delanteros

Este pequeño análisis exploratorio es uno que se puede construir sólo de forma visual al observar el grafo, pero intentaremos ir más allá con el uso de las propiedades descritas en el marco teórico. Iniciando por el grado de los jugadores obtenemos los siguientes datos.

JUGADOR	GRADO
Lewandoski	8
Pedri	6
Balde	1
Busquets	3
Dembelé	4
Frenkie	5
Gavi	5
García	0
Alba	4
Koundé	3
Raphinha	4
Torres	0
Araujo	2
Alonso	3
Pique	3
Christensen	3
Fati	4

Como se explicó en la sección de fundamentos teóricos el **grado** de un vértice (jugador) es el número de aristas que salen del mismo. En este caso podemos ver que la mayoría de jugadores presentan un grado particularmente alto, además podemos ver que aquellos con el grado más alto forman parte del subgrupo azul (armadores de juego), curiosamente no existen jugadores con un grado lo suficientemente bajo como para asumir que este jugador no está presente en la construcción de juego. Esto último nos hace creer que es un equipo que mantiene mucho la posesión, pero para estar seguros usaremos la **densidad** del grafo.

A manera de repaso recordaremos que la **densidad** de un grafo es un valor entre cero y uno que muestra que tan *completo* es un grafo, refiriéndome a completo como aquél grafo en el que todo vértice del grafo se relaciona con el resto de manera directa, es decir; mediante una única arista y el camino más corto entre un vértice y otro es siempre es uno. En este caso particular hablando de posesión del balón un equipo con un grafo completo es un equipo en el que todos los jugadores se comprometen en mantener el balón y son capaces de conectar y circular el balón fácilmente ante cualquier otro jugador independientemente de su posición. Esto en fútbol es extremadamente impro-

bable, pues no es muy común ver a un defensa central compartiendo posesión del balón con un delantero centro. Por ello podemos definir una pequeña clasificación de los equipos acorde a su **densidad**.

DENSIDAD	PRINCIPIO	EJEMPLO
[0, 0.25]	Equipos de repliegue (*Park the Bus*)	Atlético de Madrid
(0.25, 0.50]	Equipos de contrataque	Real Madrid
(0.50, 0.75]	Equipos de transiciones rápidas (*Gegenpress*)	Liverpool
(0.75, 1]	Equipos de posesión (*Tiki-Taka*)	Manchester City

La clasificación de los equipos según la densidad de su grafo se basa en la observación empírica de los estilos de juego y su correlación con las interacciones en campo. Equipos con densidades bajas tienden a adoptar tácticas defensivas profundas, donde el énfasis está en la solidez defensiva sobre la posesión; este es el caso del Atlético de Madrid. A medida que aumenta la densidad, se observan equipos como el Real Madrid que favorecen el contragolpe rápido, mientras que densidades aún más altas indican un estilo de juego centrado en transiciones veloces y presión inmediata tras pérdida, como el Liverpool. Finalmente, con densidades cercanas a uno, equipos como el Manchester City exhiben un estilo de posesión dominante, manteniendo un flujo constante de pases y control del juego, característico del *Tiki-Taka*. Esta clasificación emerge de patrones consistentes en la interacción de jugadores en el terreno de juego, reflejando directrices tácticas fundamentales en su estilo de juego.

En el caso particular de nuestro rival el FC Barçelona obtuvo una densidad de 0.7941 que lo clasifica directamente como un equipo de posesión lo cual concuerda con la tabla de grados que vimos anteriormente, pues todos los jugadores muestran una alta capacidad de circulación de balón. Ahora es cuando surge un problema, en el caso de que saliesen equipos con densidades menores a 0.50 los grados ya nos indicarían los jugadores con mayor relevancia en el juego, pues serían aquellos con el grado más alto, lo cual implicaría que el balón suele pasar en más ocasiones con dichos jugadores. Sin embargo, en el caso del FC Barçelona todos tienen un grado alto, es por ello que trabajaremos también con la **centralidad de intermediación**.

