Universidad Nacional de Córdoba

Facultad de Matemática, Astronomía, Física y Computación
Grupo de la Teoría de la Materia Condensada (GTMC)

Tesis Doctoral de Física

Procesos de memoria en sistemas con distribuciones de Zipf-Pareto

---

Lic. Ana L. Schaigorodsky

Director: Dr. Orlando V. Billoni

Córdoba, 19 de Marzo de 2018

Esta tesis está dedicada a todas las mujeres en ciencia

"Defiende tu derecho a pensar, porque incluso pensar de manera errónea es mejor que no pensar"
Hipatia de Alejandría

Agradecimientos

A mi director, Orlando, por su confianza, apoyo y motivación. Fue un privilegio contar con su guía.

Al GTMC, en especial a Juan y Nahuel, por su acompañamiento académico y humano.

A la Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba y al CONICET.

A la comisión asesora de doctorado, Paula Bercoff, Gustavo Sibona y Sergio Cannas.

Al tribunal de tesis, Karina Chattah, Adolfo Banchio y Gabriel Fabricius.

A mis amigos de la facultad, en especial a Eloisa, Bernardo, José y Marcelo.

A Carlos, Marga y Emilia, mis padres y hermana, por su eterno cariño y apoyo en todas las etapas de mi vida.

A mi compañero de vida y amigo, Matías, por tanta fuerza y amor.

Resumen

Estudios recientes realizados en una base de datos de Ajedrez cronológicamente ordenada, han mostrado que la distribución de popularidades de líneas de juego de Ajedrez se ajusta a una ley de Zipf. La ley de Zipf es común a muchos sistemas y es usualmente observada en conjunto con efectos de memoria tales como correlaciones de largo alcance y *burstiness*. Sin embargo los modelos existentes que estudian estos fenómenos no dan cuenta simultáneamente con la ley de Zipf y los efectos de memoria. En este trabajo de tesis, mediante una variante del modelo de crecimiento preferencial de Yule-Simon, introducido por Cattuto et al., se provee una explicación de la aparición simultanea de la ley de Zipf y los efectos de memoria en forma de correlaciones de largo alcance en la base de datos de Ajedrez. Se encuentra que el modelo de Cattuto et al. es capaz de reproducir ambos fenómenos, la ley de Zipf y las correlaciones de largo alcance, incluyendo además los efectos de tamaño del exponente de Hurst de las correspondientes series temporales. Más aún, se encuentra *burstiness* en la actividad de los grupos de jugadores más activos, aunque la actividad agregada del conjunto completo de jugadores presenta una distribución de tiempos entre eventos sin *burstiness*. Dado que el modelo de Cattuto et al. no es capaz de producir series temporales con comportamiento '*bursty*', se realiza una modificación al núcleo de memoria que permite lograr una dinámica *bursty*. Introduciendo un núcleo de memoria finito, se mantiene el comportamiento de ley de potencia en la distribución de popularidades y, al mismo tiempo se obtienen series temporales que presentan *burstiness* como consecuencia de una transición de fase, en la cual, en el estado crítico, la dinámica está dominada por las fluctuaciones.

Clasificación: 05.45.Tp, 01.80.+b, 05.10.-a, 02.50.Ey

Palabras Claves: Ley de Zipf-Pareto - Series Temporales - Memoria - Correlaciones de largo alcance - Burstiness - Modelos de crecimiento preferencial

Abstract

Recent works studying a chronologically sorted chess database have shown that the popularity distribution of opening lines in the game of chess follow a Zipf law. Zipf law is common to many systems and is usually observed together with memory effects, such as long-range correlations and burstiness. Nevertheless, existing models that study these phenomena do not account for the Zipf's law and memory effects simultaneously. In this thesis, using a variant of the Yule-Simon preferential growth model, introduced by Cattuto et al., we provide an explanation of the simultaneous emergence of Zipf's law and memory effects in the form of long-range correlations in the chess database. We find that Cattuto's model is able to reproduce both phenomena, Zipf's law and the long-range correlations., including the size effects displayed by the Hurst exponent of the corresponding time series. Furthermore, we find burstiness in the activity of the most active players, although the aggregated activity of all players in the database presents an interevent time distribution without burstiness. Since Cattuto's model is not able to generate times series with a bursty behavior, we made a modification to the memory kernel that allows a bursty dynamics. By introducing a finite memory kernel, we keep the power-law behavior in the popularity distribution and, at the same time, we obtain time series that present burstiness as a consequence of a phase transition in which, at the critical point, the dynamic is ruled by fluctuations.

El trabajo presentado en esta tesis ha sido publicado en los artículos científicos listados a continuación:

- J. I. Perotti, H.H. Jo, A. L. Schaigorodsky, and O. V. Billoni. Innovation and nested preferential growth in chess playing behavior. EPL (Europhysics Letters), 104(4):48005, 2013.

- Ana L. Schaigorodsky, Juan I. Perotti, and Orlando V. Billoni. Memory and long-range correlations in chess games. Physica A: Statistical Mechanics and its Applications, 394(0):304 – 311, 2014.

- Ana L. Schaigorodsky, Juan I. Perotti, and Orlando V. Billoni. A study of memory effects in a chess database. PLOS ONE, 11(12):1–18, 12 2016.

- Ana L. Schaigorodsky, Juan I. Perotti, Nahuel Almeira, and Orlando V. Billoni. Short-ranged memory model with preferential growth. Phys. Rev. E, 97:022132, Feb 2018.

Índice general

1 Introducción

En años recientes el alcance del estudio de la física estadística se ha extendido a otros campos como, por ejemplo, el comportamiento humano, a nivel individual y colectivo [1, 2, 3]. En particular, la dinámica generada al jugar juegos con un conjunto de reglas bien definidas provee un ámbito experimental conveniente para comprender algunos rasgos del comportamiento humano [4, 5, 6, 7, 8, 9, 10, 11], como por ejemplo los procesos de toma de decisiones [12, 13, 14, 15, 16]. Esto es particularmente conveniente desde el punto de vista de la física ya que, bajo las reglas de un juego, las variables del comportamiento bajo estudio están altamente acotadas.

Estudios recientes de registros de juegos [4, 5, 6, 7, 8, 9, 10] han mostrado que se pueden establecer paralelos útiles entre los patrones del comportamiento humano basado en juegos y teorías de procesos físicos bien definidos. El juego del Ajedrez, el cual es visto como un símbolo de proeza intelectual, es particularmente interesante [13, 17, 18, 19]. En el mundo hay grandes comunidades de ajedrecistas produciendo registros extensos de partidas, proveyendo una fuente de datos apropiada para análisis estadísticos en gran escala. Al explorar una base de datos de ajedrez, Blasius y Tönjes [13] observaron que la distribución de pesos de líneas de juego de Ajedrez sigue una ley de Zipf con exponente universal. Los mencionados autores explicaron este fenómeno en términos de un tratamiento analítico que corresponde a proceso multiplicativo.

La ley de Zipf, o Zipf-Pareto, ha sido ampliamente estudiada debido a que está presente en diversos sistemas empíricos, lo que sugiere la existencia de un principio universal detrás del fenómeno. En su trabajo Zipf [20, 21] propone una ley con el fin de modelar el comportamiento de las distribuciones de tamaños, la cual es esencialmente una función que decae como ley de potencia. Dentro de los sistemas que presentan una distribución de ley de potencia, los cuales resultan extraordinariamente diversos, se pueden mencionar la distribución de ingresos de compañías [22], frecuencia de palabras en textos literarios [23], energía de tormentas solares [24], géneros de especies [25], intensidad de terremotos [26], tamaños de cráteres de la luna [27], la frecuencia de ocurrencia de nombres en muchas culturas [28] y el número de citas que recibe un trabajo científico [29]. Uno de los primeros modelos capaces de explicar la aparición de leyes de Zipf-Pareto fue introducido por Yule [25], el cual fue ideado para explicar la aparición de distribuciones de ley de potencias en los tamaños de géneros biológicos. Más adelante, Simon [30] introdujo una variante similar, pero menos general del modelo, la cual se ajusta más naturalmente al contexto de la ley de Zipf observada en textos

literarios. Esta versión del modelo es conocida como modelo de Yule-Simon, y variaciones del mismo han re-aparecido en la literatura en varias ocasiones[31].

El estudio de la existencia de correlaciones de largo alcance y las distribuciones de tiempos entre eventos en sistemas en los que la ley de Zipf está presente es también de gran interés [32, 33]. De los sistemas antes mencionados, los cuerpos literarios han sido ampliamente estudiados debido al gran número de textos disponibles en formato digital, lo cual facilita su análisis [34]. Estudios en las últimas décadas han mostrado que series temporales generadas a partir de textos literarios exhiben memoria de largo alcance [32, 35] y *burstiness* [33]. Similarmente, sistemas naturales tales como secuencias de terremotos y tormentas solares han sido analizados, y se ha encontrado que los mismos muestran procesos temporales inhomogéneos [36, 37], más específicamente, las secuencias de tiempos entre eventos poseen una dinámica *bursty*, la cual está caracterizada por períodos de alta actividad seguidos de largos períodos de baja actividad. Más aún, se ha encontrado que la ley de Heaps, la cual es una ley empírica que relaciona la cantidad de palabras distintas en un texto con la longitud del mismo, está relacionada con la presencia de la ley de Zipf [38, 39, 34]. Estos fenómenos estadísticos no pueden ser explicados en términos de un proceso multiplicativo, tal como el propuesto por Blasius y Tönjes [13]. En este sentido, el mecanismo de generación de secuencias artificiales de eventos necesita nuevos ingredientes para reproducir los efectos de memoria observados en sistemas empíricos.

En este trabajo de tesis se estudiaron efectos de memoria en sistemas empíricos que muestran distribuciones de Zipf-Pareto empleando como modelo de estudio una base de datos de Ajedrez. Con el objetivo de reproducir las características encontradas, se exploraron modelos de crecimiento preferencial, tales como los modelos de Yule-Simon y de Cattuto et al., y se realizaron modificaciones a los mismos a fin de generar secuencias con una dinámica *bursty* que no es observada en las versiones originales de dichos modelos.

Más específicamente, se encontró que la base de datos de Ajedrez presenta una dinámica particular en la introducción de nuevos elementos, la cual es bien descrita por la ley de Heaps a tiempos largos. Al estudiar la dinámica del crecimiento del árbol de partidas, se muestra que la aparición de las leyes de Zipf y Heaps pueden ser explicadas en términos de un proceso de crecimiento preferencial anidado de Yule-Simon [17]. Sin embargo, los efectos de memoria y efectos de tamaño encontrados en la base de datos no pueden ser explicados mediante este proceso. Este problema es abordado siguiendo un resultado de Cattuto et al. [40], el cual introduce una modificación al proceso de Yule-Simon al incorporar un núcleo de memoria probabilístico de cola larga. El modelo de Cattuto et al. introduce memoria y al mismo tiempo preserva la distribución de frecuencias de ley de potencia característica de la ley de Zipf [32]. Se muestra que el modelo de Cattuto reproduce las propiedades estadísticas observadas en la base de datos

de Ajedrez, ya que el modelo no solo exhibe memoria, sino también correlaciones de largo alcance y efectos de tamaño. Más aún, se muestra que el comportamiento *bursty* está asociado a jugadores individuales, y a los jugadores más activos particularmente, pero desaparece al analizar el conjunto completo de jugadores, mientras que las correlaciones de largo alcance resultan más robustas y se encuentran en todos los subconjuntos de partidas analizadas. Al mismo tiempo, se muestra que el modelo de Cattuto no es capaz de reproducir la dinámica *bursty*.

Con el fin de generar una dinámica *bursty*, se reemplazó el núcleo de memoria de decaimiento lento del modelo de Cattuto por un núcleo de tamaño finito; a esta variante se la llamó *Bounded Memory Preferential Growth* (crecimiento preferencial con memoria limitada) o modelo BMPG. Esta modificación introduce *bursts* en la ocurrencia de elementos en las secuencias generadas, sin embargo elimina las correlaciones de largo alcance, aunque la longitud característica de las mismas exhibe propiedades de escala no triviales. Al mismo tiempo, el modelo preserva la distribución de ley de potencia de los modelos de Yule-Simon y Cattuto, aunque con diferente exponente. Mediante un análisis analítico y simulaciones numéricas se caracterizaron las propiedades del modelo, encontrando que la fenomenología observada está determinada por las fluctuaciones en el núcleo de memoria.

Esta tesis está organizada en tres Partes. En la Parte I Marco Teórico y Modelos, se introducen los conceptos teóricos así como los modelos estocásticos de Yule-Simon y Cattuto empleados en este trabajo. En la Parte II, Resultados, se muestran los resultados obtenidos divididos en cinco Capítulos. En el Capítulo 4, Base de Datos, se muestra una caracterización de la base de datos de Ajedrez empleada en este trabajo. En el Capítulo 5, Innovación de líneas de juego, se presenta el análisis realizado de la existencia de la ley de Heaps en la base de datos de Ajedrez y del modelo propuesto de crecimiento preferencial anidado. En el Capítulo 6, Memoria y correlaciones de largo alcance, se analiza la presencia de efectos de memoria en forma de correlaciones de largo alcance en la base de datos estudiada, así como el análisis de los modelos de Yule-Simon y Cattuto. Luego, en el Capítulo 7, Análisis de tiempo entre eventos, se analizan las secuencias de tiempo entre eventos de la base de datos como en las secuencias generadas por los modelos con el fin de determinar la presencia de *burstiness*. Finalmente, en el Capítulo 8, Crecimiento preferencial con memoria de corto alcance, se presenta la modificación propuesta de los modelos de Yule-Simon y Cattuto con el fin de producir secuencias con dinámica *bursty* así como su caracterización. Por último, en la Parte III, Conclusiones, se provee de una discusión general sobre los resultados obtenidos.

Parte I

Marco Teórico y Modelos

2 Estadística

2.1 Correlaciones en series de datos

Uno de los resultados de la estadística más utilizados de forma automática, muchas veces sin tomar cuidado de las condiciones bajo las cuales se deriva, es el que establece que *la varianza del valor medio de una muestra es igual a la varianza de una observación dividido el tamaño de la muestra*, es decir, dado un conjunto de observaciones independientes $X_1, ..., X_N$ con media $\mu = E(X_i)$ y varianza $\sigma^2 = var(X_i) = E[(X_i - \mu)^2]$, donde $E[X_i]$ representa el valor de espectación de X_i, la varianza de $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ es

$$var(\overline{X}) = \frac{\sigma^2}{N} \tag{2.1}$$

A fin de estudiar las condiciones bajo las cuales esta ecuación es válida, se considera un conjunto de observaciones realizadas aleatoriamente $\{X_i : i = 1, ..., N\}$, donde el índice i denota un orden natural, como por ejemplo tiempo o posición. De esta forma $X_1, ..., X_N$ son variables aleatorias que comparten la misma distribución marginal F.

No es complicado establecer las condiciones bajo las cuales la Ec. (2.1) es válida,

1. La media $\mu = E(X_i)$ existe y es finita.

2. La varianza $\sigma^2 = var(X_i)$ existe y es finita.

3. $X_1, ..., X_N$ son no correlacionados, es decir

$$\rho(i, j) = 0 \qquad i \neq j$$

donde

$$\rho(i, j) = \frac{\gamma(i, j)}{\sigma^2}$$

es la autocorrelación entre X_i y X_j, y

$$\gamma(i, j) = E[(X_i - \mu)(X_j - \mu)]$$

es la autocovarianza entre X_i y X_j.

Las suposiciones 1 y 2 dependen solo de la distribución marginal F y es relativamente simple verificar su cumplimiento al momento de realizar un experimento. La suposición 3 resulta ser la más problemática. En ciertas situaciones se considera que la dependencia entre las observaciones es lo suficientemente débil como para ser despreciable a los fines prácticos. Sin embargo esto no es siempre posible, ya que correlaciones significativas pueden producirse a pesar de las precauciones tomadas. Es por esto que es de importancia estudiar cómo la Ec. (2.1) es afectada cuando las observaciones están correlacionadas.

Con el fin de que $\overline{X}$ sea significativo, se asume la media $\mu = E(X_i)$ constante. La ecuación general de la varianza es

$$var(\overline{X}) = \frac{1}{N^2} \sum_{i,j=1}^{N} \gamma(i,j) = \frac{\sigma^2}{N^2} \sum_{i,j=1}^{N} \rho(i,j). \tag{2.2}$$

Si las correlaciones para $i \neq j$ suman cero, esto es,

$$\sum_{i \neq j}^{N} \rho(i,j) = 0, \tag{2.3}$$

entonces,

$$\sum_{i,j}^{N} \rho(i,j) = N$$

y la Ec. (2.1) resulta válida. Es decir, el caso donde $X_1, ..., X_N$ son no correlacionados. Si la Ec.(2.3) no se cumple, la varianza de $\overline{X}$ resulta

$$var(\overline{X}) = \frac{\sigma^2}{N}[1 + \delta_N(\rho),] \tag{2.4}$$

con un término de corrección distinto de cero

$$\delta_N(\rho) = \frac{1}{N} \sum_{i \neq j} \rho(i,j). \tag{2.5}$$

Si las correlaciones $\rho(i,j)$ solo dependen de la separación $|i-j|$ y la media $\mu = E(X_i)$ es constante, se dice que el proceso estocástico es *estacionario*, entonces es posible escribir la Ec. (2.5) de forma más simple como

$$\delta_N(\rho) = 2 \sum_{k=1}^{N} \left(1 - \frac{k}{N}\right) \rho(k). \tag{2.6}$$

También es importante estudiar el comportamiento asintótico de $var(\overline{X})$ cuando $N \to \infty$. La varianza de $\overline{X}$ es proporcional a N^{-1} siempre y cuando

$$\delta(\rho) = \lim_{N\to\infty} \delta_N(\rho) = \lim_{N\to\infty} \frac{1}{N} \sum_{i\neq j} \rho(i,j) \tag{2.7}$$

exista, sea finito y mayor a -1. Entonces se obtiene, para el comportamiento asintótico,

$$var(\overline{X}) \sim \frac{\sigma^2}{N}[1 + \delta(\rho)] = c(\rho)\frac{\sigma^2}{N}, \tag{2.8}$$

donde $\sim$ significa asintóticamente y $c(\rho) = 1 + \delta(\rho)$.

La mayoría de las series temporales en la literatura exhiben este comportamiento. Los más conocidos son los procesos ARMA (*autoregressive moving average*) y los procesos de Markov[41].

La Ec. (2.8) es una generalización de la Ec. (2.1) ya que permite una constante $c(\rho)$ distinta de 1. Sin embargo, esta generalización no es suficiente. Existen conjuntos de datos para los cuales la varianza de $\overline{X}$ difiere de la Ec. (2.1) no solo en una constante, sino en la velocidad a la cual converge a cero. La forma más simple de modelar este comportamiento es considerar un decaimiento más lento proporcional a $N^{-\alpha}$ para algún $\alpha \in (0, 1)$, es decir,

$$var(\overline{X}) \sim c(\rho)\frac{\sigma^2}{N^\alpha}, \tag{2.9}$$

donde ahora la constante $c(\rho)$ está definida como:

$$c(\rho) = \lim_{N\to\infty} N^{\alpha-2} \sum_{i\neq j} \rho(i,j). \tag{2.10}$$

La relación entre la Ec. (2.9) y la estructura de las correlaciones se observa simplemente al considerar correlaciones dependientes solamente de la distancia $|i - j|$ (proceso estocástico estacionario). Analizando las Ec. (2.6) y (2.10) se concluye que el comportamiento asintótico de la suma de todas las correlaciones con separaciones $-N + 1, ..., N - 1$ debe ser proporcional a $N^{1-\alpha}$

$$\sum_{k=-(N-1)}^{N-1} \rho(k) \sim constante \cdot N^{1-\alpha}, \tag{2.11}$$

lo que implica que $\sum_{-\infty}^{\infty} \rho(k)$ diverge, ya que $\alpha < 1$.

Específicamente, la Ec. (2.11) es válida si

$$\rho(k) \sim c_\rho |k|^{-\alpha} \tag{2.12}$$

cuando $|k| \to \infty$, y donde c_ρ es una constante positiva. En este caso, como las correlaciones decaen más lentamente que $1/n$ no existe escala característica en la cual las mismas puedan ser despreciadas. La interpretación intuitiva de la Ec. (2.12) es que el proceso tiene memoria de largo alcance. Es decir, la dependencia entre los eventos separados por una gran distancia disminuye lentamente con el aumento de $|k|$. Un proceso estacionario cuyas correlaciones decaen lentamente según la Ec. (2.12) es llamado *proceso estacionario con memoria de largo alcance* o *dependencia de largo alcance*. De otra manera, un proceso estacionario X_t es llamado estacionario con memoria de largo alcance o dependencia de largo alcance, o correlaciones de largo rango, si existe un número real $\alpha \in (0, 1)$ y una constante $c_\rho > 0$ tal que

$$\lim_{k\to\infty} \frac{\rho(k)}{c_\rho k^{-\alpha}} = 1. \tag{2.13}$$

La definición dada por la Ec. (2.13) es una definición asintótica, y como tal solo describe el comportamiento de las correlaciones cuando las distancias tienden a infinito; cada correlación individual puede ser arbitrariamente pequeña.

La densidad espectral $f(\lambda)$ de una función de autocorrelación $\rho(k)$ puede ser definida como

$$f(\lambda) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{ik\lambda},$$

donde λ es la frecuencia. Entonces, la Ec. (2.12) implica que

$$f(\lambda) \sim c_f |\lambda|^{\alpha-1} = c_f |\lambda|^{-\beta} \tag{2.14}$$

cuando $\lambda \to 0$ y donde c_f es una constante positiva.

2.2 El efecto Hurst

Desde la antigüedad el río Nilo ha sido conocido por su comportamiento a largo plazo, caracterizado por extensos períodos de sequía, durante los cuales los niveles del río tienden a ser bajos, seguidos por extensos períodos de crecidas, con niveles altos. A tiempos largos la serie temporal de niveles del Nilo resulta estacionaria. Por otro lado, al observar intervalos de tiempos reducidos, parecen surgir ciclos o tendencias locales. Sin embargo la serie completa no exhibe ciclos persistentes (Figura 2.1).

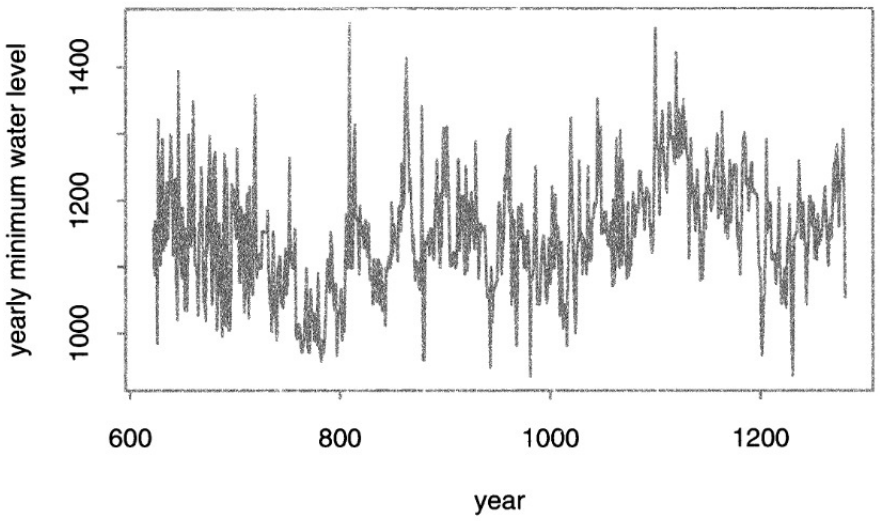

Figura 2.1.: Nivel mínimo anual del río Nilo (622-1281 d.C.). Figura extraída de [41].

El hidrólogo Harold E. Hurst advirtió este comportamiento al investigar el problema de regularización del flujo del Nilo (1951). Más específicamente descubrió que puede ser descrito de la siguiente manera: Suponiendo que se desea calcular la capacidad de un reservorio ideal en un intervalo de tiempo $(t, t+k)$, donde por ideal se refiere a que el flujo es uniforme dentro del reservorio, que el nivel al tiempo $t+k$ es igual al nivel al tiempo t y que el reservorio no desborda. A fin de simplificar el problema, se asume que el tiempo es discreto y que no existen pérdidas en el reservorio (por evaporación, derrame, etc.). Denotando al flujo entrante al tiempo i por X_i y al flujo entrante acumulado al tiempo j por $Y_j = \sum_{i=1}^{j} X_i$, la capacidad ideal es igual a

$$R(t,k) = \max_{0 \leqslant i \leqslant k} [Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)] - \min_{0 \leqslant i \leqslant k} [Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)], \quad (2.15)$$

donde $R(t,k)$ es llamado *rango ajustado*. A fin de estudiar las propiedades independientemente de la escala utilizada, $R(t,k)$ es normalizado mediante

$$S(t,k) = \sqrt{\frac{1}{k} \sum_{i=t+1}^{t+k} (X_i - \overline{X}_{t,k})^2}, \quad (2.16)$$

donde $\overline{X}_{t,k} = \frac{1}{k} \sum_{i=t+1}^{t+k} X_i$. La razón

$$R/S = \frac{R(t,k)}{S(t,k)} \quad (2.17)$$

es el *rango reescalado ajustado* o estadística R/S. Hurst observó que al graficar el logaritmo de R/S en función de k, para valores considerables de k, $\log(R/S)$ se

encontraba dispersado alrededor de una recta con pendiente mayor a $\frac{1}{2}$. En términos probabilísticos esto es

$$\log E[R/S] \approx a + H \log(k), \qquad H > \frac{1}{2}. \tag{2.18}$$

Hurst descubrió que en el caso del río Nilo, así como en muchos registros hidrológicos, geofísicos y climatológicos, R/S se comporta como una constante por k^H para algún $H > \frac{1}{2}$. Este es el llamado *efecto Hurst*.

El parámetro α en la Ec. (2.12) está relacionada con el exponente de Hurst H mediante la ecuación $\alpha = 2 - 2H$[42]. Es decir, $H > 1/2$ implica que $\alpha < 1$, y por lo tanto se puede decir que se trata de un proceso con memoria de largo alcance.

2.3 Cálculo del exponente de Hurst H

Sea un conjunto de datos $\{X_i : i = 1, ..., N\}$ en los cuales se desea estudiar las correlaciones de X_i y $X_{i+\ell}$ sobre diferentes escalas temporales ℓ a fin de determinar la presencia de memoria de largo alcance. Con el fin de librarse de un desplazamiento (*offset*) constante en los datos se acostumbra sustraer la media $\langle X \rangle = m = \frac{1}{N}\sum_{i=1}^{N} X_i$ a fin de obtener una serie centrada en cero, $\widetilde{X}_i \equiv X_i - m$. Cuantitativamente la correlación entre dos valores de X separados por ℓ está definida por la función de auto-correlación,

$$C(\ell) = \frac{1}{N-\ell} \sum_{i=1}^{N-\ell} \widetilde{X}_i \widetilde{X}_{i+\ell}.$$

Si $\{X_i\}$ son no correlacionadas, $C(\ell)$ es cero para $\ell > 0$, y, como se mencionó previamente, si existen las correlaciones de largo alcance, $C(\ell)$ decae como ley de potencia,

$$C(\ell) \sim \ell^{-\gamma}$$

con exponente $0 < \gamma < 1$. Muchas veces no es posible realizar un cálculo directo de $C(\ell)$ debido a la presencia de ruido superpuesto al conjunto de datos o bien, a tendencias subyacentes de origen desconocido cuyas escalas tampoco son conocidas [43], y por lo tanto se debe calcular el exponente γ de forma indirecta.

Los métodos más utilizados a fin de determinar la existencia de correlaciones de largo alcance en una serie temporal se centran en el cálculo del coeficiente de Hurst H, entre los cuales se pueden mencionar el método de rango reescalado o estadística R/S, *detrended fluctuation analysis* o DFA, varianza agregada, periodograma, *wavelet*

analysis y estimador local Whittel[42]. En este trabajo se emplearán los dos primeros métodos mencionados, especialmente el método DFA.

2.3.1 Método de Rango Reescalado R/S

A fin de calcular H se debe primero estimar la dependencia del rango reescalado con los rangos temporales de las observaciones. Para esto la serie temporal de N observaciones es dividida en series de menor longitud $\ell = N, N/2, N/4, ...$ no superpuestas.

Para cada subconjunto de observaciones de longitud ℓ, $X = X_1, X_2, ..., X_\ell$, se computa:

1. La media:
$$m = \frac{1}{\ell}\sum_{i=1}^{\ell} X_i$$

2. Una serie centrada en la media:
$$\widetilde{X}_t = X_t - m \qquad t = 1, 2, ..., \ell$$

3. La desviación acumulada de la serie respecto de la media:
$$Y(t) = \sum_{i=1}^{t} \widetilde{X}_i \qquad t = 1, 2, ..., \ell$$

4. El rango R:
$$R(\ell) = \text{máx}[Y(1), Y(2), ..., Y(\ell)] - \text{mín}[Y(1), Y(2), ..., Y(\ell)]$$

5. La desviación estándar S:
$$S(\ell) = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell} (X_i - m)^2}$$

Luego, se promedia el rango reescalado $R(\ell)/S(\ell)$ sobre todas las series temporales parciales de longitud ℓ, y finalmente, se estima H ajustando los datos a la ley de potencias $E\left[\frac{R(\ell)}{S(\ell)}\right] = C\ell^H$. Para esto se emplea la Ec. (2.18) y se realiza una regresión lineal a fin de calcular la pendiente H.

2.3.2 Método DFA

El DFA, *Detrended Fluctuation Analysis*[43, 44], es un método de determinación de correlaciones de largo alcance en series temporales no estacionarias, consolidado para determinar comportamiento de escala de conjuntos de datos con presencia de ruido y tendencias de origen y forma desconocida. Fue desarrollado por Peng et al. con el fin de detectar correlaciones de largo alcance en secuencias de nucleótidos, cuya estructura de mosaico genera tendencias subyacentes que dificultan el cálculo directo de la función autocorrelación. En este sentido el método resulta más adecuado que el método R/S.

El método calcula una función de fluctuación $F(\ell)$ específica a una escala temporal ℓ [45], la cual, para series temporales con correlaciones de largo alcance tiene la forma

$$F(\ell) \sim \ell^{\zeta}. \tag{2.19}$$

Dada una serie temporal $X(t)$ de longitud N, el procedimiento del DFA consiste en calcular la media $m = \frac{1}{N}\sum_{i=1}^{N} X_i$ y la serie centrada en la media $\widetilde{X}_t = X_t - m,\ t = 1, 2, ..., \ell$, y luego se siguen los siguientes pasos:

1. Se determina el perfil

$$Y(t) = \sum_{i=1}^{t} \widetilde{X}_i \qquad t = 1, 2, ..., N$$

   de la serie de longitud N.

2. Se divide la secuencia $Y(t)$ en N/ℓ segmentos de longitud ℓ.

3. Se define la tendencia local en cada segmento ajustando un polinomio $g(t)$ (generalmente de 1^{er} grado) a cada segmento.

4. Se define la caminata '*detrended*' como la diferencia entre $Y(t)$ y el ajuste del polinomio $g(t)$ en cada segmento,

$$\varepsilon(t) = Y(t) - g(t)$$

   En a Figura 2.2 se muestra un ejemplo de un perfil calculado a partir de una serie temporal, con $\ell = 200$ (panel superior) y $\ell = 100$ (panel inferior). Como se puede observar al aumentar el tamaño ℓ de los segmentos aumenta también la diferencia entre el perfil $Y(t)$ y el ajuste lineal, es decir, la varianza de $\varepsilon(t)$ crece con ℓ.

5. Se calcula la varianza de $\varepsilon_\ell(t)$ para cada intervalo.

6. Se calcula la media de las varianzas en todos los segmentos de longitud ℓ.

Finalmente, se obtiene $F(\ell)$,

$$F(\ell) = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \varepsilon_\ell(t)^2} \sim \ell^\zeta. \tag{2.20}$$

Particularmente, cuando $\zeta < 1$, la serie temporal resulta estacionaria y $\zeta = H$, donde H es el exponente de Hurst.

