

# Algunas Ideas Matematicas Aplicadas al Futbol

David Alejandro Perdomo Meza

[theopenfootball.blogspot.co.uk](http://theopenfootball.blogspot.co.uk)

September 8, 2016

# Contenido

- 1 Ideas Pasadas: TDA with Metric Learning Applied to Football Data (Tesis de Maestria)
- 2 Ideas Presentes: Passing Network Motifs
- 3 Ideas Futuras: El Futbol como un Proceso de Markov

# Contenido

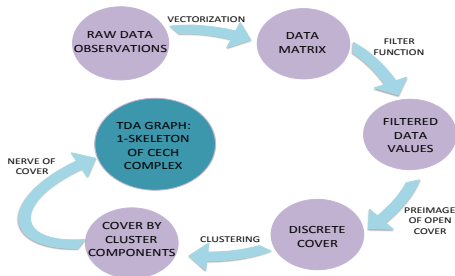
- 1 Ideas Pasadas: TDA with Metric Learning Applied to Football Data  
(Tesis de Maestria)
- 2 Ideas Presentes: Passing Network Motifs
- 3 Ideas Futuras: El Futbol como un Proceso de Markov

# El Proyecto

- **Datos:** Usamos una base de datos publicada por OPTA para la temporada 2011-2012 de la Premier League. Esta base contenia los acumulados por jugador por partido de mas de 200 estadisticas de juego (pases en el ultimo tercio, disparos de cabeza, intercepciones, etc.). Esta informacion la usamos de dos maneras: 1) Acumulados de temporada por jugador, 2) Rendimiento por partido por equipo
- **TDA:** Es una tecnica de visualizacion en dos dimensiones (el resultado es un grafo) para datos en dimensiones superiores. Su objetivo es recuperar para el usuario informacion cualitativa del “layout” de los datos en su espacio original.



# TDA Pipeline



- 1 **Metric Learning:** A la matriz de datos original se le aplica una transformacion lineal para obtener una metrica entre observaciones mas “educada”.
- 2 **Graficar el Grafo:** No es un paso trivial desde la matriz asociada, y la visualizacion depende fuertemente de como se haga.

# Metric Learning

Necesitamos una métrica significativa en el contexto del problema.

## Definición

Sean  $M, L$  dos matrices  $d \times d$  de rango completo, con  $M$  semidefinida positiva.

$$1 \quad D_M(x, y) := \sqrt{(x - y)^t M (x - y)}$$

$$2 \quad \hat{D}_L(x, y) = \| L(x - y) \|$$

## Proposición

Para cualquier matriz  $L$  de rango completo y  $M = L^t L$ , entonces  $M$  es semidefinida positiva y  $D_M = \hat{D}_L$ .

# Algoritmo de Weinberger y Saul

Debemos contar con *etiquetas* que determinen cuáles puntos deberían estar cerca entre sí.

Para un  $k$  fijo y  $x_i \in \mathbb{X}$ , definimos los **vecinos objetivo** como los  $k$  puntos en  $\mathbb{X}$  más cercanos a  $x_i$  con la misma etiqueta que este.

Si  $x_j$  es vecino objetivo de  $x_i$  escribimos  $i \rightsquigarrow j$ .

Definamos  $T := \{(i, j) : i \rightsquigarrow j\}$ , y  $I = \{(i, j, l) : (i, j) \in T \text{ y } (i, l) \notin T\}$ .

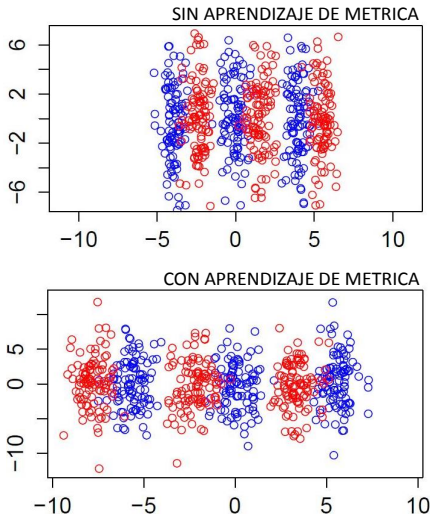
$$\min_M \mu \sum_T (D_M(x_i, x_j))^2 + (1 - \mu) \sum_I \zeta_{i,j,l} \quad \text{con restricciones:}$$

$$(1) \quad D_M(x_i, x_l)^2 - D_M(x_i, x_j)^2 \geq 1 - \zeta_{i,j,l} \quad (1)$$

$$(2) \quad \zeta_{i,j,l} \geq 0$$

$$(3) \quad 0 \preceq M.$$

# Ejemplo de Metrica Aprendida





## Clustering the Bins

Utilizamos *complete-linkage hierarchical clustering*. Para escoger la cantidad de *clusters* implementamos el siguiente indicador:

**Definimos**  $W(k) := \sum_{i=1}^k \sum_{x \in \mathfrak{C}_i} \|x - \hat{m}_i\|$

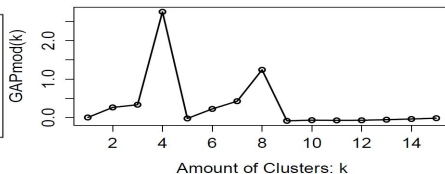
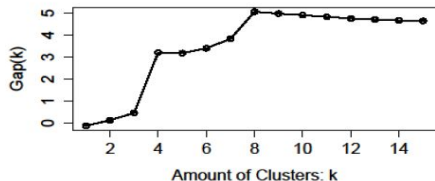
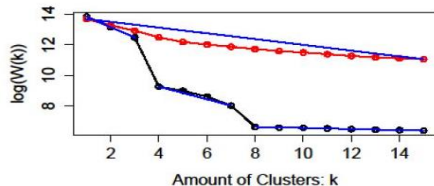
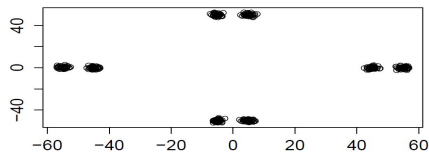
❶ **Gap Statistic:**  $Gap(k) = E_n^*(\log(W(k))) - \log(W(k))$

❷ **Modified Gap Statistic:**

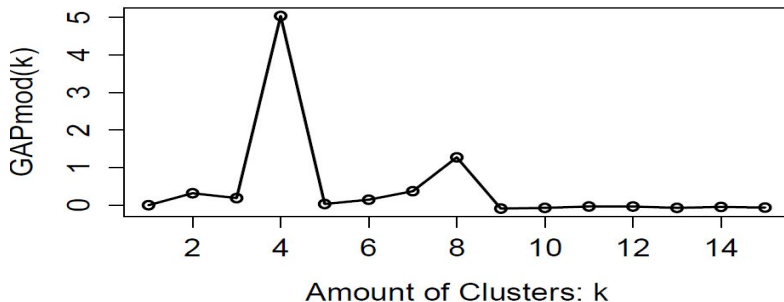
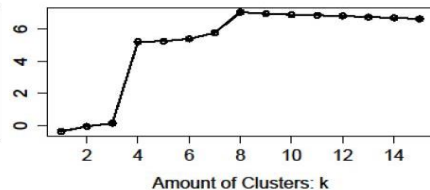
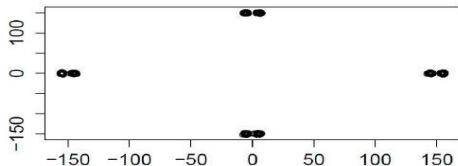
❶  $Gap_{mod}(k+1) = Gap(k+1) - Gap(k)$

❷  $Gap_{mod}(k+1) = \log\left(\frac{W(k+1)}{W(k)}\right)$

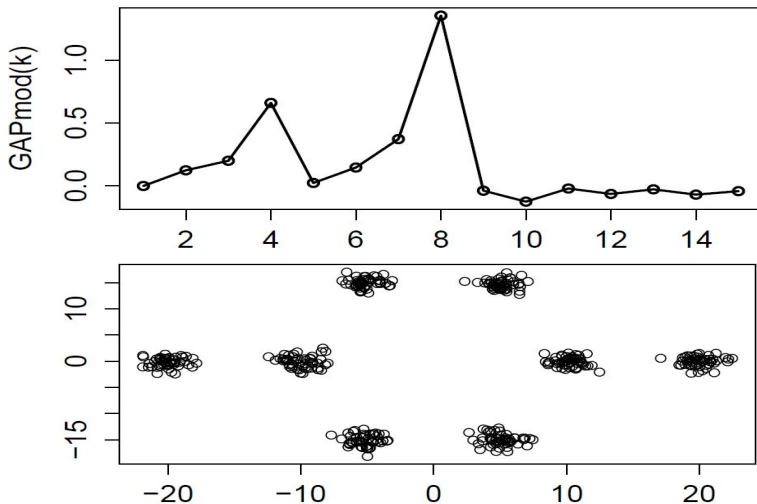
# Gap<sub>Mod</sub> vs Gap



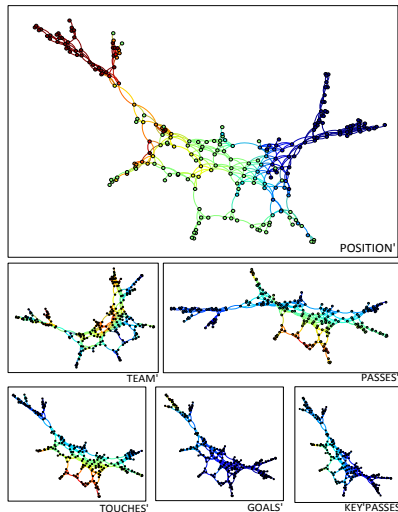
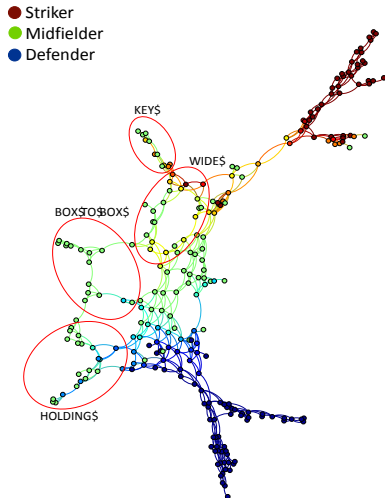
# Gap<sub>Mod</sub> vs Gap