Como se describió en su momento, la centralidad de intermediación medirá el grado en que un vértice actúa como puente o intermediario en la comuni-

cación entre otros vértices dentro de un grafo. En nuestro caso particular este nos resumirá no sólo cuanto se relacionan los jugadores entre sí, sino también resultará a aquellos jugadores por los que pasa el balón en mayor medida. En términos futboleros esto quiere decir que los jugadores con mayor centralidad de intermediación serán jugadores de construcción, jugadores a los que habrá que presionar más o tener más de cerca para evitar que puedan mantener esa posesión que tanto les gusta tener. Aqui la tabla para visualizar los valores de centralidad.

JUGADOR	CENTRALIDAD
Lewandoski	0
Pedri	4.6
Balde	4.6
Busquets	0
Dembelé	3.2
Frenkie	7.5
Gavi	0
García	0
Alba	4.6
Koundé	0
Raphinha	0
Torres	0
Araujo	0
Alonso	0
Pique	0
Christensen	0
Fati	0

Ahora las cosas son más claras, pues si bien el balón pasa por todos los jugadores, los verdaderos creadores de juego en posesión y los que más se encargan de mantener la posesión del balón son sólo los centrocampistas. Frenkie de Jong es el jugador por el que más pasa el balón, entonces ya sabemos quien es la prioridad en cuanto a defender. La posibilidad de poder delimitar de 17 jugadores a tan sólo cinco es algo precioso, pues el enfoque cambia drásticamente. A partir de estos datos sabemos que no es que se tengan que tener un marcaje personal sobre todos los jugadores, sino que basta priorizar a cinco de ellos, habilitando así un juego más competitivo.

Ahora si filtramos a los jugadores dado el análisis hecho y de forma subjetiva nos quedamos con los posibles titulares entonces ahora podemos tener un grafo que demuestra la más probable estructura en juego del equipo como se ve a continuación.

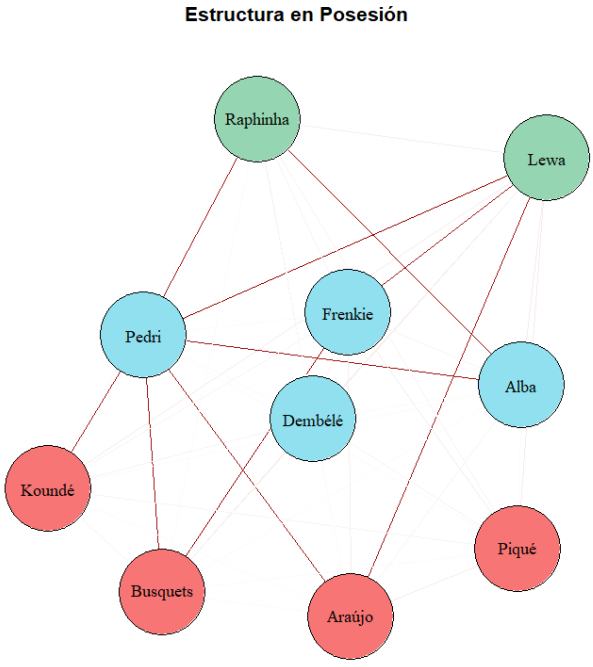

Figura 23: Alineación 4-4-2 en rombo con las interacciones más probables durante el partido

Después de ejecutar el modelo para estimar la probabilidad de interacciones en una red del equipo y filtrar los nodos no titulares, llegamos a algunas conclusiones interesantes. En primer lugar, destaca de manera evidente, incluso sin necesidad de calcular la centralidad, que Pedri emerge como el jugador más crucial en la alineación titular. Su rol de enlace entre la defensa y el ataque es fundamental para el funcionamiento del equipo.

Por otro lado, observamos que Dembelé, a pesar de su importancia en la evaluación general del equipo, parece estar algo aislado en el grafo, con conexiones limitadas con otros titulares. Esto sugiere que, aunque influyente, su

integración en el juego colectivo es practicamente inexistente. Jordi Alba se ve obligado a desempeñarse en una posición inusual debido a las circunstancias del equipo, lo que puede afectó su rendimiento y su relación con otros jugadores.