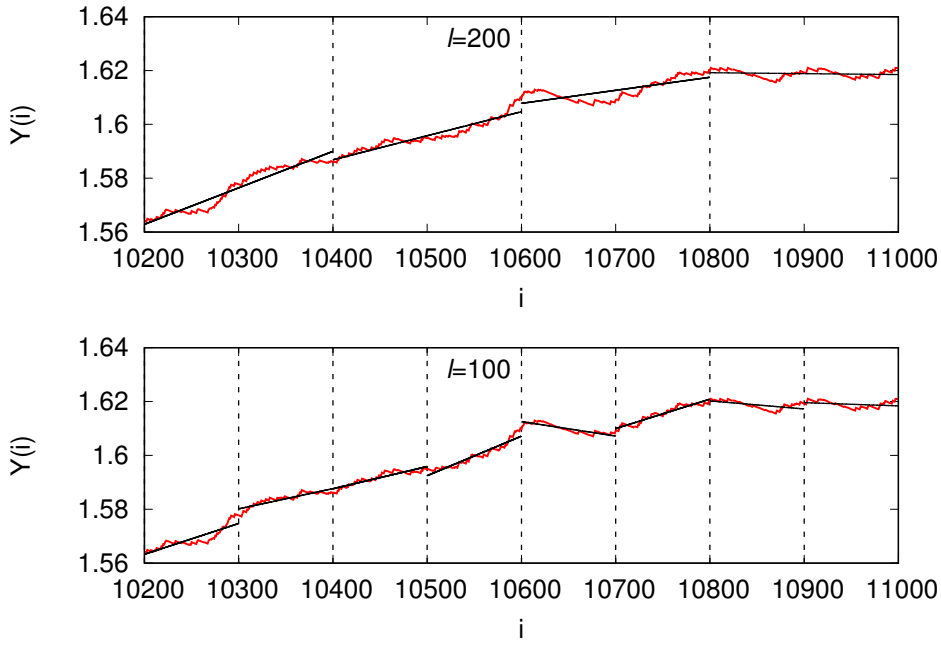

Figura 2.2.: Ejemplo de una caminata que presenta correlaciones de largo alcance. El *detrented fluctuation analysis* se aplica con una escala $\ell = 200$ (panel superior) y $\ell = 100$ (panel inferior). En cada segmento se muestran los ajustes correspondientes a cada uno de ellos.

2.4 *Burstiness*

Muchos constituyentes de sistemas sociales y naturales muestran regularidades estadísticas notables. Un efecto estadístico interesante, observado frecuentemente en sistemas empíricos, es la presencia de memoria, la cual se presenta en forma de correlaciones de largo alcance y también como *burstiness*. Al analizar los tiempos entre eventos de muchos de estos sistemas, se ha encontrado que dichas secuencias de tiempos pre-

sentan épocas de gran y baja actividad, i.e. una dinámica *bursty*. El acceso reciente a bases de datos de gran escala han permitido estudiar con mayor profundidad dicha dinámica. Dentro de los sistemas estudiados en la literatura se encuentran las secuencias de terremotos [46, 36], comunicación vía e-mail [47], tormentas solares [37], actividad neuronal [48], y comportamiento humano en general [49, 50].

En particular, la distribución de tiempo entre eventos de los elementos en una base de datos puede ser analizada de manera similar que en la ocurrencia de palabras en un texto [33]. Los elementos de una base de datos cronológicamente ordenada pueden ser enumeradas de acuerdo a su orden de aparición. Específicamente, se denota con $t \in 1, 2, ..., N$ a la secuencia de tiempos ordinales de aparición de los diferentes elementos de la base de datos. Por lo tanto, el $j-$ésimo tiempo entre eventos de un elemento g está definido como,

$$\tau_j^{(g)} = t^{(g)}(j+1) - t^{(g)}(j) \tag{2.21}$$

donde $t^{(g)}(j)$ representa la $j-$ésima aparición del elemento g. Si el elemento g ocurre con frecuencia $\nu_g = N^{(g)}/N$, se puede estimar el tiempo entre eventos promedio como $\left\langle \tau^{(g)} \right\rangle \approx 1/\nu_g$. Aquí, $N(g)$ es el número de veces que ocurre un elemento particular g en la base de datos. El tiempo entre eventos $\left\langle \tau^{(g)} \right\rangle$ es usualmente llamado longitud de onda de Zipf en el caso de análisis de textos [33], donde g representa una palabra en particular, como por ejemplo 'árbol' ó 'los'.

El proceso puntual aleatorio (*point process*) más simple para el análisis de tiempos entre eventos es el proceso de Poisson. Un elemento particular g ocurre con una probabilidad por unidad de tiempo μ_g la cual se asume constante, y como consecuencia, la distribución de tiempo entre eventos del elemento g resulta una distribución exponencial $f(\tau^{(g)}) = \mu_g \exp(-\mu_g \tau^{(g)})$. Aquí, la tasa puede aproximarse como $\mu_g \approx \nu_g$, donde la relación es aproximada por dos razones. En primer lugar, los procesos de Poisson están definidos para tiempo continuo, mientras que en las bases de datos empíricas el tiempo es discreto; y en segundo lugar, la fracción ν_g corresponde a un número finito de eventos, mientras que un proceso de Poisson describe un proceso estacionario infinitamente largo. Con esto en mente, la aproximación $\mu_g \approx \nu_g$ funciona bien siempre y cuando $\nu_g \ll 1$ y $N \gg N^{(g)} \gg 1$.

A fin de simplificar notación se escribirá τ en lugar de $\tau^{(g)}$ en las situaciones en las que no sea necesario referirse a un elemento particular g. En el análisis empírico de datos, es conveniente utilizar la densidad de probabilidad acumulada $F(\tau) = \int_\tau^\infty f(\tau')d\tau'$ en lugar de la aplicación directa de la densidad de probabilidad $f(\tau)$. Esto tiene razones prácticas; la función $F(\tau)$ es usualmente más simple que $f(\tau)$. Por ejemplo, una desviación del comportamiento exponencial en $F(\tau)$ indica la presencia de efectos de memoria. En

el caso de palabras en un texto, esta desviación es bien descrita por una exponencial estirada[1] de un solo parámetro, o función de Weibull [33],

$$f(\tau) = \frac{\beta_B}{\tau_0}\left(\frac{\tau}{\tau_0}\right)^{\beta_B - 1} e^{-\left(\frac{\tau}{\tau_0}\right)^{\beta_B}}. \tag{2.22}$$

En esta distribución $\langle\tau\rangle = \tau_0\Gamma\left(\frac{\beta_B+1}{\beta_B}\right)$, donde Γ es la función Gamma y $0 < \beta_B \leq 1$, y la distribución acumulada correspondiente es,

$$F(\tau) = e^{-\left(\frac{\tau}{\tau_0}\right)^{\beta_B}}. \tag{2.23}$$

Cuando β_B se desvía de 1, implica la presencia de *burstiness* en la serie temporal. Un '*burst*' corresponde a un incremento en los niveles de actividad por períodos cortos de tiempo, seguidos de largos períodos de inactividad [51], y cuando el valor de β_B se aproxima a cero la aparición de *bursts* en la serie aumenta. A fin de probar si la distribución de tiempos entre eventos acumulada realmente se ajusta a una exponencial estirada, es conveniente graficar $-\log(F(\tau))$ como función de τ en escala logarítmica [33, 52]. En esta gráfica, la exponencial estirada se convierte en una recta, cuya pendiente es el exponente de *burstiness* β_B.

Un indicador más claro de la presencia de *burstiness* se presenta cuando la distribución de tiempos entre eventos está mejor descrita por una ley de potencia,

$$f(\tau) = A\tau^{-\eta}, \tag{2.24}$$

donde el grado de *burstiness* está caracterizado por el exponente η de la ley de potencia; cuanto menor es el exponente, mayor será la dinámica *bursty*.

La desviación de $f(\tau)$ de un proceso de Poisson puede ser también caracterizada con el coeficiente de variación $\sigma_\tau/\langle\tau\rangle$, donde σ_τ es la desviación estándar de los tiempos entre eventos. Luego, el coeficiente de variación es utilizado para calcular el parámetro de *burstiness* B como [51],

$$B = \frac{\sigma_\tau/\langle\tau\rangle - 1}{\sigma_\tau/\langle\tau\rangle + 1} = \frac{\sigma_\tau - \langle\tau\rangle}{\sigma_\tau + \langle\tau\rangle}. \tag{2.25}$$

Cuando hay presencia de *burstiness*, por ejemplo $\beta_B < 1$ en la Ec. (2.22), este parámetro es mayor a cero ($B > 0$). Si la dinámica es regular B es menor a cero. Cuando $B = 0$, no hay ni *burstiness* ni regularidad en la serie, como es el caso en que $\beta_B = 1$.

[1] Del inglés *exponential streched*

El parámetro de *burstiness* B de la Ec. (2.25) es muy utilizado debido a su simpleza. Sin embargo, su cálculo debe realizarse cuidadosamente, ya que no resulta robusto ante efectos de tamaño finito [53].

2.5 Ley de Zipf-Pareto

Cuando la probabilidad de medir un valor particular de alguna cantidad varía inversamente como potencia de ese mismo valor, se dice que la cantidad en cuestión sigue una ley de potencia o una distribución libre de escala, también conocida como Ley de Zipf o distribución de Pareto. La frecuencia de uso de palabras en múltiples lenguas [20], las tormentas solares [24], las ciudades más extensas [54] y la magnitud de terremotos [55] pueden ser descritas en términos de la Ley de Zipf, la cual captura la relación entre la frecuencia de un set de objetos o eventos y su tamaño. En todos estos ejemplos mencionados el exponente de la distribución resulta cercano a 2, esto es, siguen una ley de potencias x^{-2}, donde x es el tamaño del evento [56].

En la Figura 2.3 se muestra, a modo de ejemplo, el histograma de tamaños de las ciudades estadounidenses [54]; en el mismo se observa la existencia de un gran número de ciudades relativamente pequeñas, y un número reducido de ciudades cuya población supera considerablemente a la media. Resulta notable al estudiar el panel derecho de la Figura 2.3 como, al graficar el histograma en escala logarítmica, su aspecto general resulta similar a una función lineal. Denotando por $p(x)dx$ a la fracción de ciudades con población entre x y $x + dx$, resulta $ln\, p(x) = -\alpha\, lnx + c$, donde α y c son constantes, lo que es equivalente a

$$p(x) = Cx^{-\alpha} \tag{2.26}$$

con exponente $\alpha = 2{,}5$.

En sistemas donde aparece la ley de Zipf es común también utilizar la función de rangos. Para calcular esta distribución, por ejemplo en textos literarios, se toma la palabra más popular y se le asigna el rango $r = 1$, a la segunda más popular el rango $r = 2$, etcétera. La función de rangos resulta,

$$x(r) = ar^{-\beta}, \tag{2.27}$$

donde $x(r)$ es la frecuencia de la palabra de rango r. Se puede ver también que los exponentes α de la Ec. (2.26) y β de la Ec. (2.27) están relacionados de acuerdo a,

$$\alpha = 1 + \frac{1}{\beta}. \tag{2.28}$$

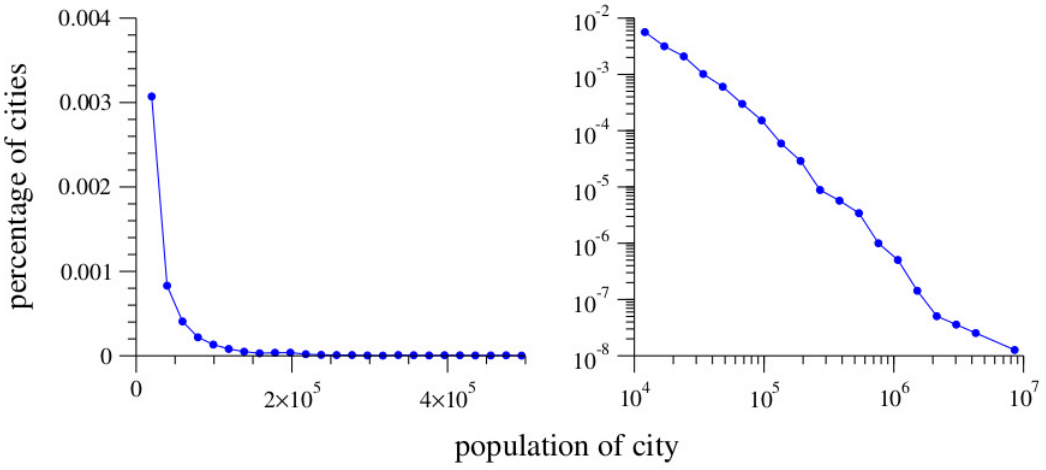

Figura 2.3.: Izquierda: histograma de la población de las ciudades estadounidenses cuya población supera 10000 habitantes. Derecha: histograma del mismo conjunto de datos en escalas logarítmicas. Figura extraída de [54].

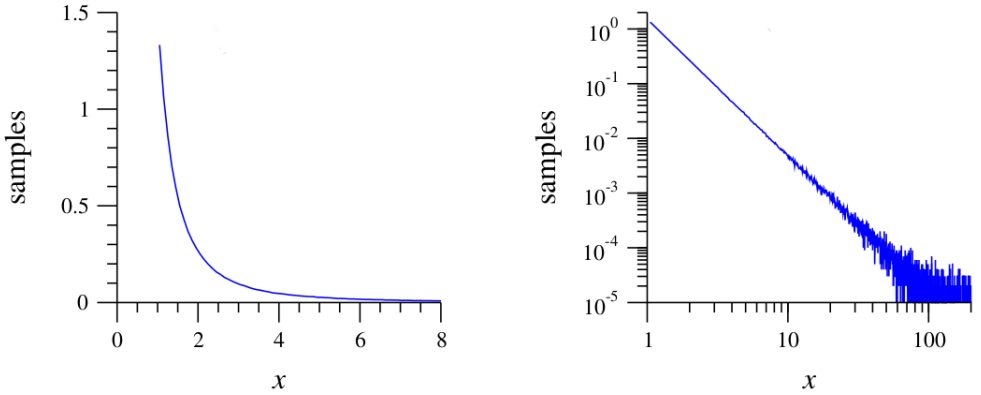

Figura 2.4.: Datos artificiales que consta de números reales aleatorios extraídos de una distribución de probabilidad de ley de potencias según la Ec. (2.26) para $\alpha = 2{,}5$. Figura extraída de [54].

La identificación del comportamiento de ley de potencia de sistemas naturales o artificiales es complicada. La estrategia estándar utilizada [54] es la mostrada en el ejemplo anterior y consiste en obtener un histograma de una cantidad que al ser graficada en escala logarítmica es muy cercana a una recta. Esta no es la mejor forma de proceder, ya que generalmente se observa ruido en la cola de la distribución a causa de que los eventos en dicha zona son menos frecuentes (Figura 2.4), lo que significa que cada intervalo ("bin") posee muy pocas mediciones. Una de las soluciones a este problema consiste en variar el ancho de los intervalos del histograma. Al realizar esto se debe normalizar, es decir, el número de elementos en un intervalo Δx debe ser dividido por la longitud Δx, a fin de que el conteo normalizado de la muestra resulte independiente de la

longitud del intervalo. La elección más usual es crear los intervalos tal que cada uno sea un múltiplo fijo más ancho que el anterior. Esto es conocido como *bining logarítmico*.

Existen múltiples mecanismos los cuales generan distribuciones con comportamiento de leyes de potencias, que han sido utilizados para el análisis de datos experimentales, entre los cuales se pueden mencionar la combinación de exponenciales, cantidades inversas, caminatas aleatorias, proceso de Yule, tolerancia altamente optimizada, ruido coherente y modelos multiplicativos modificados [54, 57].

2.5.1 La Ley de Zipf en el Ajedrez

Las partidas de Ajedrez (Apéndice A) pueden describirse como las ramas de un grafo o árbol cuya raíz es la posición inicial del juego. Cada conexión de dicho árbol representa una movida legalmente permitida por las reglas del juego, y cada nodo una de las posibles posiciones. De este modo, una partida en particular puede representarse por una secuencia de nodos $g_0, g_1, g_2, ..., g_d$ o equivalentemente por una secuencia de conexiones $l_1, l_2, ..., l_d$ en el árbol. La raíz del árbol g_0 está presente en todas las partidas posibles. El árbol de partidas posee aproximadamente 10^{120} nodos (número de Shannon[58]), correspondiendo a un factor de ramificación promedio igual a 30 ramas por nodo y a una longitud promedio de las partidas en 40 movidas. Sin embargo, a pesar de la complejidad del árbol de partidas posibles tan sólo una pequeña fracción de las partidas son exploradas en la práctica. Esta observación es de crucial importancia para entender la naturaleza de los fenómenos de tomas de decisiones.

En un trabajo reciente de Blasius y Tönjes [13] se estudia una base de datos de partidas de Ajedrez entre humanos (SCIDBASE [59]) encontrando que la popularidad de las diferentes líneas de juego satisface la ley de Zipf. Este hallazgo se relaciona a la existencia de líneas de juego que son comunes y que los jugadores tienden a elegir frecuentemente. Además, es importante ya que conecta los procesos de tomas de decisiones con un espectro de procesos complejos caracterizados por la ley de Zipf. Más precisamente, el estudio de Blasius y Tönjes se enfoca en una versión pesada del árbol de partidas (Figura 2.5). Cada nodo g en el árbol tiene asociado un número de partidas s_g, y cada conexión l una fracción r_l de partidas que continúa por la correspondiente línea de juego. De este modo, si l es la conexión que va desde la posición g hasta la posición g', luego se satisface $s_g r_l = s_{g'}$. La raíz tiene un número s_{g_0} de partidas que es igual al número N de partidas en la base de datos estudiada. En términos de los procesos de toma de decisiones, s_g denota la popularidad con la cuál es jugada la correspondiente apertura o línea de juego.

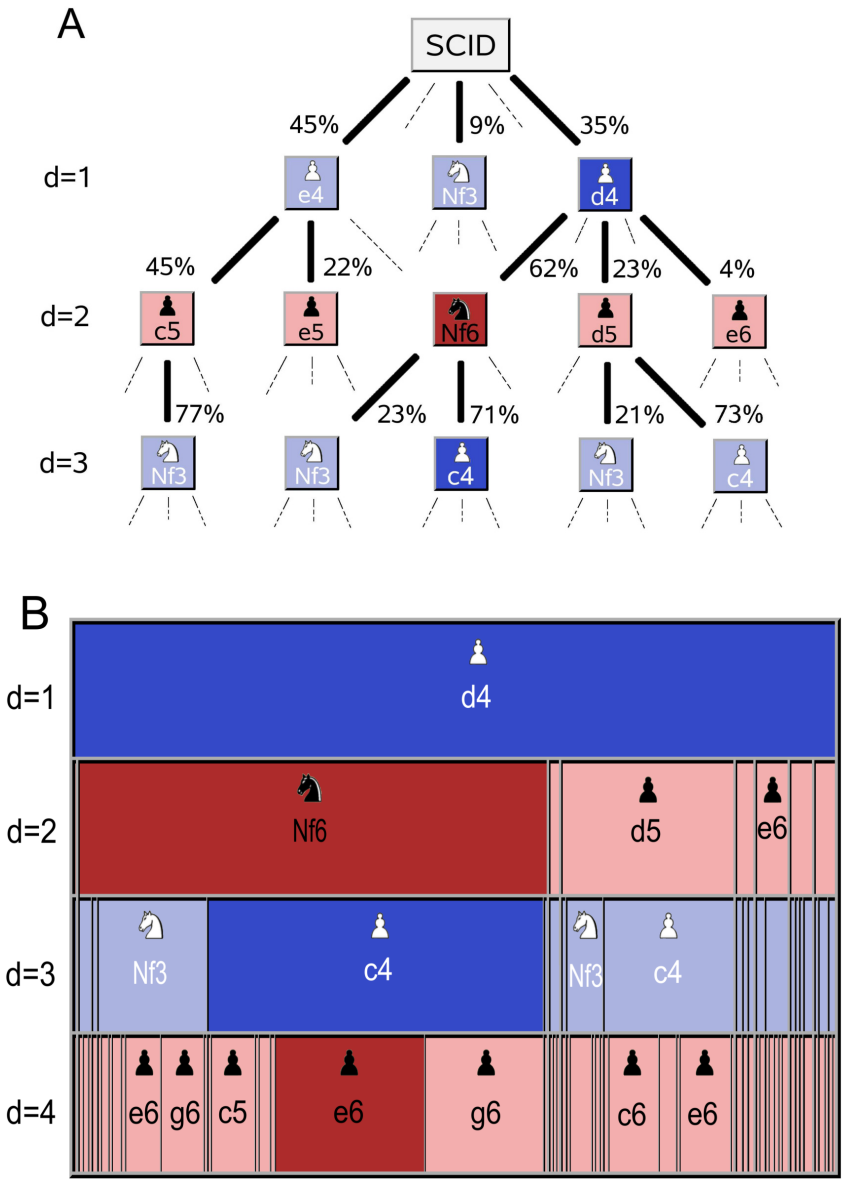

Figura 2.5.: (A) Arbol de la base SCIDBASE. Las líneas sólidas representan las posibles continuaciones del juego junto con sus correspondientes probabilidades, y las líneas de puntos, otras posibles continuaciones menos probables que no se muestran. (B) Representación alternativa que destaca la segmentación sucesiva del conjunto de partidas. Cada nodo g esta representado por un recuadro cuyo tamaño es proporcional a su frecuencia s_g. En la profundidad siguiente las partidas se dividen en sub-conjuntos de acuerdo con las posibles continuaciones del juego. Figura extraída de [13].

De acuerdo con Blasius y Tönjes, la distribución de popularidades tomando en cuenta todos los nodos del árbol satisface una ley de potencia (Figura 2.6(A))

$$P(s) \sim s^{-\alpha}$$

con exponente $\alpha = 2$, lo cuál corresponde a la ley de Zipf, que comúnmente se encuentra en textos literarios [54]. Al estudiar el fenómeno en más detalle, se encuentra que las frecuencias $P_d(s)$ de los juegos correspondientes a los primeros d movimientos, i.e., tomando todos los nodos a profundidad d son también consistentes con un comportamiento de ley de potencias (Figura 2.6(B))

$$P_d(s) \sim s^{-\alpha_d},$$

en donde los exponentes α_d no son universales sino que aumentan linealmente con d (Inset Figura 2.6(B)).

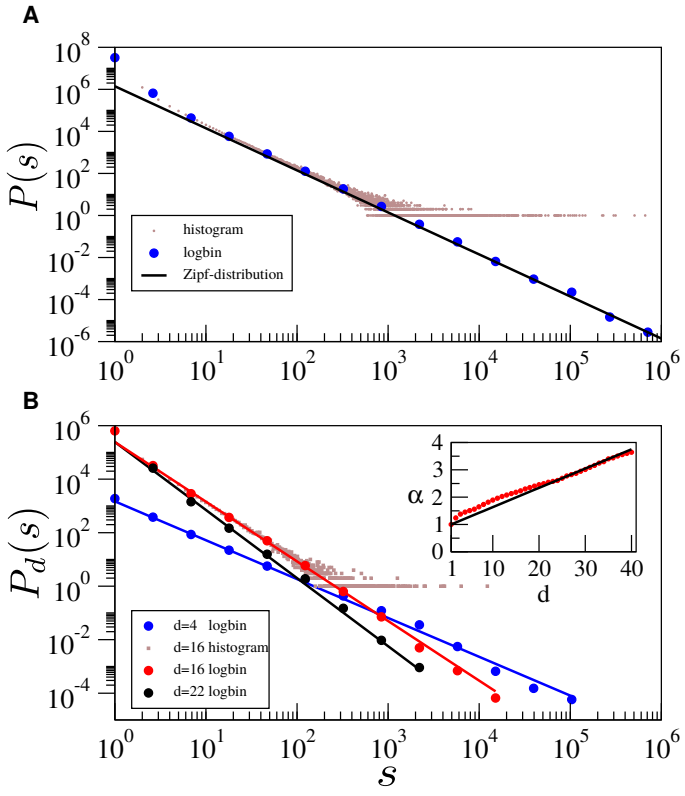

Figura 2.6.: (A) Histograma de la frecuencia pesos $P(s)$ de las aperturas hasta una profundidad $d = 40$ con bin logarítmico. Una regresión lineal resulta en un exponente de $\alpha = 2{,}05$. (B) Número de aperturas $P_d(s)$ de profundidad d con popularidad s para $d = 16$ e histogramas con bin logarítmico para $d = 4$, $d = 16$ y $d = 22$. Inset: pendiente α_d en función de la profundidad d y la estimación analítica (Ec. 2.33) utilizando $N = 1{,}4 \times 10^6$ y $\beta = 0$. Figura extraída de [13].

Las leyes de potencias con exponentes no universales pueden explicarse utilizando caminatas aleatorias multiplicativas [60, 61]. La propuesta de Blasius y Tönjes consiste en un modelo basado en este tipo de procesos que explica con gran precisión las distribuciones observadas. Más precisamente, el número s_d de partidas que tiene un dado nodo tras d movimientos viene dado por la ecuación,

$$s_d = N \prod_{i=1}^{d} r_i, \qquad s_0 = N. \tag{2.29}$$

En el trabajo de Blasius y Tönjes se asume que el árbol de partidas de la Figura 2.5 es autosimilar de manera que cada factor de ramificación $r_i \in [0, 1]$ es una variable aleatoria correspondiente a una distribución de probabilidades $q(r)$ que es independiente del nodo s en consideración. En particular $q(r)$ es independiente del número de partidas N, y de la profundidad d de la posición. La distribución $q(r)$ fue medida por Blasius y Tönjes y se encuentra que la misma está bien descrita por la expresión no paramétrica (ver Figura 2.7(A))

$$q(r) = \frac{2}{\pi\sqrt{1-r^2}}. \tag{2.30}$$

correspondiente a la distribución arcoseno. Tal distribución $q(r)$ es aproximadamente constante para valores relativamente chicos de r y diverge como $(1-r)^{1/2}$ cuando $r \to 1$.

A su vez, el trabajo provee una derivación analítica de las distribuciones $P(s)$ y $P_d(s)$ partiendo de una aproximación a la distribución $q(r)$ dada por,

$$q(r) = (1+\upsilon)r^{\upsilon}, \quad 0 \leq r \leq 1, \tag{2.31}$$

la cuál típicamente aparece en procesos de crecimiento preferencial[2] [62] derivados de modelos de crecimiento preferencial [30]. Los cálculos determinan que

$$P_d(s) = \frac{(1+\upsilon)^d}{N(d-1)!} \left(\log \frac{N}{s}\right)^{d-1} \left(\frac{N}{s}\right)^{1-\upsilon}. \tag{2.32}$$

Utilizando una expansión logarítmica en el rango $1 \ll s \ll N$ esta expresión exhibe una comportamiento tipo ley de potencias con exponente $-\alpha_d$ dado por

$$\alpha_d = (1-\upsilon) + \frac{1}{\log N}(d-1), \tag{2.33}$$

de modo que α_d crece linealmente con la profundidad d más una corrección logarítmica estando en buena concordancia con lo observado (Inset de la Figura 2.6(B)).

[2] Del inglés *preferential attachment*

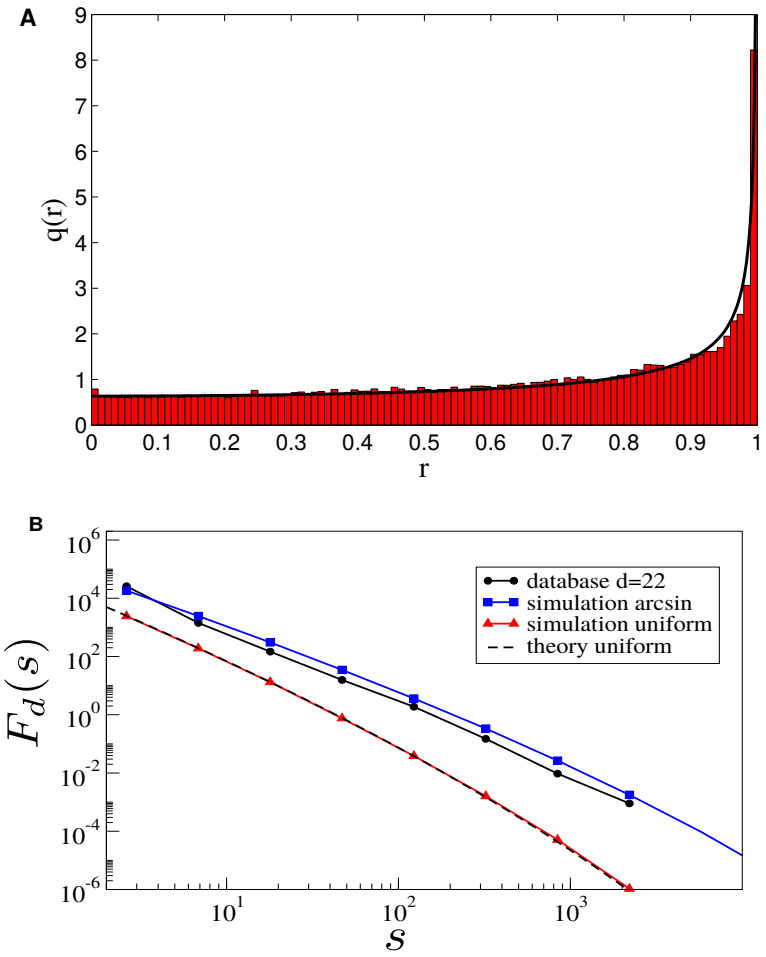

Figura 2.7.: (A) Densidad de probabilidad $q(r)$ de las razones de ramificación r medida utilizando la base de datos Scid con intervalos constantes $\Delta r = 0{,}01$, y la distribución arcoseno (Ec. 2.30). (B) Probabilidad $F_d(s) = \frac{s}{N} P_d(s)$ de que un nodo a una distancia d del nodo raíz posea popularidad s para el caso $d = 22$ en la base de datos SCIDBASE (línea negra). Comparativamente se muestran las curvas correspondientes a una simulación directa del proceso multiplicativo con la distribución $q(r)$ original (Ec. 2.29, línea azul), y una distribución $q(r)$ uniforme (Ec. 2.31 con $\upsilon = 0$, línea roja). Resultados teóricos según la Ec. 2.32 (línea a rayas). Figura extraída de [13].

Como se muestra en la Figura 2.7 las simulaciones del proceso multiplicativo (Ec. (2.29)) empleando la distribución arcoseno (Ec. (2.30)) resultan una buena aproximación de las frecuencias pesadas $P_d(s)$ de la base de datos de ajedrez. Si las tasas de ramificación son aproximadas por una distribución uniforme $q(r) = 1$, los valores de $P_d(s)$ resultan sistemáticamente pequeños, ya que esta distribución produce un mayor flujo hacia el estado absorbente $s^* = 1$ que el observado en la base de datos. Sin embargo, debido al comportamiento asintótico de $q(r)$ cuando $r \to 0$, esta aproximación produce una pendiente correcta en el gráfico log-log de forma tal que el exponente α_d puede ser estimado con la Ec (2.33), tomando $\upsilon = 0$. Posteriormente, en el trabajo

de Blasius y Tönjes mediante el uso de la teoría de los procesos de renovación,[3] se muestra que el comportamiento asintótico de $P(s) = \sum_d P_d(s)$, en el rango $s \gg 1$, puede derivarse para un amplio espectro de distribuciones $q(r)$ encontrándose que

$$\lim_{(s/N)\to 0} P(s) = \frac{N}{\mu s^2},$$

donde $\mu = \langle -\log r \rangle$, lo cuál está en excelente acuerdo con lo encontrado empíricamente (Figura 2.6). Así, el proceso multiplicativo de la Ec. (2.29) siempre lleva un 'scaling' universal asintótico para $s \ll N$, para cualquier distribución de ramificaciones $q(r)$ bien comportada. Este resultado es importante ya que muestra que los procesos que dan lugar a distribuciones Zipf del peso de los sub-árboles de un árbol autosimilar es mucho más amplia que la clase de procesos basados en conexión preferencial o procesos de crecimiento[63].

