DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

# Trabajo Práctico №1

**El Gran TP**

14 de Abril de 2017 — Métodos Numéricos



| Integrante | LU | Correo electrónico |
|---|---|---|
| Alvarez, Ezequiel | 421/13 | ezequiel.a.alvarez@gmail.com |
| Cámera, Joel Esteban | 257/14 | joel.e.camera@gmail.com |
| Sicardi, Sebastián Matías | 042/13 | smcsicardi@gmail.com |







**Facultad de Ciencias Exactas y Naturales**
Universidad de Buenos Aires
Ciudad Universitaria – Pabellón I, Planta Baja
Intendente Güiraldes 2160 – C1428EGA
Ciudad Autónoma de Buenos Aires, Rep. Argentina
Tel/Fax: +54 11 4576 3359
http://exactas.uba.ar

# Índice

**1. Introducción** **3**

**2. The Colley Matrix Method** **4**
2.1. Generación del sistema lineal . . . . . 4
2.2. La matriz de Colley es estrictamente diagonal dominante . . . . . 6
2.3. La matriz de Colley es simétrica definida positiva . . . . . 6

**3. Implementación** **7**
3.1. Estructuras de Datos . . . . . 7
3.2. Algoritmos . . . . . 7

**4. Experimentación** **8**
4.1. Análisis de Performance: Eliminación Gaussiana vs Factorización de Cholesky . . . . . 8
4.1.1. Conclusión . . . . . 11
4.2. Análisis Cualitativo del método Colley para el ranking de equipos . . . . . 12
4.2.1. Comparación de Resultados en el Ranking ATP . . . . . 12
4.2.2. ¡Go Jack Sock! Cómo optimizar los resultados para el mejor ranking posible ganando la menor cantidad de partidos . . . . . 14
4.2.3. Comparación de Resultados en el Ranking NBA . . . . . 16
4.2.4. ¿Es el método *justo*? . . . . . 19
4.3. Conclusión . . . . . 19

**5. Implementación del empate en la matriz de Colley** **21**

**6. Apéndice: Enunciado** **23**

# 1. Introducción

La confección de rankings y tablas de posiciones resultan, tanto para los aficionados como para toda persona que tiene algún interés en deportes, de vital importancia para obtener una comparación entre competidores en todo momento de la competencia y tener una definición consistente de *mejor*.

Las distintas naturalezas de los deportes al rededor del mundo hace que no siempre sea viable una competencia a modo de torneo *todos contra todos*, lo cual genera que no sea un problema sencillo comparar dos equipos que no se hayan enfrentado nunca. Es por esto que se necesitan métodos matemáticos que comparen los desempeños de los equipos a lo largo de un torneo.

Por ejemplo, la FIFA confecciona un ranking de las mejores selecciones de fútbol [1], pero como éstas no se enfrentan todas contra todas, realiza un ranking basado en los partidos, otorgándole un peso a cada uno dependiendo si es amistoso, eliminatoria o final del mundo.

En el presente trabajo analizaremos los métodos *Colley Matrix Method* y *Winning Percentage* para abordar éste problema y luego los compararemos con métodos utilizados en algunos deportes sobre casos reales para ver que comportamiento tienen y como difieren entre sí.

[1] http://www.fifa.com/fifa-world-ranking/procedure/men.html

# 2. The Colley Matrix Method

Históricamente en el fútbol Norte Americano los rankings se calcularon con encuestas a periodistas deportivos y entrenadores. Esto además de engorroso puede ser inconsistente y llevar a desacuerdos[2] . Wesley N. Colley propuso un sistema que no solo es automatizable, sino que es relativamente sencillo de entender y difícil de explotar. Principalmente:

1. no toma en cuenta historia ni tradición
2. es reproducible
3. usa un mínimo de asumpciones
4. no hace ajustes *ad hoc*
5. ajusta para la fuerza del schedule
6. ignora scores *inflados*
7. produce resultados con sentido común

## 2.1. Generación del sistema lineal

El método se provecha de *La Regla de Laplace de sucesos* y sólo necesita conocer el historial de partidos y resultados. La regla establece que, si sobre $k$ eventos se observan $s$ casos exitosos, entonces $(s+1)/(k+2)$ es un mejor estimador que calcular el simple porcentaje de ganados $s/k$. En base a esto, el problema se reformula como la resolución de un sistema de ecuaciones lineales para obtener el estimador de cada equipo.

El sistema se obtiene de la siguiente forma:

- Sea $\Gamma = \{1, 2, ..., K\}$ el conjunto de equipos que participan de la competencia.
- Dado $i \in \Gamma$, llamamos:
  - $n_i$ al total de partidos jugados.
  - $w_i$ al total de partidos ganados.
  - $l_i$ al total de partidos perdidos.
- Dados $i, j \in \Gamma$, llamamos $n_{ij}$ a la cantidad de partidos jugados entre $i$ y $j$. Por lo tanto, $n_{ij} = n_{ji}$.

Con esta notación, el estimador para la probabilidad de que el equipo $i$ gane el próximo partido es:

$$r_i = \frac{1 + w_i}{2 + n_i} = \frac{1 + w_i}{2 + w_i + l_i}$$

También, tenemos las siguientes nociones que se desprenden de lo visto hasta ahora:

- Sin información de los equipos se puede pensar que $r_i = 1/2$, $i \in \Gamma$.
- $n_i = w_i + l_i$, en donde $n_i$ incluye todos los partidos jugados (incluyendo repetidos). Llamamos $r_i^j$ al rating del $j-$ésimo oponente de $i$.

Con lo anterior se puede escribir:

$$\begin{aligned} w_i &= \frac{w_i - l_i}{2} + \frac{n_i}{2} \\ &= \frac{w_i - l_i}{2} + \sum_{j=1}^{n_i} \frac{1}{2} \\ &\approx \frac{w_i - l_i}{2} + \sum_{j=1}^{n_i} r_i^j \end{aligned}$$

Reemplazando 1/2 por los rankings correspondientes. Entonces, teniendo:

$$r_i = \frac{1 + w_i}{2 + n_i} \quad \text{y} \quad w_i = \frac{w_i - l_i}{2} + \sum_{j=1}^{n_i} r_i^j$$

Obtenemos:

[2] `https://en.wikipedia.org/wiki/Bowl_Championship_Series_controversies`

$$(2+n_i)r_i - \sum_{j=1}^{n_i} r_i^j = 1 + \frac{w_i - l_i}{2} \quad \text{para } i \in \Gamma$$

Esto lleva a un sistema lineal del tipo $Cr = b$, con $C \in \mathbb{R}^{KxK}$ y un vector $b \in \mathbb{R}^K$ tal que el ranking buscado $r \in \mathbb{R}^K$ es la solución del sistema.

Finalmente la matriz $C$, que se conoce como la matriz de Colley, se genera de la siguiente forma:

$$C_{ij} = \begin{cases} -n_{ij} & \text{si} \quad i \neq j \\ 2 + n_i & \text{si} \quad i = j \end{cases}$$

Y el vector $b$:

$$b_i = 1 + \frac{w_i + l_i}{2},\ i \in \Gamma$$

Por último, lo que queda es resolver el sistema generado para obtener el vector $r \in \mathbb{R}^K$ que contiene el rating de cada equipo.

Como ejemplo, tomamos un subconjunto de diez partidos del campeonato de NFL del 2007 extraidos del paper *Generalizing Google's PageRank to Rank National Football League Teams*[3].

Se tienen 6 equipos en donde a cada uno se lo mapea con un número:

| Carolina | Dallas | Huston | New Orleans | Philadelphia | Washington |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |

La entrada de los partidos es la siguiente:

| Nro. Equipo 1 | Goles Equipo 1 | Nro. Equipo 2 | Goles Equipo 2 |
|---|---|---|---|
| 1 | 16 | 4 | 13 |
| 2 | 38 | 5 | 17 |
| 2 | 28 | 6 | 23 |
| 3 | 34 | 1 | 21 |
| 3 | 23 | 4 | 10 |
| 4 | 31 | 1 | 6 |
| 5 | 33 | 6 | 25 |
| 5 | 38 | 4 | 23 |
| 6 | 27 | 2 | 6 |
| 6 | 20 | 5 | 12 |

Con estos datos la matriz de Colley y el vector $b$ que se generan son los siguientes:

$$C = \begin{bmatrix} 5 & 0 & -1 & -2 & 0 & 0 \\ 0 & 5 & 0 & 0 & -1 & -2 \\ -1 & 0 & 4 & -1 & 0 & 0 \\ -2 & 0 & -1 & 6 & -1 & 0 \\ 0 & -1 & 0 & -1 & 6 & -2 \\ 0 & -2 & 0 & 0 & -2 & 6 \end{bmatrix} \quad \text{y} \quad b = \begin{bmatrix} 1/2 \\ 3/2 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Si se resuelve el sistema, el vector $r$ de raitings del ejemplo queda de la siguiente forma:

$$r = \begin{bmatrix} 0{,}3597 & 0{,}616 & 0{,}6687 & 0{,}3149 & 0{,}5015 & 0{,}5392 \end{bmatrix}^t$$

Según estos resultados el ranking de equipos (del primero al último) es:

Houston Dallas Washington Philadelphia Carolina New Orleans

[3] Angela Y. Govan, Carl D. Meyer, and Rusell Albright. Generalizing google's pagerank to rank national football league teams. In Proceedings of SAS Global Forum 2008, 2008, página 2.

## 2.2. La matriz de Colley es estrictamente diagonal dominante

Esto es importante ya que permite que se pueda realizar Eliminación Gausseana para triangular la matriz sin el problema de encontrar un 0 como pivote.

La demostración se desprende de la definición casi instantáneamente ya que $C_{ij} = -n_{ij}$, donde $-n_{ij}$ es la cantidad de partidos jugados entre los equipos $i$ y $j$. Por lo tanto, la suma en modulo de toda la fila $i$ (menos el valor de la diagonal) da como resultado la cantidad de partidos del equipo $i$.

$$\sum_{j \neq i} |C_{ij}| = n_i$$

Pero $C_{ii} = 2 + n_i$, por lo que la matriz es e.d.d..