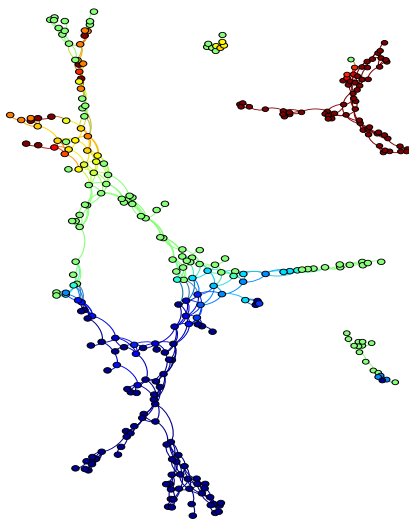
## $Gap_{Mod}$ vs $Gap$



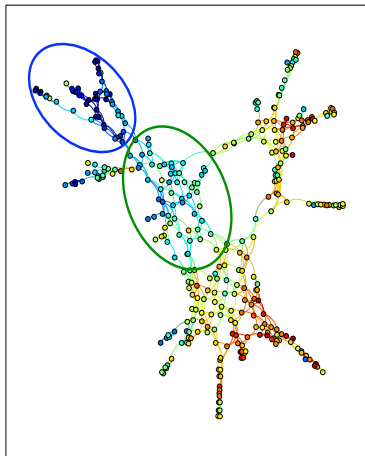
# Ejemplos de Resultados



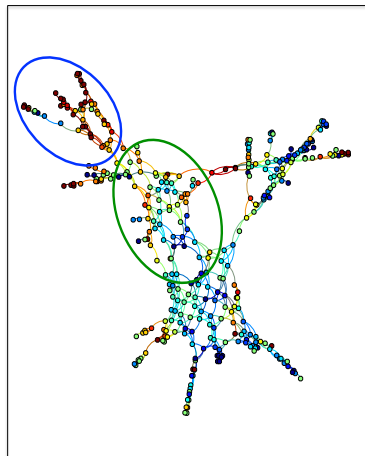
# Ejemplos de Resultados



# Ejemplos de Resultados

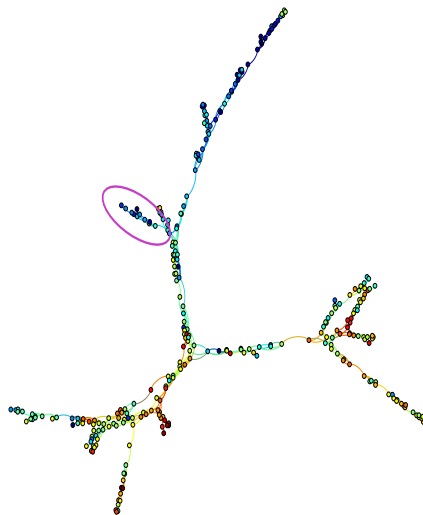


TEAM POSITION



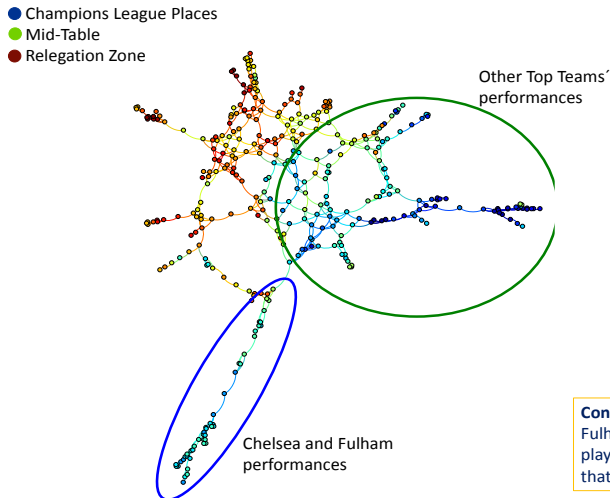
RESULTS

# Ejemplos de Resultados





# Ejemplos de Resultados



**Conclusion:** Chelsea and Fulham employ a similar playing style, distinct from that of other Top teams

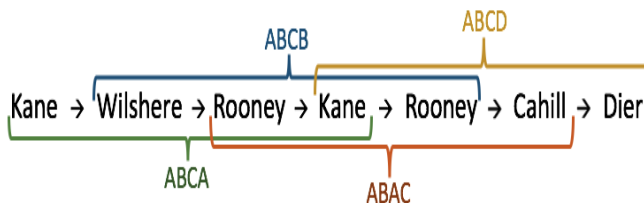
# Contenido

- 1 Ideas Pasadas: TDA with Metric Learning Applied to Football Data (Tesis de Maestria)
- 2 Ideas Presentes: Passing Network Motifs
- 3 Ideas Futuras: El Futbol como un Proceso de Markov