Una de las consecuencias de la teoría de Blasius y Tönjes es que, asociando una secuencia de d movimientos a un proceso de d decisiones mutuamente excluyentes, la distribución de las secuencias de decisiones, o estrategias, que toman lugar s veces, $P_d(s) \sim s^{-\alpha_d}$, pone en evidencia una transición desde exponentes $\alpha_d \leq 2$, donde existen unas pocas estrategias que son muy comunes, a exponentes elevados $\alpha_d > 2$, donde todas las estrategias resultan igualmente dominantes[4]. Esta transición es causada por la divergencia del primer momento en leyes de potencias con exponentes mayores a -2 [54]. El número crítico de decisiones d_{cr} para el cual ocurre la transición es calculado a partir de la Ec. (2.33),

$$d_{cr} = 1 + (1 + \beta) \log\, N.$$

Para el caso de SCIDBASE en donde $N = 1{,}4 \times 10^6$, se tiene que $d_{cr} = 15$. Esto separa a la base en dos regímenes diferentes: en la fase inicial ($d < d_{cr}$) la mayor parte de las partidas de ajedrez están distribuidas entre un pequeño número de aperturas populares, mientras que más allá de la profundidad de juego crítica, las secuencias raramente utilizadas son las dominantes de modo que al considerarlas todas juntas comprenden la mayoría de las partidas. Es importante resaltar que este resultado es un efecto de la estadística y no indican un cambio de comportamiento de los jugadores al incrementarse la profundidad del juego.

2.6 Ley de Heaps

En lingüística la ley de Heaps es una ley empírica que describe el número de palabras distintas en un texto en función de la longitud de dicho texto. Esta ley se utiliza para

[3] Del inglés *renewal processes*.
[4] Cualquier estrategia presenta una popularidad s bien aproximada por la media $\langle s \rangle$

caracterizar el procesamiento de lenguaje natural, de acuerdo a la misma el tamaño del vocabulario crece de forma sub-lineal,

$$n(t) \sim t^{\lambda}, \tag{2.34}$$

donde n es la cantidad de palabras distintas, t la longitud total del texto y $\lambda < 1$ [39]. En la Figura 2.8 se muestran ejemplos de este comportamiento en diversos textos literarios.

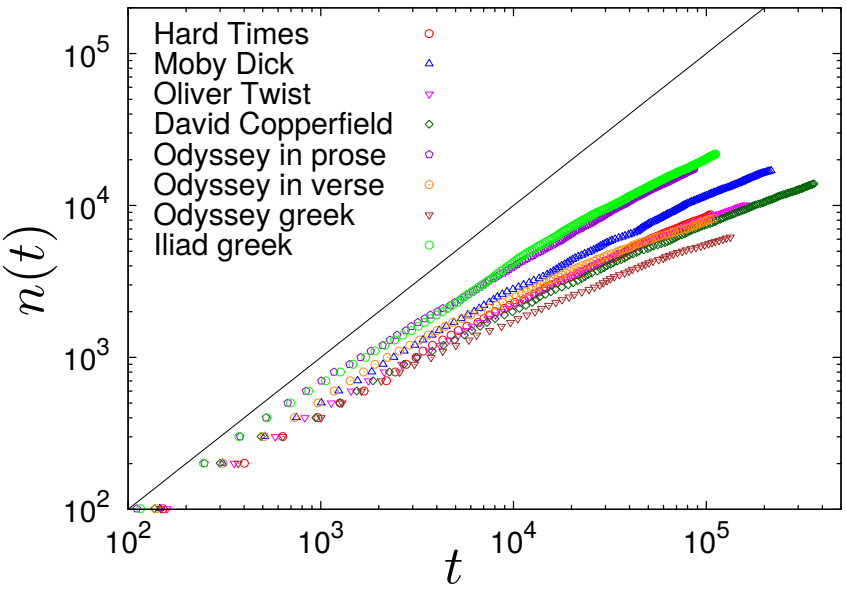

Figura 2.8.: Ejemplos de la ley de Heaps calculada para diversos textos. Figura extraída de [64].

Un fenómeno particular e interesante es la coexistencia de las leyes de Zipf y Heaps, las que han sido observadas simultáneamente en textos de diversos lenguajes tales como inglés, ruso y español [65, 66]. En la Sección 2.5 se introdujeron la función de rango y la distribución de popularidad,

$$x(r) \sim r^{-\beta} \qquad p(x) \sim x^{-\alpha},$$

donde r es el rango de la palabra de frecuencia x. Es interesante estudiar la forma de la ley de Heaps teniendo en cuenta estas distribuciones y la relación entre los exponentes de las mismas $\alpha = 1 + 1/\beta$. Tomando en cuenta que $(r-1)$ es la cantidad de palabras distintas de frecuencia mayor a $x(r)$, y $n(t)$ es la cantidad de palabras distintas,

$$r - 1 = n(t) \int_{x(r)}^{x_{max}} p(x')dx',$$

donde x_{max} es la frecuencia máxima. Utilizando la condición de normalización $\int_1^{x_{max}} p(x')dx' = 1$ y el hecho que,

$$t = \sum_{r=1}^{n(t)} x(r) \approx \int_1^{n(t)} x(r)dr,$$

se llega a que

$$\frac{n(t)^\beta (n(t)^{1-\beta} - 1)}{1-\beta} = t. \tag{2.35}$$

Claramente esta relación no es simplemente una ley de potencia, ya que en realidad la ley de Heaps es un resultado aproximado el cual es posible derivar a partir de la Ec. (2.35). Particularmente, si β es mucho más grande que 1, $n(t)^{1-\beta} \ll 1$ y

$$n(t) \approx (1-\beta)^{1/\beta} t^{1/\beta}.$$

En cambio, si $\beta \ll 1$, $n(t)^{1-\beta} \gg 1$ y $n(t) \approx (1-\beta)t$. Este resultado [66] puede ser escrito como una relación entre los exponentes de la ley de Heaps λ y de la ley de Zipf α,

$$\lambda = \begin{cases} \frac{1}{\beta} \left(= \frac{1}{\alpha - 1} \right) & \beta > 1 \, (\alpha < 2) \\ 1 & \beta < 1 \, (\alpha > 2) \end{cases} \tag{2.36}$$

Resulta evidente que la relación entre los exponentes obtenida en la Ec. (2.36) no es válida para $\beta = 1$ ($\alpha = 2$), que es el caso de la ley de Zipf encontrada en textos literarios y Ajedrez. Para tomar en cuenta este caso particular se toma el límite $\beta \to 1$ en la Ec. (2.35), en el cual $n(t)^\beta \approx n(t)$ y $n(t)^{1-\beta} \approx 1 + (1-\beta)\ln(n(t))$ con lo que se obtiene,

$$n(t)\ln(n(t)) = t \qquad \beta = 1 \, (\alpha = 2). \tag{2.37}$$

3 Modelos

3.1 El modelo de Yule-Simon

Uno de los primeros modelos capaces de explicar la aparición de leyes de Zipf-Pareto fue introducido por Yule [25], el cual fue ideado para explicar la aparición de distribuciones de ley de potencias en los tamaños de géneros biológicos. Más adelante, Simon [30] introdujo una variante similar, pero menos general del modelo, la cual se ajusta mas naturalmente al contexto de la ley de Zipf observada en textos literarios. Esta versión del modelo es conocida como modelo de Yule-Simon, y variaciones del mismo han re-aparecido en la literatura en varias ocasiones. La variante más reciente, conocida como *preferential attachment* (crecimiento preferencial), se convirtió en una de las ideas más importantes en los comienzos del desarrollo de las teorías sobre redes complejas [31].

El modelo de Yule-Simon aplicado a la generación de secuencias de clases de elementos funciona de la siguiente manera. Se comienza con un estado inicial de n_0 elementos, por ejemplo n_0 números aleatorios sorteados de una distribución uniforme. A cada paso temporal t se presentan dos opciones: i) introducir un elemento de una nueva clase a la serie de datos con probabilidad p; o ii) copiar un elemento ya existente de la serie de elementos con probabilidad $\bar{p} = 1 - p$. En el último caso se debe determinar cuál de los elementos ya existentes será copiado. En el modelo de Yule-Simon todos los elementos existentes tienen la misma probabilidad de ser copiados. Ya que a cada paso temporal un elemento es agregado, ya sea nuevo o copiado, el número de elementos total en la base de datos construida a tiempo t es $N(t) = t + n_0 \approx t$. La probabilidad de escoger un elemento de una clase i particular que ya ha ocurrido $s_i(t)$ veces a tiempo t es,

$$\Pi(i, t) = \frac{s_i(t)}{N(t)}. \tag{3.1}$$

Esto significa que en el modelo de Yule-Simon copiar un cierto elemento no depende de la distribución temporal de las ocurrencias de dicho elemento, sino de cuan popular es la clase a la cual pertenece.

La derivación analítica de la ley de potencia producida por el modelo de Yule-Simon se realiza en la aproximación de valores medios, donde se supone un límite continuo en

el tiempo de la ocurrencia de los elementos de la base. Dado un número de ocurrencias s_i^* a tiempo t, al tiempo $t + \Delta t$ se tendrá,

$$s_i^*(t + \Delta t) = s_i^*(t) + (1 - p)\Pi(i, t)\Delta t.$$

Despejando y tomando el límite cuando $\Delta t \to 0$ se obtiene,

$$\frac{ds_i^*}{dt} = \lim_{\Delta t \to 0} \frac{s_i^*(t + \Delta t) - s_i^*(t)}{\Delta t} = \bar{p}\Pi(i, t) \qquad \Pi(i, t) = \frac{s_i^*}{N(t)} = \frac{s_i^*}{t}$$

$$\frac{ds_i^*}{dt} = \bar{p}\frac{s_i^*}{t}$$

$$\frac{ds_i^*}{s_i^*} = \bar{p}\frac{dt}{t}.$$

Entonces en forma integral se tiene que:

$$\int_1^{s_i^*} \frac{ds_i^{*'}}{s_i^{*'}} = (1 - p) \int_{t_i}^{t} \frac{dt}{t},$$

y finalmente se obtiene,

$$s_i^*(t) = \left(\frac{t}{t_i}\right)^{1-p} \qquad s_i^*(t_i) = 1. \tag{3.2}$$

donde t_i es el tiempo de la primera aparición de elementos de la clase i.

Debido a que este desarrollo se realiza en la aproximación de valores medios no tiene en cuenta las fluctuaciones, es decir la ocurrencia de una clase de elementos individual se desviará de este valor esperado. Hay clases que ocurrirán con mayor frecuencia que el valor esperado, y otras con menos frecuencia. Por lo tanto, resulta natural plantear la pregunta: ¿qué forma tiene la distribución de probabilidad de la ocurrencia de una clase i? [67]

Se define $u_i = t_i + \Delta t$, donde t_i es el tiempo en el cual la clase de elementos i ocurrió por primera vez, y Δt el tiempo transcurrido desde entonces. A fin de calcular la distribución de probabilidades de las fluctuaciones se calcula la probabilidad $P(s_i(t) = s)$ de que el número de ocurrencias de la clase i a tiempo t sea $s_i(t) = s$.

Realizando los cálculos apropiados (Apéndice C) y tomando el límite $p \to 0$ ($\bar{p} \to 1$) se obtiene,

$$P(s_i(u_i) = s) \overset{p \to 0}{\sim} t_i u_i^{-s} (\Delta t - s + 2)^{s-1}. \tag{3.3}$$

Para el caso de $p > 0$ no es posible una resolución analítica, por lo cual se debe acudir al cálculo numérico.

Con estos resultados es posible calcular la escala de la desviación del valor esperado del número acumulado de ocurrencias de elementos individuales. Como se mencionó anteriormente, el valor esperado es $s_i^*(t) = (t/t_i)^{1-p}$, el cual aumenta más lentamente en el caso de t_i mayores, entonces la desviación de elementos con t_i mayores será menor que para elementos cuyos valores de t_i sean menores si Δt fuese el mismo en ambos casos. Por lo tanto se normaliza el tamaño de la desviación con la dependencia de t_i utilizando distintos períodos de observación Δt_i,

$$u_i = t_i + \Delta t_i = \lambda t_i,$$

donde λ es una constante, pero distinta para cada clase de elementos de la secuencia. Entonces, para el valor de referencia con el que se mide la escala de la desviación para cada clase de elementos se tiene que,

$$\left(\frac{\lambda t_i}{t_i}\right)^{1-p} = \lambda^{1-p} \overset{p\to 0}{=} \lambda.$$

Reemplazando $s = x\lambda^{\bar{p}}$ (x veces el valor de referencia) en la Ec. (3.3),

$$\begin{aligned} P(s_i(u_i) = x\lambda) &\overset{p\to 0}{\sim} t_i(\lambda t_i)^{-x\lambda}[(\lambda-1)t_i - x\lambda + 2)^{x\lambda-1} \\ &= \lambda^{-1}\left(1 - \frac{1}{\lambda} - \frac{x - 2/\lambda}{t_i}\right) \end{aligned}.$$

Ya que x es la escala de la desviación, $s = x\lambda^{\bar{p}} = x\lambda^{1-p} = xs_i^*(u_i)$. Por otro lado, $s_i(u_i) = s$, entonces $s_i(u_i) = xs_i^*(u_i)$, y se obtiene,

$$x = \frac{s_i(u_i)}{s_i^*(u_i)}.$$

Si se asume que $t_i \gg x \sim 1 \Rightarrow \frac{x-2/\lambda}{t_i} \sim 0$, entonces se obtiene la distribución de probabilidad de la fluctuación para todos los elementos de la base,

$$P(s_i(u_i) = x\lambda) \overset{p\to 0}{\sim} \frac{1}{\lambda - 1}\left(1 - \frac{1}{\lambda}\right)^{x\lambda},$$

la cual no depende de t_i y decae exponencialmente.

3.2 El modelo de Cattuto

En la sección anterior se presentó el modelo de Yule-Simon. En el mismo, la probabilidad de copiar un elemento de una clase ya existente en la secuencia en construcción no depende de la distribución temporal de las ocurrencias de dicha clase, y por lo tanto el proceso no exhibe efectos de memoria. Cattuto et al. [40] en su trabajo proponen una modificación del proceso de Yule-Simon mediante la introducción de un núcleo de memoria. En el modelo de Cattuto el proceso de generación comienza de igual manera que en el modelo de Yule-Simon. Se comienza con un estado inicial de n_0 elementos, y a cada paso temporal t se introduce un elemento de una nueva clase a la base de datos con probabilidad p, o se copia un elemento ya existente con probabilidad $\bar{p} = 1 - p$; pero a diferencia del modelo de Yule-Simon, la probabilidad de copiar un elemento que ha ocurrido previamente depende de la lejanía temporal en la cual éste ocurrió, tomando así en cuenta la 'edad' del elemento. Si el elemento apareció a tiempo $t - \Delta t$, la probabilidad de copiado al tiempo t está definida por,

$$Q(t, \Delta t) = \frac{C(t)}{\kappa_c + \Delta t}, \tag{3.4}$$

donde κ_c es la escala temporal en la cual elementos recientes tienen probabilidades asociadas similares, y puede ser considerado como una medida de la extensión del núcleo de memoria; $C(t)$ es un factor de normalización logarítmico, pues este núcleo de memoria es de largo alcance.

Con el fin de calcular la distribución de popularidad de los elementos de la serie temporal artificialmente generada por este proceso, se comienza por calcular la probabilidad $P(\Delta t)$ de que el elemento i, habiendo ocurrido por primera vez a tiempo t, ocurra nuevamente a $t + \Delta t$ por segunda vez.

Para $\Delta t = 1$ se tiene que,

$$\begin{aligned} P(1) &= \bar{p}\frac{C}{\kappa_c + 1} \\ &= \frac{\alpha}{\kappa_c + 1}, \end{aligned}$$

donde se considera $C(t) = C$ constante ($\Delta t \ll t$), ya que esta constante tiene una dependencia logarítmica, y $\alpha \equiv C\bar{p}$. Para el caso de $\Delta t > 1$,

$$\begin{aligned} P(\Delta t) &= \left\{ \prod_{\tau=1}^{\Delta t-1} \left[p + \bar{p}(1 - Q(\tau)) \right] \right\} \bar{p} Q(\Delta t) \\ &= \left(\frac{\alpha}{\kappa_c + \Delta t} \right) \prod_{\tau=1}^{\Delta t-1} \left(1 - \frac{\alpha}{\kappa_c + \tau} \right). \end{aligned}$$

Usando la fórmula,

$$\prod_{\tau=1}^{\Delta t-1}\left(1-\frac{\alpha}{\kappa_c+\tau}\right)=\frac{(-\alpha-\kappa_c+1)_{\Delta t-1}}{(\kappa_c+1)_{\Delta t-1}},$$

donde $(x)_n \equiv \frac{\Gamma(x+n)}{\Gamma(x)}$ es el símbolo de Pochhammer, se obtiene,

$$P(\Delta t)=\bar{p}\frac{C}{\kappa_c+\Delta t}\frac{\Gamma(-\alpha+\kappa_c+\Delta t)\Gamma(\kappa_c+1)}{\Gamma(-\alpha+\kappa_c+1)\Gamma(\kappa_c+\Delta t)},$$

de lo cual, asumiendo que $t \gg \Delta t \gg 1$, se obtiene,

$$P(\Delta t)=\alpha(\Delta t+\kappa_c)^{-\alpha-1}(\kappa_c+1)^{\alpha}. \tag{3.5}$$

Por simplicidad, se toma $\kappa_c = 0$. A cualquier tiempo t, el tiempo característico de retorno $\langle \Delta t \rangle$ puede ser calculado con la Ec. (3.5),

$$\begin{aligned}\langle \Delta t \rangle &= \sum_{\Delta t=1}^{t} \Delta t P(\Delta t) \\ &\simeq \frac{\alpha}{1-\alpha} t^{1-\alpha}\end{aligned} \tag{3.6}$$

En una descripción continua, la frecuencia s_i de una dada clase de elementos i varía de acuerdo a la ecuación,

$$\frac{ds_i}{dt}=\bar{p}\Pi_i, \tag{3.7}$$

donde Π_i es la probabilidad de escoger una ocurrencia previa de los elementos de la clase i. Considerando el núcleo de memoria del modelo de Cattuto con $\kappa_c = 0$,

$$\Pi_i = C\sum_{j=1}^{j=s_i}\frac{1}{t-t_j^{(i)}},$$

donde $t_j^{(i)}$ $(j = 1, 2, ..., s_i)$ son los tiempos anteriores en los cuales ocurrió el elemento i.

Usando la aproximación de valores medios,

$$\Pi_i = C\sum_{j=1}^{j=s_i}\frac{1}{t-t_j^{(i)}} \simeq C s_i \left\langle \frac{1}{t-t_j} \right\rangle_j, \tag{3.8}$$

donde $\langle\rangle_j$ denota el promedio sobre las s_i ocurrencias de i. Si se asume que el promedio está dominado por la contribución de la ocurrencia más reciente del elemento i a tiempo t_{s_i}, y utilizando el resultado de la Ec. (3.6), se obtiene,

$$\left\langle \frac{1}{t-t_j} \right\rangle_j \simeq \frac{1}{t-t_{s_i}} \simeq \frac{1}{\langle \Delta t \rangle} = \frac{1-\alpha}{\alpha} \frac{1}{t^{1-\alpha}}. \tag{3.9}$$

Esta expresión captura correctamente la dependencia temporal del promedio $\langle (t-t_j)^{-1} \rangle$ para una dada frecuencia s_i si se introduce un factor constante Ω de corrección [40],

$$\left\langle \frac{1}{t-t_j} \right\rangle_j \simeq \frac{1}{\Omega} \frac{1-\alpha}{\alpha} \frac{1}{t^{1-\alpha}}. \tag{3.10}$$

Combinando las Ec. (3.10), Ec (3.8) y Ec. (3.7), se obtiene,

$$\frac{ds_i}{dt} \simeq \alpha s_i \left\langle \frac{1}{t-t_j} \right\rangle_j = \frac{s_i}{\Omega}(1-\alpha)t^{\alpha-1} \tag{3.11}$$

$$\int_1^{s_i} \frac{ds_i'}{s_i'} = \frac{1-\alpha}{\Omega} \int_{t_i}^{t} t'^{\alpha-1} dt' \tag{3.12}$$

$$s_i = \exp\left[\frac{1-\alpha}{\alpha\Omega} t^\alpha\right] \exp\left[-\frac{1-\alpha}{\alpha\Omega} t_i^\alpha\right] \tag{3.13}$$

$$= A e^{-K t_i^\alpha}, \tag{3.14}$$

donde $K \equiv \frac{1-\alpha}{\Omega\alpha}$ y $A \equiv e^{Kt^\alpha}$.

Finalmente, la densidad de probabilidad de popularidades o frecuencias $P(s)$ puede ser calculada como [62] (para mayor detalle ver Apéndice D),

$$P(s) = \frac{p}{(n_0+pt)(K\alpha)s} \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha-1} \tag{3.15}$$

Parte II

Resultados

4 Base de Datos

4.1 Descripción de la base de datos

Un partido de Ajedrez es usualmente dividido en tres etapas, apertura, juego medio y final. Existen muchas aperturas específicas –secuencias de movimientos en la etapa inicial del juego– las cuales están bien documentadas ya que son consideradas buenas jugadas en un sentido competitivo. Como consecuencia de esto, los primeros movimientos de muchas partidas en las bases de datos de Ajedrez coinciden.

Las aperturas evolucionan continuamente y la complejidad de las posiciones determina su extensión. El conocimiento de líneas de aperturas es parte de los antecedentes teóricos de los jugadores de Ajedrez, y la habilidad de los mismos está ciertamente relacionada a cuantas líneas de apertura conocen.

En la práctica, la extensión de la etapa de apertura no puede ser definida precisamente y depende del tipo de apertura, es por esto que se utilizarán los términos 'líneas de apertura' y 'líneas de juego' para hacer referencia a secuencias con igual cantidad de movimientos.

La base de datos utilizada, SCIDBASE [59], cuenta con más de $3{,}5 \times 10^6$ partidas de Ajedrez, desde el año 206 dC al 2007, y fue convertida al formato PGN (*portable game notation*) utilizando una variación de la SCIDBASE llamada Scid vs Pc [68].

El formato en el cual se encuentran registradas las partidas es el siguiente: #(indicando una nueva partida), número de partida (orden en la base), año, día, mes, jugador de las blancas, jugador de las negras, Elo de las blancas, Elo de las negras, resultado del partido, evento (por ejemplo si la partida fue jugada en un torneo); y luego se encuentran registrados los movimientos realizados en dos columnas, la primera correspondiente a los movimientos de las blancas y la segunda a los movimientos de las negras.

El Elo (Apéndice A.3) es una calificación dinámica en el cual cada jugador posee un puntaje numérico, el cual no es calculado de manera absoluta, sino que es estimado a partir de las victorias, empates y derrotas en enfrentamientos con otros jugadores, y por lo tanto, cambia luego de que un jugador juega una partida. A pesar de esto, en la base de datos, los Elos de los jugadores son aproximados y permanecen constantes a través

del tiempo. Además, como el sistema de puntuación Elo fue implementado en 1970, para los partidos que tomaron lugar antes de 1970, los Elos de los jugadores son una estimación.

Del total de las partidas registradas sólo $1{,}5 \times 10^6$ poseen todos los datos completos. Particularmente, en algunos casos no se conoce la fecha exacta en la cual la partida fue jugada, o no se conoce el puntaje de Elo de los jugadores. Es por esto que en este trabajo estas partidas con "datos corruptos" fueron eliminadas y se trabajó solamente con las partidas cuyos datos están completos, lo que resulta en un período temporal desde el año 1998 al 2007.

4.2 Análisis de la estructura de las partidas

En el juego del Ajedrez cada secuencia de movimientos posible puede ser mapeado a un 'árbol de partidas' (Figura 4.1(a)), donde el nodo raíz es la posición inicial de las piezas en el tablero. En el árbol de jugadas cada movimiento es representado por una arista o conexión, y hay una correspondencia uno-a-uno entre las líneas de juego y los nodos. La distancia topológica entre la raíz y un nodo es la profundidad d de la línea de juego correspondiente.

Un nodo, o línea de juego en el árbol, se denota por g, y la popularidad de la línea de juego g –i.e. el número de veces que g aparece en la base de datos– es denotada por s_g. En la (Figura 4.1(a)) se muestra un árbol de jugadas parcial donde la *popularidad* está representada por el tamaño del nodo. El número de ramas que salen de un nodo g se denota por b_g, la profundidad de g por d_g, y el número de nodos a profundidad d por n_d, lo que corresponde al número de líneas de juego diferentes que pueden ser encontradas en la base de datos a profundidad d. Similarmente, N_d es el numero total de juegos en la base de datos que han alcanzado una profundidad $d_g = d$.

A cada profundidad d se puede calcular un factor de ramificación, o cociente de ramificación, usando la fórmula:

$$\langle b_d \rangle = \frac{1}{n_d} \sum_{g:d_g=d} b_g = \frac{n_{d+1}}{n_d}, \tag{4.1}$$

donde la suma se realiza sobre todos los nodos g a profundidad d. En la práctica, la base de datos de Ajedrez crece continuamente, i.e., nuevas partidas son incorporadas a la base de datos con el paso del tiempo, por lo tanto estas cantidades cambian con el tiempo. Por razones prácticas, no se utiliza el tiempo real, se utiliza un tiempo ordinal denotado por t; en este sentido, $g(t)$ es la línea de juego asociada al t-ésimo partido que aparece en la base de datos. Similarmente, $s_g(t)$ es el número de las t partidas que han

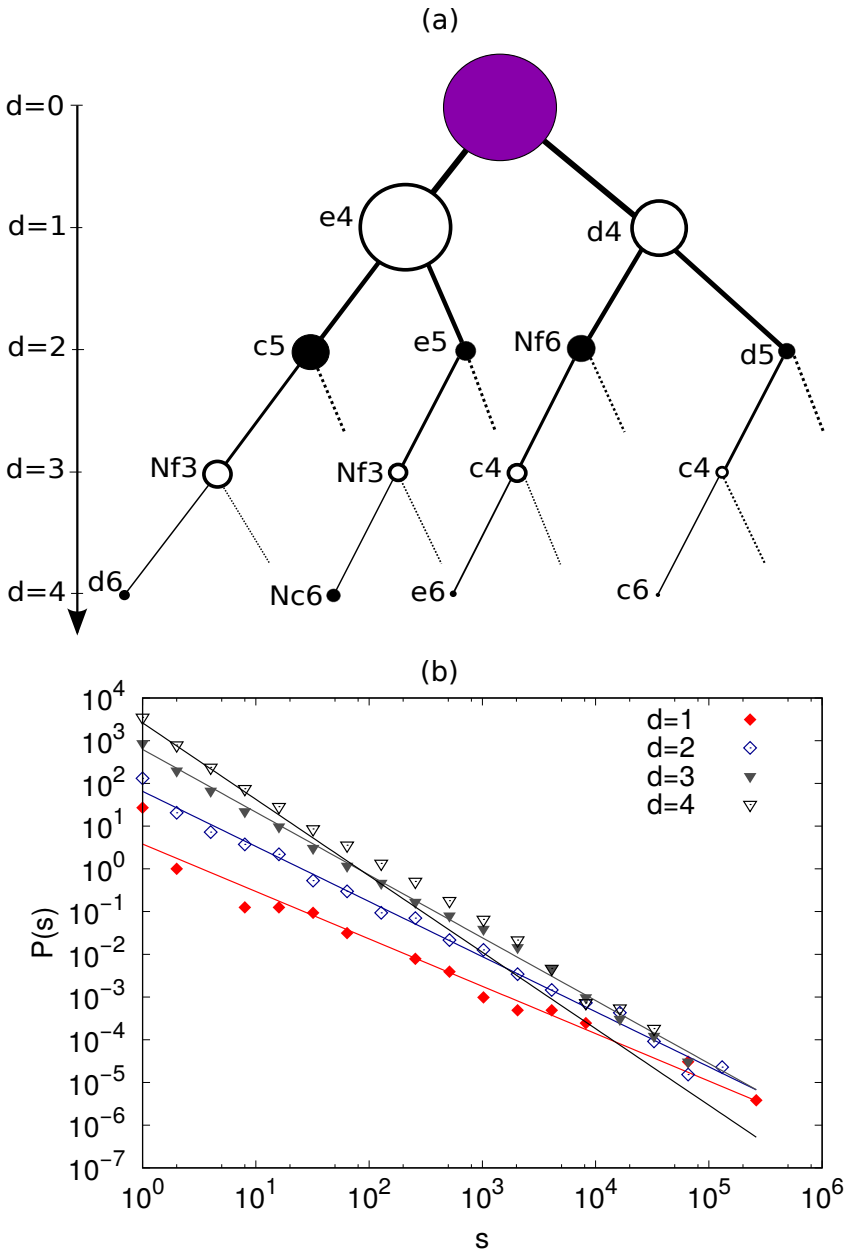

Figura 4.1.: (a) Árbol de jugadas correspondiente a las principales líneas de apertura hasta profundidad $d = 4$. El tamaño de los nodos es proporcional a su popularidad. Solo se muestran las líneas principales. (b) Distribución de popularidades de los nodos a profundidad $d = 1, 2, 3$ y 4; estas distribuciones son bien ajustadas por una ley de potencia $P(s) \propto s^{-\alpha}$ con $\alpha = 1{,}10 \pm 0{,}05$, $1{,}29 \pm 0{,}03$, $1{,}47 \pm 0{,}02$ y $1{,}59 \pm 0{,}02$ ($R^2 = 0{,}972$, $0{,}993$, $0{,}996$ y $0{,}997$) respectivamente. Los errores son estimados por el ajuste.

alcanzado al nodo g, N_d es el número de partidas que han alcanzado una profundidad d y $n_d(t)$ es el número de líneas de juego diferentes dentro de los $N_d(t)$ juegos [17].

Desde el punto de vista estadístico, la popularidad de una línea de juego dada depende del número de movimientos considerados, i.e., la profundidad d del juego. Como se

mencionó en la Sección 2.5.1, Blasius y Tönjes encontraron que la distribución de popularidades es una ley de potencia con exponente dependiente de d, lo que significa que hay pocas líneas de aperturas que son muy populares, mientras que el resto raramente se juegan. Estos resultados han sido reproducidos y se muestran en la Figura 4.1(b), donde la distribución de popularidad se muestra para $d = 1, 2, 3$ y 4, y las curvas fueron ajustadas utilizando regresión lineal. Claramente, el exponente aumenta con d, como fue reportado en [13]. Una secuencia específica de movimientos hasta una cierta profundidad puede ser pensada como una palabra, una cadena en notación algebraica, y a la base de datos como un cuerpo literario donde el t-ésimo juego correspondería a la $t-$ésima palabra. De esta forma, analizar la base de datos a distintas profundidades es análogo a analizar distintos textos, todos extraídos de la misma base de datos y todos con distintos exponentes de Zipf.