## 2.3. La matriz de Colley es simétrica definida positiva

Esto se deduce de la definición

$$C_{ij} = \begin{cases} -n_{ij} & \text{si} \quad i \neq j \\ 2 + n_i & \text{si} \quad i = j \end{cases}$$

donde claramente $-n_{ij} = -n_{ji}$ para todo $i, j \in \Gamma$.

Para ver que la matriz es definida positiva repasamos la demonstración del paper original[4]. Primero, vemos que la matriz de Colley podemos representarla de la siguiente manera,

$$C = 2I + \sum_{k}^{\text{todos los juegos}} G^k$$

donde $I$ es la matriz identidad y $G^k$ es la matriz operador que se agrega por cada uno de los $k$ partidos. $G^k$ va a ser de la forma $G^k_{ii} = G^k_{jj} = 1$, ya que en la diagonal siempre se suma uno a la cantidad de partidos jugados, y donde $G^k_{ij} = G^k_{ji} = -1$ representa $-n_{ij}$. El resto de las celdas están en 0. Finalmente la matriz $G^k$ queda:

$$C = \begin{bmatrix} 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 1 & \dots & -1 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & -1 & \dots & 1 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix}$$

Recordamos, que ser definida positiva implica

$$\forall v \in \mathbb{R}^k, v \neq 0 \quad \text{y} \quad v^t C v > 0$$

Por lo tanto, podemos reemplazar $C$ en $2I + \sum_k G^k$. Además, utilizando que la multiplicación de matrices es distributiva, $(A + B)v = Av + Bv$, por lo que separamos la inecuación para examinarla. La matriz $2I$ es trivialmente definida positiva, lo que nos deja por ver las matrices $G^k$.

Por la forma que tienen, multiplicar $G^k v$ da como un resultado un vector de ceros excepto en los indices $i$ y $j$ que tienen $r_i - r_j$ y $r_j - r_i$ respectivamente. Entonces, el producto $v^t(G^k v)$ queda

$$v^t(G^k v) = r_i(r_i - r_j) + r_j(r_j - r_i) = (r_i - r_j)^2 \geq 0$$

y con esto, como $(r_i - r_j)^2 \geq 0$ y $v^t(2Iv) > 0$, probamos que la matriz de Colley $C$ es definida positiva.

Finalmente, como $C$ es simétrica definida positiva (s.d.p.) es inversible y por lo tanto el sistema a resolver es compatible determinado (en otras palabras, tiene una única solución). También ser s.d.p. implica que tiene factorización de Cholesky, lo que nos permite usar este método junto con Eliminación Gausseana para resolverlo.

[4] `http://colleyrankings.com/matrate.pdf`

# 3. Implementación

Desde el principio decidimos acotarnos a un subset de C++ mínimo, estándar y sin dependencias externas. Básicamente solo clases de la *C++ Standard Library* y funciones propias (sin objetos, templates, etc.). Para poder facilitar el proceso de desarrollo, construimos un *Makefile* y una estructura de directorios sencilla. Esto nos permitió mas adelante generar varios ejecutables que compartieran la mayoría del código, específicamente *main.cpp* para el resultado y *tests.cpp* para el proceso de benchmarking.

## 3.1. Estructuras de Datos

Para simplificar los tipos y evitar repetir código similar decidimos que las matrices y vectores sean del tipo *vector<vector<double>>*[5] al que definimos con el alias *matriz*.

Los partidos los guardamos en un vector de la estructura *Partido*, que contiene la *id* los equipos que participaron, cuantos puntos hizo cada uno y la fecha. Esto es apropiado dado la naturaleza secuencial del parseo de los datos y el posterior procesamiento.

Luego, para la Matriz de Colley utilizamos un *map<int, Equipo >* [6] donde el *int* como clave es el número (*id*) del equipo en el csv y *Equipo* es una estructura que tiene el *índice* del equipo en la matriz (dado por el orden de aparición en el archivo de entrada), la cantidad de partidos jugados, la cantidad de partidos ganados y la cantidad de perdidos.

Dado que la *id* de los equipos no siempre es secuencial, vimos necesario asignarles el índice recién mencionado. Por esto es que usamos un mapa asociativo: nos permite pasar de *id* de equipo (lo que tiene el vector de *Partido*) a índices de la matriz de Colley.

Aprovechamos también que *map* mantiene las *keys* ordenadas[7], para luego brindar los resultados en el orden numérico original que tenían los equipos.

## 3.2. Algoritmos

Para resolver el sistema lineal de la matriz de Colley $Ax = b$, implementamos dos algoritmos: Eliminación Gausseana y factorización de Cholesky.

La eliminación Gausseana consiste en realizar operaciones entre filas de manera que la matriz quede triangular superior. Luego, es cuestión de hacer sustitución hacia atrás para despejar $b$.

Por otro lado, la factorización de Cholesky resuelve el sistema descomponiendo la matriz de Colley en un producto entre una matriz triangular inferior $L$ y su transpuesta $L^t$. Luego resolver el sistema consiste en resolver dos sistemas: $L^t x = y$ y $Ly = b$.

Como sabemos ambos algoritmos son aproximadamente $O(n^3)$, con el detalle de que Cholesky hace aproximadamente la mitad de operaciones que Eliminación Gausseana. Vale aclarar también, que no tomamos en cuenta pivotear ni otros métodos que ayuden con la estabilidad numérica.

[5] `http://www.cplusplus.com/reference/vector/vector/`
[6] `http://www.cplusplus.com/reference/map/map/`
[7] `http://www.cplusplus.com/reference/map/map/begin/`

# 4. Experimentación

## 4.1. Análisis de Performance: Eliminación Gaussiana vs Factorización de Cholesky

Gracias a que la matriz de Colley es simétrica definida positiva, podemos utilizar la factorización de Cholesky en contrapartida con la Eliminación Gaussiana. Como la factorización de Cholesky genera menor cantidad de operaciones (siempre hablando en términos de complejidad, aunque las complejidades de ambas son iguales: $\mathcal{O}(n^3)$ ), tenemos como hipótesis que debería funcionar más rápido que la eliminación gaussiana.

Para el cálculo del tiempo, no tomamos en cuenta la creación de la matriz de Colley ni la matriz $b$, sólo medimos los algoritmos para la resolución. También, consideramos como *outliers* los valores que están por arriba de tres veces la desviación estándar y los descartamos. Algo más que vale la pena destacar es que se hicieron 10000 corridas de cada algoritmo por cada instancia medida.

Para empezar corrimos los tests sobre los datos provistos por la cátedra, es decir, la temporada de la NBA y los resultados del circuito ATP.

Tomando como entrada los datos de la temporada de la NBA, obtuvimos el histograma de la figura 1. En éste se puede observar claramente como la eliminación Gausseana fue más eficiente en promedio que la factorización de Cholesky.

| | media | desviación estándar |
|---|---|---|
| eliminación Gausseana | 29.57e3 | 7.35e3 |
| factorización de Cholesky | 40.93e3 | 14.88e3 |

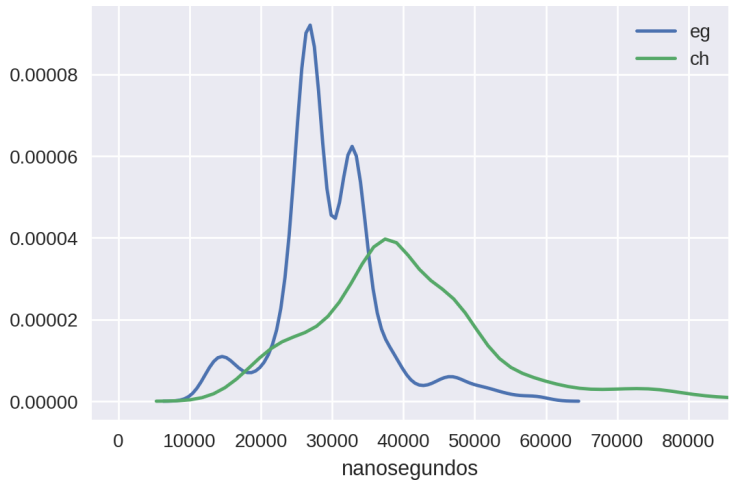



Figura 1: Histograma para los datos de la NBA provistos por la cátedra

Observamos que Cholesky no sólo tardó más en correr en promedio sino que los tiempos encima fueron más dispersos que los de la eliminación Gausseana. Esto contradice nuestra hipótesis de que la factorización de Cholesky iba a resultar más rápida. Pero esto no nos detuvo, ya que es un solo caso, y observamos los datos del circuito ATP.

| | media | desviación estándar |
|---|---|---|
| Eliminación Gausseana | 2.09e+7 | 0.36e7 |
| Factorización de Cholesky | 1.45e+7 | 0.12e7 |

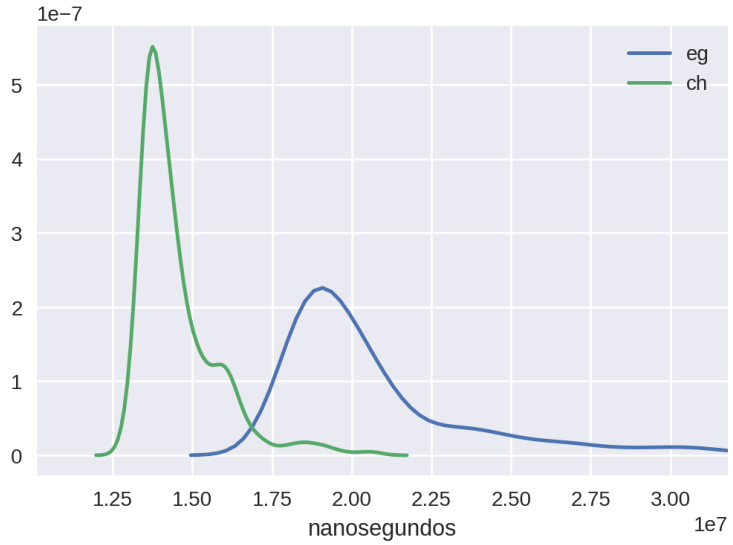



Figura 2: Histograma para los datos del circuito ATP provistos por la cátedra

Como se pude observar en la figura 2, los resultados con los datos del circuito ATP dieron al revés que con el torneo de la NBA: la factorización de Cholesky no sólo fue más rápida que la Eliminación Gausseana sino que también fue menos dispersa.