Por su parte, Gerard Piqué no presenta interacciones significativas en el grafo, lo cual es comprensible dado su rol como defensor. Los defensores tienden a tener menos interacciones en comparación con centrocampistas o delanteros, cuya presencia en el equipo es más pronunciada debido a su participación en la construcción y finalización de jugadas. Sin embargo esto demarca el declive de un central cómo el que en sus tiempos de éxito demarcaba una capacidad de asociación mucho mayor.

Sergio Busquets parece desempeñar un papel clave como enlace defensivo para facilitar la salida de balón, mientras que Frenkie de Jong se destaca por su participación en la generación de oportunidades de gol, evidenciando un patrón claro de distribución de roles dentro del equipo.

A modo de complemento, resulta interesante agregar un detalle al análisis táctico del rival. Dada la vasta cantidad de posibles grafos generados por la predictiva, se hace necesario establecer una manera de consolidar un análisis general para múltiples grafos generados. En este contexto, cobra relevancia la noción de **isomorfismo**. Si, después de varios análisis, se identifican diversos oponentes cuyos grafos son **isomorfos** entre sí, es posible generalizar ciertos patrones tácticos en cuanto a la estructura de posesión para estos adversarios.

En el ámbito del fútbol, esto implica que la alineación específica mostrada por un equipo es independiente de su estructura durante la posesión. Por ejemplo, si un equipo se alinea inicialmente en un $4-3-3$ y otro en un $4-2-3-1$, pero sus grafos estimados por la predictiva son isomorfos entre sí, entonces la disposición inicial en el campo carece de importancia. Ambos equipos siguen los mismos patrones tácticos al encontrarse en posesión del balón.

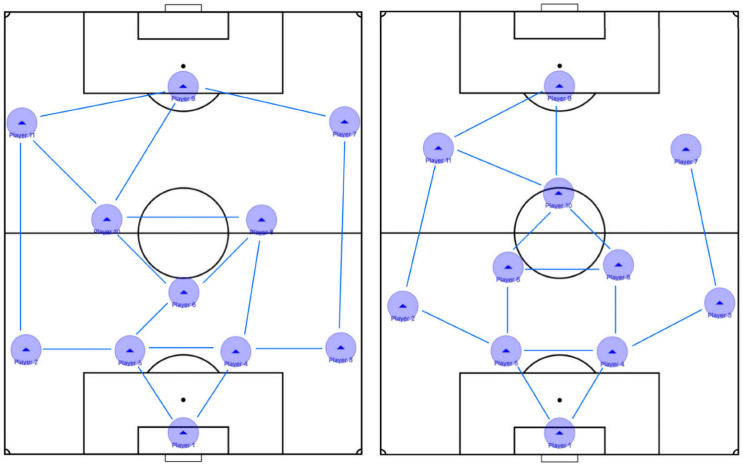

Figura 24: Ejemplo de isomorfismo en alineación 4-3-3 vs. 4-2-3-1

La última propiedad a trabajar para este grafo es el **camino** más corto entre un vértice y otro. En este caso nos interesa saber cuál es el camino más corto que tiene el equipo para que el balón llegue a su máximo goleador desde la defensa, particularmente uno de sus jugadores más reconocidos Ronald Araujo. Sin embargo hay que recordar lo que representan las interacciones y eso es probabilidad, lo cual quiere decir que el camino más corto en este caso es el camino con la menor probabilidad de llegar al delantero estrella Lewandoski. En este caso particular el camino mas corto es Araujo $\rightarrow$ Dembelé $\rightarrow$ Lewandoski, camino que tiene mucho sentido, pues Dembelé es el peor creador de juego para el equipo.

Este caso de uso es uno particular usando únicamente el grafo generado por posesión. Pero mediante otros grafos estimados se pueden usar otras propiedades que nos da la teoría de gráficas para la evaluación del juego rival.

5. Conclusiones

5.1. Conclusiones generales

Esta tesis ha culminado con el desarrollo de un modelo que ha demostrado una buena capacidad predictiva para el análisis táctico del fútbol. El enfoque se basa en representaciones espaciales y se apoya en métodos de inferencia Bayesianos, lo que ha resultado en un marco robusto y práctico para comprender y cuantificar la estrategia en el campo de juego.