La estructura del árbol de partidas también depende de la profundidad d. En la Figura 4.2 se muestra la media del cociente de ramificación en función de la profundidad d. El cociente ramificación cuantifica la complejidad del juego y la memoria de los jugadores de Ajedrez al seguir las líneas de juego. El cociente de ramificación $\langle b_r \rangle$ alcanza un valor ≈ 1 para $d = 25$, lo que significa que la generación de nuevas ramas es despreciable a partir de dicha profundidad, marcando el comienzo de la etapa conocido como juego medio. En la Figura 4.2 se muestra también el número de líneas de juego diferentes, n_d, en función de la profundidad d. Al comienzo de las partidas, e.g. $d = 4$, el número de líneas de juego que siguen los jugadores es relativamente pequeño, y un número significativo de jugadores siguen la línea de juego más popular. La complejidad estadística del juego está reflejada en el cociente de ramificación $\langle b_d \rangle$. Además, $\langle b_d \rangle$ depende del tamaño de la base de datos, ya que durante el crecimiento de la base de datos, nuevas ramas son creadas, y al mismo tiempo el rango de popularidad depende de d. Entonces, a $d = 4$ se captura la memoria y la complejidad del juego, ya que las líneas de juego más importantes pueden ser identificadas a esta profundidad y el cociente de ramificación es aún mayor a 1 ($\langle b_d \rangle \approx 3{,}5$). Además, a esta profundidad, el exponente de la distribución de popularidad es $\alpha < 2$ y por lo tanto el rango de popularidades es más extenso que para mayores d. Al calcular la distribución de número de ramas generadas por cada nodo b_g para diferentes valores de d se encuentra que, para profundidades menores ($d \leq 19$), la distribución es exponencial, mientras que para profundidades más allá de $d = 20$ una ley de potencia provee un mejor ajuste. De todas formas, se debe notar que el rango del ajuste es sólo un orden de magnitud, y por lo tanto el ajuste de la ley de potencia no es preciso. Las fluctuaciones (Figura 4.2 inset) decaen exponencialmente al incrementarse d. Se pueden identificar dos regímenes y la transición entre ellos esta relacionado al cambio de régimen observado en $\langle b_d \rangle$ y n_d. Por lo tanto, la mayor parte de los análisis realizados están restringidos a las 6279 líneas de juego de longitud $d = 4$ de la base de datos. En particular se presta especial atención a la línea de apertura más popular a esta profundidad: **1 e4 e5 2 ♘f3 ♘c6** – ya que representa cerca del $7{,}8\,\%$ de los juegos de la base de datos. La razón de esto es que

muchas aperturas populares comparten los cuatro primeros movimientos. Por ejemplo: **3 ♗b5** (Ruy Lopez, la más popular por amplio margen), **3 ♗c4** (Giuoco piano), **3 d4** (Scotch opening), por mencionar algunas.

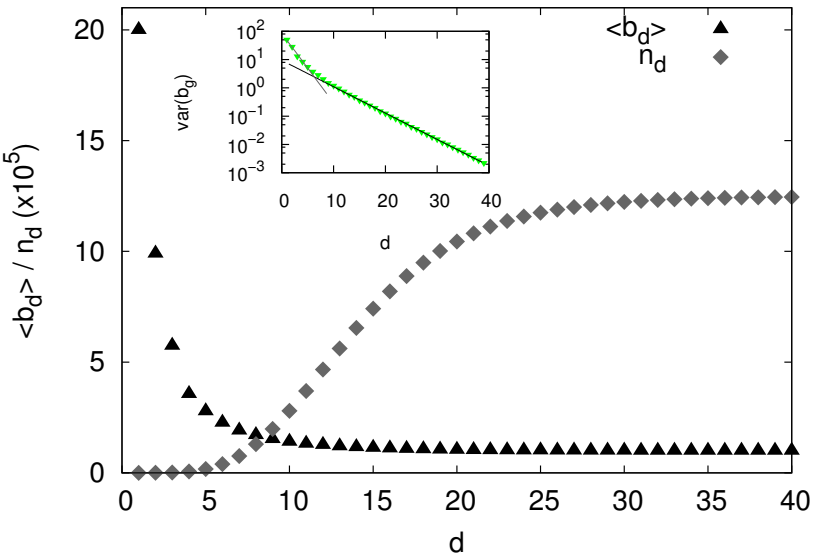

Figura 4.2.: Cociente de ramificación promedio, $\langle b_d \rangle$ y número n_d en función del nivel de profundidad d en la base de datos. Inset: varianza de la distribución de ramas por nodo b_g como función de la profundidad d y ajuste lineal de dos regímenes exponenciales, con exponente $0{,}57$ para $1 < d \leq 9$ y $0{,}21$ para $d \geq 10$.

5 Innovación de líneas de juego

En esta sección se presentan los resultados obtenidos a partir del estudio de la innovación de líneas de juego en la base de datos de Ajedrez, i.e. la dinámica que describe la introducción de nuevos elementos a la base de datos con el tiempo. Para describir esta dinámica, se introduce un modelo basado en un crecimiento preferencial anidado el cual es capaz de reproducir la fenomenología encontrada en la base de datos, en particular la ley de Heaps.

5.1 La ley de Heaps en el Ajedrez

Como se introdujo en la Sección 2.6, la dinámica de innovación de muchos sistemas sigue la ley de Heaps. En particular, para el caso de la base de datos de Ajedrez, un evento innovador ocurre cuando una partida genera una nueva rama en el árbol de partidas. En el árbol de partidas, el t-ésimo juego puede generar una nueva rama a profundidad $d_b(t)$ y finalizara a profundidad $d_e(t) \geq d_b(t)$, donde nuevamente t juega el rol de tiempo ordinal.

Por cada evento innovador que ocurre a profundidad $d_b(t)$, el ancho del árbol de jugadas o número de nodos n_d a profundidad d evoluciona como:

$$n_d(t) = \begin{cases} n_d(t-1)+1 & \text{si } d_b(t) < d \leq d_e(t) \\ n_d(t-1) & \text{otros casos} \end{cases}.$$

Si t_d es el número de partidas que alcanzan al menos profundidad d luego de t partidas, se encuentra que (Figura 5.1):

$$n_d(t_d) = \begin{cases} t_d, & t_d \ll t_d^* \\ t_d^*(t_d/t_d^*)^{\lambda_d}, & t_d \gg t_d^* \end{cases} \tag{5.1}$$

donde t_d^* es un valor característico donde se produce un cambio de régimen en la innovación de líneas de juego de profundidad d. El exponente λ_d, que caracteriza la tasa de innovación, satura exponencialmente con d como:

$$\lambda_d = 1 - B^d \tag{5.2}$$

con $B \simeq 0{,}85$ (Figura 5.1, inset). La relación de escala $n_d \sim t_d^{\lambda_d}$ para $t_d \gg t_d^*$ corresponde a la Ley de Heaps [34] (Sección 2.6), comúnmente encontrada en el crecimiento del vocabulario en cuerpos literarios y lenguajes. Más aún, se encontró que la intersección t_d^* crece exponencialmente con el exponente de Heaps,

$$t_d^* \sim \exp A\lambda_d \tag{5.3}$$

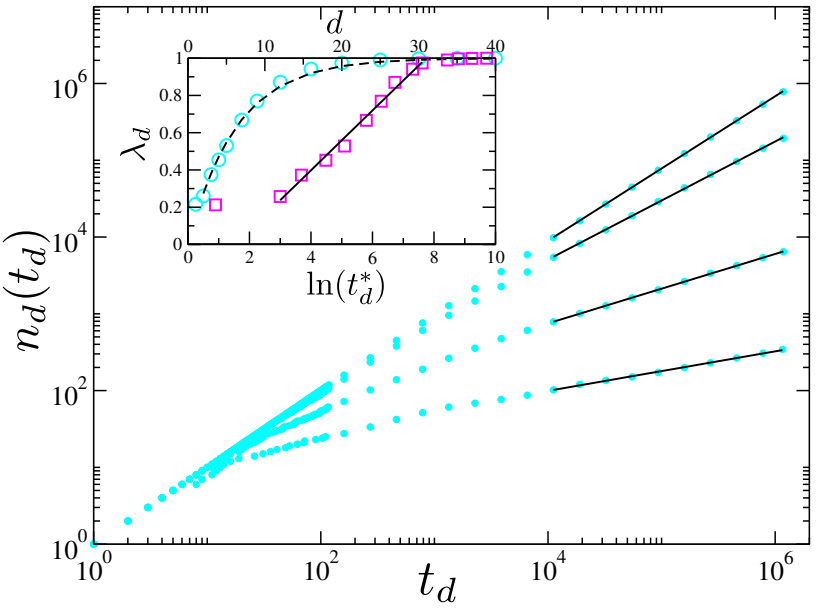

Figura 5.1.: Ancho del árbol a profundidad d, n_d, en función del número de partidas t_d que han alcanzado dicha profundidad. Sólo se muestran los casos para $d = 2, 4, 9$ y 16 y los datos están logaritmicamente *binneados* para $t_d > 100$. Para valores grandes de t_d se muestran los ajustes con la función $n_d(t) = t_d^*(t/t_d^*)^{\lambda_d}$. Inset: valores estimados de λ_d en función de d (círculos) y en función de los valores estimados de $\ln t_d^*$ (cuadrados). La línea a trazos corresponde al ajuste con $\lambda_d = 1 - B^d$, donde $B = 0{,}854 \pm 0{,}002$ ($R^2 = 0{,}998$), y la línea sólida al ajuste con $\lambda_d = a + (1/A) \ln t_d^*$ (Ec. (5.3)), donde $1/A = 0{,}160 \pm 0{,}006$ ($R^2 = 0{,}99$).

5.2 Innovación y crecimiento preferencial anidado

A fin de comprender el mecanismo subyacente de la evolución del árbol de partidas, se caracterizaron los procesos de innovación y re-utilización de líneas existentes [17]. Sea $s(t)$ la frecuencia de ocurrencia de una dada secuencia de movimientos luego de que t partidas han sido jugadas. Se caracterizó el proceso de innovación mediante la probabilidad $p(s)$ de que una partida que ha alcanzado un nodo con frecuencia s genere una nueva rama que parte de ese nodo; esta probabilidad está dada por,

$$p(s) \simeq s^{-\nu} \tag{5.4}$$

donde $\nu \simeq 0{,}88$ (Figura 5.2(a)). Para caracterizar el proceso de re-utilización se estudió la probabilidad condicional, $\pi(s|s')$, que una partida siga una arista ya existente desde un nodo con frecuencia s' a uno de sus nodos hijos de frecuencia s. Se midió $\pi(s|s')$ a medida que las partidas cruzan las aristas del árbol en crecimiento. Por ejemplo, en la Figura 5.2(b) (línea magenta) se muestra $\pi(s|s' = 100)$, la cual fue obtenida para movimientos con profundidad $d \geq 5$ para poder asegurar una estadística adecuada. Resultados similares fueron encontrados para otros valores de s'. Se encontró que $\pi(s|s')$ es una función inhomogénea que satisface:

$$\pi(s|s') = \frac{1}{s'} q\left(r\right), \tag{5.5}$$

donde $q(r)$ es la función densidad de probabilidad de tasa de frecuencias, $r = s/s'$ (Figura 5.2(b) histograma cian). La densidad de probabilidad, $q(r)$, es medida a lo largo del proceso de crecimiento del árbol. Los movimientos que finalizan en nodos con $s < 100$ o $d < 5$ son descartados, para evitar efectos de discretización en $q(r)$. La forma funcional $q(r) = 2/[\pi\sqrt{1-r^2}]$ (Figura 5.2(b) línea negra a rayas) fue previamente determinada por Blasius y Tönjes [13] (Ec. (2.30)) midiendo los valores de r a medida que las partidas cruzan las aristas del árbol ya completo. Ya que $q(r)$ es una función creciente de r las ramas de mayor frecuencia son jugadas con mayor preferencia, y por lo tanto el proceso corresponde a un crecimiento preferencial. Sin embargo, es diferente del típico caso donde la probabilidad del crecimiento preferencial crece linealmente con la frecuencia [25, 30, 69, 31, 70]. Se debe notar que los comportamientos de escala de las Ec. (5.4) y Ec. (5.5) han sido medidos en toda la extensión del árbol, y por lo tanto corresponden al comportamiento de un proceso independiente de la profundidad. Tanto lo mencionado, como la forma funcional de $\pi(s|s')$ evidencian la naturaleza auto-similar de la evolución del árbol de partidas.

Luego se caracterizó la estructura del árbol en la etapa de crecimiento mediante las propiedades estadísticas de la profundidad de juego d_e. Se encontró que la fracción $S_t(d_e)$ de partidas que no finalizaron hasta profundidad d_e, sigue una distribución de Gumbel para máximos [71],

$$F_t(d_e) = 1 - S_t(d_e) = \exp(-\exp(-(d_e - \mu_e)/\beta_e)), \tag{5.6}$$

y es independiente del número de partidas jugadas hasta el momento, i.e. del tiempo t; en el inset de la Figura 5.2(c) se muestra el correspondiente ajuste. Por otro lado, la fracción $S_t(d_b)$ de partidas que, a profundidad d_b, no se han ramificado si depende de t (Figura 5.2(c)). Por lo tanto, resulta que las profundidades de juego son prácticamente independientes de la etapa de crecimiento del árbol, no así de la ramificación del árbol.

Con el fin de comprender el mecanismo de generación del árbol de juego, se introducen algunas consideraciones teóricas. Si se asume que las partidas tienen longitud infinita,

$t_d = t \ \forall \ d$. Esta suposición está justificada ya que, como se mencionó antes, las longitudes (profundidades) de juego y el crecimiento del árbol son estadísticamente independientes. Además, se considera el comportamiento asintótico para t grande en lo que resta de la Sección. De acuerdo con el enfoque de campo medio dependiente de la profundidad, el factor de ramificación b_d a profundidad d satisface (ver Ec. (4.1)):

$$b_d(t) = \frac{n_{d+1}(t)}{n_d(t)} \sim t^{B^d(1-B)}, \tag{5.7}$$

y la frecuencia o popularidad media por nodo a profundidad d está dada por:

$$s_d = \frac{t}{n_d(t)} \sim t^{B^d}. \tag{5.8}$$

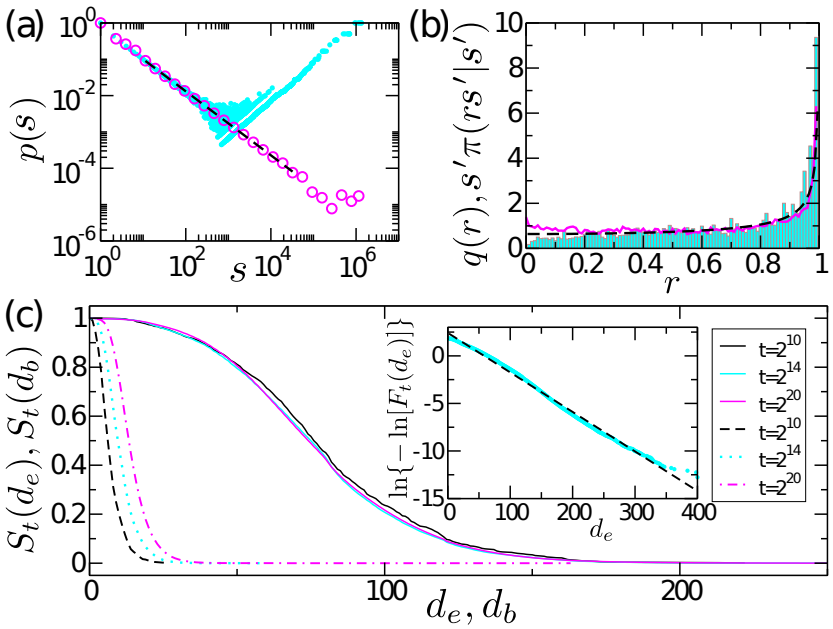

Figura 5.2.: (a) Probabilidad de generar una nueva rama $p(s)$, como función de la frecuencia del nodo s (círculos cian). Los datos con *binneado* logarítmico (círculos magenta) fueron ajustados con $p(s) \sim s^{-\nu}$ (línea negra a rayas) con $\nu = 0{,}881 \pm 0{,}009$ ($R^2 = 0{,}9998$). (b) Función densidad de probabilidad empírica, $q(r)$, de la tasa de frecuencia $r = s/s'$ (histograma cian) y la correspondiente expresión analítica, $q(r) = 2/[\pi\sqrt{1-r^2}]$ (línea negra a rayas), comparado con la forma reescalada $s'\pi(rs'|s')$ de la distribución de probabilidad condicional de crecimiento $\pi(s|s')$ (línea continua magenta). (c) Fracción $S_t(d_e)(S_t(d_b))$ de partidas con profundidad de juego (profundidad de ramificación) mayor o igual a d_e (d_b) luego de que se hayan jugado t partidas (línea solida (línea a trazos)). Colores diferentes indican distintos valores de t. Las curvas $S_t(d_e)$ colapsan en una sola, mientras que las curvas de $S_t(d_b)$ dependen de t. Inset: Ajuste a una distribución de Gumbel para máximos $F_t(d_e) = 1 - S_t(d_e) = \exp(-\exp(-(d_e - \mu_e)/\beta_e))$ (línea negra) con $\mu_e = 58{,}2 \pm 0{,}2$ y $\beta_e = 24{,}1 \pm 0{,}1$ ($R^2 = 0{,}995$) para t grande (círculos cian).

Combinando las Ec. (5.7) y Ec. (5.8) se obtiene,

$$b_d \sim {s_d}^{1-B}. \tag{5.9}$$

El factor de ramificación a profundidad d crece sub-linealmente con la frecuencia o popularidad media de los nodos s_d. La derivada con respecto a t del factor de ramificación lleva a,

$$\frac{db_d}{dt} \sim {s_d}^{-B} \frac{ds_d}{dt}, \tag{5.10}$$

y por otra parte, usando la Ec. (5.4),

$$\frac{db_d}{dt} \sim p(s_d) \frac{ds_d}{dt}, \tag{5.11}$$

En la última expresión se asume que la tasa de ramificación aumenta cada vez que una partida que llega a un nodo genera una nueva rama. La llegada de nuevas partidas al nodo a profundidad d ocurre a una tasa ds_d/dt, y la generación de nuevas ramas con probabilidad $p(s_d)$, por lo tanto se obtiene la relación aproximada $\nu \sim B$ (aproximación de campo medio). El hecho que el exponente ν sea independiente de d es consistente con un proceso de crecimiento auto-similar donde opera el mismo mecanismo estocástico en cada nodo independientemente de la profundidad. La tasa de innovación por nodo, db_d/dt, no es constante, lo que es opuesto a la tasa de crecimiento constante de la formulación estándar del proceso de crecimiento preferencial de Yule-Simon [25, 30].

En base a las consideraciones teóricas, se propone un mecanismo generativo que consiste en un proceso de crecimiento preferencial anidado. Se generan cien árboles de partidas, cada uno conteniendo 10^6 partidas, y luego se calcula λ_d en función de d, y también t_d^*, a partir de la media de $n_d(t)$, a fin de comparar los resultados de las simulaciones con el caso empírico (Figura 5.3). Cada simulación comienza con un nodo raíz, y las partidas son agregadas una a una. Si la $(t+1)-$ésima partida incorporada alcanza un nodo v con frecuencia $s_v(t)$, el nodo v genera un nuevo hijo con probabilidad $p(s_v(t))$. De otro modo, el juego continúa hacia uno de los hijos ya existentes de v con probabilidad $1 - p(s_v(t))$. En este caso, se utilizan dos tipos de probabilidades de crecimiento preferencial, un crecimiento preferencial no lineal de acuerdo con la Ec. (5.5), y un crecimiento preferencial lineal dado por $\pi(s|s') \propto s$. Más específicamente, cuando no ocurre un evento de ramificación, un movimiento desde un nodo v a uno de sus nodos hijos u se realiza con probabilidad:

$$\pi(s_u|s_v) = \frac{q(n_u/n_v)}{\sum_{u'} q(n_{u'}/n_v)}, \tag{5.12}$$

para el caso no lineal, y con probabilidad:

$$\pi(s_u|s_v) = \frac{n_u}{n_v}, \tag{5.13}$$

para el caso lineal. Después que las partidas atraviesan el árbol, las frecuencias de los nodos en el camino correspondiente aumentan en uno. En las simulaciones se eligió el valor para el parámetro ν que provee la mejor predicción para $B = 0{,}85$, a fin de reproducir el caso empírico. Para el caso del crecimiento preferencial no lineal,

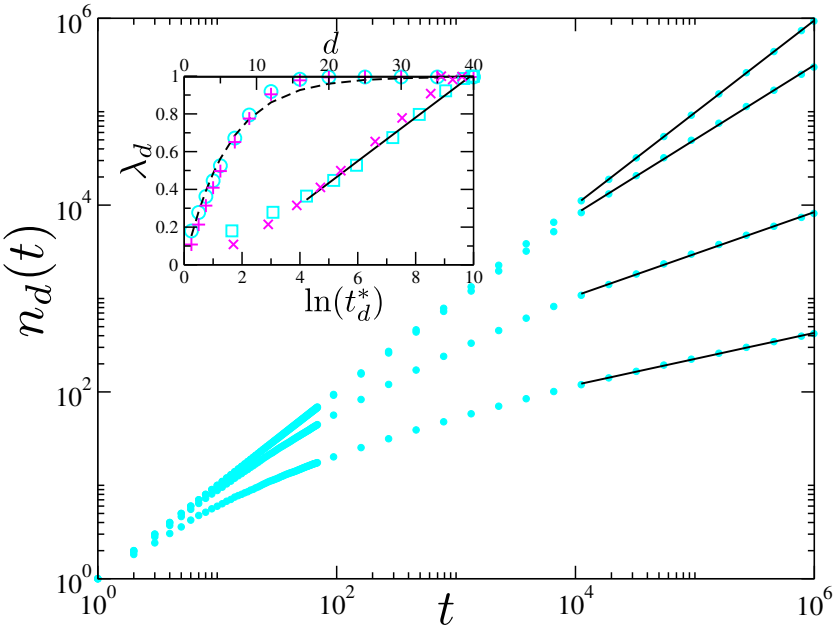

Figura 5.3.: Resultados de las simulaciones para los modelos donde se implementa el mecanismo de crecimiento preferencial anidado. Se grafica, para el modelo con probabilidad de crecimiento preferencial lineal, el ancho del árbol, n_d, en función del número de partidas, y profundidades $d = 2, 4, 9$ y 16. Los datos están 'binneados' logaritmicamente para $t > 100$. Las líneas solidas negras son ajustes correspondientes a $n_d = t_d^*(t/t_d^*)t^{\lambda_d}$. Inset: Valores estimados de λ_d en función de d para una probabilidad de crecimiento preferencial no lineal (lineal) graficados con signos más color magenta (círculos cian), y en función de los valores estimados de $\ln(t_d^*)$ con cruces color magenta (cuadrados cian). La línea negra a trazos corresponde a $\lambda_d = 1 - B^d$ con $B = 0{,}846 \pm 0{,}005$ ($R^2 = 0{,}98$) y la línea negra sólida corresponde a $\lambda_d = a + (1/A)\ln(t_d^*)$, donde $a = -0{,}15 \pm 0{,}02$ y $1/A = 0{,}116 \pm 0{,}003$ ($R^2 = 0{,}996$). Ambos son ajustes para el caso de crecimiento preferencial lineal.

la mejor predicción de B ocurre usando $\nu = 0{,}95$, y en el caso lineal usando $\nu = 0{,}85$ (Figura 5.3, inset, signos más y círculos respectivamente). En ambos casos la aproximación $B \simeq \nu$ es válida. Más aún, el comportamiento de escala entre el punto de cruce t_d^* y λ_d reproduce apropiadamente el caso empírico (Figura 5.3, inset, signos cruces y cuadrados respectivamente). En ausencia de preferencia en la elección de los nodos, en otras palabras el caso anidado no preferencial, i.e. $\pi(s|s')$ independiente de s, los resultados de las simulaciones se desvían significativamente del caso empírico.

El mecanismo propuesto resulta robusto ante variaciones en los detalles. Por ejemplo, si se reemplaza la Ec. (5.4) por un proceso de fragmentación [13] o si se introduce ruido en la selección de los nodos hijos [72], se llega a los mismos resultados.

5.3 Conclusiones parciales

La naturaleza autosimilar del árbol de partidas y el mecanismo que lo genera implica una ausencia de escalas típicas en el fenómeno de innovación en el Ajedrez. La Ec. (5.4) indica que no existen vértices con frecuencias o popularidades particularmente altas luego de los cuales la innovación se vuelve imposible. En otras palabras, en el Ajedrez no existen estrategias ganadoras, y siempre hay una posibilidad para introducir soluciones innovadoras.

Los resultados presentados en este capítulo muestran similitudes con el crecimiento del vocabulario en la evolución de los lenguajes. La intersección en la ley de Heaps en la Ec. (5.3) tiene una interpretación directa en el contexto del crecimiento de vocabulario. De acuerdo con Gerlach y Altmann [38] dicha intersección tiene origen en la existencia de dos tipos de palabras, palabras núcleo y no-núcleo, dado por una separación de las escalas temporales en el proceso de evolución del lenguaje. Esto sugiere la existencia de secuencias de movimientos núcleo y no-núcleo en el árbol de partidas. Sin embargo, se debe notar que la base de datos no contiene partidas jugadas en el principio del desarrollo del juego, sólo a partir del año 1998. Por lo tanto, el crecimiento inicial lineal de $n_d(t)$ puede ser consecuencia de innovaciones no realistas debidas a fluctuaciones aleatorias para valores pequeños de t. No obstante, en las simulaciones realizadas (Figura 5.1) no se implementó un comienzo retrasado en el proceso de medición del crecimiento del árbol, y aún así exhiben una transición entre el punto de t_d^* y λ_d (Ec. (5.3)). Para el comportamiento a tiempos largos, se ha encontrado que el exponente de Heaps es mayor en lenguajes con un grado mayor de inflexiones, donde, a partir de palabras raíz, se generan muchas otras palabras mediante declinaciones y conjugaciones [72]. Esto es consistente con los resultados presentados en este capítulo si se realiza una analogía entre el grado de inflexión y la profundidad en el árbol, ya que λ_d crece con d. En resumen, aquí se provee de más evidencia sobre el origen de exponentes de Zipf no-universales y sobre el origen de la ley de Heaps.

6 Memoria y correlaciones de largo alcance

En este capítulo se presentan los resultados obtenidos a partir del análisis de la existencia de correlaciones de largo alcance en secuencias de partidas de Ajedrez. Con ese fin se construyeron series temporales discretas utilizando diferentes reglas de asignación. Para asegurar la confiabilidad de los resultados se utilizaron cuatro de estas reglas. A su vez se emplearon distintas técnicas de detección de correlaciones de largo alcance, el análisis de rango reescalado y DFA, las cuales fueron introducidas en la Sección 2.3. Además, se estudia el problema de la emergencia de efectos de memoria en la base de datos a partir de la generación de secuencias temporales empleando los modelos de Yule-Simon y Cattuto (Capítulo 3).

6.1 Correlaciones de largo alcance en el Ajedrez

El primer paso del análisis consiste en generar, a partir de la base de datos de Ajedrez cronológicamente ordenada, una serie temporal discreta. La base de datos contiene aproximadamente 380 partidas por día uniformemente distribuidas en 9 años (≈ 3650 días). Las partidas que han sido jugadas en el mismo día no están cronológicamente ordenadas, y por lo tanto la resolución temporal mínima en la serie temporal estudiada es un día. Para generar las series temporales $X(t)$ se emplean varias reglas de asignación, que no deben introducir correlaciones espurias [32, 35]. Se utilizaron dos reglas de asignación:

- **SAR** (Regla de Asignación de Similitud): Dado un subconjunto de $\tau_{SAR}+1$ partidas consecutivas, se evalúa el grado de similitud de la última de ellas con el resto. Se elige comparar la partida $t-$ésima, la más reciente, con el resto de las partidas utilizando la siguiente expresión,

$$X_{SAR}(t) = \sum_{t'=t-\tau_{SAR}}^{t-1} S_{\tilde{d}}(t,t'), \tag{6.1}$$

  donde $S_d(t,t') = d_{eq} \in \{0,1,2,3,...,\tilde{d}\}$ si, y solo si, la $t-$ésima partida, de profundidad d, y la $t'-$ésima, de profundidad d', son idénticas hasta el movimiento d_{eq} incluido, donde $\tilde{d} = \min(d,d')$ es el número máximo de movimiento a ser

considerados. En la base de datos el valor medio de las longitudes de las partidas es ≈ 60 (5.3), y la partida más larga corresponde a $d = 400$. De acuerdo con la Ec. (6.1), cuando $\tau_{SAR} = 1$, $X_{SAR}(t) = S_{\tilde{d}}(t, t-1)$, lo que significa que cada punto de la serie temporal utiliza dos partidas consecutivas. Por ejemplo, si la secuencia de movimientos de dos partidas a tiempos t y $t-1$ son, en notación algebraica, **1 e4 e5 2 ♘f3 ♘c6 3 ♗b5 a6**... y **1 e4 e5 2 ♘f3 ♘c6 3 d3 d6**..., respectivamente, entonces $X_{SAR}(t) = 4$, ya que estas dos partidas tiene en común los primeros cuatro movimientos.

- **PAR** (Regla de Asignación de Popularidad): Cada elemento de $X_{PAR}(t)$ de la serie temporal corresponde a la popularidad a profundidad d de la $t-$ésima línea de juego sobre toda la base de datos. Esta regla de asignación es similar a la introducida por Montemurro y Pury [32] para el estudio de cuerpos literarios. Los juegos que evolucionan de la misma forma hasta profundidad d determina un subconjunto de secuencias de movimientos con la misma popularidad. Se recuerda que la popularidad de una dada partida es igual al número de elementos del subconjunto al que pertenece. Esta regla de asignación puede no resultar apropiada para ciertos sistemas, ya que grandes fluctuaciones en la serie temporal pueden llevar a efectos de memoria de largo alcance espurios [73].

Las reglas de asignación SAR y PAR son aproximadamente equivalentes, ya que las partidas con secuencias de movimientos populares tienen mayor probabilidad de coincidir con otras partidas. Las partidas populares están relacionadas a líneas de juego, las cuales alcanzan distintas profundidades. Del sistema de clasificación de aperturas de ajedrez se sabe que las líneas de juego se ramifican a diferentes profundidades dependiendo del sistema de apertura. Sin embargo, en una caracterización global [13] el árbol de partidas de ajedrez es auto-similar hasta cierto grado, lo que significa que la topología del árbol es casi independiente de la profundidad.

Para analizar la presencia de correlaciones de largo alcance, se midió el exponente de Hurst (H) de las series temporales generadas por las reglas de asignación previamente mencionadas. El exponente de Hurst se calculó utilizando el método de DFA lineal (ver Sección 2.3.2). Los paneles superiores de la Figura 6.1 muestran fragmentos de las series temporales obtenidas utilizando las reglas de asignación SAR y PAR. En los paneles de la derecha se muestra la asignación PAR para una profundidad $d = 4$ y en los paneles de la izquierda la asignación SAR con $\tau_{SAR} = 1$ en la Ec. (6.1). Estos parámetros son apropiados para comparar las dos reglas de asignación mencionadas. Como se mostrará, el valor medio de la serie SAR (con $\tau_{SAR} = 1$) es pequeña, lo que significa que la asignación SAR examina los primeros movimientos de las secuencias de las partidas. *Por lo tanto es conveniente utilizar $d = 4$ en el caso de la regla PAR, ya que las dos reglas de asignación resultan comparables*. Por otro lado, se ha comprobado que el exponente de

Hurst es aproximadamente independiente de d hasta $d = 20$ en la asignación PAR. Para $d > 20$ las correlaciones se ven afectadas, probablemente a causa de que el exponente de la ley de potencia que ajusta la distribución de popularidades es mayor a 2, por lo tanto las fluctuaciones en la serie generada con PAR se tornan relevantes [13]. En el caso de la regla SAR el exponente de Hurst es independiente de τ_{SAR} para $\tau_{SAR} \leq 500$, más allá de dicho valor el valor de H disminuye, y las correlaciones de largo alcance desaparecen. Se puede ver que en el caso de la regla SAR, aumentar τ_{SAR} implica promediar sobre tiempos cortos. Como se mencionó previamente, los puntos de la serie obtenidos por la asignación PAR están distribuidos de acuerdo a una distribución libre de escala. En el inset correspondiente se muestra esta distribución, el exponente obtenido por el ajuste de la ley libre de escala es el mismo que el obtenido por Blasius y Tönjes [13], tal como se mencionó en la Sección 4. La serie obtenida con la regla SAR muestra valores pequeños, ya que la mayor parte de las coincidencias entre partidas están restringidas a los primeros movimientos, de hecho la distribución es bien ajustada por una función exponencial de rápida caída (Figura 6.1(a), inset). Esto significa que la mayor parte de las partidas consecutivas son similares en sus primeros movimientos y pocos de ellos coinciden más allá de $d = 10$. En los paneles inferiores de la Figura 6.1 se muestran las series acumuladas completas para los casos de SAR (izquierda) y PAR (derecha). Se puede apreciar persistencia a partir del comportamiento monótono de la curva sobre tiempos largos.