Por lo tanto, los resultados obtenidos resultan contradictorios. Esto hizo que nos preguntemos: Ya que las entradas y los tipos de torneos entre la NBA y el ATP son distintos, ¿cómo cambia la performance de los algoritmos a medida que cambiamos los parámetros de entrada?. Como vimos en el caso del torneo de la NBA, eliminación Gausseana obtuvo mejores resultados, pero, ¿para qué configuraciones de entrada la eliminación Gausseana es mejor? ¿Qué pasa con matrices con muchos equipos y pocos partidos (matriz rala)? ¿Qué pasa cuando la matriz es completa?.

Para responder a estas preguntas hicimos un script en python que genera instancias artificiales pseudoaleatorias con cantidad de equipos y partidos definidas como parámetros de entrada. Éste se llama *generate.py* y se puede encontrar en la carpeta *tests*. Para obtener números pseudoaleatorios utilizamos la librería **random**[8].

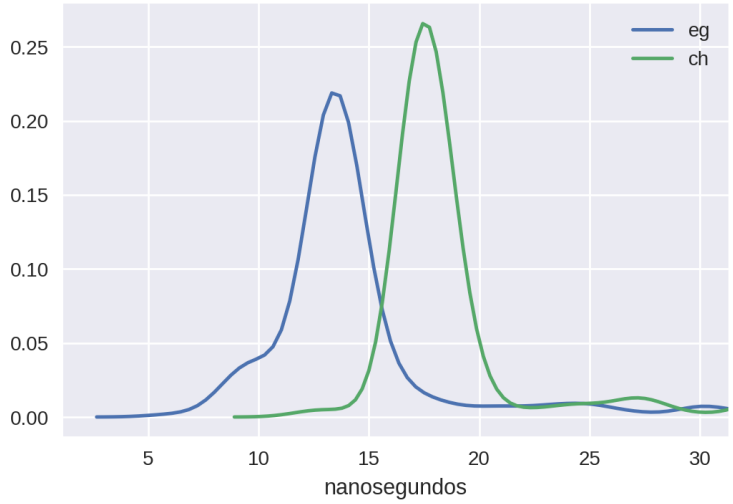



Figura 3: Histograma para una entrada generada aleatoriamente con 20 equipos y 1000 partidos

Cómo primera muestra, generamos una instancia de 20 equipos y 1000 partidos. Los resultados se pueden apreciar en la figura 3. Observamos como en un caso con pocos equipos y muchos partidos la factorización de Cholesky puede tardar mucho más que la Eliminación Gausseana, así como pasó con los datos de la NBA.

[8] `https://docs.python.org/2/library/random.html`

Al correr una instancia con 200 equipos y 1000 partidos en cambio vemos lo contrario. En la figura 4 vemos como la eliminación Gausseana tiene la media corrida hacia la derecha de la de Cholesky, en otras palabras en promedio tardo más.

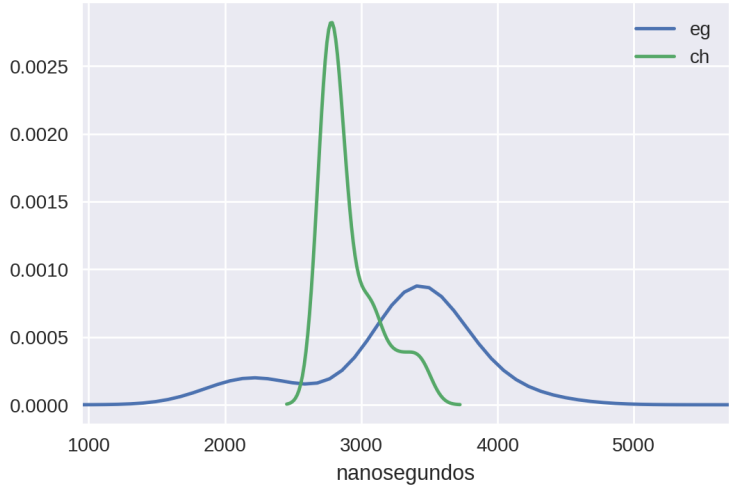



Figura 4: Histograma para una entrada generada aleatoriamente con 200 equipos y 1000 partidos

Para una instancia de 500 equipos y 1000 partidos la diferencia entre ambos algoritmos se hace más notoria aún. En la figura 5 es evidente como la factorización de Cholesky es siempre más rápida.

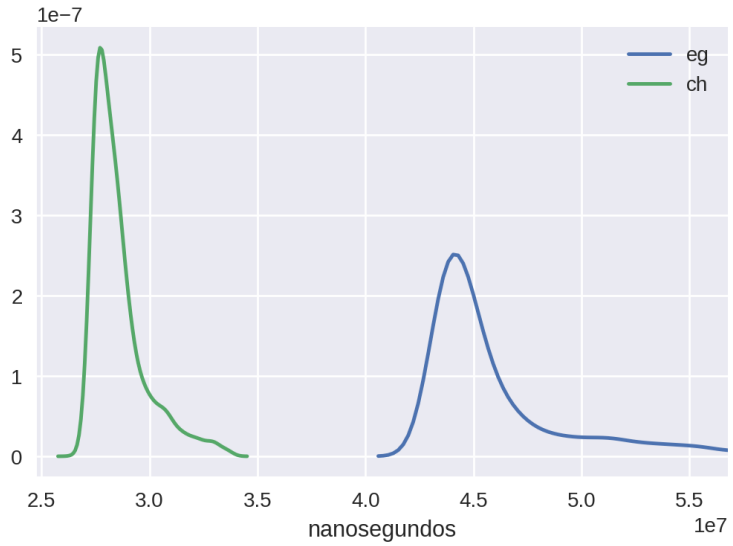



Figura 5: Histograma para una entrada generada aleatoriamente con 500 equipos y 1000 partidos

Estas ultimas mediciones nos hacen pesar que si los casos son lo suficientemente chicos Eliminación Gausseana tiene una mejor performance que factorización de Cholesky. Por ello, para poder observar mejor este comportamiento, decidimos experimentar con instancias donde la cantidad de equipos es creciente y donde la cantidad de partidos es creciente. Elegimos representarlo en un gráfico del estilo heatmap donde cada celda tiene el resultado de dividr el promedio de eliminación Gausseana sobre el de Cholesky, o sea que cuanto más rojo más rápido fue Cholesky y viceversa el azul.

En la figura 6, podemos observar el resultado del experimento. Si la cantidad de equipos va en aumento la performance de la factorización de Cholesky va mejorando con respecto a la de eliminación Gausseana. Este comportamiento puede deberse a que el algoritmo de Cholesky toma raíces cuadradas y no sólo operaciones

elementales como el otro (llamamos operaciones elementales a la suma, resta, producto y división), lo que produce que en matrices chicas no se amortice tanto el costo temporal del cálculo. Por otra parte, vemos que al aumentar la cantidad de partidos los algoritmos se mantienen estables en promedio, algo que no nos sorprende ya que lo único que cambia al aumentar la cantidad de partidos son los valores de la matriz pero no el tamaño de la misma.

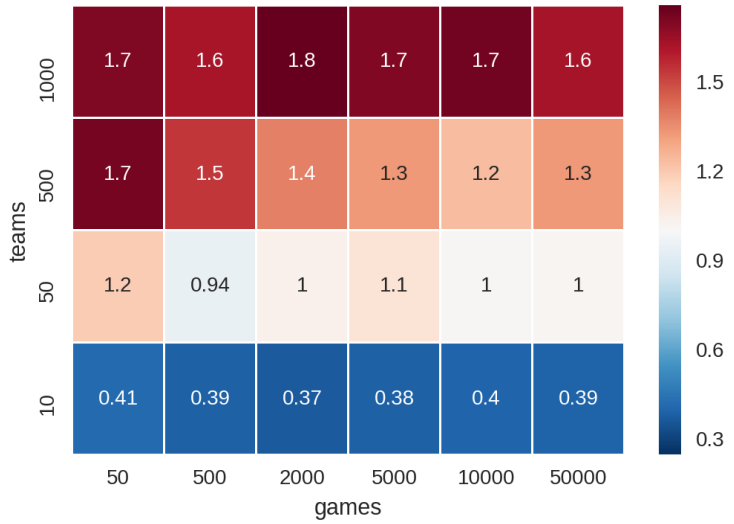



Figura 6: Heatmap de la relación entre los tiempos promedio de factorización de Cholesky y eliminación Gausseana con respecto a distinta cantidad de equipos y partidos.

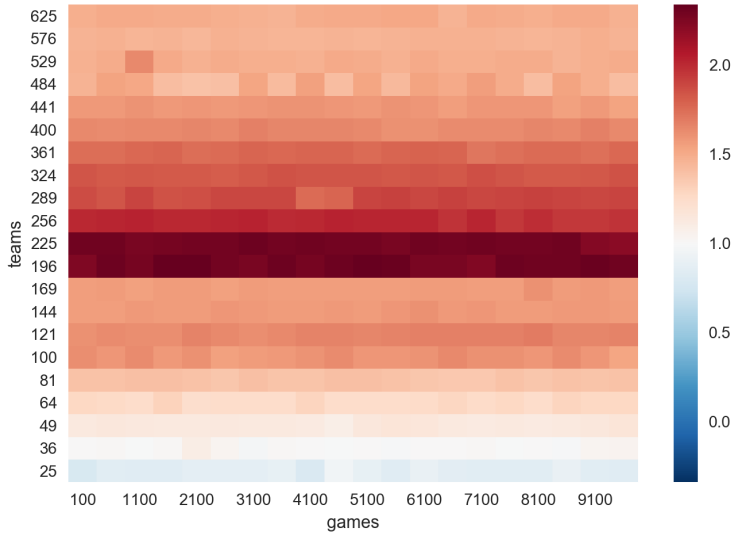



Figura 7: Heatmap de la relación entre los tiempos promedio de factorización de Cholesky y eliminación Gausseana con respecto a distinta cantidad de equipos y partidos.