# Passing Motifs

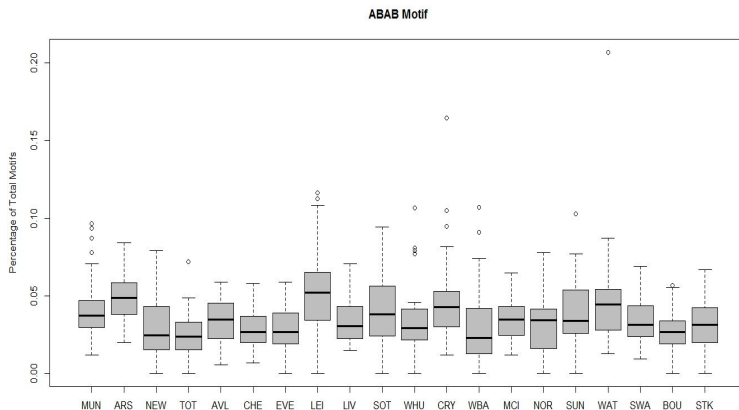
Relajando la identidad de los jugadores involucrados, una secuencia de 3 pases consecutivos tiene 5 “motifs” posibles:

- 1 ABAB
- 2 ABAC
- 3 ABCA
- 4 ABCB
- 5 ABCD

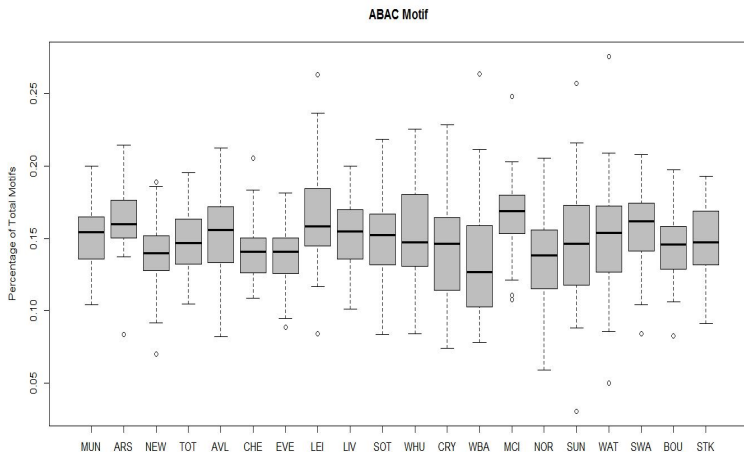


# Motif ABAB

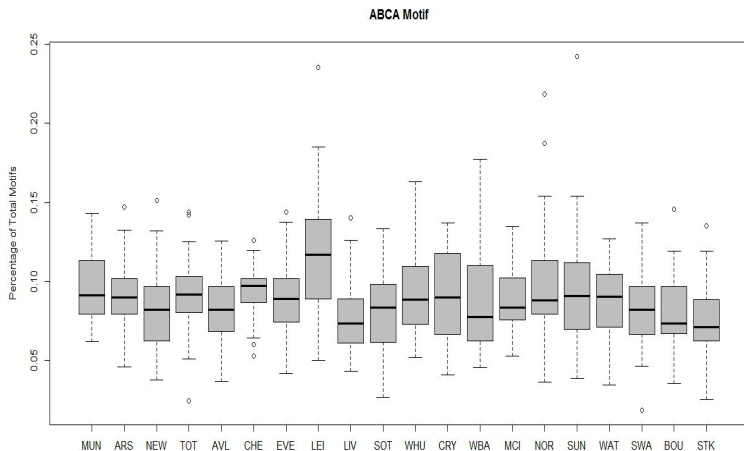
Miremos la distribucion de las frecuencias relativas de cada *motif* versus el total de *motifs* completados en cada partido por cada equipo. (Datos de la temporada 2015-16 de la Premier League)