En la Figura 6.2 se muestran los datos correspondientes a los análisis R/S (panel superior) y DFA (paneles medio e inferior) para las series temporales generadas mediante las reglas SAR y PAR. En todos los casos es posible ajustar una línea recta en un gráfico log-log, como puede observarse. El exponente de Hurst obtenido por ambos análisis son mayores a $H = 0{,}5$ y similares entre sí; sin embargo los exponentes resultan mayores para el caso de la regla PAR, $H \simeq 0{,}75$, comparado con la serie SAR $\simeq 0{,}67$. De acuerdo con estos valores, todas las series generadas presentan correlaciones de largo alcance. Ya que SAR es una regla de asignación local y PAR esta basada en una caracterización global de la base de datos, la serie SAR debería ser más ruidosa que la serie PAR. De acuerdo a esto, se espera que el exponente de Hurst de la serie SAR sea menor que en el caso de PAR. El hecho que ambos métodos, R/S y DFA, arrojen valores muy similares indica que las no-estacionalidades son despreciables en las series temporales. Cuando todas las series son aleatoriamente mezcladas (Figura 6.2, panel inferior), el exponente de Hurst en ambos análisis resulta cercano a $0{,}5$, lo que corresponde a series sin correlaciones de largo alcance. Esto significa que las reglas de asignación utilizadas no introducen correlaciones artificiales.

A fin de comparar las correlaciones obtenidas de distintos subconjuntos de la serie temporal completa se estudió la dependencia del exponente de Hurst H con la longitud de la serie. En el panel izquierdo de la Figura 6.3 se muestra el exponente de Hurst como función de la longitud N de las series temporales generadas con las reglas de asignación

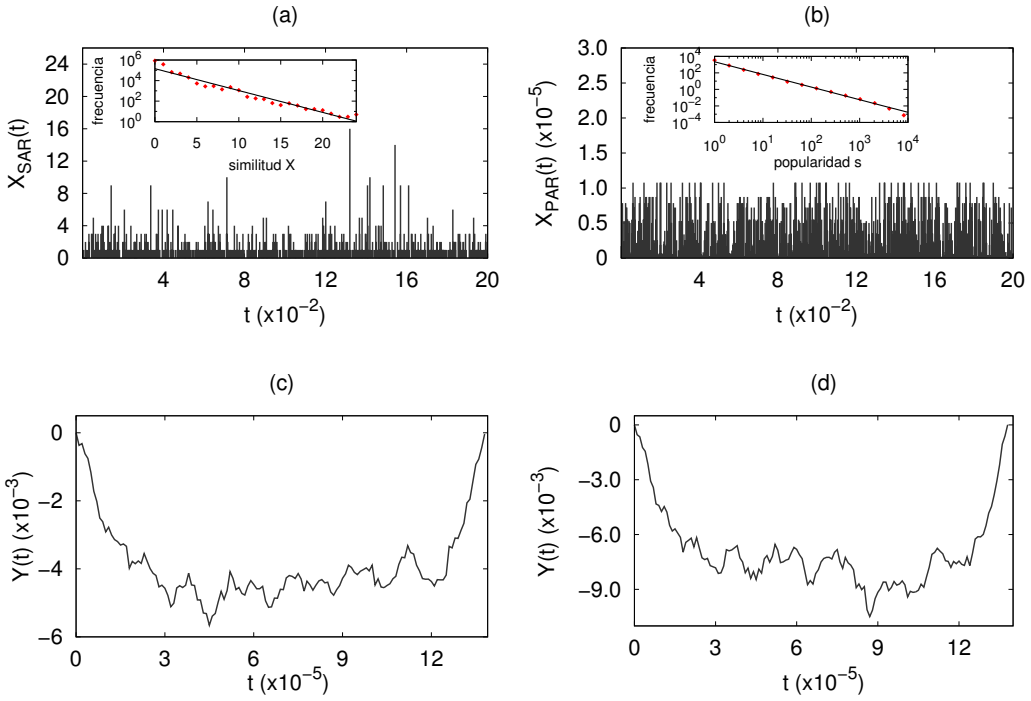

Figura 6.1.: Los paneles superiores muestran fragmentos de las series temporales $X(t)$ (2000 puntos) obtenidos utilizando SAR (a) y PAR (b). Los paneles inferiores muestran la serie integrada $Y(t)$ completa, SAR (c) y PAR (d). Insets: distribución de similitudes (a) y popularidad (b). Las líneas rectas corresponden a ajustes utilizando: $f(x) \sim e^{-\alpha x}$ y $g(x) \sim x^{-\beta}$, (a) y (b) respectivamente, donde $\alpha = 0{,}49$ y $\beta = 1{,}53$.

SAR y PAR. En el análisis se utilizó el método DFA, y se promedió sobre todos los posibles subconjuntos de partidas consecutivas de longitud N. Se puede observar una cierta correlación entre los valores de H de las asignaciones. El exponente H de la serie PAR es mayor que en las otras series. En los casos de las series SAR y PAR, el valor de H crece cuando se aumenta la longitud de la serie, alcanzando un valor máximo en aproximadamente 2×10^5 y luego se estabilizan alrededor de $H \simeq 0{,}70$ en la serie PAR y $H \simeq 0{,}65$ en la serie SAR. Para las series aleatoriamente mezcladas (Figura 6.3, panel derecho) no se observan efectos de tamaño y H fluctúa alrededor de $0{,}5$ en todos los casos, como se espera. Cabe mencionar que en este análisis, un cambio en el tamaño de la serie es equivalente a un cambio en la extensión total del tiempo real.

Ya que el nivel de los jugadores de Ajedrez puede ser clasificado de acuerdo a su Elo,que es un sistema de puntuación introducido por el físico Apard Elo [74] (ver Apéndice A.3), se repitió el análisis filtrando la base de datos de acuerdo al rango de Elo de los jugadores. Se utilizaron los siguientes rangos de Elo [16]: $[1, 2199]$, $[2200, 2399]$ y superior a 2400. Estos intervalos corresponden aproximadamente a: no expertos y candidatos a maestros, maestros y grandes maestros, respectivamente. En particular, esta partición también asegura que cada intervalo contiene aproximadamente la misma cantidad de partidas. En

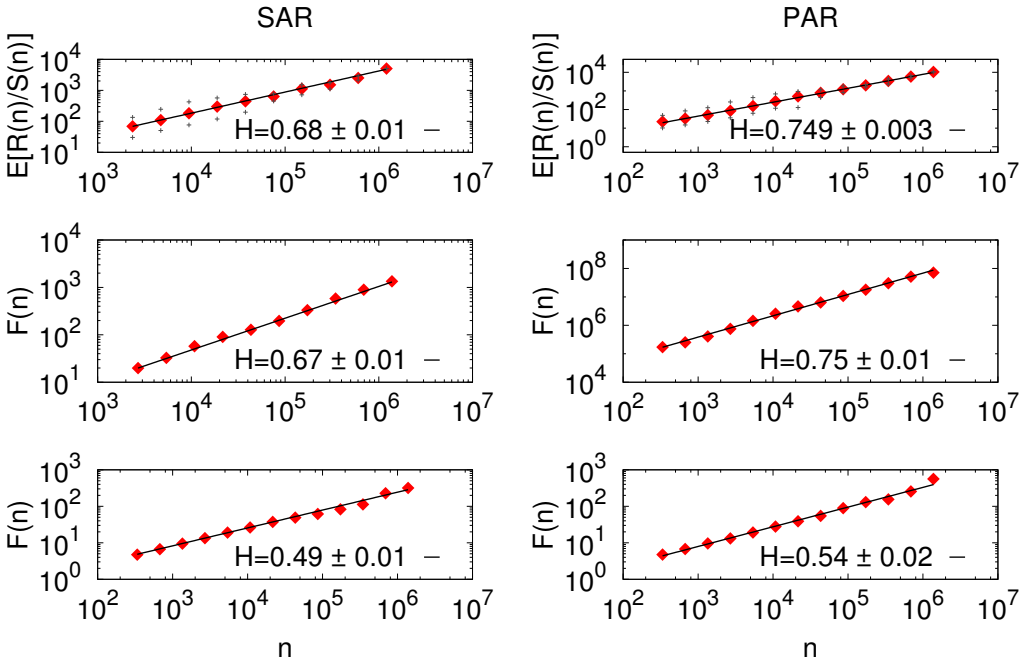

Figura 6.2.: Análisis R/S (panel superior) y análisis DFA (paneles medio e inferior) para las asignaciones SAR y PAR. Las líneas rectas corresponden a ajustes lineales de los datos en log-log.

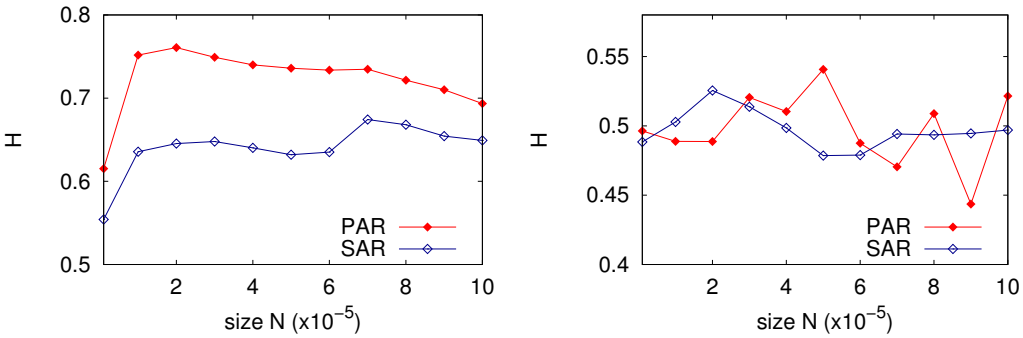

Figura 6.3.: Panel izquierdo: Exponente de Hurst obtenido con el método de DFA como función de la longitud de la serie usando las reglas de asignación SAR y PAR. Panel derecho: Exponente de Hurst obtenido a partir de las series temporales 'shuffled'.

la Figura 6.4 se muestra el exponente de Hurst como función de la longitud de la serie temporal asociada a los intervalos de Elo definidos anteriormente. En el panel superior de la figura se muestra la distribución de Elo de la base de datos. Como se espera, la distribución es bien ajustada por una función Gaussiana $\exp\left(\frac{-(x-x_0)^2}{2a^2}\right)$ (línea continua), con valores de ajuste $x_0 = 2303$ y $a = 203$. En la izquierda se muestra la serie SAR; y en la derecha, la serie PAR. Al igual que para la base de datos completa, el comportamiento de H para las dos reglas de asignación son similares, y los exponentes de la serie generada con la asignación PAR son levemente mayores que para la serie generada con la regla

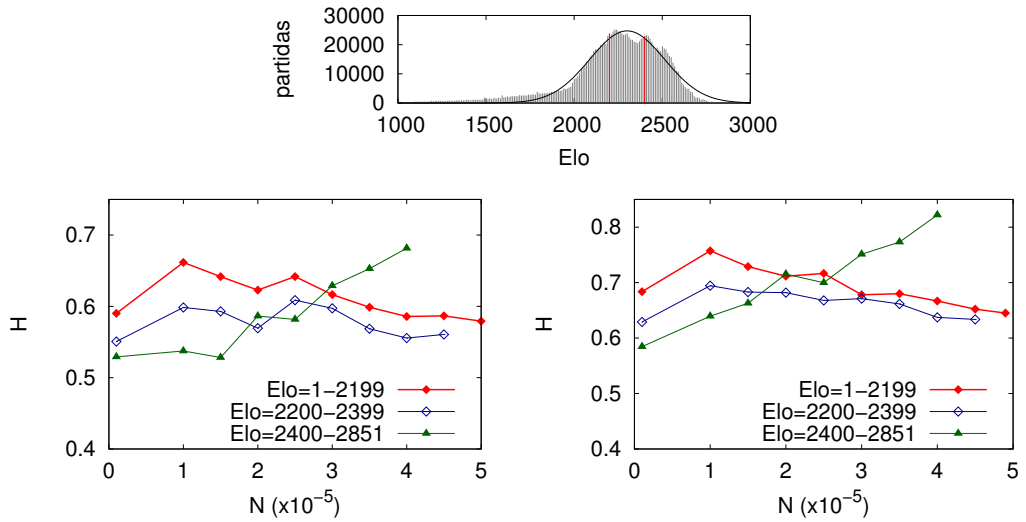

Figura 6.4.: Exponentes de Hurst obtenidos por el método de DFA computado en diferentes intervalos de Elo como función de la longitud de la serie, para las asignaciones SAR (panel inferior izquierdo) y PAR (panel inferior derecho). Panel superior: Distribución de Elo de la base de datos completa, las líneas verticales rojas indican los intervalos de Elo. La línea continua corresponde a un ajuste a una función Gaussiana, $f(x) = \exp\left(\frac{-(x-x_0)^2}{2a^2}\right)$, con $x_0 = 2303$ y $a = 203$.

SAR. A tiempos cortos los exponentes de Hurst en los tres intervalos de Elo es cercano a $0{,}5$, i.e. en esta escala temporal las partidas no están correlacionadas, y no es posible detectar persistencia. En tiempos intermedios, los intervalos de Elo menores muestran correlaciones más fuertes; y a tiempos largos, el orden se revierte y las partidas de jugadores de mayor nivel son los que muestran más correlación en sus líneas de juego.

6.2 Modelos de Ley de Zipf

Ajuste de los parámetros de los modelos

Se comienza por ajustar los parámetros de los modelos presentados en el capítulo 3, a fin de poder reproducir algunas propiedades básicas estadísticas de la base de datos. Los parámetros que se deben ajustar son: p en el caso del modelo de Yule-Simon, y p y κ_c en el modelo de Cattuto. Se generan series temporales artificiales de $N = 10^6$ elementos al aplicar las reglas de evolución de ambos modelos una cantidad N de veces.

El valor apropiado para el valor del parámetro p, que es la tasa de introducción de nuevas líneas de juego, puede ser directamente estimado de la base de datos utilizando la relación,

$$p \approx \frac{n_d(t_{total})}{t_{total}}. \tag{6.2}$$

Esta estimación es válida sólo como primera aproximación ya que implícitamente se asume que p es una función constante de t, cuando en realidad es una función del tiempo, ya que el número de líneas de juego diferentes crece con el tiempo de acuerdo a la ley de Heaps [17] (sección 5.1) y no linealmente como en la Ec. (6.2). Sin embargo, a fin de mantener el análisis simple, se escoge trabajar dentro de la aproximación de p constante, como en el caso de los modelos de Yule-Simon y Cattuto. Para el caso de $d = 4$ el valor estimado es $p = 0{,}005$. Se debe mencionar que para valores mayores de d la aproximación de p constante no es apropiada [17] ya que el valor de p crece rápidamente a medida que d aumenta (Figura 6.5).

A fin de obtener un valor apropiado para el parámetro κ_c –un parámetro solo del modelo de Cattuto– se varía κ_c hasta que el modelo de Cattuto sea capaz de reproducir el tiempo entre eventos promedio $\left\langle \tau^{(g^*)} \right\rangle$ de la línea de juego más popular, g^*, en la base de datos a profundidad $d = 4$. Entonces, con p dado por la Ec. (6.2), la mejor aproximación del modelo de Cattuto ocurre para $\kappa_c = 96$, valor para el cual el modelo de Cattuto arroja el valor $\left\langle \tau^{(g^*)} \right\rangle = 12{,}41$, y en la base de datos se obtiene $\left\langle \tau^{(g^*)} \right\rangle = 12{,}82$. Aún más, la línea de juego más popular generada por el modelo de Cattuto representa el $8{,}1\,\%$ de todas las líneas de juego, valor cercano al empírico, el cual es $7{,}8\,\%$.

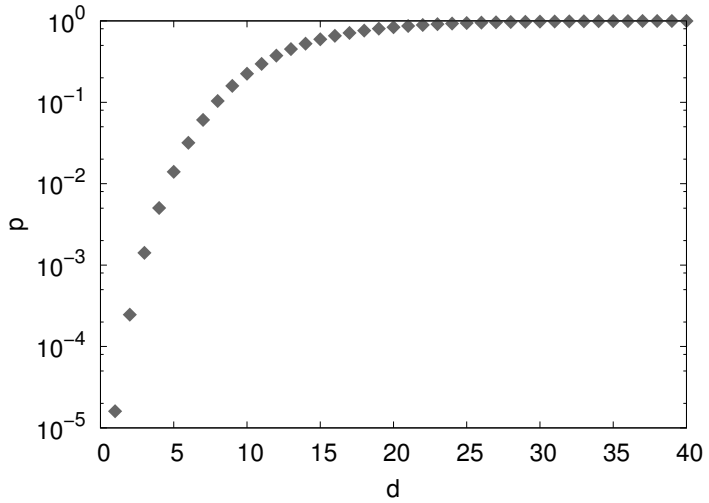

Figura 6.5.: Valor de p calculado en la base de datos de Ajedrez calculado con la Ec. (6.2) en función de d.

El modelo de Yule-Simon también provee una predicción para $\left\langle \tau^{(g^*)} \right\rangle$. Sin embargo, cuando se calcula p con la Ec. (6.2), la predicción resulta $\left\langle \tau^{(g^*)} \right\rangle = 7{,}68$, un valor considerablemente menor que el medido en la base de datos. La predicción correcta puede

Tabla 6.1.

Data	p	$\left\langle \tau^{(g^*)} \right\rangle \approx \tau_P$
Base de datos	$0{,}005$	$12{,}82$
Cattuto	$0{,}005$	$12{,}41$
Yule-Simon	$0{,}005$	$7{,}68$
Yule-Simon	$0{,}1$	$14{,}12$

ser obtenida de todas formas si se toma $p = 0{,}1$, el cual es un valor considerablemente mayor al obtenido con la Ec. (6.2). En otras palabras, el modelo de Yule-Simon no es capaz de ajustar simultáneamente los valores empíricos de p y $\left\langle \tau^{(g^*)} \right\rangle$, mientras que el modelo de Cattuto si es capaz. Esto es esperable, ya que el modelo de Cattuto posee un parámetro, o grado de libertad, extra que puede ser ajustado.

Para simplificar la comparación, en la Tabla 6.1 se resumen los valores obtenidos de $\left\langle \tau^{(g^*)} \right\rangle$ para diferentes valores de p en la base de datos y en los modelos.

Comparación de los modelos

Luego de ajustar los parámetros de los modelos, p y κ_c, se comparan los modelos con algunas propiedades estadísticas complementarias medidas en la base de datos de ajedrez, tales como las antes mencionadas distribución de popularidades y los efectos de memoria de largo alcance. En lo que continúa, los parámetros de los modelos son fijados en los valores obtenidos en la sub-sección previa.

En la Figura 6.6 se muestran las distribuciones de popularidades $P(s)$ de los modelos de Yule-Simon y Cattuto, y la base de datos para líneas de juego de profundidad $d = 4$. El modelo de Yule-Simon produce una distribución de ley de potencia con exponente muy cercano a 2, lo que es esperado en este proceso para valores pequeños de $p(= 0{,}005)$ [30]. La distribución obtenida a partir del modelo de Cattuto muestra una leve curvatura, y es muy bien ajustada por la expresión teórica de la Ec. (3.15). La distribución $P(s)$ de la base de datos es mucho más cercana a la obtenida con el modelo de Cattuto que a la obtenida con el modelo de Yule-Simon. Es posible obtener una distribución de popularidades similar con el modelo de Yule-Simon solo si se relaja la restricción en la cual el valor de p está determinado por la Ec. (6.2).

A fin de analizar la presencia de correlaciones de largo alcance se midió el exponente de Hurst (H) de las series temporales generadas con los modelos y de los datos empíricos con el método de DFA. Como la regla de asignación SAR está basada en la comparación de

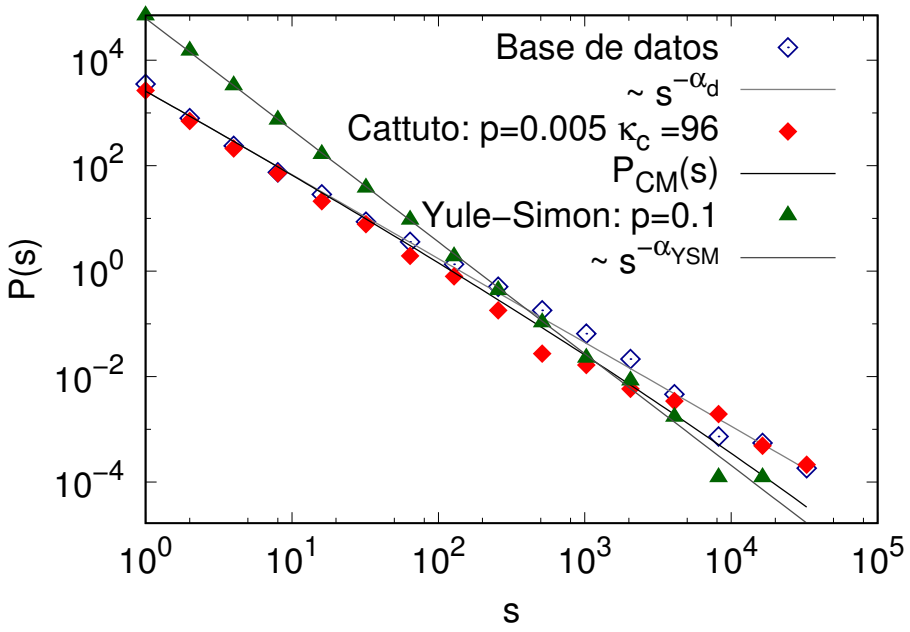

Figura 6.6.: Gráfico log-log de la distribución de popularidades: medido en la base de datos para $d = 4$ (diamantes azules), ajustado con $P(s) \sim s^{-\alpha}$ con $\alpha_d = 1{,}59 \pm 0{,}02$ ($R^2 = 0{,}997$) (línea gris clara a rayas); medido en la base generada con el modelo de Cattuto, $p = 0{,}005$ y $\kappa_c = 96$ (diamantes rojos) ajustada con $P_{CM}(s)$ (ver Ec. (3.15)) con parámetro $\Omega = 1{,}5 \pm 0{,}3$ (línea negra); y generada con el modelo de Yule-Simon, $p = 0{,}1$ (triángulos verdes) y ajustado con $P(s)$ y exponente $\alpha_{YS} = 2{,}12 \pm 0{,}03$ ($R^2 = 0{,}997$) (línea gris oscura a rayas y puntos).

las secuencias de movimientos consecutivas, no se puede utilizar esta regla de asignación en las series generadas por los modelos. Las series temporales son entonces obtenidas utilizando las tres asignaciones: PAR, GAR y UAR, y el exponente de Hurst es calculado con el método DFA lineal. Las reglas GAR y UAR funcionan de la siguiente forma:

- **GAR** (Regla de Asignación Gaussiana) y **UAR** (Regla de Asignación Uniforme): Ambas son reglas de asignación aleatorias, donde se asigna a las distintas líneas de juego a una dada profundidad d, $g_d(t)$, un número aleatorio $Y_{GAR(UAR)}$ tomado de una función distribución de probabilidad, Gaussiana para GAR y uniforme para UAR, y de esta forma la serie temporal es $Y_{GAR(UAR)} = X_{g_d(t)}$. Se nota que esta regla de asignación está basada en una caracterización global de la base de datos.

Nuevamente, los parámetros de los modelos son aquellos obtenidos en la sección 6.2. Se recuerda que, como se mencionó en la Sección 6.1, las series temporales construídas a partir de la base de datos presenta correlaciones de largo alcance para profundidades menores y mayores a $d = 4$ [18].

Consecuentemente con la falta de memoria del modelo de Yule-Simon, el exponente de Hurst correspondiente a las series generadas con el modelo de Yule-Simon es cercano

a $0{,}5$, este resultado es independiente de p y de la regla de asignación empleada (Figura 6.7).

En la Figura 6.7(a) se muestra el exponente de Hurst como función de la longitud de la serie temporal, utilizando la regla de asignación PAR en la base generada con el modelo de Cattuto y en la base de datos empírica. Las series temporales generadas con el modelo de Cattuto exhiben correlaciones de largo alcance y efectos de tamaño, con un comportamiento similar al de la base de datos. El valor de H crece hasta alcanzar un valor de $0{,}69$ en la base de datos, y hasta un valor similar $(0{,}65)$ en el caso del modelo de Cattuto. La tendencia es diferente en ambos casos, mientras que en la base de datos el exponente de Hurst alcanza un valor estacionario a tiempos cortos, crece regularmente en el modelo de Cattuto.

Grandes fluctuaciones en la serie temporal $X(t)$ pueden introducir efectos de memoria de largo alcance espurias, i.e. valores de H significativamente distintos de $0{,}5$. Como la distribución de popularidades es de cola larga –en el modelo y en la base de datos empírica– la regla de la asignación PAR lleva a grandes fluctuaciones en los valores de $X(t)$. Es conveniente entonces analizar el caso en que las fluctuaciones en los valores de las series temporales $X(t)$ están acotadas. Para tal propósito, se utilizan otras reglas de asignación, GAR y UAR, las que generan series temporales con varianza finita.

En la Figura 6.7 se muestra el exponente de Hurst H como función del tamaño de la serie temporal analizada para las otras dos reglas de asignación; GAR (Figura 6.7(b)) y UAR (Figura 6.7(c)). Los exponente de Hurst H son mayores a $0{,}5$ en la mayor parte de los puntos de la gráfica. Además, a diferencia de las regla PAR, las series GAR y UAR crecen casi monotonamente hasta alcanzar un valor de $H \simeq 0{,}72$ (GAR) y $H \simeq 0{,}63$ (UAR) para la serie completa. Por lo tanto, la presencia de las correlaciones de largo alcance es robusta ante la elección de la regla de asignación. En particular, se obtiene un muy buen acuerdo entre la serie generada por el modelo de Cattuto y la base de datos para la regla GAR. Cabe mencionar que se ha encontrado que el método DFA tiende a ser más robusto para procesos Gaussianos [75].

Para las mediciones en la base de datos, las barras de error en la Figura 6.7 (a) resultan del ajuste lineal de $F(l)$ en el DFA, mientras que para el modelo de Cattuto se computaron 10 realizaciones del modelo, y las barras de error reflejan la dispersión de los valores calculados de H. Sin embargo, en los paneles (b) y (c), los que corresponden a las reglas GAR y UAR, los errores se estimaron utilizando 10 asignaciones aleatorias diferentes para cada regla, para el modelo de Cattuto y la base de datos. Las diferentes reglas de asignación llevan a distintos valores de H, en la base de datos empírica y en el modelo de Cattuto. Esto implica que la existencia de correlaciones de largo alcance es una característica robusta desde un punto de vista cualitativo, pero que los valores que

toma H no son independientes de la regla de asignación utilizada para generar las series temporales.

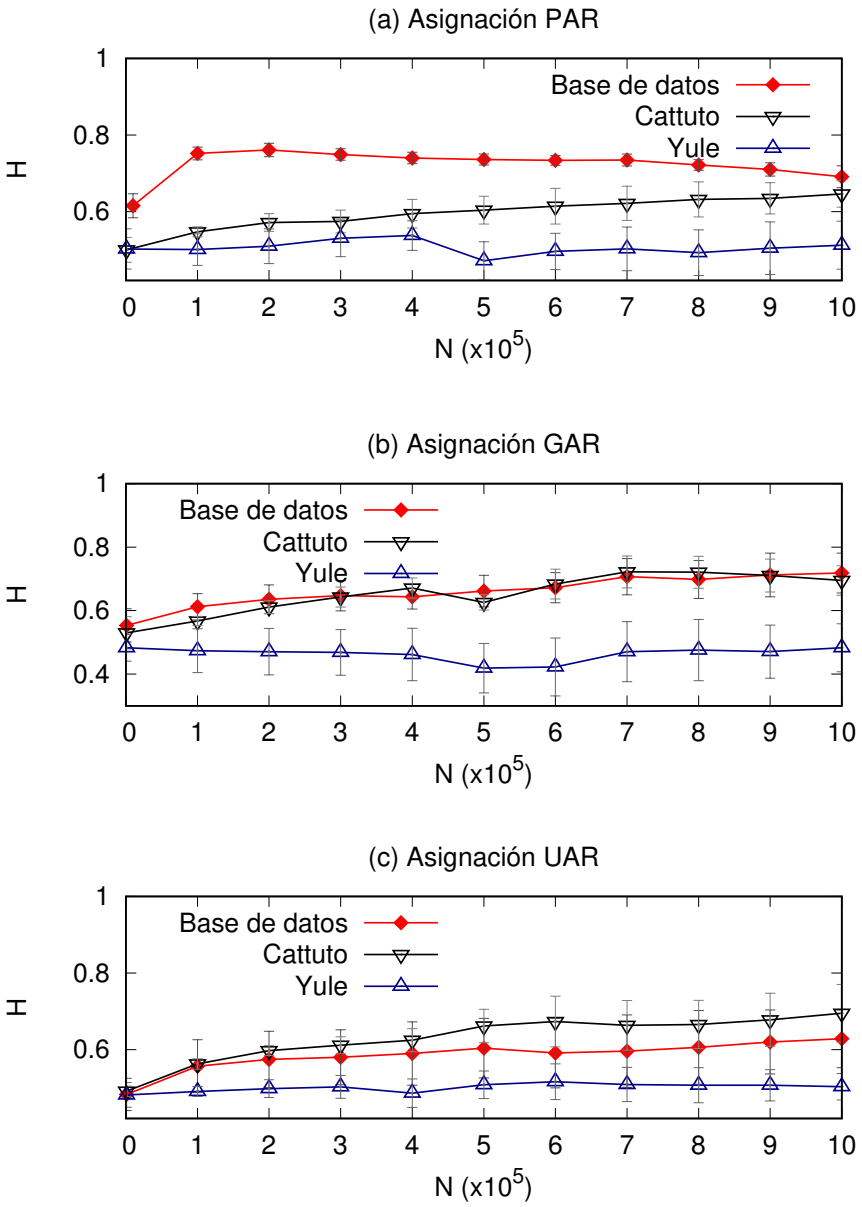

Figura 6.7.: Exponente de Hurst obtenido con el método de DFA como función de la longitud de la serie temporal en la base de datos (línea de puntos con triángulos) y generada con el modelo de Cattuto, $p = 0{,}005$ y $\kappa_c = 96$ (línea completa con diamantes rojos) utilizando las reglas de asignación: (a) PAR, (b) GAR y (c) UAR.

6.3 Conclusiones parciales

Los exponentes de Hurst obtenidos mediante los métodos R/S y DFA son similares, lo cual indica que las series temporales analizadas son estacionarias para las reglas de

asignación PAR y SAR. Para el caso de la serie generada con la regla PAR y analizada con el método R/S, los resultados obtenidos son similares a aquellos reportados para cuerpos literarios [32]. El valor $H \simeq 0{,}5$ que resulta del mezclado aleatorio de las series temporales implica que las reglas de asignación empleadas para construir dichas series no introducen correlaciones espurias. El análisis de las series en función del tiempo muestra que existe un umbral en la longitud de las series temporales para el cual las correlaciones de largo alcance comienzan a surgir. Este umbral depende de la regla de asignación y del nivel de los jugadores. Cuando la base de datos es filtrada por rangos de ELO, los resultados obtenidos indican que las correlaciones de largo alcance observadas a escalas temporales largas están relacionadas a la presencia de jugadores de alto nivel. En contraste, las partidas correspondientes a jugadores de nivel intermedio y bajo nivel muestran correlaciones a tiempos cortos.

Los efectos de tamaño en la detección de correlaciones de largo alcance han sido estudiados previamente en el trabajo realizado por Coronado et al. [76] en el cual se analizan series temporales generadas por el algoritmo de Makse [77]. En el trabajo mencionado fue observado que los efectos de tamaño resultan despreciables cuando se analizan series temporales reducidas en longitud, tan pequeñas como 10^3, con el método de DFA. En los resultados presentados en este capítulo se puede observar la presencia de efectos de tamaño en las series temporales analizadas con el método DFA, las cuales poseen longitudes mayores a 10^3 elementos, por lo tanto se puede decir que los efectos observados reflejan propiedades intrínsecas del sistema.