Finalmente, para tener una versión más granular de la figura 6, podemos ver en la figura 7 cómo se comportan los algoritmos ante más casos de test. Claramente la cantidad de partidos no afecta la relación entre la performance de los algoritmos, pero si vemos que cambia cuando modificamos la cantidad de equipos. Vemos que la diferencia máxima sucede para matrices de doscientos equipos, para luego tender a 1.5.

### 4.1.1. Conclusión

Por último, como conclusiones finales del experimento, podemos decir que nuestra hipótesis no era completamente cierta. A pesar que teóricamente el algoritmo de eliminación Gausseana tenga más operaciones

que el de factorización de Cholesky, para una cantidad pequeña de equipos el primero funciona un poco mejor que el segundo. Pero si la cantidad de equipos es lo suficientemente grande, la factorización de Cholesky tiene una performance superior y más estable que su contraparte. Creemos que puede ser el tipo de operaciones y su costo lo que hace esta diferencia, como factorización de Cholesky utiliza menos operaciones que su contraparte puede amortizar sus operaciones más costosas (como la raíz cuadrada) cuando la matriz es muy grande, en cambio cuando la matriz es pequeña éstas operaciones se hacen más decisivas en lo temporal dandole ventaja a eliminación Gausseana. Además, creemos que esta diferencia de performance en matrices pequeñas pueda deberse a elecciones de implementación propias.

## 4.2. Análisis Cualitativo del método Colley para el ranking de equipos

A continuación procedemos a analizar el resultado final del método de Colley, es decir, el ranking generado. Como es muy difícil definir la calidad de un ranking isolado, procedemos a hacer varias comparaciones con torneos oficiales.

### 4.2.1. Comparación de Resultados en el Ranking ATP

Al comparar los resultados en el circuito ATP tomamos el ranking al 28 de Diciembre de 2015[9], para así tener en cuenta todos los partidos del año.

El ranking ATP se conforma por una serie de torneos de tipo *eliminación directa* los cuáles otorgan una cierta cantidad de puntos para cada jugador dependiendo de la fase en la que fue elminado, el campeón es el que obtiene la máxima cantidad de puntos (por un amplio margen, lo que genera que sea muy importante ganar un torneo). Los puntos dependen de la categoría del torneo, siendo los de mayor calibre los que más puntos otorgan (Grand Slams).

De esta forma, resulta interesante la comparación de los dos rankings, aunque hay que tener en cuenta que Colley considera cada partido de igual importancia, es decir, solo depende del rival, mientras que el ranking ATP considera la importancia del torneo. A continuación mostramos los dos Top 20 finales de cada ranking, junto con las diferencias de posición.

Ranking ATP

| N° | Jugador |
|---|---|
| 1 | Novak Djokovic |
| | Andy Murray |
| | Roger Federer |
| | Stan Wawrinka |
| 5 | Rafael Nadal |
| | Tomas Berdych |
| | David Ferrer |
| | Kei Nishikori |
| | Richard Gasquet |
| 10 | Jo Wilfried Tsonga |
| | John Isner |
| | Kevin Anderson |
| | Marin Cilic |
| | Milos Raonic |
| 15 | Gilles Simon |
| | David Goffin |
| | Feliciano López |
| | Bernard Tomic |
| | Benoit Paire |
| 20 | Dominic Thiem |

Ranking Colley

| N° | Jugador | Dif. ATP-Colley |
|---|---|---|
| 1 | Novak Djokovic | 0 |
| | Roger Federer | ↑1 |
| | Andy Murray | ↓1 |
| | Stan Wawrinka | 0 |
| 5 | Kei Nishikori | ↑3 |
| | Rafael Nadal | ↓1 |
| | Tomas Berdych | ↓1 |
| | David Ferrer | ↓1 |
| | Richard Gasquet | 0 |
| 10 | Milos Raonic | ↑4 |
| | Jo Wilfried Tsonga | ↓1 |
| | Kevin Anderson | 0 |
| | Jack Sock | ↑13 |
| | John Isner | ↓3 |
| 15 | Gael Monfils | ↑11 |
| | Gilles Simon | ↓1 |
| | Marcelo Arevalo | ↑266 |
| | Marin Cilic | ↓5 |
| | David Goffin | ↓3 |
| 20 | Roberto Bautista Agut | ↑5 |

Si bien son distintos y Djokovic se mantiene como el mejor jugador, cuesta encontrar otras coincidencias entre los dos órdenes. Por ejemplo, el oriundo de Lincoln, Nebraska, Jack Sock, escala 13 puestos en el ranking cuando se cambia de ATP a Colley, o el sorprendente caso de Marcelo Arévalo que se ve beneficiado ampliamente con Colley, ganando 266 puestos.

*El caso Marcelo Arévalo* resulta muy interesante, ¿cómo es que escala más de doscientos puestos solamente cambiando el tipo de ranking? Investigando un poco acerca de su historial en 2015 vemos que ganó 44

[9] `http://www.espn.com/tennis/rankings/_/year/2015`

partidos de 66 jugados[10] sin embargo en la base de datos proveída por la cátedra solo aparecen cuatro partidos, los cuatro ganados contra David Souto (puesto 124 en Colley), Ricardo Rodriguez (155), Gabriel Flores (229) y Alex Llompart (126).

Primero que nada, vale aclarar que los partidos no aparecen por ser torneos locales de bajo calibre, con lo cual los únicos que aparecen en los datos son los jugados por Copa Davis. Más allá del origen de los datos esto nos hace ver que Colley suele ser bastante volátil ante una baja cantidad de partidos, disparando injustamente el ranking de este jugador que no ha ganado partidos importantes.

Por otro lado tenemos *el caso Jack Sock*, que escala unos más razonables trece puestos en Colley, luego de corroborar que todos sus partidos están en la base de datos[11], podemos ver que el nebraskense tiene un record de 35 victorias y 18 derrotas, habiendo ganado un solo torneo, el ATP de Houston, el cual otorga muy pocos puntos (250, la más baja cantidad). Esto genera que no gane tantas posiciones en el ranking ATP pero sí en Colley que no se fija en los torneos y solo se preocupa por los partidos ganados y perdidos. De esta forma podemos ver que podríamos "engañar" a Colley haciéndole creer que Jack Sock es un mejor jugador de lo que realmente es.

Aparte de esto, la pregunta es, ¿qué pasa cuando nos alejamos del Top 20? Para visualizar un poco esto, se nos ocurrió agarrar el Top 100 de cada ranking y hacer *chunks* de a 5 jugadores en el orden que aparecen, y luego comparar en un heatmap cuantos jugadores en común tienen los *chunks* entre sí.

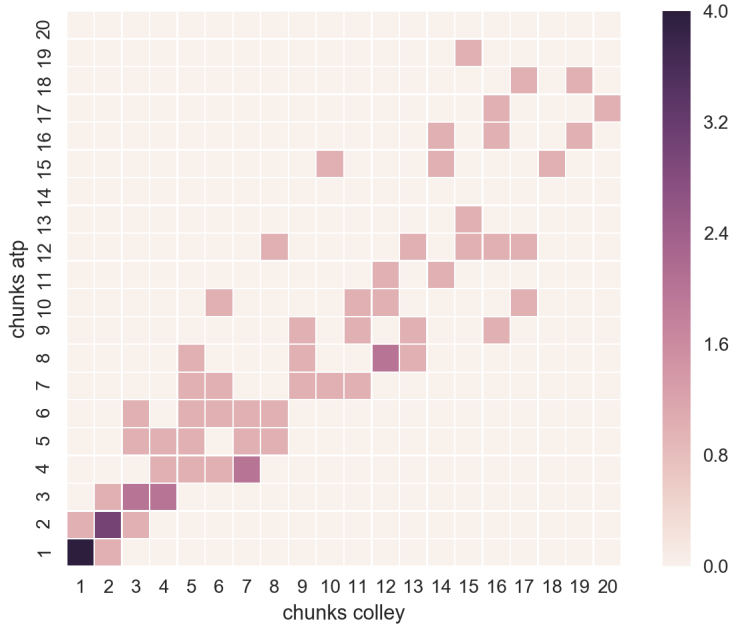



Figura 8: Heatmap de la cantidad de jugadores en común entre chunks de los rankings de Colley y ATP

En la figura 8 se ve mejor la diferencia entre los rankings, y como mientras al principio son muy parecidos, a medida que nos alejamos de los primeros puestos la diferencia tiende a ser mayor.

**¿Qué podemos aprender de todo esto?** El método de Colley puede ser interesante si aplicado en deportes de tipo *liga*, y es conveniente cuando no se puede organizar un *round-robin* por la cantidad de equipos o la indisponibilidad de fechas. Por lo tanto, en torneos que poseen *eliminación directa* como el circuito ATP, el método Colley tiene poco sentido. Como dijimos anteriormente, es difícil comparar Colley con el ranking ATP debido a que puede existir el caso de jugadores que lleguen a numerosas finales de torneos pero pierdan todas, por ende gana pocos puntos y no tiene un buen ranking en ATP, pero Colley lo "premiaría" excesivamente por ganar tantos partidos, por más que haya perdido las finales, como sucede en el caso del jugador de Lincoln, Nebraska, Jack Sock.

[10] `http://www.tennislive.net/atp/marcelo-arevalo/?y=2015`
[11] `http://www.tennislive.net/atp/jack-sock/?y=2015`

### 4.2.2. ¡Go Jack Sock! Cómo optimizar los resultados para el mejor ranking posible ganando la menor cantidad de partidos

A la hora de realizar este experimento elegimos a este jugador porque nos resultó interesante la diferencia entre su ranking ATP y el Colley, además de que al ver su calendario[12] vimos que tenía mucho espacio para ascender, ya que se ha enfrentado con varios jugadores en el Top 10, y había perdido.