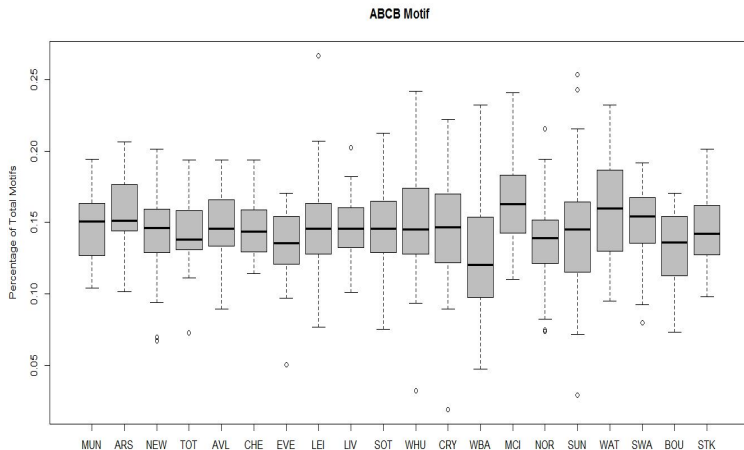
# Motif ABAC



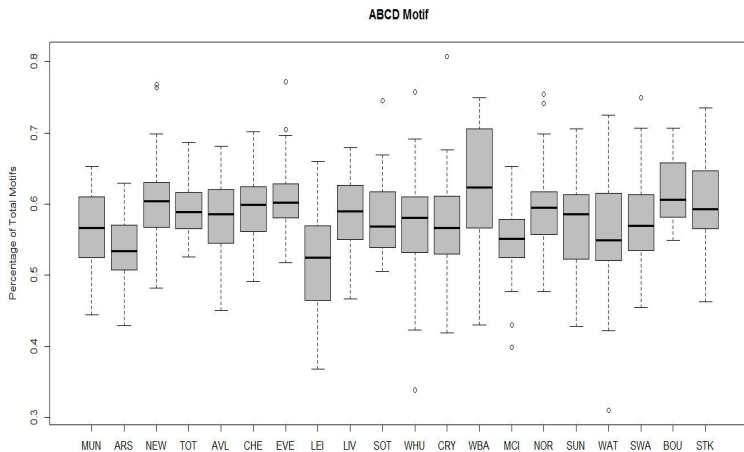
# Motif ABCA



# Motif ABCB



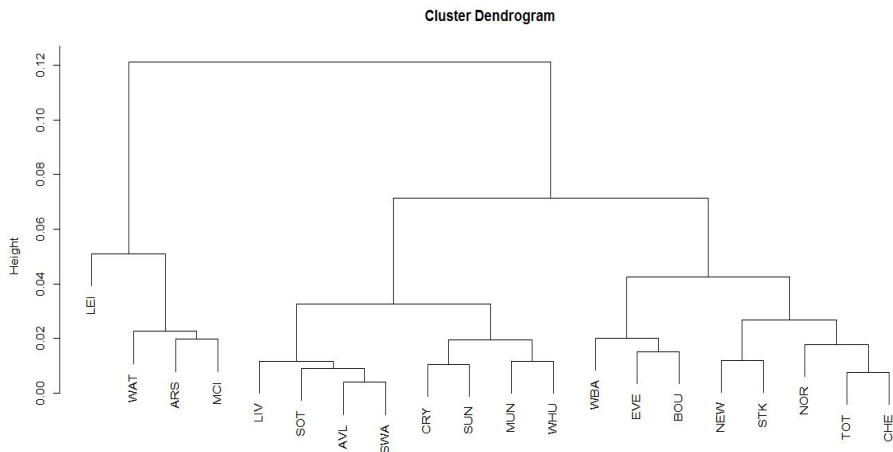
# Motif ABCD





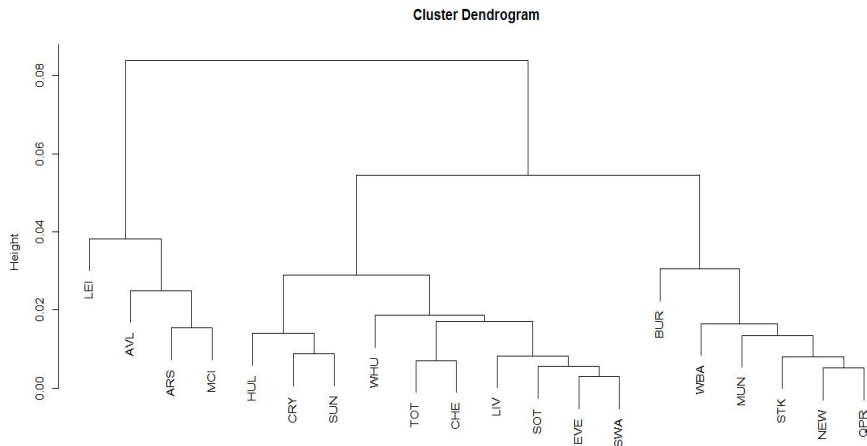
# Hierarchical Clustering of Motif Vectorization

Un equipo puede ser visto como un vector en  $\mathbb{R}^5$ , donde cada entrada corresponde al promedio de las frecuencias relativas de cada *motif*.



# Hierarchical Clustering of Motif Vectorization

Temporada 2014-15.



# Que mas se puede hacer con esto?

- Hacer TDA con los datos de los equipos en  $\mathbb{R}^5$ .
- Involucrar variables espacio-temporales (el angulo y la distancia del pase, y el tiempo transcurrido entre cada pase).
  - Si se hace PCA en las distancias de cada pase, los resultados tienen sentido; p.e. la varianza en las secuencias de tipo ABAC se concentra en el ultimo pase.
- **A nivel del jugador:** A nivel individual hay 15 motifs posibles; y tambien se pueden cuantificar las frecuencias (relativas o totales) de los jugadores para verlos como un vector en  $\mathbb{R}^{15}$ .