Para jugadores de bajo nivel e intermedio nivel la persistencia alcanza un máximo en dos o tres años y luego se estabiliza. Este comportamiento es independiente del origen temporal utilizado en el análisis. De hecho, los resultados mostrados en este capítulo fueron obtenidos al promediar sobre intervalos de tiempo equivalentes disjuntos. Para la mayor parte de los jugadores especializados las correlaciones de largo alcance comienzan a ser significativas luego de uno o dos años. De acuerdo a los resultados, los jugadores de alto nivel utilizan estrategias diferentes a los jugadores de bajo nivel. Al parecer, los jugadores varían más a menudo las líneas de juego que emplean en tiempos cortos. Esto se puede deber a que los jugadores sobresalientes conocen una mayor cantidad de líneas de apertura en profundidad y son menos influenciados por sus oponentes. En particular, en el caso de los jugadores de alto nivel las correlaciones de largo alcance en escalas temporales largas resultan más fuertes.

Por otro lado, se analizó la base de datos de ajedrez dentro del marco de dos modelos basados en un mecanismo de crecimiento preferencial, los modelos de Yule-Simon y Cattuto, ya que ambos modelos permiten generar base de datos artificiales de líneas de aperturas con distribución de ley de potencia. En particular, el modelo propuesto por Cattuto et al. incluye un núcleo de memoria a este proceso. El análisis mostrado en este capítulo demuestra que ambos modelos son capaces de reproducir, hasta cierto punto,

la distribución de popularidades obtenida a partir de la base de datos. Sin embargo, el modelo de Cattuto resulta más realista, ya que reproduce la distribución de ley de potencia de las líneas de apertura utilizando el valor de la probabilidad de introducción de una nueva línea medido en la base de datos. Más aún, debido al núcleo de memoria, el modelo de Cattuto es también capaz de reproducir las correlaciones de largo alcance y los efectos de tamaño observados en la base de datos, mientras que el modelo de Yule-Simon carece de memoria.

Específicamente, en este capítulo se muestra que el modelo de Cattuto describe correctamente la distribución de popularidades de las líneas de apertura a profundidad $d = 4$ para el valor del parámetro p medido en la base de datos, y κ_c determinado con el ajuste del tiempo entre eventos medio de la línea de apertura más popular. El modelo de Yule-Simon reproduce la distribución de popularidad, pero para un valor del parámetro p considerablemente más grande que el medido. Más aún, el modelo de Cattuto exhibe correlaciones de largo alcance y efectos de tamaño, independientemente de la regla de asignación utilizada para construir la serie temporal, aunque el grado de persistencia sí depende de la asignación. En particular, el modelo de Cattuto reproduce muy bien los efectos de tamaño y persistencia observados en la base de datos sólo para el caso de la regla de asignación Gaussiana (GAR). Esto sugiere que existen correlaciones subyacentes relacionadas con la popularidad de las líneas de juego de ajedrez y la serie temporal correspondiente, que no son capturadas por el modelo de Cattuto.

7

Análisis de los tiempos entre eventos

En este capítulo se estudia otro aspecto de la existencia de efectos de memoria no triviales, que concierne a la ocurrencia de líneas de juego específicas en la base de datos de Ajedrez. En este sentido, se analiza la existencia de *burstiness* en la ocurrencia de la línea de apertura más popular, considerando la base de datos de Ajedrez completa y el sub-conjuntos de partidas correspondientes al jugador más activo de la base. Además, se analiza la existencia de este efecto de memoria en las series temporales generadas por los modelos de Yule-Simon y Cattuto. El estudio de la existencia de una dinámica *bursty* en la secuencia de tiempo entre eventos se realiza a través del cálculo del parámetro (B) y exponente β_B de *burstiness*.

Se comienza con el estudio de *burstiness*, analizando la actividad de la línea de juego más popular tanto en la base de datos completa como en la actividad de un solo jugador. Para esto se calcula en ambos casos la distribución acumulada de tiempos entre eventos $F(\tau)$ (Sección 2.4) correspondiente a la línea de juego más popular a profundidad $d = 4$, en la base de datos y del jugador más activo. En la Figura 7.1 se muestra la gráfica de $-\ln F(\tau)$ como función de τ en doble escala logarítmica para la base de datos completa y para el jugador. Análogamente, en la Figura 7.2 se muestran los gráficos de la distribución acumulada de tiempos entre eventos de las secuencias generadas con los modelos de Cattuto y Yule-Simon. En todos casos las curvas muestran un régimen lineal permitiendo un buen ajuste de la distribución de Weibull (Ec. (2.23)). El régimen lineal se estimó eliminando puntos de ambos extremos de la curva $-\ln F(\tau)$ y realizando un ajuste lineal en cada subconjunto hasta que el valor del coeficiente de determinación R^2 alcanzase un valor estable con tolerancia 10^{-4}.

En el caso de la base de datos completa y las secuencias generadas por los modelos de Yule-Simon y Cattuto, el exponente de Weibull y el parámetro de *burstiness* resultan $\beta \approx 1$ y $B = 10^{-2}$, respectivamente, indicando ausencia de *burstiness*. Dado que no se detecta *burstiness*, el parámetro τ_0 de la distribución de Weibull debería coincidir con el tiempo característico $\tau_P \simeq \left\langle \tau^{(g)} \right\rangle$ del proceso de Poisson asociado (Sección 2.4). Esto se corrobora haciendo una aproximación derivada del modelo de Yule-Simon, aunque también es válida para el modelo de Cattuto. En el modelo de Yule-Simon, la probabilidad P_t que una línea de juego g ya existente sea repetida entre los tiempos t y $t + \delta t$, puede aproximarse por $P_t \approx ((1-p)s_d^{(g)}(t)/N_d(t))\delta t$, donde $s_d^{(g)}$ es el número de líneas de juego

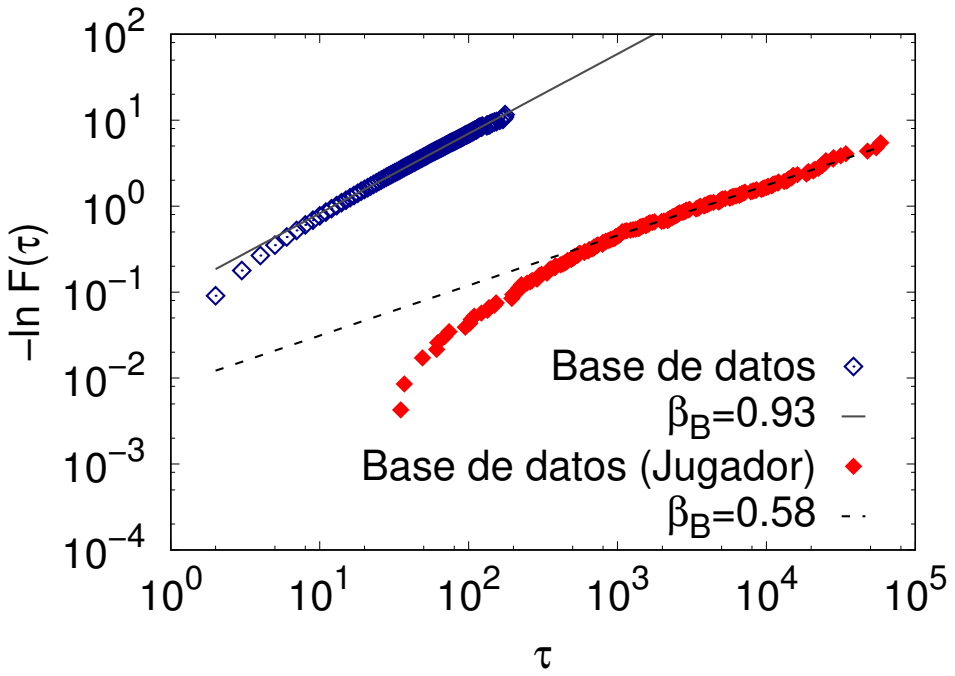

Figura 7.1.: Distribución acumulada de tiempos entre eventos de la partida más popular medida en la base de datos completa (diamantes azules) y para el jugador más activo (diamantes rojos). Las líneas corresponden a ajustes de la Ec. (2.23) con $\beta_B = 0{,}927 \pm 0{,}003$ $(R^2 = 0{,}999)$ para la base de datos completa (línea gris) y $\beta_B = 0{,}583 \pm 0{,}004$ $(R^2 = 0{,}993)$ para el jugador (línea negra rayas).

g hasta profundidad d que han sido empleadas hasta tiempo t (ver Sección 4). Luego de un tiempo transitorio, se espera que $s_d^{(g)}(t)/N_d(t)$ sea aproximadamente constante –más precisamente una cantidad de variación lenta–. Por lo tanto, $P_t = \mu\delta t$, donde $\mu := 1/\tau_P \approx (1-p)s_d^{(g)}/N_d$ es la tasa de eventos del proceso de Poisson, y

$$\tau_P \approx \frac{s_d(t_{total})}{N_d^{(g)}(t_{total})(1-p)} \tag{7.1}$$

es el correspondiente tiempo característico. Aquí, t_{total} es el número de partidas en toda la base de datos. En la Tabla 7.1 se resumen estos resultados, y es claro que hay un buen acuerdo entre el valor de ajuste τ_0 y la estimación τ_P.

De acuerdo a estos resultados el tiempo entre eventos de la línea de juego más popular indica una leve tendencia a una dinámica inhomogénea, i.e., *bursty*, cuando se considera el conjunto completo de jugadores. Con el fin de esclarecer esta tendencia se analizó la existencia de *burstiness* en la actividad de un solo jugador. Para tal análisis se escogió el jugador más activo de toda la base de datos, el cual cuenta con 1377 partidas jugadas, y se mantiene el índice de tiempo ordinal. El valor medido para la probabilidad p de introducción de una nueva línea de juego para este jugador es igual a $0{,}09$. En la Figura 7.1 se muestra la distribución de tiempos entre eventos acumulada de la línea de

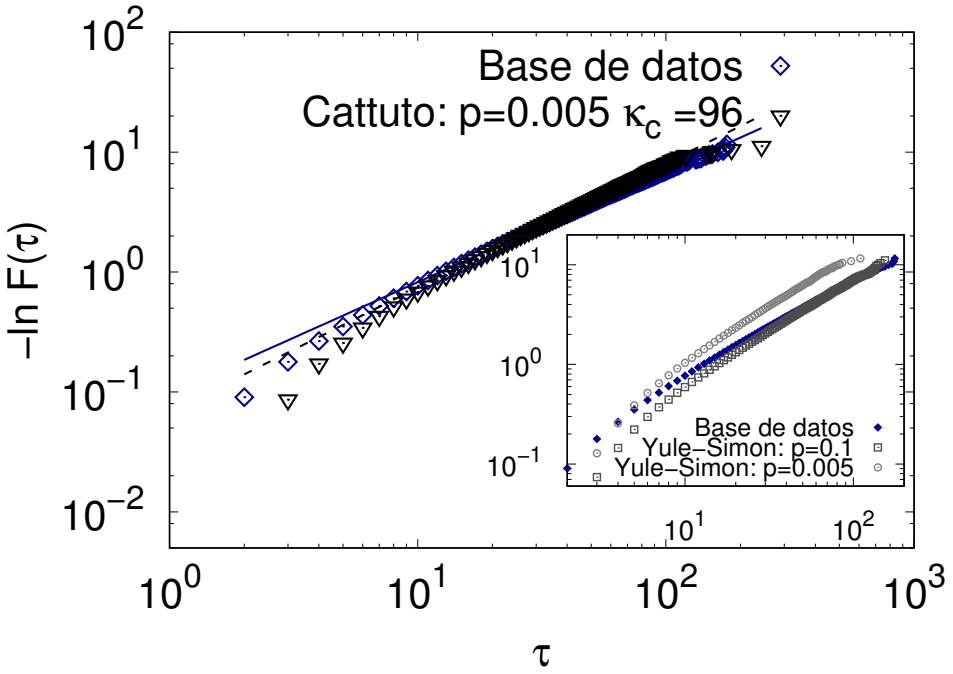

Figura 7.2.: Distribución acumulada de tiempos entre eventos de la partida más popular medida en la base de datos completa (diamantes azules) y en la serie temporal generada por el modelo de Cattuto (triángulos negros) para $p = 0{,}005$ y $\kappa_c = 96$. Las líneas corresponden a ajustes de la Ec. (2.23) con $\beta_B = 0{,}927 \pm 0{,}003$ ($R^2 = 0{,}999$) para la base de datos completa (línea azul) y $\beta_B = 1{,}035 \pm 0{,}003$ ($R^2 = 0{,}999$) para el modelo de Cattuto (línea negra a trazos). Inset: Distribución acumulada de tiempos entre eventos de la partida más popular medida en la base de datos (diamantes azules) y en las secuencias generadas por el modelo de Yule-Simon para $p = 0{,}1$ (cuadrados) y $p = 0{,}005$ (círculos).

juego más popular de este jugador. La pendiente de la recta que ajusta el régimen lineal es $\beta_B = 0{,}583 \pm 0{,}004$ y su parámetro de *burstiness* es $B = 0{,}22$, indicando la presencia de una actividad *bursty*. Más aún, los valores obtenidos para τ_0 a partir del ajuste de la Ec. (2.23) y τ_P de la Ec. (7.1) son muy diferentes (ver Tabla 7.1). Como es de esperar, un proceso de Poisson no resulta ser una buena aproximación para la distribución de tiempos entre eventos para un solo jugador. Además, como se mostró previamente, los modelos de Yule-Simon y Cattuto no generan *burstiness*, por lo tanto no son capaces de reproducir los resultados obtenidos de la dinámica de un solo jugador.

Con el fin de analizar la influencia del nivel de los jugadores en la actividad de la partida más popular se repitieron los cálculos anteriores para diferentes subconjuntos de partidas de la base de datos. En esta parte del análisis la base de datos es dividida en tres grupos, al igual que en la sub-sección 6.1, cada uno correspondiente a un intervalo de Elo: $[1, 2199]$, $[2200, 2399]$ y superior a 2400, donde todos los intervalos contienen aproximadamente la misma cantidad de partidas.

Tabla 7.1.: Resumen de los resultados correspondientes al análisis de tiempos entre eventos en la base de datos y en las series generadas por los modelos de Yule-Simon y Cattuto.

Data	p	β_B	τ_0	$\left\langle \tau^{(g^*)} \right\rangle \approx \tau_P$
Base de datos	$0{,}005$	$0{,}927 \pm 0{,}003$	$13{,}0 \pm 0{,}5$	$12{,}82$
Cattuto	$0{,}005$	$1{,}036 \pm 0{,}002$	$12{,}9 \pm 0{,}8$	$12{,}41$
Yule-Simon	$0{,}005$	$1{,}059 \pm 0{,}005$	$8{,}2 \pm 0{,}6$	$7{,}68$
Yule-Simon	$0{,}1$	$1{,}031 \pm 0{,}003$	$15{,}4 \pm 0{,}6$	$14{,}12$
Jugador individual	$0{,}09$	$0{,}583 \pm 0{,}004$	4297 ± 200	$89{,}75$

Al igual que en el análisis previo, se calculó la distribución de tiempos entre eventos acumulada para cada intervalo de Elo (Figura 7.3). Los ajustes de los correspondientes regímenes lineales resultan en $\beta_B = 0{,}89$, $\beta_B = 0{,}93$ y $\beta_B = 1{,}02$ para los intervalos $[1, 2199]$, $[2200, 2399]$ y superior a 2400 respectivamente. Cuando la serie temporal es mezclada aleatoriamente el régimen lineal de la distribución de tiempos entre eventos acumulada tiene una pendiente $\beta \approx 1$. Al parecer, existe cierta correlación entre los valores de β_B y el rango de Elo, sin embargo estas correlaciones son débiles y los resultados no son concluyentes, ya que en todos los casos los valores de β_B son relativamente cercanos a 1.

Se hace notar que la línea de juego más popular es exactamente la misma en la base de datos completa y en los dos rangos de Elo menores, mientras que difiere para el rango de Elo superior, para el cual la secuencia más popular a profundidad $d = 4$ resulta **1 e4 c5 2 ♘f3 d6**, también conocida como la Defensa Siciliana.

A fin de analizar la ausencia de *burstiness* en la base de datos completa, se estudió también el comportamiento *bursty* durante el proceso de agregado de datos. Se hizo crecer la base de datos de dos formas diferentes, por agregación de jugadores y por agregación cronológica de partidas. En el primer caso, se ordenaron los jugadores de acuerdo al número de partidas jugadas por ellos en la base de datos, de más activo a menos activo. Se hizo crecer la base de datos agregando grupos de jugadores, cada grupo conteniendo 5×10^4 partidas, hasta completar la base de datos. De esta forma, los jugadores más activos son incluidos primero durante el proceso de crecimiento. En el segundo caso, de agregación cronológica de partidas, las líneas de juego fueron agregadas en subconjuntos de 5×10^4 partidas, también hasta completar la base de datos. Para ambos procesos de crecimiento, se calculó el parámetro de *burstiness* B y el exponente de *burstiness* β_B (sección 2.4). Los valores resultantes como función del tamaño de la base de datos se muestran en la Figura 7.4. En la figura es notable que los puntos del proceso de agregación de jugadores no están uniformemente distribuidos, esto se debe a que, al agregar nuevos jugadores, muchas de las partidas jugadas por ellos ya fueron incluidos en la base de datos en crecimiento si los oponentes involucrados en las

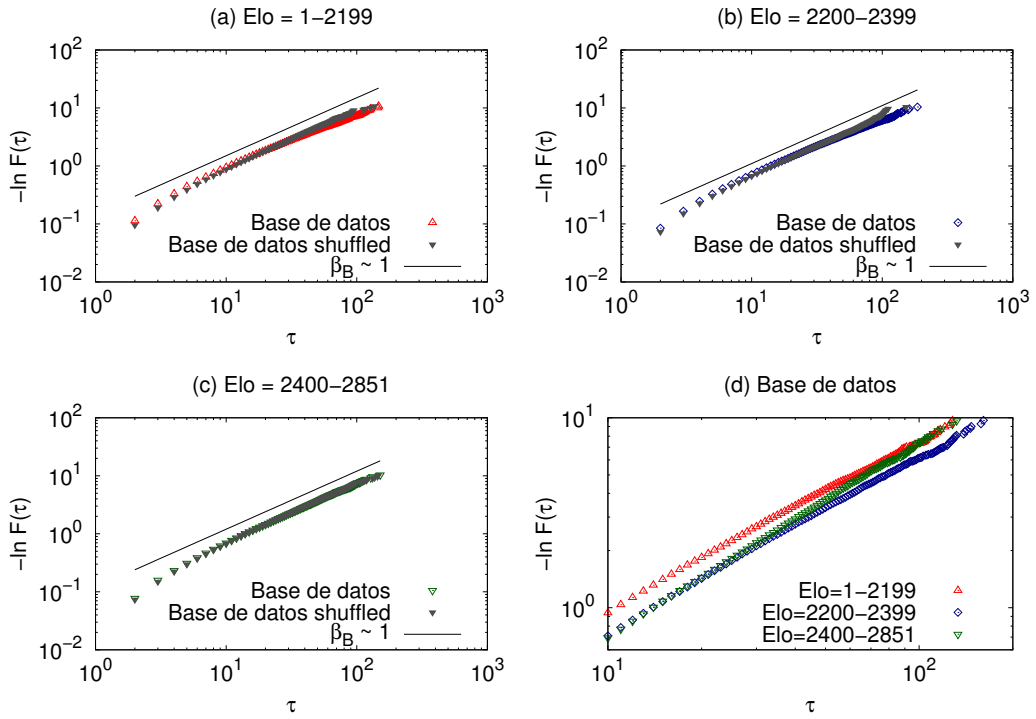

Figura 7.3.: Distribución de tiempos entre eventos acumulada para diferentes rangos de Elo: (a) rango [1,2199] (triángulos rojos), (b) rango [2200,2399] (diamantes azules), (c) rango [2400,2851] (triángulos vedes) y (d) los tres rangos superpuestos. Los puntos de colores corresponden a la distribución medida en la base de datos y los triángulos invertidos grises a la distribución de la serie aleatoriamente mezclada.

partidas ya habían sido agregados. Este efecto es más significativo cuando la base de datos alcanza un tamaño considerable.

El análisis de la Figura 7.4 revela que para ambos procesos de crecimiento, agregación de jugadores y partidas, β_B y B alcanzan un valor estable correspondiente al de la base de datos completa, lo que significa que en ambos casos el *burstiness* desaparece a un dado tamaño. Sin embargo, los valores finales son alcanzados en un número diferente de partidas. Bajo el proceso de agregación de partidas, los valores de β_B y B se estabilizan en aproximadamente 2×10^5 partidas, mientras que bajo el proceso de agregación de jugadores los valores se estabilizan mucho después. Este es un nuevo indicio de que la dinámica *bursty* en la base de datos completa esta relacionado al comportamiento de jugadores individuales. De hecho, durante la agregación cronológica de partidas, la incorporación de un número relativamente pequeño de líneas de juego (las primeras 5×10^4) ya es suficiente como para incluir una fracción considerable de los jugadores de la base de datos (alrededor del $12\,\%$ del total). Esta es la razón por la cual el comportamiento *bursty* desaparece en una escala de tiempo corta bajo la agregación de partidas. En la agregación de jugadores, en cambio, se tienen sólo 57 jugadores en el primer subconjunto de partidas agregadas, y tanto los valores del parámetro de *burstiness*

B como los del exponente de *burstiness* β_B se estabilizan cuando el número de jugadores agregados alcanza los ≈ 1000, lo que corresponde a 10^6 partidas. Las restantes 4×10^5 líneas de juego corresponden a jugadores con pocas partidas –la mayoría menos de 10–. Estos jugadores no pueden exhibir *burstiness* por sí mismos y por lo tanto eliminan el comportamiento *bursty* en la base de datos. Finalmente, se ha comprobado que todos los jugadores que poseen un registro extensivo de partidas exhiben una dinámica *bursty*.

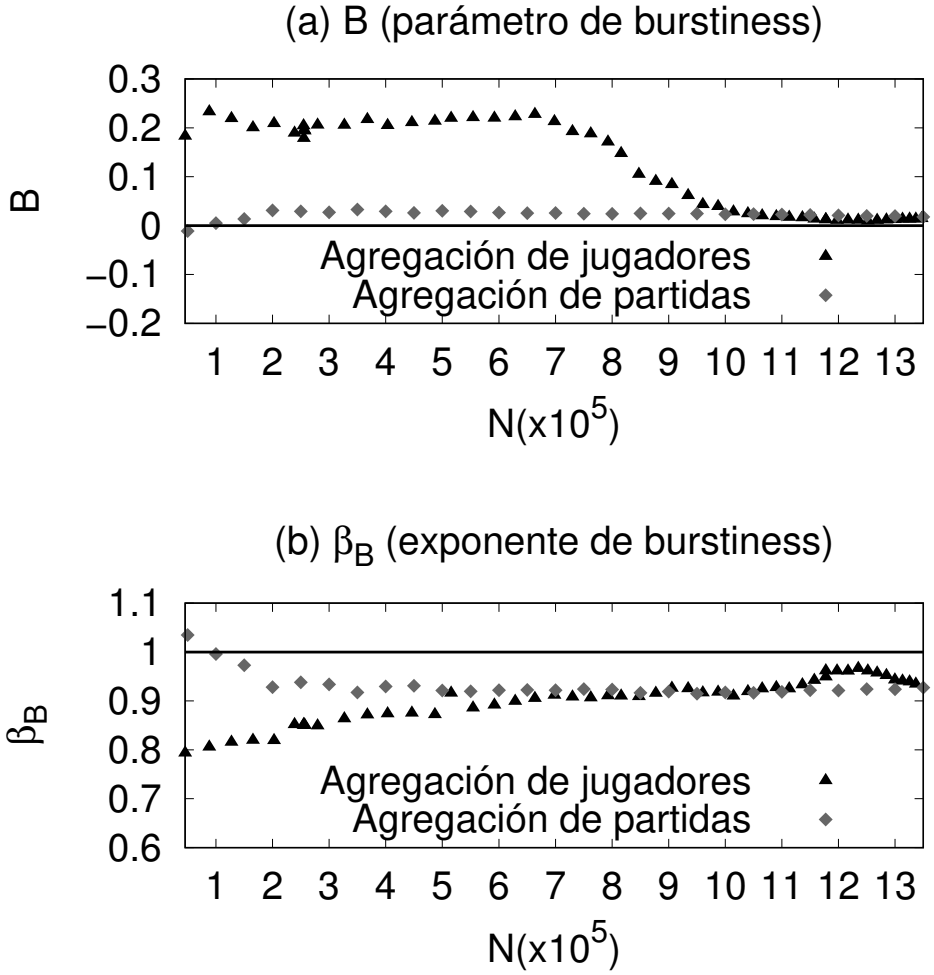

Figura 7.4.: (a) Parámetro de *burstiness* B y (b) exponente de *burstiness* β_B como función de la longitud de la serie temporal de la base de datos en crecimiento para los procesos de agregación de jugadores (triángulos negros) y agregación cronológica de partidas (diamantes grises). Las barras de error no son mostradas ya que tienen un tamaño comparable al de los puntos.

7.1 Conclusiones parciales

En este capítulo se mostraron los resultados obtenidos a partir del análisis de los tiempos entre eventos de la partida más popular, tanto en el caso de la base de datos

completa como para el jugador más activo. Este análisis resulta complementario al expuesto en el Capítulo anterior, ya que permite explorar la presencia de otro tipo de efecto de memoria, no necesariamente relacionado a las correlaciones de largo alcance, llamado *burstiness*.

A partir del ajuste de la distribución acumulada de tiempos entre eventos de la línea de apertura más popular de la base de datos, se encontró que no se puede asegurar la presencia de una dinámica *bursty*. El exponente de la distribución de Weibull resulta esencialmente igual a uno ($\beta = 0{,}927$) y, por lo tanto, la dinámica puede ser aproximada por un proceso de Poisson. Este resultado es validado a través del cálculo del parámetro de *burstiness*, el cual resulta $B \simeq 10^{-2}$. Aún más, el valor de ajuste de τ_0 y el valor calculado τ_P son muy similares. Ya que los tiempos entre eventos de las secuencias producidas por los modelos de Yule-Simon y Cattuto se pueden explicar a través de un proceso de Poisson, los tiempos entre eventos de la base de datos pueden ser reproducidos por ambos modelos. La ausencia de *burstiness* en el modelo de Cattuto sugiere que el fenómeno de *burstiness* –el cual puede explicar en algunos casos la aparición de efectos de memoria de largo alcance– juega un rol marginal en la presencia de correlaciones de largo alcance.

Aunque el comportamiento del conjunto completo de jugadores de la base de datos no exhibe *burstiness*, el análisis de los tiempos entre eventos de jugadores individuales, siempre que posean un número suficiente de partidas jugadas, sí presenta un comportamiento *bursty*. Esto indica que la ausencia de *burstiness* a nivel grupal es una consecuencia de la agregación de datos. Esto fue confirmado por el empleo de dos mecanismos de agregación de datos; agregación de jugadores y agregación de partidas, ya que el primer mecanismo preserva la dinámica *bursty* en una escala temporal considerablemente más larga que el segundo. Esto puede deberse a correlaciones subyacentes entre jugadores, e.g. los dos jugadores en una partida tienden a tener Elos similares, y por lo tanto las partidas están necesariamente correlacionadas.

Finalmente, los modelos de Yule-Simon y Cattuto no generan una dinámica *bursty*, por lo tanto el modelo de Cattuto resulta apropiado para modelar la base de datos completa pero no para analizar la dinámica de jugadores individuales.

8 Crecimiento preferencial con memoria de corto alcance

Si bien el modelo de Cattuto introduce correlaciones de largo alcance en las series generadas, las mismas no presentan *burstiness*. En este capítulo se analiza una variante del modelo de Cattuto la cual llamamos *Bounded Memory Preferential growth* (crecimiento preferencial con memoria limitada) o modelo BMPG desarrollado con el fin de producir una dinámica *bursty* en las secuencias generadas. Se caracterizaron las propiedades del modelo, tales como la distribución de popularidades, tiempos entre eventos y la dinámica dentro del núcleo de memoria, combinando análisis analítico con simulaciones numéricas extensas. En particular, esta variante del modelo presenta *burstiness* y preserva la distribución heterogénea de popularidades [78].

Se hace notar que, a pesar de la gran cantidad de trabajos en los cuales se estudian el mecanismo de crecimiento preferencial y sus variantes [79, 80, 81, 82], los que incluyen la dependencia no lineal del núcleo [83] y efectos de envejecimiento [40, 84, 85, 86], esta variante del modelo de Yule-Simon no ha sido estudiada hasta el momento.

8.1 El modelo

Como se mencionó en la sección anterior los modelos de Yule-Simon y Cattuto fallan a la hora de describir la distribución de tiempos entre eventos de las líneas de juego de jugadores individuales. Con el objetivo de generar una dinámica heterogénea, se reemplazó el núcleo de memoria de decaimiento lento en el modelo de Cattuto por un núcleo de memoria finito. El mecanismo de generación resulta idéntico al presentado en la Sección 3.2, excepto por la distribución $Q(\Delta t)$. Es decir, la diferencia reside en el núcleo de memoria, el cual está definido por una función escalón,

$$Q(\Delta t) = \begin{cases} \frac{1}{\kappa} & \Delta t \leq \kappa \\ 0 & \Delta t > \kappa \end{cases}, \tag{8.1}$$

donde κ es la extensión del núcleo. A esta variante se la llamó *Bounded Memory Preferential growth* (crecimiento preferencial con memoria limitada) o modelo BMPG. En la Figura 8.1 se ilustra el proceso del modelo BMPG.

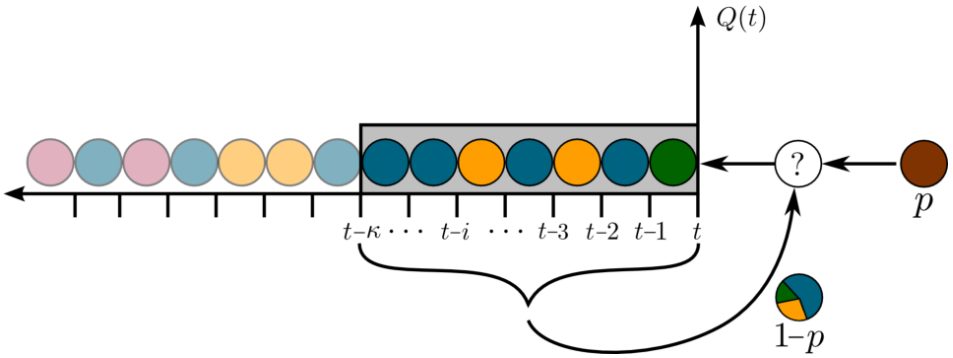

Figura 8.1.: Esquematización del proceso del modelo BMPG ilustrando como un elemento es incorporado a la serie temporal a tiempo t. Con probabilidad p el elemento que se agrega pertenece a una nueva clase, i.e., se agrega una bola de un nuevo color. Con probabilidad complementaria $1-p$ un elemento ya existente se copia, elegido uniformemente entre los últimos elementos de la serie. Los elementos que se encuentran a más de κ pasos temporales –círculos sombreadas en la figura– no pueden ser copiadas.

8.2 Ecuación maestra

Para comenzar con el análisis del modelo BMPG se describe la dinámica de los elementos dentro del núcleo de memoria. En cada paso temporal t, el último elemento en el núcleo abandona el mismo en el siguiente paso temporal. Como consecuencia de esto, una clase de elementos puede extinguirse si el elemento que abandona el núcleo es el último de su clase dentro del mismo, e.g., los círculos rosas sombreadas en la Figura 8.1. En este contexto, es interesante estudiar la dinámica de extinción de los elementos del núcleo a través del análisis de las distribuciones de tiempos de vida de las clases de elementos.