Rápidamente podemos recomendar que una estrategia ganadora para el ranking Colley sería *ganarle a los mejores* si ya tenemos los resultados de todos los demás juegos. Sabiendo que Jack Sock había perdido contra los mejores jugadores exclusivamente, ¿qué pasaría si invertimos sus resultados? Esto implicaría que le ganó a todos los mejores jugadores, ¿sería una buena estrategia?

| N° | Jugador |
|---|---|
| 142 | Luis David Martinez |
| 143 | Jurgen Melzer |
| 144 | Jack Sock |
| 145 | Michael Venus |
| 146 | Illya Marchenko |

Los resultados no son nada esperanzadores, nuestro jugador pierde 131 lugares en el ranking, lo que implica que perder contra jugadores de muy bajo ranking naturalmente reduce nuestro puntaje, de esta forma debemos encontrar una estrategia mejor que nos garantize que el jugador.

Para una mayor facilidad a la hora de generar caso lo que hicimos fue crear un experimento sintético de la misma manera que se hizo la experimentación cuantitativa, generamos un torneo de 5 equipos y 20 partidos, y separamos el equipo 0 (que llamaremos más cariñosamente Jack) que jugó 9 partidos de esos 20. Luego, procedimosa ver cual es la mejor estrategia para mejorar el ranking de Jack dentro de este torneo.

Lo que hicimos fue generar todas las combinaciones posibles de partidos ganados y perdidos para Jack dentro de este torneo, y al no ver relación a simple vista hicimos una regresión:

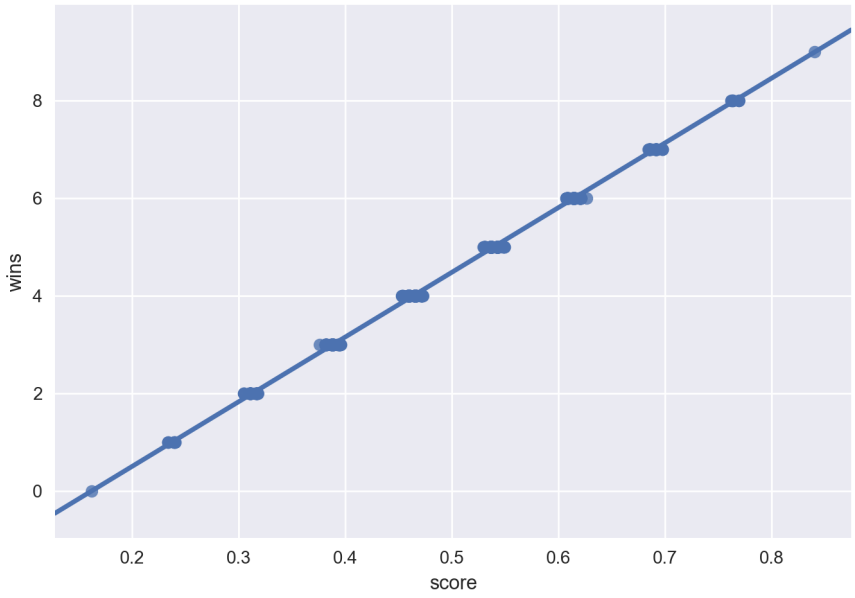



Figura 9: Relación de todas las combinaciones posibles de partidos ganados y perdidos para Jack.

Lamentablemente los resultados no son los que esperábamos, vemos que indefectiblemente no hay forma de ganar menos partidos y tener el mismo (o mejor) ranking. Esto se relaciona con los resultados obtenidos en el ATP, no encontramos una configuración que nos mejore el ranking ganando menos cantidad de partidos, incluso no había muchas configuraciones de igual cantidad de partidos ganados que cambien el ranking positivamente, solamente se logró esto agregando partidos ganados a rivales importantes, lo cual no consideramos relevante para esta sección.

[12] `http://www.tennislive.net/atp/jack-sock/?y=2015`

Por lo tanto, estos resultados resultan contradictorios con lo que estuvimos viendo de Colley, donde importaba mucho el ranking o la dificultad del partido a la hora de calcular el score. Esto es entendible en este caso ya que no estamos variando el *strenght of schedule* entre dos equipos, caso el cual sí podría pasar que un equipo que haya ganado menos esté mejor rankeado. Ante un mismo calendario de partidos lo importante es ganar la mayor cantidad posible.

De todas formas, podemos ver que Colley es bastante consistente en ese sentido, es exclusivamente mejor ganar más partidos que sólo ganarle a rivales importantes, lo cual consideramos algo bastante lógico, ya que, como dice los propios autores, lo que importa es ganar[13].

[13] `http://colleyrankings.com/advan.html`

### 4.2.3. Comparación de Resultados en el Ranking NBA

Antes de pasar a realizar comparaciones y para obtener una mejor observación de las mismas veamos como funciona el torneo de la NBA[14]. El mismo se realiza en dos etapas: La primera que llaman *Regular Season* y luego los *Playoffs*.

La NBA tiene una división particular de los equipos para el torneo. Cuenta con dos conferencias (la del este y la del oeste) donde cada una de estas tiene 3 divisiones con 5 equipos cada una.

En la *Regular Season*[15] cada equipo juega 82 partidos: enfrenta a equipos de su división 4 veces (16 partidos), enfrenta a equipos de su conferencia 4 veces (24 partidos) y los restantes 4 equipos 3 veces (12 partidos). Finalmente enfrenta cada equipo de la conferencia contraria 2 veces (30 partidos).

En esta Regular Season se generan dos rankings: uno para la conferencia del este y otro para la del oeste donde a los equipos se los rankea por el método de *Winning Percentage*.

Cuando termina la Regular Season, se toma los ocho mejores de cada ranking y se juegan los *Playoffs*[16] que es un torneo de eliminación donde para pasar de ronda hay que ganar al mejor de 7 partidos y de éste se desprende el campeón del torneo. Algo a destacar es que sólo hay victorias y derrotas, no hay empates en este torneo.

El funcionamiento de este torneo nos dice algo importante, el que más partidos gana puede que no salga campeón y ésta diferencia la dan los Playoffs ya que la Regular Season es por Winning Percentage. También podemos desprender que todos los equipos juegan entre sí en la Regular Season y algunos varias veces, esto hace que el Schedule de cada equipo no sea demasiado diferente de los demás, lo que difiere es en la etapa de Playoffs donde juegan sólo los mejores 16 equipos de la Regular Season.

Ahora bien, ¿que pasaría si utilizamos el método de Colley? Cómo el raiting de un equipo depende de los raitings de los equipos con los que jugó y de los partidos ganados esto puede hacer que se mantengan bastante parejos los resultados entre Colley y Winning Percentage en la Regular Season, pero que se modifique esta diferencia durante los Playoffs donde un equipo que no tuvo el mejor raiting en la primera parte del torneo puede salir campeón gracias a la segunda, o sea, le gane a equipos que tengan mejor raiting que él y aumente el suyo considerablemente.

Ya que los datos de los partidos de la NBA que la catedra brindó son de la temporada 2015/16, veamos como resultó ese torneo[17]:

**Regular Season**

**Conferencia del Este**

| N° | Jugador | W | L | % |
|---|---|---|---|---|
| 1 | Cleveland Cavaliers | 57 | 25 | .695 |
| 2 | Toronto Raptors | 56 | 26 | .683 |
| 3 | Miami Heat | 48 | 34 | .585 |
| 4 | Atlanta Hawks | 48 | 34 | .585 |
| 5 | Boston Celtics | 48 | 34 | .585 |
| 6 | Charlotte Hornets | 48 | 34 | .585 |
| 7 | Indiana Pacers | 45 | 37 | .549 |
| 8 | Detroit Pistons | 44 | 38 | .537 |
| 9 | Chicago Bulls | 42 | 40 | .512 |
| 10 | Washington Wizards | 41 | 41 | .500 |
| 11 | Orlando Magic | 35 | 47 | .427 |
| 12 | Milwaukee Bucks | 33 | 49 | .402 |
| 13 | New York Knicks | 32 | 50 | .390 |
| 14 | Brooklyn Nets | 21 | 61 | .256 |
| 15 | Philadelphia 76ers | 10 | 72 | .122 |

**Conferencia del Oeste**

| N° | Jugador | W | L | % |
|---|---|---|---|---|
| 1 | Golden State Warriors | 73 | 9 | .890 |
| 2 | San Antonio Spurs | 67 | 15 | .817 |
| 3 | Oklahoma City Thunder | 55 | 27 | .671 |
| 4 | Los Angeles Clippers | 53 | 29 | .646 |
| 5 | Portland Trail Blazers | 44 | 38 | .537 |
| 6 | Dallas Mavericks | 42 | 40 | .512 |
| 7 | Memphis Grizzlies | 42 | 40 | .512 |
| 8 | Houston Rockets | 41 | 41 | .500 |
| 9 | Utah Jazz | 40 | 42 | .488 |
| 10 | Sacramento Kings | 33 | 49 | .402 |
| 11 | Denver Nuggets | 33 | 49 | .402 |
| 12 | New Orleans Pelicans | 30 | 52 | .366 |
| 13 | Minnesota Timberwolves | 29 | 53 | .354 |
| 14 | Phoenix Suns | 23 | 59 | .280 |
| 15 | Los Angeles Lakers | 17 | 65 | .207 |

Como se puede apreciar en este ranking hecho con Winning Percentage, los equipos que más porcentage de ganados son Golden State Warriors con .809, seguidos de San Antonio Spurs con .817. Pero esto no hace que salgan campeones ya que como se puede ver en la figura 10 San Antonio Spurs no pasa las semifinales y Golden State logra llegar a la final pero pierde contra Cleveland que en la Regular Season tenía un porcentaje mucho menor (.695 contra .890).