# Quien puede reemplazar a Xavi?

Articulo de Javier Lopez Peña y Raul Sanchez Navarro

| Player           | Distance | Similarity (%) |
|------------------|----------|----------------|
| Yaya Touré       | 4.495    | 18.199         |
| Thiago Alcántara | 5.835    | 14.631         |
| Sergio Busquets  | 6.494    | 13.345         |
| Andrés Iniesta   | 7.038    | 12.441         |
| Cesc Fàbregas    | 7.377    | 11.938         |
| Jordi Alba       | 7.396    | 11.910         |
| Toni Kroos       | 7.853    | 11.296         |
| Mikel Arteta     | 8.257    | 10.802         |
| Michael Carrick  | 8.505    | 10.521         |
| Santiago Cazorla | 8.515    | 10.509         |
| Daley Blind      | 9.154    | 9.849          |
| Paul Scholes     | 9.240    | 9.765          |
| Gerard Piqué     | 9.524    | 9.502          |
| David Silva      | 9.640    | 9.398          |
| Marcos Rojo      | 9.671    | 9.371          |
| Angel Rangel     | 9.675    | 9.368          |
| Samir Nasri      | 9.683    | 9.360          |
| Leon Britton     | 9.797    | 9.261          |
| Aaron Ramsey     | 9.821    | 9.241          |
| Martín Montoya   | 9.846    | 9.220          |

# Quien puede reemplazar a Xavi?

Artículo de Javier Lopez Peña y Raul Sanchez Navarro

| Size | Players                                                                                                                                                                                                                                 |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Xavi                                                                                                                                                                                                                                    |
| 2    | Dani Alves, Jordi Alba                                                                                                                                                                                                                  |
| 4    | David Silva, Lionel Messi, Samir Nasri, Santiago Cazorla                                                                                                                                                                                |
| 4    | Andrés Iniesta, Cesc Fàbregas, Thiago Alcántara, Yaya Touré                                                                                                                                                                             |
| 5    | Daley Blind, Gerard Piqué, Javier Mascherano, Sergio Busquets, Toni Kroos                                                                                                                                                               |
| 6    | Adriano, Angel Rangel, Bacary Sagna, Gaël Clichy, Marcelo, Martín Montoya                                                                                                                                                               |
| 8    | Emre Can, Isco, James Rodríguez, Juan Mata, Maicon, Mesut Özil, Michael Ballack, Ryan Mason                                                                                                                                             |
| 10   | Ashley Williams, Carles Puyol, Chico, Marc Bartra, Marcos Rojo, Michael Carrick, Mikel Arteta, Nemanja Matic, Paul Scholes, Sergio Ramos                                                                                                |
| 12   | Dejan Lovren, Garry Monk, John Terry, Jonny Evans, Ki Sung-yueng, Matija Nastasic, Michael Essien, Morgan Schneiderlin, Nabil Bentaleb, Per Mertesacker, Roberto Trashorras, Vincent Kompany                                            |
| 12   | Aaron Ramsey, Alexandre Song, Fernandinho, Gareth Barry, Jerome Boateng, Jonathan de Guzmán, Leon Britton, Luka Modric, Mahamadou Diarra, Mamadou Sakho, Steven Gerrard, Xabi Alonso                                                    |
| 15   | Ander Herrera, Eric Dier, Frank Lampard, Ivan Rakitic, Jamie O'Hara, Jordan Henderson, Michael Krohn-Dehli, Rafael van der Vaart, Rafinha, Sascha Riether, Scott Parker, Seydou Keita, Steven Davis, Vassiriki Abou Diaby, Wayne Rooney |

# Network Motifs en Otros Contextos

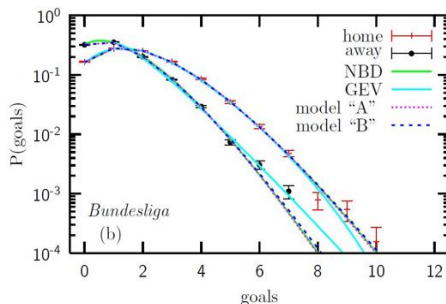
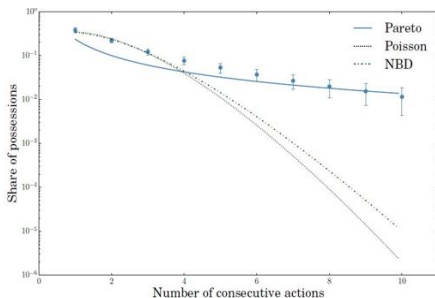
- Otros Deportes
- Colaboración Arquitectonica en un Archivo Central
- Comentarios o Participación en Redes Sociales