Uno de los logros más destacados de esta tesis ha avanzado en la mejora de modelos existentes al permitir la cuantificación de la incertidumbre estadística en un espacio social. Este avance es significativo porque brinda una comprensión más precisa de las dinámicas tácticas en el fútbol. Al tomar en cuenta las incertidumbres en el análisis, se obtienen resultados más robustos en términos de cuantificación de la incertidumbre y, por tanto, más confiables.

La flexibilidad de este modelo es otra característica que merece atención. Puede aplicarse a diversas relaciones en el fútbol, lo que lo convierte en una herramienta versátil y adaptable. Además, la capacidad de trabajar con datos faltantes es esencial, dado que el análisis de eventos deportivos en tiempo real a menudo involucra la falta de datos. Este modelo permite a los analistas aprovechar al máximo la información disponible, lo que tiene un valor incalculable en el análisis táctico.

Finalmente, el modelo demostró ser inherentemente transitivo, lo que significa que es capaz de capturar las relaciones de dependencia en la red de jugadores de manera efectiva. Esta característica es esencial para el análisis táctico, ya que refleja la realidad de cómo interactúan los jugadores en el campo de juego. En conjunto, estas conclusiones generales señalan que este modelo es un paso significativo en la comprensión y optimización de las estrategias tácticas en el fútbol.

5.2. Respuesta a las preguntas específicas

La **relevancia de un jugador** dentro de la dinámica de un equipo de fútbol se cuantifica mediante la matriz de probabilidades de interacción, que

refleja las conexiones potenciales en el campo. Este método integra las habilidades individuales y la incertidumbre del deporte para evaluar la importancia de cada jugador. Utilizando métricas como el *grado* y la *centralidad*, se determina el impacto de los jugadores en la posesión del balón. Estas métricas, que asignan una calificación numérica basada en la actividad y posición del jugador en el espacio latente, no son exclusivas de este enfoque, pero su aplicación en este contexto proporciona una comparación valiosa dentro del equipo, revelando así el papel crucial de los jugadores en la estrategia colectiva del equipo.

En el análisis de la **metodología para identificar el perfil táctico** de un equipo en posesión, se enfatiza la discusión previa sobre la importancia de conocer las *posiciones latentes*. Tras la identificación de una clasificación en tres posibles *clústers*, se identifican en la distribución posterior *subgrafos* por posición en el campo, que equivalen a subespacios en el *espacio latente*. Este enfoque proporciona una visión general del perfil táctico del equipo, siendo la *densidad* y la presencia de posibles *isomorfismos* los elementos distintivos que definen diversos perfiles tácticos y permiten la identificación de perfiles similares, respectivamente.

Finalmente, se lleva a cabo una evaluación del cálculo del camino menos probable para que el balón alcance el tercer cuarto del campo desde la defensa. Este análisis ayuda a comprender la dinámica del juego, se sustenta en la utilización de la *matriz de probabilidades de interacción*, que asigna pesos al grafo, y en la propiedad de los *caminos* según la teoría de grafos. La integración de estos elementos resulta de gran utilidad para obtener una perspectiva más precisa y completa de los movimientos en el terreno de juego.

5.3. Implicaciones de los resultados al Análisis Táctico de Fútbol

Los resultados obtenidos en esta investigación tienen varias implicaciones para el desarrollo táctico del deporte. En primer lugar, este enfoque proporciona a los clubes una herramienta poderosa para comprender y optimizar su estrategia táctica. Al utilizar representaciones espaciales y técnicas estadísticas, los clubes pueden analizar de manera más profunda cómo se desarrollan las interacciones entre jugadores en el campo y ajustar sus tácticas en consecuencia.

Además, la capacidad de manejar datos faltantes es crucial en un deporte dinámico como el fútbol, donde no todos los eventos pueden registrarse con precisión. Este modelo permite a los clubes aprovechar al máximo los datos disponibles, lo que puede conducir a una toma de decisiones tácticas más informada.

En términos de mejoras, es esencial seguir trabajando en la aplicación práctica de este modelo en el deporte. Esto incluye la recopilación de una mayor cantidad de datos con métricas más particulares y la validación continua del modelo en situaciones de juego reales. Lo anterior se comenta, pues a pesar de haber definido un modelo generalizado con covariables, estas a la hora de la aplicación no fueron utilizadas por que los datos no se tenían. Además, se podrían explorar aún más las posibilidades de adaptación del modelo a situaciones específicas de diferentes equipos y estilos de juego.