[14] https://en.wikipedia.org/wiki/National_Basketball_Association
[15] https://en.wikipedia.org/wiki/National_Basketball_Association#Regular_season
[16] https://en.wikipedia.org/wiki/National_Basketball_Association#Playoffs
[17] Datos sacados de wikipedia. https://en.wikipedia.org/wiki/2015%E2%80%9316_NBA_season

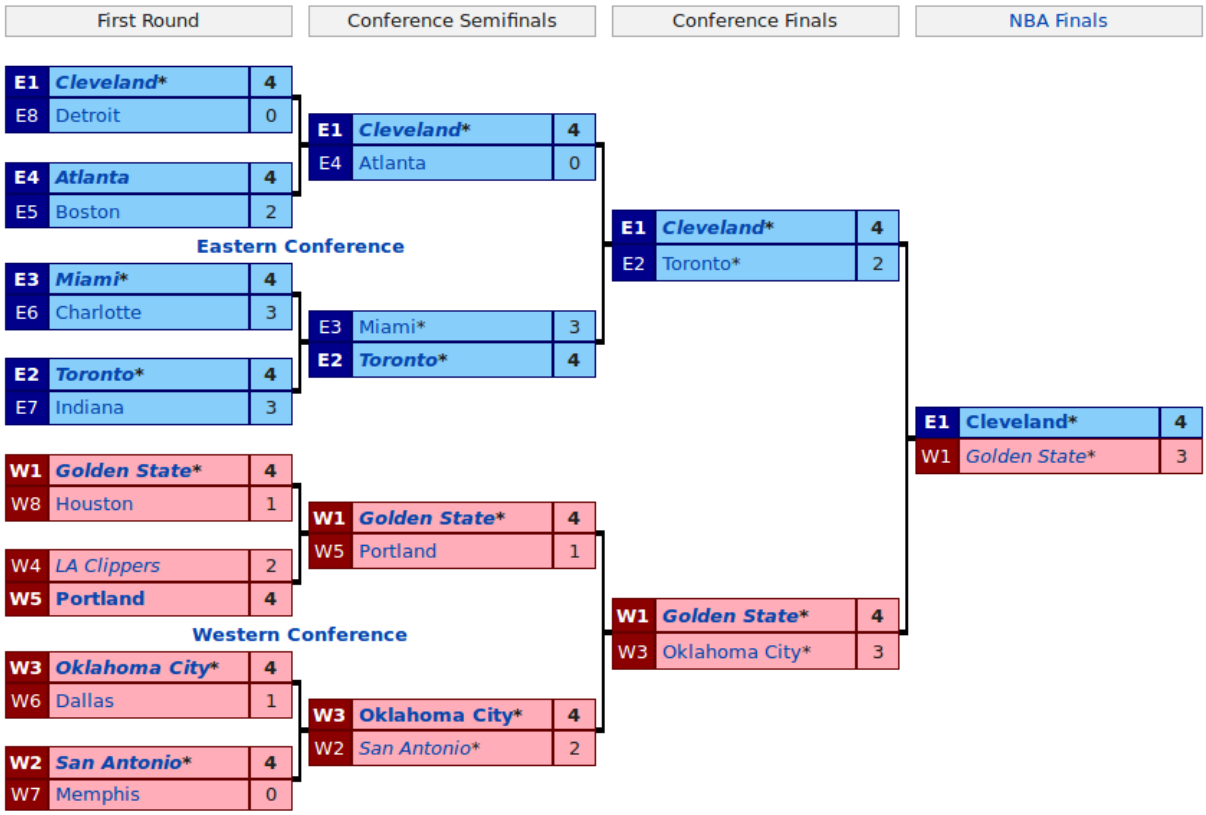



Figura 10: Playoffs de la NBA, temporada 2015/16

Cuando utilizamos los metodos de Colley y Winning Percentage para obtener el ranking de la misma temporada los resultados fueron los siguientes:

**Metodo de Colley**

| N° | Jugador |
|---|---|
| 1 | Golden State |
| | San Antonio |
| | Cleveland |
| | Toronto |
| 5 | Oklahoma City |
| | LA Clippers |
| | Miami |
| | Boston |
| | Memphis |
| 10 | Atlanta |
| | Charlotte |
| | Indiana |
| | Chicago |
| | Portland |
| 15 | Houston |
| | Dallas |
| | Detroit |
| | Utah |
| | Washington |
| 20 | Milwaukee |
| | Orlando |
| | Denver |
| | New York |
| | Sacramento |
| 25 | New Orleans |
| | Minnesota |
| | Phoenix |
| | Brooklyn |
| | LA Lakers |
| 30 | Philadelphia |

**Winning Percentage**

| N° | Jugador | Diferencia Colley/WP |
|---|---|---|
| 1 | Golden State | 0 |
| | San Antonio | 0 |
| | Cleveland | 0 |
| | Toronto | 0 |
| 5 | Oklahoma City | 0 |
| | LA Clippers | 0 |
| | Miami | 0 |
| | Boston | 0 |
| | Memphis | 0 |
| 10 | Atlanta | 0 |
| | Charlotte | 0 |
| | Indiana | 0 |
| | Portland | ↑1 |
| | Chicago | ↓1 |
| 15 | Houston | 0 |
| | Dallas | 0 |
| | Detroit | 0 |
| | Utah | 0 |
| | Washington | 0 |
| 20 | Orlando | ↑1 |
| | Milwaukee | ↓1 |
| | Denver | 0 |
| | New York | 0 |
| | Sacramento | 0 |
| 25 | New Orleans | 0 |
| | Minnesota | 0 |
| | Brooklyn | ↑1 |
| | Phoenix | ↓1 |
| | LA Lakers | 0 |
| 30 | Philadelphia | 0 |

Resulta interesante observar que en el torneo oficial de la temporada termina como campeón Cleveland y en los rankings de Colley y Winning Percentage solo llega al puesto tres. También, los equipos Golden State y San Antonio fueron los equipos que hicieron una mejor temporada y no salieron campeones ninguno de los dos pero si salió campeón Cleveland donde su temporada no fue tan buena como la de los dos primeros.

Por otra parte, sobre los rankings del metodo de Colley y Winning Percentage, las diferencias no fueron tan significativas entre ambos. Esto puede deberse a que en la Regular Season la cantidad de partidos y el schedule es muy similar para todos pero luego en los Playoffs, como algunos juegan más y son los que mejor temporada hicieron estos quedan en las primeras posiciones. También, el haber hecho una excelente temporada en la Regular Season hace que en estos rankings se mantenga en las primeras posiciones como es el caso de San Antonio.

Si queremos observar las tres diferencias de los rankings notamos que sólo fueron numericas entre equipos que no tuvieron una buena temporada y no pasaron a los Playoffs salvo el caso de Chicago/Portland. Portland pasa a los Playoffs y llega a la semi final donde pierde con Golden State, mientras que Chicago no pasa a los Playoffs. Es interesante ver que en el método de Colley Chicago tiene más raiting que Portland, sin embargo en Winning Percentage tiene sentido que Portland este arriba ya que jugó mas partidos y ganó mas en proporción. Lo que que pensamos que puede ser el de esta diferencia en el método de Colley sea el schedule de ambos equipos, que Chicago haya tenido un schedule un poco más difícil y por ello esté más arriba. También, cabe destacar que la diferencia de raitings de ambos es mínima: Chicago posee un raiting de 0.516007 mientras que Portland fue de 0.515941.

Por último, cabe destacar que para este tipo de torneo ambos métodos funcionan bastante parecidos mientras que en comparación con el Ranking oficial se comportan muy distinto. Entre Colley y Winning Percentage básicamente muestran que el que mejor temporada tuvo (ganó más de lo que perdió) termina primero pero en el Ranking oficial es más incierto ya que un equipo al que no le fue tan bien en la temporada de Regular Season pero que le alcanza para entrar a los Playoffs puede salir campeón.

#### 4.2.4. ¿Es el método *justo*?

Una pregunta válida que podría hacerse es: ¿Se puede afectar el "score" final de un equipo indirectamente?

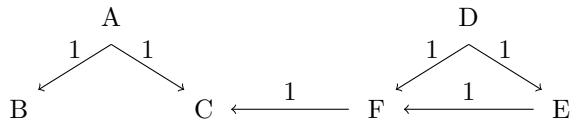
El tema de la *justicia* del método es complejo, y la analizamos mas adelante, pero ahora nos acotamos a analizar como cambia el "score" de un equipo en distintos casos.

Primero, veamos que pasa cuando no hay conexión entre dos equipos, en este caso $A$ y $D$.


La figura representa un "schedule" donde $A$ jugó con $B$ y $C$, y por otro lado $D$, $E$ y $F$ jugaron entre sí.

Si suponemos que $A$ ganó ambos partidos, entonces su "score" es 0,7 y el de $B$ y $C$ es 0,4, y **ninguna configuración** de ganados o perdidos entre $D$, $E$ y $F$ puede cambiar esto. Para ver esto usamos el archivo "justicia.in" ubicado en la carpeta "tests" y probamos varios casos.

Ahora si suponemos una conexión, por ejemplo un torneo con estos resultados,




donde las flechas representan la cantidad de partidos ganados (por ejemplo, F le ganó 1 partido a C), tenemos un "score" de 0,664 para $A$. Claramente fue afectado.

Para nosotros esto es algo positivo, ya que mas allá de como queden las posiciones finales, me dice que el "score" de un equipo depende de como juegue este y los equipos con los que juega. Y nada más.

### 4.3. Conclusión

Para finalizar, nos interesa responder las siguientes preguntas: ¿qué implica que Colley devuelva un ranking distinto al oficial del deporte?, ¿es Colley justo para estos deportes? (es decir, que no haya un jugador evidentemente peor mejor rankeado que otro muy bueno, con toda la subjetividad de estas palabras) y ¿qué implicaría implementar el método Colley en vez del oficial?

Inicialmente podemos ver que el método de Colley no implica un cambio radical en el orden en comparación con los oficiales, con algunas que otras excepciones, más allá de algunos cambios de puestos menores los jugadores o equipos más importantes tienden a mantenerse en los primeros puestos. Es decir, no vemos rankings radicalmente opuestos, como es de esperarse ya que en definitiva todo ranking busca que los equipos que más ganen sean considerados los mejores. Teniendo en cuenta esto, nos parece que Colley solamente ofrece otro tipo de ranking.