# Contenido

- 1 Ideas Pasadas: TDA with Metric Learning Applied to Football Data (Tesis de Maestria)
- 2 Ideas Presentes: Passing Network Motifs
- 3 Ideas Futuras: El Futbol como un Proceso de Markov

# Estadística Tradicional en el Futbol

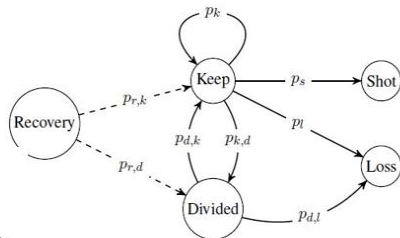
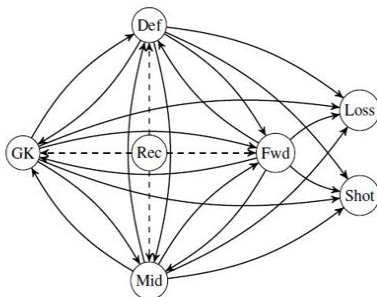
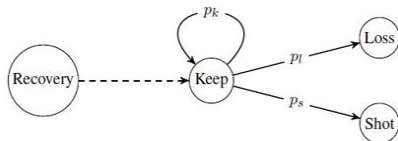
Tradicionalmente, principalmente ha habido interes en ajustar modelos para eventos sobre la cancha. En los 60's, Reep y Benjamin propusieron un modelo de Poisson para los pases consecutivos antes de perder el balon. Bittner et al. tambien intentaron diferentes modelos para la distribución de goles.





# Diferentes Acercamientos para un Modelo Markoviano

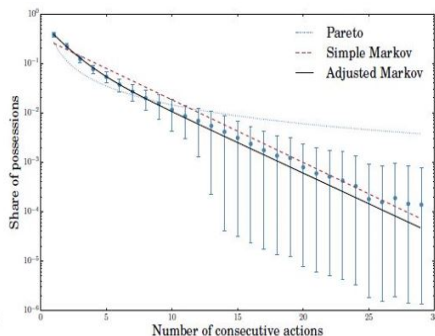
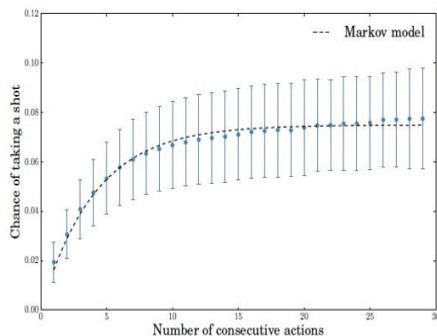
**Lopez Peña:** Propone ajustar modelos markovianos para las secuencias de posesion de un equipo.



# Diferentes Acercamientos para un Modelo Markoviano

## Lopez Peña:

- $P(X = x) = (1 - p_k)p_k^{x-1} \simeq Ce^{-\lambda x}$
- $P(X = x) \simeq C(e^{-\lambda x} + b^x)$
- $P(\text{Shot} \leq x) = \sum_{i=0}^{x-1} p_k^i p_s = p_s \frac{1 - p_k^x}{1 - p_k}$



# Diferentes Acercamientos para un Modelo Markoviano

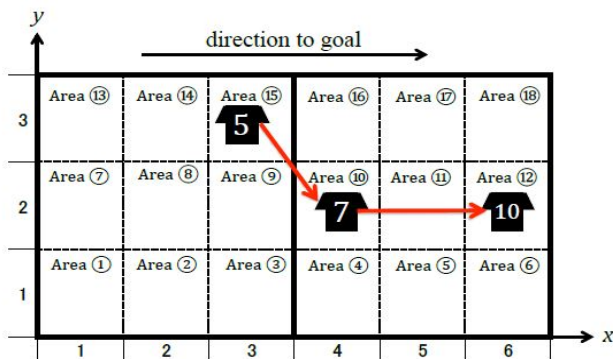
López Peña

Datos de la Temporada 2012-2013 de la Premier League

|    | Team                 | $p_k$        | $p_s$        | $b$          |
|----|----------------------|--------------|--------------|--------------|
| 1  | Manchester United    | 0.794        | 0.017        | 0.685        |
| 2  | Manchester City      | 0.794        | 0.016        | 0.683        |
| 3  | Chelsea              | 0.785        | 0.018        | 0.663        |
| 4  | Arsenal              | <b>0.797</b> | 0.015        | <b>0.690</b> |
| 5  | Tottenham Hotspur    | 0.782        | 0.016        | 0.653        |
| 6  | Everton              | 0.772        | 0.017        | 0.631        |
| 7  | Liverpool            | 0.788        | <b>0.019</b> | 0.672        |
| 8  | West Bromwich Albion | 0.771        | 0.018        | 0.630        |
| 9  | Swansea City         | 0.794        | 0.018        | 0.684        |
| 10 | West Ham United      | 0.756        | 0.018        | 0.589        |
| 11 | Norwich City         | 0.764        | 0.014        | 0.601        |
| 12 | Fulham               | 0.783        | 0.018        | 0.657        |
| 13 | Stoke City           | 0.752        | 0.015        | 0.575        |
| 14 | Southampton          | 0.772        | 0.015        | 0.630        |
| 15 | Aston Villa          | 0.771        | 0.016        | 0.619        |
| 16 | Newcastle United     | 0.771        | <b>0.019</b> | 0.624        |
| 17 | Sunderland           | 0.763        | 0.017        | 0.601        |
| 18 | Wigan Athletic       | 0.783        | 0.017        | 0.660        |
| 19 | Reading              | 0.749        | 0.016        | 0.558        |
| 20 | Queens Park Rangers  | 0.767        | 0.016        | 0.618        |