5.4. Limitaciones y Sugerencias

A pesar de los avances logrados en esta investigación, hay algunas limitaciones que deben abordarse. Por ejemplo, la representación espacial del modelo puede requerir una mayor precisión en la recopilación de datos para obtener una matriz de adyacencia inicial que represente mejor las conexiones intrínsecas de los jugadores. Además, la interpretación de los resultados en términos tácticos puede ser un desafío, lo que sugiere la necesidad de desarrollar herramientas de visualización más efectivas.

En cuanto a sugerencias para futuras investigaciones, sería beneficioso explorar aún más la aplicabilidad de este modelo en situaciones reales de partidos de fútbol, trabajando en estrecha colaboración con clubes y entrenadores. También se podría investigar la integración de datos de rendimiento físico y otros factores en el modelo para obtener una imagen más completa del rendimiento táctico de los equipos.

A su vez valdría la pena hacer un estudio más profundo sobre la probabilidad de interacción basal que al tener un resultado tan alto llama la atención. Particularmente ser capaz de identificar las características particulares para que se *garantice* la interacción. En términos de fútbol esto nos daría las con-

diciones y características puntuales de jugadores que han de tener para que generemos dicha interacción con una alta probabilidad de ocurrencia.

En conclusión, este modelo representa un avance en el análisis táctico del fútbol, pero existen desafíos y oportunidades para futuras investigaciones que podrían llevar a una comprensión más profunda del juego y, en última instancia, a un mejor rendimiento táctico en el campo.

Referencias

[1] Hoff, Peter D. & Raftery, Adrian E. & Handcock, Mark S. (2002). *Latent Space Approaches to Social Network Analysis*. Journal of the American Statistical Association, https://www.jstor.org/stable/3085833

[2] Rastelli, Riccardo, Friel, Nial & Raftery, Adrian E. (2016). *Properties of latent variable network models*. Cambridge University Press.

[3] Gollini, Isabella & Murphy, Thomas B. (2016). *Joint Modeling of Multiple Network Views*. Journal of Computational and Graphical Statistics, Vol. 25, No. 1. pp. 246-265

[4] Turnball, Kathryn R. (2019). *Advancements in Latent Space Network Modelling*. Submitted for the degree of Doctor of Philosophy at Lancaster University. pp. 10-59, 80-101.

[5] Lagos, Jesús. (2019). *Uso de Redes (grafos) en el fútbol*. De Medium Sitio web: https://medium.com/@jesslm_48641/uso-de-redes-grafos-en-el-ftbol-655d2dfd8cb1

[6] Jesús Lagos. (2019). *Métricas de grafos en el fútbol*. Barcelona 2004-2012. 1 de octubre 2021, de - Sitio web:https://www.linkedin.com/pulse/m%C3%A9tricas-de-grafos-en-el-f%C3%BAtbol-barcelona-2004-2012-jes%C3%BAs-lagos-milla/?originalSubdomain=es

[7] Rio, A. Q. del. (n.d.). 4.17 Notas históricas. La estadística Bayesiana — Estadística Básica Edulcorada. Recuperado 5 de enero 2024, de - Sitio web:https://bookdown.org/aquintela/EBE/notas-historicas-la-estadistica-Bayesiana.html

[8] Robert, C., & Casella, G. (2010). *Monte Carlo Statistical Methods*. Springer.

[9] Gelman, A. (2013). *P Values and Statistical Practice*. Lippincot Williams & Wilkins, Vol. 24(No. 1), pp.69-72.

[10] Bernardo, J. M., & Smith, A. F. M. (2009, September 25). *Bayesian Theory*. John Wiley & Sons.