Sea $P_n(t),\ (n = 0, 1, ..., \kappa)$ la probabilidad a tiempo t que una dada clase de elementos dentro del núcleo tenga popularidad n, y $\mathbf{P}(t)$ el vector con componentes $P_n(t)$ $(n = 0, 1, 2, ..., \kappa)$. La probabilidad $P_n(t)$ puede ser descrita por una cadena de Markov en tiempos discretos considerando que: un elemento de una cierta clase puede ser copiado incrementando su popularidad, $n \to n+1$; puede ser removido disminuyendo su popularidad, $n \to n-1$; o puede ser copiado y, otro elemento de la misma clase, removido en el mismo paso temporal, manteniendo su popularidad. En el proceso BMPG el elemento más viejo en el núcleo es removido, sin embargo para obtener la cadena de Markov se debe realizar una aproximación, en la cual en cada paso temporal el elemento removido se escoge aleatoriamente entre todos los elementos dentro del núcleo. En esta

aproximación se deducen las siguientes probabilidades de transición para un núcleo con n elementos de una clase particular,

$$\begin{aligned}
\Pi_{n,n} &= (1-p)\left[\left(\frac{n}{\kappa}\right)^2 + \left(1-\frac{n}{\kappa}\right)^2\right] + p\left(1-\frac{n}{\kappa}\right), \\
\Pi_{n,n+1} &= (1-p)\frac{n}{\kappa}\left(1-\frac{n}{\kappa}\right) \quad (n \neq k), \\
\Pi_{n,n-1} &= (1-p)\frac{n}{\kappa}\left(1-\frac{n}{\kappa}\right) + p\frac{n}{\kappa} \quad (n \neq 0),
\end{aligned} \tag{8.2}$$

donde $n = 0, 1, 2, ..., \kappa$. Entonces, si se conoce la distribución de probabilidad a tiempo t, es posible computar la distribución de probabilidad a tiempo $t+1$ mediante la relación $\mathbf{P}(t+1) = \mathbf{P}(t)\mathbf{\Pi}$, donde $\mathbf{\Pi}$ es la matriz con elementos $\{\Pi_{n,m}\}$ en la cual los elementos no nulos corresponden a $m = n, n-1, n+1$ y están dados por las Ecs.(8.2). En particular, utilizando $\frac{d\mathbf{P}}{dt} \approx \mathbf{P}(t+1) - \mathbf{P}(t) = \mathbf{P}(t)(\mathbf{\Pi} - \mathbf{1})$ es posible obtener la ecuación maestra,

$$\begin{aligned}
\frac{dP_n(t)}{dt} &= \Pi_{n-1,n}P_{n-1}(t) + \Pi_{n+1,n}P_{n+1}(t) \\
&\quad -(\Pi_{n,n} - 1)P_n(t),
\end{aligned} \tag{8.3}$$

donde la matriz de transición se extiende para incluir los casos especiales $\Pi_{-1,0} = \Pi_{\kappa+1,\kappa} = 0$. Expresando el tiempo en unidades de κ ($\tilde{t} = t/\kappa$) y acomodando términos, la ecuación maestra puede escribirse como:

$$\begin{aligned}
\frac{dP_n(\tilde{t})}{d\tilde{t}} &= b^{(n-1)}P_{n-1}(\tilde{t}) + d^{(n+1)}P_{n+1}(\tilde{t}) \\
&\quad -(b^{(n)} + d^{(n)})P_n(\tilde{t}),
\end{aligned} \tag{8.4}$$

donde $b^{(-1)} = d^{(\kappa+1)} = 0$ y para los casos restantes,

$$\begin{aligned}
b^{(n)} &= n\left[(1-p) - (1-p)\frac{n}{\kappa}\right] = n\beta(n) \\
d^{(n)} &= n\left[1 - (1-p)\frac{n}{\kappa}\right] = n(p + \beta(n)).
\end{aligned} \tag{8.5}$$

Aquí $\beta(n) = (1-p)(1-\frac{n}{\kappa})$. Cuando β es constante esta ecuación maestra corresponde a un proceso de Galton-Watson [87]. Sin embargo, ya que el número de estados es limitado en este modelo–i.e., la popularidad dentro del núcleo de un dado elemento no puede superar el tamaño del mismo– β resulta una función decreciente de n. En el contexto de los modelos de ecologías adaptativas, usualmente se asume que el tamaño de la población total es fija y está determinada por la capacidad de carga del ambiente [88]. Resulta entonces interesante el análisis de la distribución de tiempos de vida de los elementos dentro del núcleo, ya que estudios anteriores han utilizado este abordaje para modelar la distribución de tiempos de vida de especies en ecología [87]. Como se mostrará más adelante, distintos regímenes surgen en la dinámica del modelo BMPG dependiendo

del valor del producto $p\kappa$. En la Figura 8.2 se muestran las distribuciones de tiempos de vida de los elementos en el modelo BMPG para tres distintos regímenes $p > p_c$, $p = p_c$, y $p < p_c$, donde $p_c = 1/\kappa$. Estas distribuciones, como todos los resultados presentados de ahora en más, fueron obtenidos mediante simulaciones numéricas, generando secuencias de $N = 10^7$ elementos. En estas series se computan los tiempos de vida de cada clase de elemento g, y se promedia sobre 20 realizaciones para cada valor de los parámetros. En los tres casos, los tiempos de vida de los elementos está distribuido de acuerdo a una distribución de cola larga, que puede ser aproximada por una ley de potencia con exponente $\approx -1{,}9$. Este exponente está entre $-3/2$, el cual es el tiempo de salida en un problema de caminata aleatoria, y -2, el proceso de ramificación de Galton-Watson crítico [87]. Más aún, se ha sugerido [89] que los tiempos de vida de grupos taxonómicos en el registro de fósiles poseen una distribución con una cola de ley de potencia con exponentes en el rango entre $-3/2$ a -2.

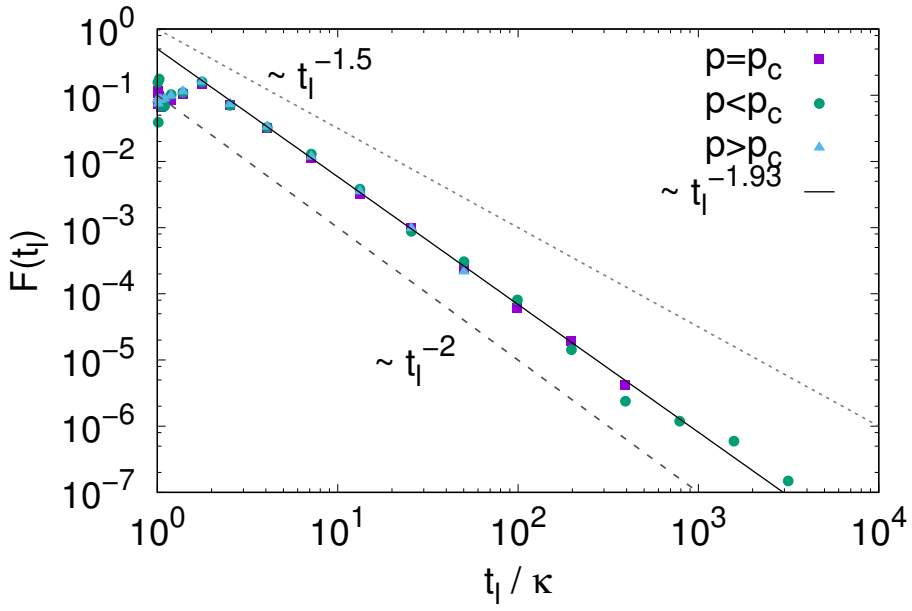

Figura 8.2.: Distribución de tiempos de vida calculadas para el modelo BMPG con $\kappa = 500$ para $p = p_c = 1/\kappa$ (cuadrados violetas), $p = 0{,}01p_c$ (círculos verdes) y $p = 10p_c$ (triángulos azules). Las líneas corresponden a leyes de potencia con exponentes $-1{,}93$ (línea completa), $-3/2$ (línea de puntos) y -2 (línea a rayas).

8.3 Distribución de Popularidad

A fin de caracterizar más profundamente el modelo, se calculó la distribución de popularidad $P(s)$ de las secuencias generadas para el modelo BMPG utilizando diferentes valores de los parámetros p y κ. En la Figura 8.3 (panel superior) se muestran los resultados obtenidos para $\kappa = 100$, 300 y 500, donde en cada caso $p = p_c = 1/\kappa$. Para

todas las $P(s)$ calculadas se encontró que las distribuciones se ajustan bien a una ley de potencia,

$$P(s) \sim s^{-\alpha}, \tag{8.6}$$

con $\alpha \simeq 3/2$. Se debe notar que el núcleo de tamaño finito mantiene la cola de ley de potencia encontrado en los modelos originales de Cattuto y Yule-Simon, pero modifica el exponente α en una forma particular, como se discutirá más adelante.

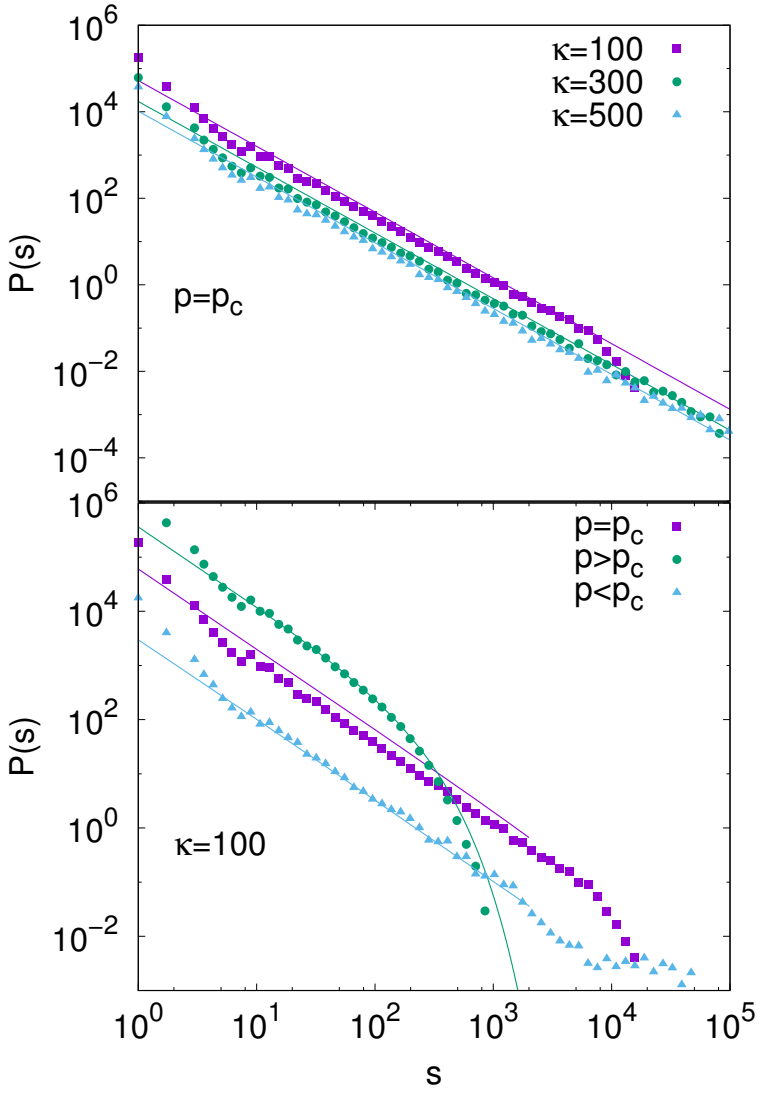

Figura 8.3.: Panel superior: Distribuciones de popularidad calculadas para las secuencias generadas con el modelo BMPG para $\kappa = 100$, 300 y 500, y $p = p_c$. Las líneas rectas corresponden a ajustes de la Ec. (8.6). En todos los casos los exponentes del ajuste corresponden a $\alpha \simeq 3/2$. Panel inferior: Distribución de popularidad calculada para las secuencias generadas con el modelo BMPG para $\kappa = 100$ y $p = 0{,}1p_c$, p_c y $10p_c$; las líneas completas corresponden a la estimación de la Ec. (8.10).

Con el fin de analizar la distribución de popularidad de los elementos en el modelo BMPG, se relaciona la generación de los elementos de cada clase a un proceso de ramificación correspondiente. Un proceso de ramificación puede ser asociado a un árbol

con un nodo de partida o raíz. El proceso comienza con un nodo raíz el cual crea hijos, y esos hijos crean hijos propios, etcétera. El número de hijos que genera cada nodo es representado por una variable estocástica. Más aún, cada nodo pertenece a una generación específica definida por su distancia al nodo raíz del árbol correspondiente. En este contexto, la generación de elementos de una dada clase en el modelo BMPG puede aproximarse por la evolución de un proceso de ramificación. Cada copia de un elemento de una dada clase corresponde a uno de los nodos en el árbol en su correspondiente generación, con excepción de la raíz, la que corresponde a la primera aparición de esta clase de elementos. De esta forma, cada vez que la raíz es copiada, nace un hijo de la primera generación y, de manera similar, si un nodo de la primera generación es copiado, un hijo de la segunda generación nace, etcétera, y de esta forma se construye el árbol.

Cada elemento puede tener a lo sumo κ hijos, ya que sólo puede ser copiado mientras esté dentro del núcleo de memoria. En este proceso, la probabilidad que un dado nodo tenga i hijos puede ser aproximado por una distribución binomial,

$$p_i = \binom{\kappa}{i} \theta^i (1-\theta)^{\kappa - i}, \tag{8.7}$$

donde $\theta = \frac{1}{\kappa}(1-p)$ es la probabilidad que un dado elemento en el núcleo sea copiado en cada paso temporal. En un proceso de ramificación reglado por una distribución binomial, cada conexión en el árbol correspondiente se genera con probabilidad θ. Ya que el número de conexiones en un árbol es $s-1$, donde s es el número de nodos, o elementos, la probabilidad de generar $s-1$ conexiones es θ^{s-1}. El número de conexiones ausentes para completar un árbol κ-ario (de κ ramas) con s nodos internos es $\kappa s - (s-1)$, donde internos significa que los nodos no son hojas del árbol. Por lo tanto, la probabilidad que este proceso genere un árbol κ-ario particular con s nodos internos y $\kappa s - (s-1)$ conexiones ausentes es [55] $\theta^{s-1}(1-\theta)^{(\kappa-1)s+1}$. En la Figura 8.4 se muestran dos ejemplos de árboles κ-arios con $\kappa = 3$ y tres nodos internos ($s = 3$).

El número de árboles κ-arios con s nodos internos es [90],

$$N^{(\kappa)}(s) = \frac{1}{(\kappa-1)s+1} \binom{\kappa s}{s}. \tag{8.8}$$

Entonces la probabilidad de tener un árbol κ-ario con s nodos internos y $\kappa s - (s-1)$ conexiones ausentes, o en el contexto del modelo BMPG, un elemento de popularidad s, es:

$$P(s) = \frac{1}{(\kappa-1)s+1} \binom{\kappa s}{s} \theta^{s-1}(1-\theta)^{(\kappa-1)s+1}. \tag{8.9}$$

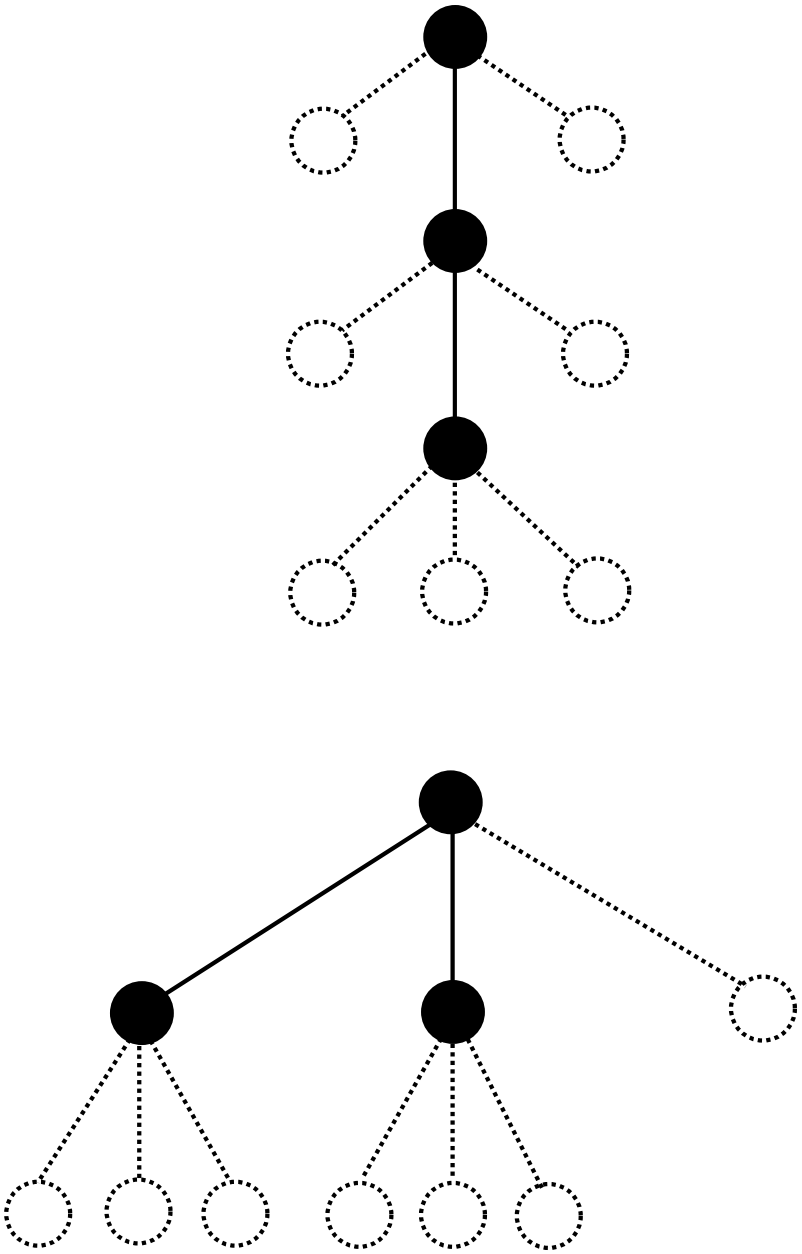

Figura 8.4.: Ilustración de dos árboles κ-arios con $\kappa = 3$ y $s = 3$. Los nodos negros corresponden a los nodos internos y los marcados en líneas de puntos son los conexiones y nodos ausentes.

Usando la aproximación de Stirling para $\kappa \gg 1$ y $s \gg 1$ se obtiene el comportamiento asintótico de la Ec. (8.9),

$$P(s) \sim e^{-s/s_0(\kappa,\theta)} s^{-3/2}, \tag{8.10}$$

con $s_0(\kappa,\theta) \approx \frac{\kappa-1}{\kappa}\frac{2}{(1-\kappa\theta)^2} = \frac{\kappa-1}{\kappa}\frac{2}{p^2}$.

En la Figura 8.3 (panel inferior) se muestran las distribuciones resultantes $P(s)$ calculadas para la secuencia generada con el modelo BMPG, junto con las predicciones teóricas de la Ec. (8.10). Se puede observar que las predicciones teóricas ajustan de

buena manera las distribuciones calculadas para las secuencias generadas con el modelo, cuando los valores de κ son lo suficientemente grandes; esto confirma que un proceso de ramificación con un distribución binomial es una buena aproximación en este escenario. Más aún, cuando $p = p_c = 1/\kappa$ la cola exponencial de la distribución $P(s)$ (Ec. (8.10)) se aleja como $s_0 \propto \kappa^2$. Por lo tanto, para valores muy grandes de κ, como los utilizados en la Figura 8.3 (panel superior), la distribución tiende a una ley de potencia, $P(s) \sim s^{-3/2}$ en un rango amplio de valores de s como se observa en la figura.

8.4 Correlaciones

Como se mencionó previamente, las secuencias generadas por el modelo de Cattuto original tienen correlaciones de largo alcance. Ya que el modelo BMPG tiene un núcleo de memoria de tamaño finito, no se esperan correlaciones de largo alcance en las series temporales generadas $x[t]$. En esta sección se analiza el tipo de correlaciones presentes en las series temporales. Para construir las series temporales se utilizó la regla de asignación GAR (Sección 6.1) [19]. A fin de observar como se comporta la longitud o tiempo de correlación en este núcleo, se calculó la función autocorrelación $C(\Delta t, t)$ de la serie como,

$$C(\Delta t, t) = \frac{\langle (x[t] - \mu(t))(x[t + \Delta t] - \mu(t + \Delta t)) \rangle}{\sigma(t)\sigma(t + \Delta t)}, \tag{8.11}$$

donde $\mu(t) = \langle x[t] \rangle$ y $\sigma^2(t) = \langle (x[t] - \mu)^2 \rangle$ y $\langle ... \rangle$ significa tomar valor de expectación. Considerando que la serie temporal es estacionaria, entonces $\mu(t) = \mu(t + \Delta t) = \mu_0$, $\sigma(t) = \sigma(t + \Delta t) = \sigma_0$, y la función autocorrelación $C(\Delta t, t) = C(\Delta t)$, la cual puede ser calculada como,

$$C(\Delta t) = \frac{1}{N - \Delta t} \sum_{i=1}^{N-\Delta t} \hat{x}[t_i]\hat{x}[t_i + \Delta t], \tag{8.12}$$

donde $\hat{x}[t_i] = (x[t_i] - \mu_0)/\sigma_0$, $\mu_0 = \frac{1}{N}\sum_i^N x[t_i]$ y $\sigma_0^2 = \frac{1}{N}\sum_i^N (x[t_i] - \mu_0)^2$. Se encontró que $C(\Delta t)$ decae exponencialmente, $C(\Delta t) \sim e^{\frac{-\Delta t}{R}}$, como puede verse en la Figura 8.5. Se calculó la función autocorrelación para varios valores de κ y $p = p_c = 1/\kappa$, y se obtuvo la tiempo de correlación R ajustando la función exponencial. En el inset de la Figura 8.5 se muestra la longitud de correlación R como función de κ. Se encontró que R crece como ley de potencia con exponente $\gamma = 1{,}9$. La Figura 8.5 también muestra el colapso de las curvas de autocorrelación cuando el tiempo es medido en unidades de la longitud $R \sim \kappa^{1,9}$.

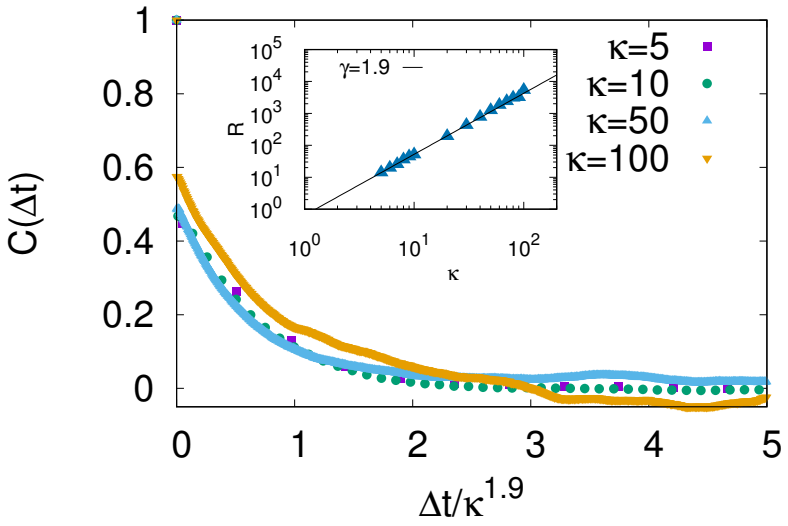

Figura 8.5.: Función autocorrelación $C(\Delta t)$ calculada para las secuencias generadas con el modelo BMPG para $\kappa = 5, 10, 50,$ y 100 y $p = p_c$, reescalado por κ^{γ} con $\gamma = 1{,}9$. Inset: Tiempo característico de decaimiento, o longitud de correlación, R como función de la extensión del núcleo de memoria κ. La línea recta corresponde al ajuste lineal en escala logarítmica con exponente $\gamma = 1{,}9$.

8.5 Análisis de tiempo entre eventos

Como se mencionó en la sección 2.4, una desviación del comportamiento exponencial en la distribución de tiempo entre eventos es un claro indicio de la presencia de efectos de memoria. En el caso de palabras en un texto, o en la ocurrencia de líneas de juego de un jugador individual, la desviación es bien descrita por una exponencial estirada. Un indicio más claro de la presencia de *burstiness* se presenta cuando la distribución de tiempos entre eventos es descrita por una ley de potencia,

$$f(\tau) = A\tau^{-\eta}, \tag{8.13}$$

donde el grado de *burstiness* está caracterizado por el exponente η de la ley de potencia; donde cuanto menor es el exponente, más marcada resulta la dinámica *bursty*.

Utilizando lo descrito anteriormente, junto con el cálculo del coeficiente de *burstiness* B (sección 2.4), se analizó la dinámica de los tiempos entre eventos de varios elementos en series generadas por el modelo BMPG.

Para comenzar, se definió el nivel de actividad de un dado elemento g como la inversa de la longitud de onda de Zipf, como $a_g = 1/\langle\tau^{(g)}\rangle$ donde $\langle\tau^{(g)}\rangle$ nuevamente es la media de los tiempos entre eventos de este elemento. Luego, analizando la serie temporal se encontró que, en promedio, la actividad de los elementos depende de su popularidad

a través de la relación $a \sim s^{1/2}$, i.e., los elementos más activos son también los más populares.

Se analizaron distribuciones de tiempos entre eventos para diferentes conjuntos de clases de elementos. Cada conjunto se construyó agregando elementos en un orden decreciente de popularidad hasta alcanzar un número específico de tiempos entre eventos. En esta forma se construyeron cuatro conjuntos de $\sim 2 \times 10^6$ tiempos entre eventos con niveles de actividad decreciente.

En el panel superior de la Figura 8.6 se muestra la distribución de tiempos entre eventos $f(\tau)$ para el conjuntos de elementos más activo, donde se utilizó $\kappa = 500$ y tres valores del parámetro p ($p = 0{,}1p_c$, $p = p_c$, y $p = 10p_c$). Como se puede observar, $f(\tau)$ tiene un decaimiento de ley de potencia en los tres casos con un exponente $\eta = 3$, lo cual indica una clara presencia de *burstiness*. Para $p = p_c$ la distribución completa es bien descrita por una ley de potencia. No obstante, la desviación de p de p_c afecta la distribución a tiempos cortos. En $p = 0{,}1p_c$ la serie es homogénea, i.e. compuesta casi en su totalidad por elementos de una sola clase, incrementando el número de tiempos entre eventos cortos. Al crecer p ($p = 10p_c$), el número de clases diferentes también aumenta, y los tiempos entre eventos son más largos, resultando en una distribución plana para valores pequeños de τ, como se puede apreciar en la figura. También se observó que, independientemente del valor de κ utilizado en el modelo, el exponente de decaimiento permanece igual. Por comparación, en la figura también se muestra la distribución obtenida con el modelo de Cattuto para $\kappa_C = 500$ y $p = 1/\kappa_C$, así como el ajuste correspondiente de una distribución de Weibull (Ec.(2.22)) con $\beta = 1$. Como se mencionó, en este caso la distribución $f(\tau)$ se puede explicar mediante un proceso de Poisson.

En el panel inferior de la Figura 8.6 se muestran las distribuciones de tiempos entre eventos $f(\tau)$ ($\kappa = 500$ y $p = p_c$) para varias clases de elementos agrupados en conjuntos correspondientes a distintos niveles de popularidad promedio. Todas las distribuciones presentan un decaimiento de ley de potencia con exponente $\eta = 3$, y colapsan cuando los ejes coordenados son reescalados con el factor $\langle \tau \rangle = \langle s \rangle^{-1/2}$, donde $\langle s \rangle$ es la popularidad promedio del conjunto correspondiente.

También se calculó el parámetro de *burstiness* para diferentes valores de κ y p, usando en este caso el conjunto más activo (inset Figura 8.6 –panel inferior–). Para todos los valores de κ existe un rango alrededor de $\kappa p \sim 10$ en el cual se evidencia la dinámica *bursty*, mientras que para valores pequeños de κp, $B < 0$, ya que las series resultan más regulares.

Como la distribución de tiempos entre eventos para $p = p_c$ se ajusta muy bien a una ley de potencia con exponente $\eta = 3$, es posible calcular los valores correspondientes de

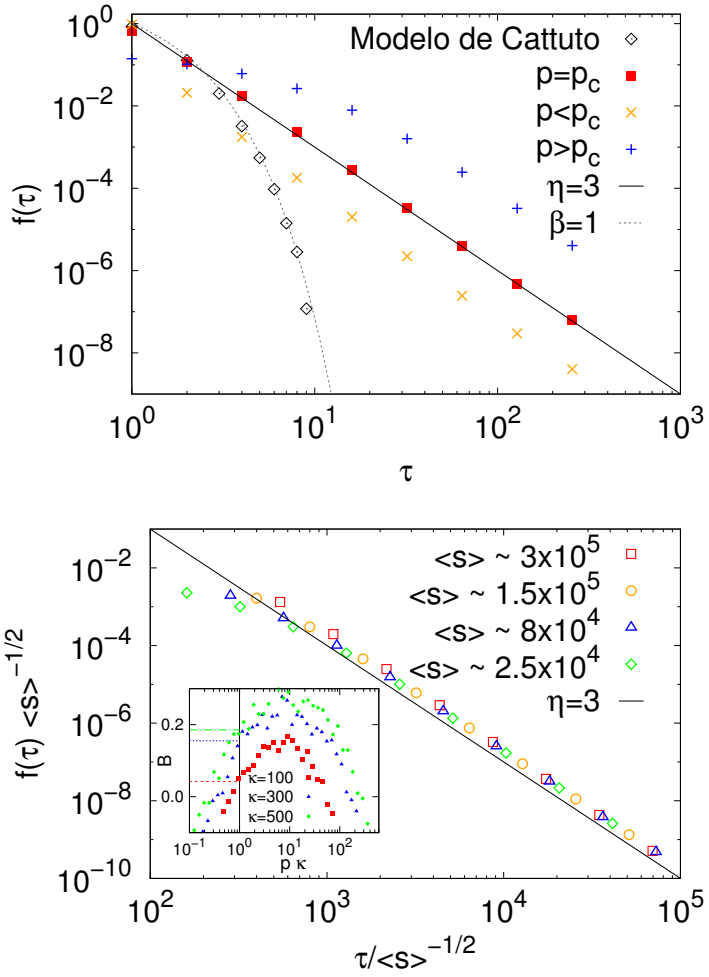

Figura 8.6.: Panel superior: distribución de tiempos entre eventos medida en las secuencias generadas con el modelo BMPG para $\kappa = 500$ y tres valores de p; $p = p_c$ (cuadrados violetas), $p = 0{,}1p_c$ (círculos verdes) y $p = 10p_c$ (triángulos celestes). También se muestran los ajustes de acuerdo a la Ec. (8.13) (línea negra) para el caso de $p = p_c$. Además, se muestra la distribución de tiempos entre eventos para las secuencias generadas con el modelo de Cattuto para $\kappa_C = 500$ y $p = 1/\kappa_C$ (diamantes negros vacíos) y el ajuste de acuerdo a la Ec. (2.22) (línea gris a rayas). Panel inferior: Distribución de tiempos entre eventos para diferentes niveles de actividad, reescalados con $\langle s \rangle^{-1/2}$ y una ley de potencia de exponente 3 para comparar. Inset: Parámetro de burstiness B calculado para el modelo BMPG para varios valores de κ y p como función de κp; la línea negra vertical corresponde a $p\kappa = 1$ y las líneas horizontales corresponden a los valores calculados de B a partir de la distribución $f(\tau)$–Ecs. (8.14) y (8.15)–.

B para distintos valores de κ mediante el cálculo directo de $\langle \tau \rangle$ y $\langle \tau^2 \rangle$ a partir de la distribución, apropiadamente normalizada, resultando en,

$$\langle \tau \rangle = \int_1^{\kappa} \tau f(\tau) d\tau = \frac{2\kappa}{\kappa + 1}, \tag{8.14}$$

y

$$\left\langle \tau^2 \right\rangle = \int_1^{\kappa} \tau^2 f(\tau) d\tau = \frac{2\kappa^2 \ln(\kappa)}{\kappa^2 - 1}. \tag{8.15}$$

Utilizando estas expresiones se calculó B (Eq. (2.25)) para diferentes valores de κ. Estos resultados se muestran en el inset de la Figura 8.6 (panel inferior) como líneas horizontales. Como se puede observar, los valores calculados de B coinciden con los valores medidos en las series generadas con el modelo BMPG para $p = p_c$.