No es raro que haya deportes en los que se tienen en cuenta varios rankings, como el promedio en el futbol argentino o el Campeón Mundial No-Oficial[18]. Esto nos lleva a la segunda pregunta: ¿es Colley justo? rápidamente podríamos decir que no, como hemos visto en el ejemplo de Jack Sock que se lo rankeaba mejor de lo que realmente era, pero, en definitiva, ¿que sería que el ranking sea realmente *justo*? ¿No podría considerarse justo si las reglas son las mismas para todos y están establecidas desde el principio?

Es interesante notar que en la temporada de la NBA analizada, Golden State logró el récord de mayor cantidad de victorias en la temporada regular, para luego perder el título en la final contra Cleveland, ¿fue esto justo si la temporada de Golden State es considerada una de las mejores en la historia de la NBA y el deporte en general?[19] De nuevo, creemos que en definitiva el método de Colley es tan bueno como cualquier otro, mientras las reglas estén definidas de antemano. Es por esto que nadie reprocha la no consagración de Golden State, ya que era sabido que tenían que ganar los *playoffs*.

Finalmente, ¿qué pasaría si se implementase Colley para estos deportes? En el caso del circuito ATP generaría una revolución, ya que implicaría que ganar un torneo no sería cosa importante y ningún torneo sería de mayor calibre que otro (es normal que los grandes tenistas no jueguen torneos menores para prepararse para un Grand Slam) solamente importaría ganarle al mejor rankeado.

El caso de la NBA sería aún más interesante, ya que no generaría grandes cambios en la temporada regular. Esto implica que podría aplicarse fácilmente y resolver el problema del *strength of schedule*, y así

[18] `http://www.ufwc.co.uk/`

[19] `http://www.rollingstone.com/sports/features/are-the-golden-state-warriors-the-greatest-team-ever-20160414`

saber quién es el mejor equipo sin que se tengan que enfrentar indefectiblemente. Pero en definitiva, ¿cual es la gracia de un deporte si no se puede ver a los mejores equipos enfrentarse entre sí?

# 5. Implementación del empate en la matriz de Colley

Es importante tener en cuenta que el Método de Colley solamente considera deportes en los que hay un resultado binario *victoria* y *derrota*. A la luz de esto debemos considerar como seria una posible implementación del método en el que se pueda considerar un tercer resultado: el empate.

Primero que nada, veamos que pasa si agregamos el empate directamente a las ecuaciones de Colley. Supongamos que $d_i$ representa la cantidad de empates de $i$, y digamos que estos valen la mitad para el ranking $r_i$, entonces tenemos:

$$r_i = \frac{1 + w_i + \frac{d_i}{2}}{2 + n_i} \quad \text{y} \quad w_i = \frac{w_i - l_i - d_i}{2} + \sum_{j=1}^{n_i} r_i^j$$

donde reemplazando,

$$(2 + n_i)r_i - \sum_{j=1}^{n_i} r_i^j = 1 + \frac{w_i - l_i - d_i}{2} + \frac{d_i}{2} \quad \text{para } i \in \Gamma$$

vemos que claramente $d_i$ se cancela. O sea, nos queda el mismo sistema.

Por otro lado, el problema resulta interesante porque a diferencia de la victoria (que siempre es positiva para el equipo) y la derrota (que es negativa), puede suceder que un equipo considere positivo empatar contra otro de mayor "ranking", y este equipo que obtuvo un empate contra otro de menor calibre lo considerase algo negativo.

Siguiendo la deducción matemática que hicimos, decidimos que los empates simplemente sumen un partido jugado. Esto no modifica las propiedades de la matriz de Colley ya que se mantiene estrictamente diagonal dominante y simétrica definida positiva. Lo que modifica son los ratings de los equipos que empataron ya que el rating del equipo $i \in \Gamma$ tendría el denominador más grande.

Esto podría generar que un equipo con un rating alto lo disminuya en cambio uno con un rating bajo lo aumente, dependiendo del caso.

Para ver como se comporta en la práctica, generamos el siguiente caso artificial de 4 equipos en donde los círculos significan partidos ganados y las cruces partidos perdidos. También esta su respectivo ranking generado por la tabla de Colley. Todos los equipos jugaron un partido entre sí.

**Resultados**

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | - | O | O | O |
| 1 | X | - | O | X |
| 2 | X | X | - | X |
| 3 | X | O | O | - |

**Ranking**

| equipo | rating |
|---|---|
| 0 | 0.75 |
| 3 | 0.583333 |
| 1 | 0.416667 |
| 2 | 0.25 |

Si cambiamos un partido por un empate entre el mejor equipo (0) y el peor equipo (2) vamos a notar como esto perjudicaría al primero y mejoraría al segundo.

**Resultados**

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | - | O | E | O |
| 1 | X | - | O | X |
| 2 | E | X | - | X |
| 3 | X | O | O | - |

**Ranking**

| equipo | rating |
|---|---|
| 0 | 0.666667 |
| 3 | 0.583333 |
| 1 | 0.416667 |
| 2 | 0.333333 |

Es interesante ver como sólo modificó los ratings de los equipos involucrados y que perjudica mucho al mejor equipo mientras que beneficia mucho al peor aunque no lo suficiente como para modificar sus posiciones en esta tabla. También, la cantidad que se le agrega al equipo 2 es la misma que se le quita al equipo 0 (0.083333).

Ahora, ¿Que pasaría si un equipo empata todos los partidos?

**Resultados**

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | - | O | E | O |
| 1 | X | - | E | X |
| 2 | E | E | - | E |
| 3 | X | E | O | - |

**Ranking**

| equipo | rating |
|---|---|
| 0 | 0.666667 |
| 3 | 0.5 |
| 2 | 0.5 |
| 1 | 0.333333 |

Se puede ver como el equipo 2 le fue mejor y pudo subir posiciones en la tabla. Algo que cabe destacar es que la diferencia que se está modificando es un múltiplo de 0.083333.

Raiting del equipo $1 = 0{,}416667 - 0{,}083333 = 0{,}333333$

Raiting del equipo $2 = 0{,}333333 + 2 * 0{,}083333 = 0{,}5$

Raiting del equipo $3 = 0{,}583333 - 0{,}083333 = 0{,}5$

Entonces, reemplazar un partido por un empate hace que el valor 0.083333 se le quite al que ganó y se le agregue al que perdió.

Ahora bien, si a la tabla anterior le agregamos un empate entre el equipo 0 y el 1 (o sea, agregamos un partido más) podemos observar como se comporta el ranking:

**Resultados**

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | - | O/E | E | O |
| 1 | X/E | - | E | X |
| 2 | E | E | - | E |
| 3 | X | E | O | - |

**Ranking**

| equipo | raiting |
|---|---|
| 0 | 0.625 |
| 3 | 0.5 |
| 2 | 0.5 |
| 1 | 0.375 |

Podemos ver que afecta negativamente al equipo 0 y positivamente al 1. Esto tiene sentido ya que le estaría quitando peso a las victorias mientras que, como el equipo 1 tiene mas derrotas, este empate hace que no le pesen tanto. Ahora bien, si agregamos un empate entre el equipo 2 y el 3 que tienen un mismo raiting ocurre lo siguiente:

**Resultados**

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | - | O/E | E | O |
| 1 | X/E | - | E | X |
| 2 | E | E | - | E/E |
| 3 | X | E/E | O | - |

**Ranking**

| equipo | rating |
|---|---|
| 0 | 0.625 |
| 3 | 0.5 |
| 2 | 0.5 |
| 1 | 0.375 |

Se puede notar que los ratings de ambos equipos se mantienen iguales.

Como conclusión, creemos que la repercusión de los empates en el método Colley puede ser útil ya que un empate sólo afecta a los equipos que empataron si tienen ratings distintos entre ambos con un efecto negativo al que tenía el mejor rating y uno positivo al peor. Si los ratings son iguales entre los equipos que empatan, un empate sólo los mantiene sus ratings iguales entre sí.

# 6. Apéndice: Enunciado

**Métodos Numéricos**
Primer Cuatrimestre 2017
**Trabajo Práctico 1**

*Departamento de Computación*
*Facultad de Ciencias Exactas y Naturales*
*Universidad de Buenos Aires*

**El gran TP**

---

**Contexto y motivación**

Las competencias deportivas, en todas sus variantes y disciplinas, requieren casi inevitablemente la comparación entre competidores mediante la confección de *Tablas de Posiciones* y *Rankings* en base a resultados obtenidos en un período de tiempo determinado. Estos ordenamientos de equipos están generalmente (aunque no siempre) basados en reglas relativamente claras y simples, como proporción de victorias sobre partidos jugados o el clásico sistema de puntajes por partidos ganados, empatados y perdidos. Sin embargo, estos métodos simples y conocidos por todos muchas veces no logran capturar la complejidad de la competencia y la comparación. Esto es particularmente evidente en ligas donde, por ejemplo, todos los equipos no juegan la misma cantidad de veces entre sí.

A modo de ejemplo, la NBA y NFL representan dos ligas con fixtures de temporadas regulares con estas características. En los últimos tiempos, el Torneo de Primera División de AFA se sumó a este tipo de competencias, ya que la incorporación de la *Fecha de Clásicos* parece ser una interesante idea comercial, pero no tanto desde el punto de vista deportivo ya que cada equipo juega contra su *clásico* más veces que el resto. Como contraparte, éstos rankings son utilizados muchas veces como criterio de decisión, como por ejemplo para determinar la participación en alguna competencia de nivel internacional. En el caso de competencias en los Estados Unidos, las posiciones finales determinan cuál es la prioridad entre los equipos para la elección de los nuevos jugadores que ingresan a la liga mediante el conocido proceso de *Draft*. Luego, la confección de los rankings finales de los equipos constituye un elemento sensible, afectando intereses deportivos y económicos de gran relevancia.

En un contexto de extremada desconfianza respecto a los manejos a nivel local, regional e internacional de las confederaciones de fútbol, en este trabajo nos proponemos estudiar el comportamiento de otras métricas para la generación de rankings en competencias deportivas con el fin de brindar mayor transparencia y nivelar la competitividad, en un futuro, de nuestras ligas locales.