A. Apéndice I

Algorithm 1: Algoritmo de Metropolis-Hastings

Entrada: Estado inicial Z_0, número de iteraciones k
Salida : Muestras $Z_1, Z_2, \ldots, Z_k$ de la distribución de probabilidad objetivo

for $k = 1$ hasta k **do**
  `Proposición de un nuevo estado`
  Generar propuesta Z^* a partir de Z_k utilizando $J(Z^* \mid Z_k)$
  `Cálculo de la Aceptación`
  Calcular $A = \text{mín}\left(1, \frac{\pi(Z^*)\cdot \mathbb{P}(Z_k \mid Z^*)}{\pi(Z_k)\cdot \mathbb{P}(Z^* \mid Z_k)}\right)$
  `Aceptación o Rechazo`
  Generar número aleatorio $U \sim \text{Uniforme}(0, 1)$
  if $U \leq A$ **then**
    $Z_{k+1} \leftarrow Z^*$ `Aceptar la propuesta`
  end
  else
    $Z_{k+1} \leftarrow Z_k$ `Rechazar la propuesta`
  end
end

Definición 12 (Cadena de Markov)**.** *Una* ***cadena de Markov*** *es un proceso estocástico en el que la probabilidad condicional de que el sistema se encuentre en un estado futuro depende únicamente del estado presente y no de los estados anteriores. Formalmente, una cadena de Markov es descrita por la propiedad de Markov, la cual establece que para cualquier sucesión de estados $S_1, S_2, \ldots, S_n$ en el espacio de estados E, la probabilidad condicional de que el sistema esté en el estado S_{n+1}, dado el conocimiento de todos los estados anteriores, es simplemente la probabilidad de que el sistema esté en el estado S_{n+1}, dado el estado actual S_n:*

$$\mathbb{P}(S_{n+1}|S_1, S_2, \ldots, S_n) = \mathbb{P}(S_{n+1}|S_n)$$

Esto puede expresarse de manera más concisa utilizando la notación de probabilidad condicional:

$$P(S_{n+1}|S_1, S_2, \ldots, S_n) = P(S_{n+1}|S_n)$$

En esta expresión, $P(S_{n+1}|S_n)$ representa la probabilidad de transición entre los estados S_n y S_{n+1}, y esta probabilidad es constante independientemente de la historia completa de los estados anteriores.

Definición 13 (Effective Sample Size)**.** *El Tamaño de Muestra Efectivo (ESS) se define como el tamaño de una muestra de datos independientes que contendría la misma cantidad de información que la muestra original. En otras palabras, el ESS ajusta el tamaño de la muestra real teniendo en cuenta la autocorrelación entre los datos.*

Consideremos una muestra aleatoria $x = (x_1, x_2, ..., x_N)$ de tamaño N de una distribución de probabilidad desconocida $f(x)$. El ESS se calcula mediante la fórmula:

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{K} \hat{\rho}_k}$$

Donde:

- *$\hat{\rho}_k$ es el estimador de la autocorrelación en el desfase k de la muestra x.*
- *K es el número de desfases considerados en el cálculo.*

La autocorrelación ($\hat{\rho}_k$) mide la relación entre observaciones separadas por k intervalos de tiempo o espacio. Un valor alto de autocorrelación indica una dependencia fuerte entre las observaciones, mientras que un valor cercano a cero indica independencia.

El término $1+2\sum_{k=1}^{K} \hat{\rho}_k$ representa la corrección por autocorrelación. Cuanto menor sea la autocorrelación en la muestra, mayor será el valor de ESS, lo que indica que la muestra contiene más información independiente.

El desfase se refiere a la separación entre observaciones dentro de la cadena del MCMC. En este caso un desfase de k implica un intervalo de interaciones de k unidades entre dos puntos medidos.

Por ejemplo, en una serie temporal diaria, un desfase de 1 día (desfase 1) significaría la comparación de los datos de hoy con los de mañana. Un desfase de 2 implicaría una comparación entre los datos de hoy y los del día después de mañana, y así sucesivamente.

El análisis de desfase se usa para calcular la autocorrelación, que mide la correlación de una cadena con ella misma en diferentes desfases. Si la cadena es completamente aleatoria, se esperaría ver una autocorrelación cercana a cero en todos los desfases no cero. Si hay algún tipo de patrón o dependencia entre iteraciones, como en los casos de tendencias o ciclos estacionales, se verían autocorrelaciones significativamente diferentes de cero para algunos desfases.