# Diferentes Acercamientos para un Modelo Markoviano

Narizuka, Yamamoto, Yamazaki:



Al probar su modelo de grafo con datos reales, encontraron que el coeficiente de clustering es de una orden de magnitud mayor a  $C_{rand}$ , exhibiendo *small world property*.

# Diferentes Acercamientos para un Modelo Markoviano

## Narizuka, Yamamoto, Yamazaki:

Si  $i$  y  $j$  corresponden al mismo jugador,  $P_{i \rightarrow j} = 0$  (esto es necesario?)

$$P_{i \rightarrow j} \simeq \frac{Q_{\alpha}(r_{ij}) \times R_{\beta, \xi}(L_j)}{Z}$$

Donde  $r_{ij}$  corresponde a la distancia entre las zonas de los nodos  $i$  y  $j$ , y  $Q_{\alpha}$  es una funcion que modela la probabilidad de hacer un pase de esa distancia.  $L_j$  representa la distancia de la zona de  $j$  a la “zona base” del jugador de  $j$ .  $R_{\beta, \xi}$  (el “fitness”) modela la probabilidad de que un jugador este a esa distancia de su “zona base”. Los autores la modelan asi:

$$R_{\beta, \xi}(L_j) = \begin{cases} 1, & \text{si } L_j \leq \xi \\ e^{-\beta(L_j - \xi)}, & \text{si } L_j > \xi \end{cases}$$

# Diferentes Acercamientos para un Modelo Markoviano

**Narizuka, Yamamoto, Yamazaki:**

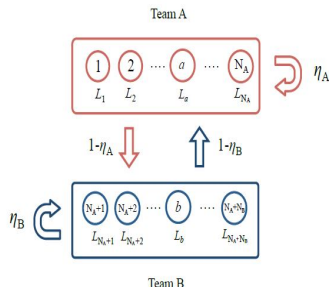
Involucrando a otro equipo:

$$P_{a' \rightarrow a} = \eta_A R_A(L_a) / Z_A$$

$$P_{b' \rightarrow b} = \eta_B R_B(L_b) / Z_B$$

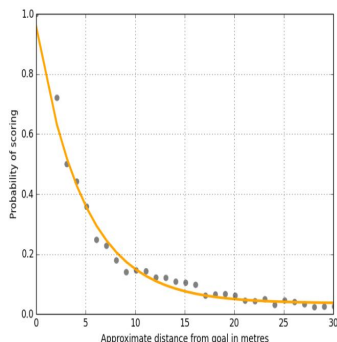
$$P_{a' \rightarrow b} = (1 - \eta_A) R_B(L_b) / Z_B$$

$$P_{b' \rightarrow a} = (1 - \eta_B) R_A(L_a) / Z_A$$



# Expected Goals Models

Hay una serie de modelos para valorar la opcion de marcar un gol dada una oportunidad o disparo. Eastwood ajusta una modelo sencillo de decrecimiento exponencial para la probabilidad de anotar con la distancia del arco.



Hay modelos mas complejos que involucran muchas mas variables predictoras.

# Podemos juntar estas ideas de forma convincente?

- Particionar la cancha en zonas/jugadores
- “Zona base” de jugador
- Probabilidad de trasladar de una zona/jugador a otro con un pase, involucrando “fitness” (por jugador?)
- Involucrar la precision de cada jugador (por zona?)
- Involucrar probabilidad de perderla y los estados del oponente
- Involucrar la probabilidad de disparar y “Expected Goals” (por jugador?)



# El Algoritmo de Metropolis-Hastings

Tenemos una funcion objetivo clara: Sea  $t_{max}$  el promedio de “acciones” que tiene un partido de futbol. Queremos optimizar el valor de *Expected Goals* para  $t = t_{max}$ . Que podemos controlar?

- **Táctica:** “Zonas bases” de los jugadores
- **Convocatoria Nacional:** Identidad de los jugadores
- **Fantasy Team:** Identidad de los jugadores

El algoritmo de Metropolis-Hastings suele converger rapido para este tipo de problemas.

## Pregunta: Y las contribuciones defensivas?

Estamos entrenados para cuantificar lo que está mas que lo que no está. Como cuantificamos las contribuciones defensivas?