**El problema**

Existen en la literatura distintos enfoques para abordar el problema de determinar el *ranking* de equipos de una competencia en base a los resultados de un conjunto de partidos. En Govan et al. [5] se hace una breve reseña de varios de ellos, e incluso los autores proponen uno nuevo.[1] Entre los métodos presentados se encuentra el denomiando *Colley Matrix Method* (CMM) [1, 5]. El método se basa en la *Regla de Laplace de sucesos* y solo requiere conocer el historial de partidos y los respectivos resultados (básicamente, quién ganó) de los mismos. Esta regla permite aproximar las probabilidades de eventos *booleanos*, en nuestro caso que un

[1] Remarcamos que este no es el método involucrado en el TP1. Será visto en el segundo tercio de la materia.

equipo gane o pierda un partido. En particular, si sobre $k$ eventos observamos $s$ casos exitosos, la regla establece que $(s+1)/(k+2)$ es un mejor estimador que el porcentaje estándar, $s/k$. En base a esta idea, el problema se reformula como la resolución de un sistema de ecuaciones lineales, que permite obtener estos estimadores y, por lo tanto, el ranking deseado.

Extendiendo la notación introducida en Govan et al. [5], sea $\Gamma = \{1, 2, \ldots, T\}$ el conjunto de participantes de la competencia. Para cada equipo $i \in \Gamma$ llamamos $n_i$ al número total de partidos jugados por el equipo $i$, $w_i$ al número de partidos ganados por el equipo $i$ y, análogamente, $l_i$ al número de partidos perdidos por el equipo $i$. Definimos también dados $i, j \in \Gamma$, $i \neq j$, $n_{ij}$ al número de enfrentamientos entre $i$ y $j$. Es importante destacar que el modelo asume que el empate no es un resultado posible.

El método CMM propone construir una matriz $C \in \mathbb{R}^{T \times T}$ y un vector $b \in \mathbb{R}^T$, tal que el ranking buscado $r \in \mathbb{R}^T$ es la solución del sistema $Cr = b$. Para el armado del sistema, se define

$$C_{ij} = \begin{cases} -n_{ij} & \text{si } i \neq j, \\ 2 + n_i & \text{si } i = j. \end{cases}$$

y $b_i = 1 + (w_i - l_i)/2$, $i \in \Gamma$.

Los detalles respecto a la formulación del sistema pueden ser consultados en Colley [1]. Este método puede ser aplicado a una gran variedad de deportes y tipos de competencias, incluyendo información de conferencias, divisiones, etc. El objetivo central de este trabajo práctico consiste en estudiar el comportamiento del mismo, en conjunto con el análisis de algunos de los métodos que pueden ser utilizados para su resolución.

Como punto de comparación, se considerará (al menos) un método alternativo para generar rankings. Una opción es considerar el *porcentaje de victorias* (WP), donde el puntaje asignado al equipo $i \in \Gamma$ está dado por $w_i/(w_i+l_i)$. En caso de ser factible, es posible también incorporar el método que se aplique en la competencia elegida.

**Enunciado**

Se debe implementar un programa en `C` o `C++` que tome como entrada el detalle de los partidos de la competencia y calcule el ranking en función de los métodos mencionados en la sección anterior (CMM, WP ó el método elegido por el grupo). El formato de los archivos se detalla en la siguiente sección.

La matriz resultante del sistema planteado por el método CMM es *Simétrica y Definida Positiva* (ver, e.g., [1]) y, por lo tanto, es posible encontrar la Factorización de Cholesky para resolver el sistema. Luego, como parte obligatoria en relación a los métodos de resolución de sistemas de ecuaciones lineales se pide implementar:

- el método de Eliminación Gaussiana clásico (EG), y
- el método de Cholesky (CL).

Es importante incluir en el informe del TP, en la sección desarrollo, aquellas decisiones tomadas en función de la estructuras de datos utilizadas y las alternativas consideradas y descartadas durante el proceso. Además, sabemos que existen casos donde el algoritmo EG

no puede encontrar una solución. Se pide incluir en el desarrollo una justificación sobre por qué el algoritmo funciona correctamente en el caso del método CMM.

La experimentación será divida en dos partes, cada una con sus respectivos ejes. En primer lugar, buscamos hacer una evaluación cuantitativa de los métodos de resolución de sistemas lineales considerados, i.e., EG y CL, en términos del tiempo de cómputo y el tamaño de los sistemas a resolver. En particular, se pide comparar, para distintos tamaños de matrices, el tiempo de cómputo requerido para cada método en el contexto donde la información de la matriz del sistema ($C$) se mantiene invariante, pero varía el término independiente ($b$). Si las instancias obtenidas de datos reales no permiten notar diferencias significativas, se puede reformular el experimento utilizando instancias artificiales generadas convenientemente. Justificar cómo se generan estos datos y por qué es posible tomar esta decisión para este aspecto del análisis.

La segunda parte de la experimentación se centra en el análisis cualitativo respecto del comportamiento de los métodos CMM, WP o el elegido por los integrantes del grupo. Entre los experimentos a realizar, se pide como mínimo analizar los siguientes aspectos e intentar responder las siguientes preguntas:

- Utilizar principalmente datos de competencias reales que permitan identificar características distintivas de los métodos, y relacionarlas con eventos que ocurren en los mismos. Comparar los rankings obtenidos por cada uno de los métodos considerados.
- El método CMM es *justo*? Es decir, es posible que el resultado de un partido entre dos equipos afecte indirectamente el ranking de un tercero?
- Dados los resultados de todos los partidos considerados en la competencia, y un equipo particular. Determinar una estrategia que permita obtener la mayor posición posible, buscando minimizar el número de partidos ganados.[2]

En todos los casos es obligatorio fundamentar los experimentos planteados, proveer los archivos e información necesarios para replicarlos, presentar los resultados de forma conveniente y clara y analizar los mismos con el nivel de detalle apropiado. En caso de ser necesario, es posible también generar instancias artificiales con el fin de ejemplificar y mostrar un comportamiento determinado.

Como puntos opcionales para incluir en el desarrollo y/o experimentación, se considera lo siguiente:

1. Proponer y discutir (al menos) una forma alternativa de modelar el empate entre equipos en CMM.

### Parámetros y formato de archivos

El programa deberá tomar por línea de comandos tres parámetros. El primero de ellos contendrá el path al archivo de entrada con los partidos y resultados de la competencia; el segundo la salida con el ranking correspondiente, y el tercero indicando el método a considerar (0 CMM-EG, 1 CMM-CL, 2 WP o alternativo).

[2]No es necesario que la cantidad de partidos ganados sea la mínima, pero sí que la estrategia planteada trate de minimizar este aspecto.

El archivo de entrada contiene primero una línea con información sobre la cantidad de equipos ($n$), y la cantidad de partidos totales a considerar ($k$). Luego, siguen $k$ liñeas donde cada una de ellas representa un partido y contiene la siguiente información: identificador de fecha (es un dato opcional al problema, pero que puede ayudar a la hora de experimentar, un `string`), equipo $i$, goles equipo $i$, equipo $j$, goles equipo $j$.

A continuación se muestra el archivo de entrada con la información del ejemplo utilizado en Govan et al. [5]:

```
6 10
1 1 16 4 13
1 2 38 5 17
1 2 28 6 23
1 3 34 1 21
1 3 23 4 10
1 4 31 1 6
1 5 33 6 25
1 5 38 4 23
1 6 27 2 6
1 6 20 5 12
```

Es importante destacar que, en este último caso, los equipos son identificados mediante un número. Opcionalmente podrá considerarse un archivo que contenga, para cada equipo, cuál es el código con el que se lo identifica.

Una vez ejecutado el algoritmo, el programa deberá generar un archivo de salida que contenga una línea por cada equipo ($n$ líneas en total), acompañada del puntaje obtenido por el algoritmo CMM/WP/método alternativo.

Para instancias correspondientes a resultados entre equipos, la cátedra provee algunas opciones con información real de resultados en distintas competencias. Desde ya que cada grupo puede buscar/generar sus propios conjuntos de datos en caso que así lo considere. En [3] se provee un extenso set de datos con resultados históricos de la liga ATP de tenis profesional, divididos por año. Si bien los archivos contienen estadísticas detalladas sobre los partidos del circuito, en nuestro caso solo se necesitan un subconjunto muy reducido de los mismos. Por otro lado, en [4] se proveen resultados detallados para disintas ligas, profesionales y universitarias, de los Estados Unidos. Si bien es facil interpretar los archivos, la cátedra provee junto con este enunciado scripts en python para poder traducir los archivos obtenidos en cada uno de estos repositorios al formato requerido por el TP[3]. Finalmente, otra alternativa es considerar el repositorio DataHub [2], que contiene información estadística y resultados para distintas ligas y deportes de todo el mundo. En este caso, no se proveen herramientas adicionales para su pre-procesamiento.

Junto con el presente enunciado, se adjunta una serie de scripts hechos en `python` y un conjunto instancias de test que deberán ser utilizados para la compilación y un testeo básico de la implementación. Se recomienda leer el archivo `README.txt` con el detalle sobre su utilización.

[3]Los mismos son opcionales. En caso de encontrar algun error/bug en los mismos, por favor comunicarlo a la brevedad a la lista de docentes de la materia.

**Fechas de entrega**

- *Formato Electrónico:* Viernes 14 de Abril de 2017, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP1]` seguido de la lista de apellidos de los integrantes del grupo.
- *Formato físico:* Lunes 17 de Abril de 2017, a las 18 hs. en la clase de laboratorio.

**Importante:** El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.

# Referencias


[1] Colley rankings. `http://colleyrankings.com`.

[2] Datahub. `http://datahub.io`.

[3] Jeff sackmann atp tennis rankings. `http://github.com/JeffSackmann/tennis_atp`.

[4] Massey ratings. `http://masseyratings.com/data.php`.

[5] Angela Y. Govan, Carl D. Meyer, and Rusell Albright. Generalizing google's pagerank to rank national football league teams. In *Proceedings of SAS Global Forum 2008*, 2008.