Prior Sensible para los Hiperparámetros

El modelo de posición latente requiere especificar varios hiper-priors, y el modelo es sensible a su elección. Por ejemplo, una varianza intra-cluster a priori demasiado alta conduce a que los clústers se difuminen entre sí, mientras que una varianza intra-cluster a priori demasiado baja crea una distribución posterior en el que todos los clústers se concentran en un punto, causando que el ajuste colapse. Para la conveniencia del usuario, *latentnet* implementa una heurística que produce resultados adecuados en una variedad de redes. La heurística utilizada es la siguiente:

$$\psi = \frac{C_\beta}{\left(\frac{1}{Y}\sum_{i,j\in Y} x^2\right)}$$

$$\sigma^2 = C_{\sigma^2}(\sqrt[d]{n}/G)^d$$

donde Y es la matriz de adyacencia, G es el número de clústers, n el número de vértices y C_k una constante asociada al parámetro k.

En una red típica, incluso los lazos dentro de un clúster particular forman lejos de un grafo completo, por lo que la cantidad de espacio ocupado por un clúster tiende a aumentar con el número de nodos en el clúster, más rápido que una distribución normal con una varianza fija.

Los nodos dentro de cada clúster necesitan cierta cantidad de espacio. El volumen de una hiperesfera d-dimensional es proporcional a su radio elevado a la potencia d, por lo que tiene sentido hacer que la varianza a priori sea

proporcional a la raíz d-ésima del tamaño del clúster.

Para los coeficientes de covariables, la heurística representa una creencia a priori de que los coeficientes de covariables con magnitudes altas tenderían, ellos mismos, a tener magnitudes más bajas para compensar. Por lo tanto, sus respectivas varianzas a priori se dividen por el cuadrado medio de la covariable.

B. Apéndice II

Configuración:

```
Cargar bibliotecas necesarias
(latentnet, jsonlite, knitr, tidyverse, ggplot2,
and many more)
```

Cargar datos:

```
ruta_archivo = "Possession.csv"
possession = leer_csv(ruta_archivo)
X = preprocesar_datos(possession)
```

Extracción del nombre de jugadores:

```
nombres = extraer_nombres(possession_clean)
```

Asignar nombres a nodos.

Matriz de Adjacencia:

Algorithm 2: Preprocesamiento de datos para grafo inicial

```
for todo jugador i do
    for todo otro jugador j do
        p1 ← X_i
        p2 ← X_j
        r ← 0
        if p1, p2 ≠ 0 then
            maxit ← máx{p1, p2}
            p ← (p1+p2)/maxit
            r ← r + p
            Y_{i,j} ← r
            Y_{j,i} ← r
        end
    end
end
```

Formulación de la red: Creación de red y grafo a partir de la matriz de adjacencia.

```
grafo <- graph_from_adjacency_matrix(
        Y,
        weighted = TRUE,
        mode = "undirected")

set.vertex.attribute(grafo,
                     "name",
                     value = nombres)
```

Latent Net Modelling (Hoff & Raftery): Ajustar modelo LNM y realizar diagnósticos (red_fit).

```
set.seed(1234)

red_fit <- lnm(grafo ~ euclidean(d=2))
summary(red_fit)
mcmc.diagnostics(red_fit)
plot(red_fit, labels = TRUE)
```

Modelo euclidiano con 3 clústers: Ajustar modelo con 3 clústers y realizar diagnósticos (red_fit2)

```
matriz_distancia <- as.dist(Y)

arbol_jerarquico <- hclust(matriz_distancia,
                           method = "average")

plot(arbol_jerarquico,
    main = "Dendrograma␣Jerarquico",
    xlab = "Indice",
    ylab = "Altura")

grupos <- cutree(arbol_jerarquico,
                 k = 3)

print(grupos)
```

Heatmap de probabilidades de interacción: Calcular y visualizar probabilidades entre jugadores (red_fit).

Grafo final: Definir clústers manualmente tras observar el dendograma de la agrupación jerárquica y visualizar el grafo.

Análisis Táctico: Calcular camino más corto, grado, densidad y centralidad de intermediación. Imprimir resumen de datos y densidad del grafo.