8.6 Análisis del núcleo

A fin de comprender la emergencia de *burstiness* en los núcleos de tamaño finito se analizó el estado dentro del núcleo en el modelo BMPG en función de los parámetros del modelo. En analogía con las transiciones de fase, como primer enfoque se analizó el estado del núcleo mediante la definición de un parámetro de orden $0 \leq \phi \leq 1$ de la siguiente manera,

$$\phi = \frac{1}{\kappa} \max_g n_g^{(\kappa)}, \tag{8.16}$$

donde $n_g^{(\kappa)}$ es la popularidad de la clase de elementos g dentro del núcleo de longitud κ. De acuerdo con esta definición, cuando el núcleo está lleno de elementos de una sola clase, el parámetro de orden es igual a uno. Por otro lado, cuando todos los elementos en el núcleo pertenecen todos a distintas clases $\phi = 1/\kappa$, tomando su mínimo valor. En la Figura 8.7 (panel superior) se muestra el valor medio del parámetro de orden como función de $p\kappa$, para diferentes valores de la extensión del núcleo κ, donde ϕ se calcula dentro de N/κ segmentos disjuntos de longitud κ a lo largo de la serie generada (N es la longitud total de la serie), y la media $\langle \phi \rangle$ se calcula promediando sobre los N/κ segmentos. Se puede notar que todas las curvas colapsan cuando se las grafica en función de κp, a su vez se puede observar, a medida que κp crece, una transición continua de un estado desordenado $\phi \sim 1/\kappa \approx 0$ a un estado ordenado $\phi \approx 1$. La inflexión de las curvas ocurre en $p\kappa \approx 1$, lo que sugiere una transición en este punto.

Con el fin de poner a prueba la existencia de cierta forma de criticalidad en la transición, se estudiaron las fluctuaciones del parámetro de orden al calcular la varianza $\sigma_\phi^2 = \langle \phi^2 \rangle - \langle \phi \rangle^2$. La varianza como función de $p\kappa$ muestra un pico en $p\kappa \approx 1$ –inset Figura 8.7 (panel inferior)– indicando que las fluctuaciones alcanzan un máximo en el punto de inflexión del parámetro de orden. Ya que todas las curvas colapsan cuando se grafican en función de $p\kappa$, la magnitud de las fluctuaciones es independiente del tamaño del núcleo. Utilizando un argumento eurístico se puede realizar una analogía con la mecánica estadística asociando p a la temperatura –ya que esta variable introduce desorden en el núcleo– y la extensión del núcleo κ al tamaño del sistema. Si se piensa que la probabilidad de introducir un nuevo elemento resulta de un proceso activado, entonces $p \propto \exp(-\Delta E/T)$, donde ΔE es la energía de activación. De esto se obtiene

que $T \propto -1/\ln(p)$, o de manera simplificada $T = -1/\ln(p)$, lo que introduce todo el rango de temperaturas. En este contexto es posible también definir la susceptibilidad como $\chi = \frac{\kappa}{T}\sigma_\phi^2 = -\ln(p)\kappa\sigma_\phi^2$. En la Figura 8.7 (panel inferior) se muestra un gráfico de la susceptibilidad χ en función de $T = -1/\ln(p)$, donde se puede observar que el pico de χ crece con el tamaño del núcleo, y se espera que diverja en $T = 0$ en el límite termodinámico $\kappa \to \infty$, asemejándose al comportamiento de, por ejemplo, el modelo de Ising unidimensional.

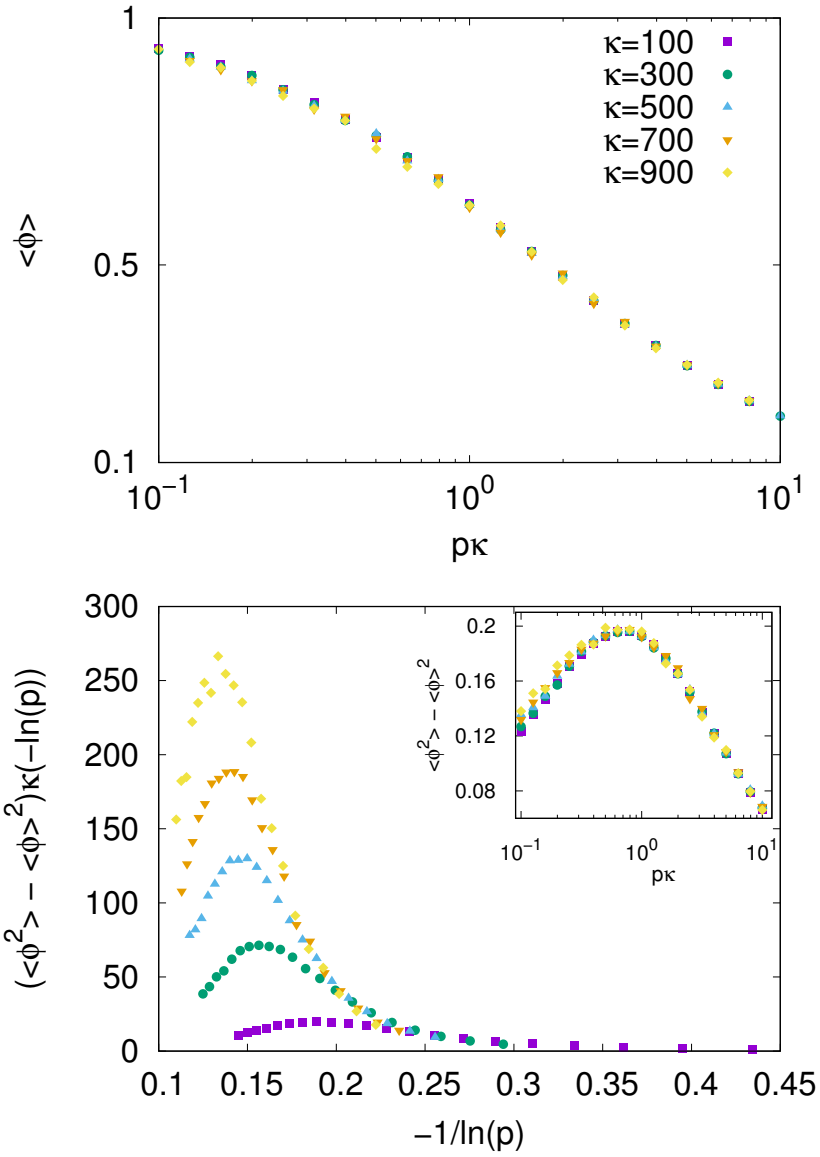

Figura 8.7.: La media del parámetro de orden $\langle\phi\rangle$ (panel superior) y las fluctuaciones $(\langle\phi^2\rangle - \langle\phi\rangle^2)\kappa/p$ (panel inferior) calculados para el modelo BMPG para varios valores de κ en función de p y reescalado por $p_c = 1/\kappa$ (Inset).

En el panel superior de la Figura 8.8 se muestran las configuraciones del núcleo de memoria para los tres estados principales, super-crítico ($p > p_c$), crítico ($p = p_c$), y

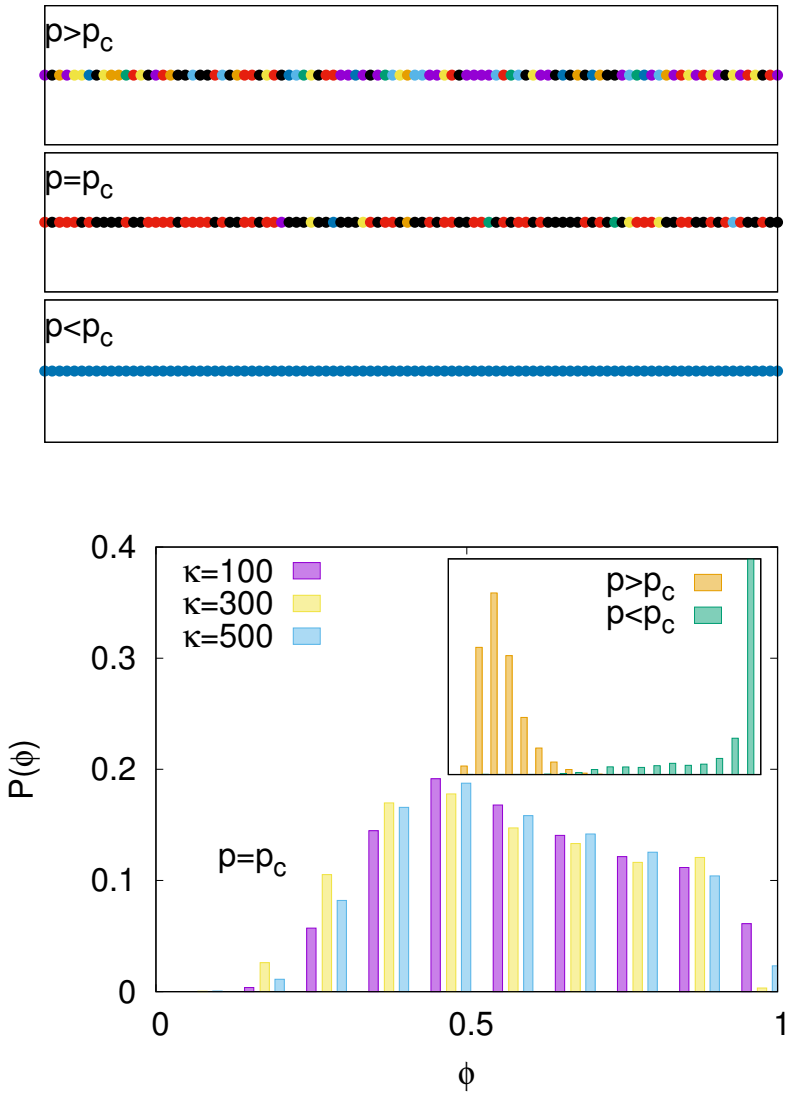

Figura 8.8.: Panel superior: configuraciones del núcleo para los estados super-crítico ($p > p_c$), crítico ($p = p_c$) y sub-crítico ($p < p_c$). Panel inferior: Distribución del parámetro de orden calculado para el modelo BMPG para distintos valores de κ en el caso crítico. Inset: Ejemplos de distribuciones del parámetro de orden en los casos $p > p_c$ y $p < p_c$.

sub-crítico ($p < p_c$). En el estado sub-crítico el núcleo se encuentra lleno de elementos de una sola clase ($\phi \sim 1$), en el super-crítico, muchos elementos de popularidades similares coexisten en el núcleo, y en el caso crítico es posible distinguir el elemento más popular mediante simple inspección. Finalmente, una mejor descripción de las fluctuaciones de ϕ dentro del núcleo se pueden obtener calculando la distribución del parámetro de orden. Se calcularon las distribuciones en el punto donde las fluctuaciones son máximas, i.e., $p = p_c = 1/\kappa$, para distintos valores de κ. En el panel inferior de la Figura 8.8 se muestran las distribuciones correspondientes para $\kappa = 100, 300$ and 500. Las distribuciones no dependen del valor de κ, y se ensanchan alrededor de $\phi = 0{,}5$,

confirmando los resultados obtenidos al medir la varianza. En el inset se muestran las distribuciones para los casos sub-crítico ($p < p_c$) y super-crítico $p > p_c$ para $\kappa = 500$ y, $p = 0{,}1p_c$ y $p = 10p_c$. Ambas distribuciones son estrechas, la primera mostrando un pico cerca de 1, y la segunda cerca de 0, tal como se esperaba, dado que en esos puntos las fluctuaciones son bajas.

La distribución de los elementos dentro del núcleo pueden ser estudiada mediante la entropía del núcleo,

$$S = -\sum_g \frac{n_g^{(\kappa)}}{\kappa} \ln\left(\frac{n_g^{(\kappa)}}{\kappa}\right), \tag{8.17}$$

donde, nuevamente, $n_g^{(\kappa)}$ es la popularidad de la g-ésima clase de elementos dentro del núcleo. En la Figura 8.9 se muestra el valor medio de la entropía, μ_S (panel superior), y sus fluctuaciones, σ_S (panel inferior), en función de $p\kappa$ para distintos valores de κ. Similarmente a lo observado para el parámetro de orden, todas las curvas de la entropía colapsan cuando se grafican en función de $p\kappa$. La varianza de la entropía se comporta de la misma forma que las fluctuaciones del parámetro de orden, alcanzando su valor máximo en $p \approx p_c$, indicando nuevamente la existencia de una transición. Sin embargo, las mismas parecen aumentar con el tamaño del núcleo. Las fluctuaciones de la entropía están directamente relacionadas a la presencia de *burstiness*, ya que implican que la tasa con la cual una clase de elementos es copiado es una cantidad variable.

Es interesante notar que la actividad *bursty* de las tormentas solares tiene una distribución de tiempos entre eventos con una cola de ley de potencia con exponente ≈ 3, en conformidad con los resultados obtenidos para el modelo BMPG. La distribución de tormentas solares ha sido satisfactoriamente explicado utilizando un proceso de Poisson dependiente del tiempo, que resulta de una superposición de procesos de Poisson constantes de a trozos [37]. En dicho modelo se utiliza un método de la estadística Bayesiana [91] en el cual se toma un conjunto de datos y determina la descomposición en procesos de Poisson constantes de a trozos. Más específicamente, el método toma un conjunto de tiempos (t_i, t_f), y devuelve un arreglo de tasas de Poisson $(\mu_1, ..., \mu_n)$, y un conjunto de tiempos de cambio $(t_i, t_1, ..., t_{n-1}, t_f)$ los que corresponden a los puntos donde cambia la tasa del proceso. Los intervalos temporales en los cuales la tasa es constante se llaman "bloques Bayesianos". El proceso consiste en comparar la expectativa[1] que el conjunto de datos, o sub-conjunto de datos cuando se tiene más de un intervalo, sea generado por un proceso de Poisson con una tasa constante, o un proceso de Poisson con dos tasas. Este mecanismo se repite para cada intervalo, y en cada paso se decide sub-dividir el mismo o no. El proceso de segmentación se detiene cuando el sistema cumple una condición impuesta que depende del mismo, por ejemplo, el requerimiento que los intervalos tengan un número mínimo de eventos.

[1] Del inglés *likelihood*.

Dentro de este enfoque, el proceso es descompuesto en intervalos de tiempo, en los cuales los tiempos entre eventos son consistentes con un proceso de Poisson de tasa constante. En analogía con las tormentas solares, la tasa de copiado de una dada clase de elementos varía en el modelo BMPG cerca de la transición, evidenciado por las fluctuaciones de la entropía y parámetro de orden. El mecanismo mencionado es robusto frente a variaciones de la distribución de tasas de Poisson. Por lo tanto, puede ser fácilmente adaptado para explicar la distribución de tiempos entre eventos del modelo BMPG.

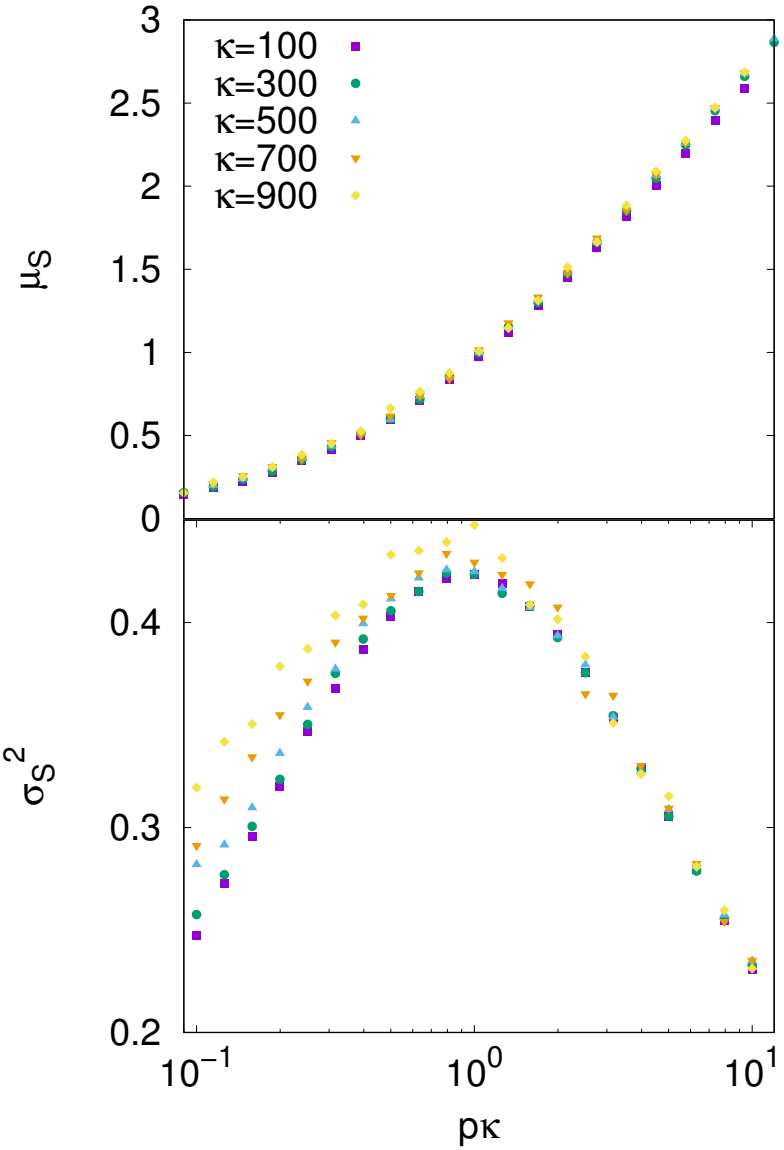

Figura 8.9.: Media de la entropía μ_S (panel superior) y fluctuaciones σ_S (panel inferior) calculado para el modelo BMPG para varios valores de κ en función de $p\kappa$.

Finalmente, se repitió el análisis utilizando un núcleo de memoria finito con decaimiento exponencial,

$$Q(\Delta t) = \frac{1}{\kappa_e} \exp(-\Delta t/\kappa_e). \tag{8.18}$$

Los resultados del análisis con este núcleo de memoria se muestran en la Figura 8.10. Comparando estos resultados con los mostrados en las Figuras 8.3, 8.6, 8.7 y 8.9 se puede ver que la elección particular de la forma funcional del núcleo de memoria no afecta a los resultados obtenidos de manera significativa. Esto sugiere que las propiedades estadísticas encontradas son independientes de la forma funcional específica del núcleo de memoria, y que las mismas dependen del comportamiento asintótico para valores grandes de Δt.

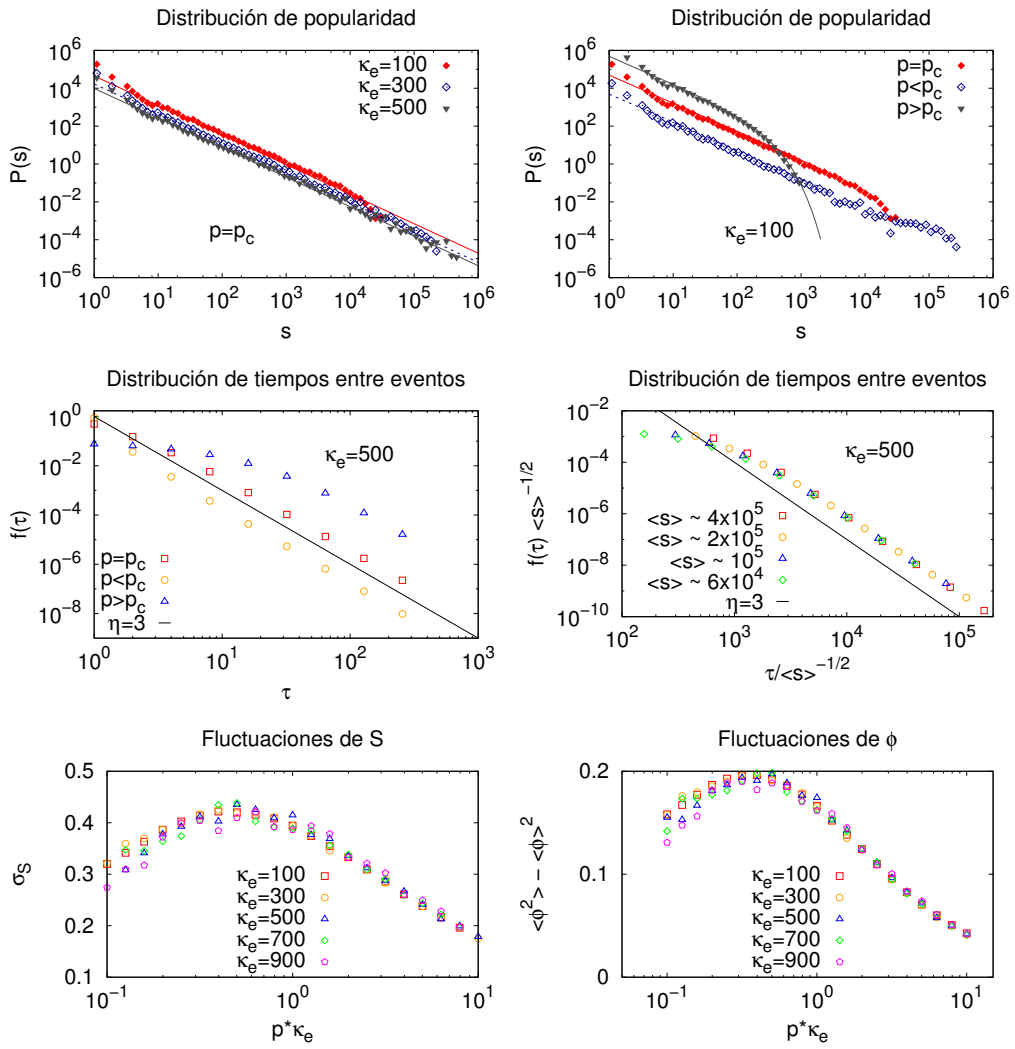

Figura 8.10.: Resultados obtenidos con el núcleo de memoria exponencial.

8.7 Conclusiones parciales

En este capítulo se presentaron los resultados del estudio y caracterización de un modelo estocástico de crecimiento preferencial con núcleo de memoria acotado. Específicamente, se modificó el proceso estocástico de Yule-Simon introduciendo un núcleo de memoria de tamaño finito de extensión κ, y a este modelo se lo llamó *Bounded Memory Preferential growth* (crecimiento preferencial con memoria limitada) o modelo BMPG. Se estudiaron diversas propiedades estadísticas de las series de elementos generadas por el modelo BMPG mediante simulaciones numéricas y herramientas estándar de procesos estocásticos.

Se encontró que las distribuciones de tiempos de vida de las distintas clases de elementos en las series siguen una ley de potencia. Se derivó la ecuación maestra que gobierna la probabilidad de tener s copias de una dada clase de elementos dentro del núcleo a tiempo t, y se encontró que esta ecuación es similar a las propuestas para modelos de tiempos de vida de especies en ecologías [87], los que consisten en procesos de nacimiento y muerte. Más aún, los exponentes de las distribuciones obtenidas para el modelo BMG, $\approx 1{,}9$, están dentro del rango de los valores reportados en sistemas empíricos [89].

El modelo BMPG genera también elementos cuyas popularidades están distribuidas de acuerdo a una ley de potencia. Esto es consistente con las distribuciones altamente sesgadas y de cola larga del modelo original de Yule-Simon y de la versión modificada por Cattuto et al., la cual incluye un núcleo de memoria de largo alcance. En particular, el exponente de la distribución $(3/2)$ en el caso del núcleo de tamaño finito puede ser explicado en términos de un proceso de ramificación con una distribución binomial, donde el número máximo de hijos que puede generar un nodo es igual al tamaño del núcleo. Este enfoque funciona bien para núcleos lo suficientemente grandes y es también capaz de explicar la cola del decaimiento exponencial observado en las distribuciones obtenidos con este modelo.

Las correlaciones observadas en las series de elementos producidas por el modelo BMPG son de corto alcance, como se esperaba, diferente a las correlaciones de largo alcance observadas en el modelo de Cattuto. Sin embargo, la longitud de correlación del núcleo finito crece casi cuadraticamente con la extensión del núcleo.

Un efecto interesante observado en la presencia de un núcleo finito, el cual no se observa en los modelos de Yule-Simon o Cattuto, es que las distribuciones de tiempos entre eventos de los elementos de la serie, con un nivel de actividad definido, decae como ley de potencia con exponente ~ 3, y es independiente del nivel de actividad del conjunto de elementos. Esto significa que las secuencias generadas de un conjunto de

elementos muestras períodos de alta actividad seguidos por períodos de latencia, en un dado un rango de los parámetros del modelo. Más aún, las distribuciones de tiempos entre eventos colapsan cuando se reescalan de acuerdo al nivel de actividad medio del conjunto de elementos, y además se encuentra que el nivel de actividad aumenta con la popularidad de los elementos de la serie. Además, el parámetro de *burstiness* es mayor a cero en un rango de valores de los parámetros del modelo, reforzando la evidencia de presencia de *burstiness* en la serie generada. El mecanismo de está dinámica *bursty* puede ser asociada a la superposición de procesos de Poisson con una distribución particular de tasas de Poisson. Particularmente este mecanismo fue usado para explicar la distribución de tiempos entre evento de tormentas solares [92] las que, al igual que en el modelo BMPG, siguen una ley de potencia con exponente 3. En contraste, los resultados de los estudios del modelo de Yule-Simon y Cattuto en los capítulos anteriores muestran que las distribuciones de tiempos entre eventos de las series generadas con estos modelos pueden ser explicadas con un proceso de Poisson con una sola tasa.

Con el fin de explicar la presencia de la dinámica *bursty* en las secuencias generadas por el modelo BMPG, se caracterizó el estado del núcleo de memoria mediante la definición de un parámetro de orden que mide la fracción de ocupación del núcleo por la clase de elementos más popular. Además, como medida de la distribución de las diferentes clases de elementos dentro del núcleo, se calculó la entropía del núcleo. Mediante el estudio del valor medio y las fluctuaciones del parámetro de orden y la entropía, se encontró que el estado del núcleo sufre una transición de un estado ordenado (valor pequeños de p) a un estado desordenado (valores grandes de p), y que existe un punto crítico en $p = p_c = 1/\kappa$ donde las fluctuaciones de ambas cantidades alcanzan un máximo. Particularmente, las fluctuaciones de la entropía están relacionadas a la aparición de *burstiness* en el modelo BMPG ya que implican que la tasa a la cual un elemento es copiado es una cantidad variable. Más aún, como todas las curvas colapsan cuando se grafican en función de $p\kappa$, la magnitud de las fluctuaciones del parámetro de orden y la entropía resultan independientes del tamaño del núcleo.

Parte III

Conclusiones

En este trabajo de tesis se estudiaron diversos efectos estadísticos encontrados en sistemas empíricos que exhiben la ley de Zipf-Pareto, en particular se tomó como sistema modelo una base de datos de Ajedrez. Específicamente, se estudió la presencia de la ley de Heaps en la aparición de nuevas líneas de juego, la ley de Zipf en la distribución de popularidades de líneas de apertura. Además, se exploró la existencia de efectos de memoria en la forma de correlaciones de largo alcance y en la distribución de tiempos entre eventos.

A partir de la naturaleza autosimilar del árbol de partidas, se observó que el mecanismo que lo genera implica una ausencia de escalas típicas en el fenómeno de innovación en el Ajedrez; no existen nodos con frecuencias o popularidades particularmente altas luego de los cuales la innovación se vuelve imposible. En otras palabras, en el Ajedrez no existen estrategias ganadoras, y siempre hay una posibilidad para introducir soluciones innovadoras. Por otra parte, los resultados obtenidos sobre la innovación de líneas de juego muestran similitudes con el crecimiento del vocabulario en textos literarios, los que sugieren la existencia de secuencias de movimientos núcleo y no-núcleo en el árbol de partidas de Ajedrez, debido a una diferencia de las escalas temporales en el proceso de evolución del cuerpo de partidas. Particularmente, el comportamiento a tiempos largos muestra que el exponente de Heaps aumenta con la profundidad del árbol, lo cual es una característica encontrada en lenguajes que tienen un alto grado de inflexión.

A partir del estudio de las series temporales generadas empleando la base de datos empírica, usando diversas reglas de asignación, se encontró que las secuencias de partidas de Ajedrez poseen correlaciones de largo alcance y muestran efectos de tamaño. El hecho que dichas correlaciones desaparecen cuando se realiza un mezclado aleatorio comprueba que estos efectos de memoria encontrados no son producto de las reglas de asignación utilizadas, sino una característica intrínseca del sistema. Más aún, al estudiar la base de datos filtrada por rangos de ELO, los resultados obtenidos indican que las correlaciones

de largo alcance observadas a escalas temporales largas están relacionadas a la presencia de jugadores de alto nivel. En contraste, las partidas correspondientes a jugadores de nivel intermedio y bajo nivel muestran correlaciones significativas a tiempos cortos. Esto se puede deber a que los jugadores sobresalientes conocen una mayor cantidad de líneas de apertura en profundidad y son menos influenciados por sus oponentes. En particular, en el caso de los jugadores de alto nivel las correlaciones de largo alcance en escalas temporales largas resultan más fuertes.

Por otro lado, se analizó la base de datos de Ajedrez dentro del marco de dos modelos basados en un mecanismo de crecimiento preferencial, los modelos de Yule-Simon y Cattuto et al., ya que ambos modelos permiten generar secuencias artificiales con distribuciones de ley de potencia, o Zipf-Pareto, similares a las encontradas en la serie empírica. En particular, el modelo propuesto por Cattuto et al. agrega un núcleo de memoria al proceso de Yule-Simon. El análisis realizado demuestra que ambos modelos son capaces de reproducir, hasta cierto punto, la distribución de popularidades obtenida a partir de la base de datos. Sin embargo, el modelo de Cattuto resulta más realista, ya que reproduce la distribución de ley de potencia de las líneas de apertura usando valores de los parámetros del modelo medidos en la base de datos. Más aún, debido al núcleo de memoria, el modelo de Cattuto es también capaz de reproducir las correlaciones de largo alcance y los efectos de tamaño observados en la serie empírica, mientras que el modelo de Yule-Simon carece de memoria. Particularmente, el modelo de Cattuto exhibe correlaciones de largo alcance y efectos de tamaño, independientemente de la regla de asignación utilizada para construir la serie temporal, aunque el grado de persistencia sí depende de la regla de asignación. Específicamente, el modelo de Cattuto reproduce los efectos de tamaño y persistencia observados en la base de datos para el caso de la regla de asignación Gaussiana (GAR). Esto sugiere que existen correlaciones subyacentes relacionadas con la popularidad de las líneas de juego de ajedrez y la serie temporal correspondiente, que no son capturadas por el modelo de Cattuto et al..

Además, se realizó un análisis complementario al estudio de las correlaciones de largo alcance, que consiste en el estudio de la distribución de tiempos entre eventos que permite explorar otros efectos de memoria tal como la presencia de *burstiness*. En este análisis se estudió la secuencia de tiempos entre eventos de la partida más popular, tanto en el caso de la base de datos completa como para el jugador más activo. A partir del cálculo de la distribución acumulada de tiempo entre eventos y el parámetro de *burstiness*, usando la base de datos completa, se encontró que no se puede asegurar la presencia de una dinámica *bursty*, y además que esta dinámica puede ser aproximada por un proceso de Poisson. Aunque el comportamiento del conjunto completo de jugadores de la base de datos no exhibe *burstiness*, el análisis de los tiempos entre eventos de las partidas más populares empleadas por jugadores individuales muy activos sí presenta un comportamiento *bursty*. Esto indica que la ausencia de *burstiness* a nivel grupal es una consecuencia de la agregación de datos. Se corroboró además que los modelos de

Yule-Simon y Cattuto no generan una dinámica *bursty*. Además, las secuencias generadas por los mismos se pueden explicar a través de un proceso de Poisson y, en consecuencia, el modelo de Cattuto resulta apropiado para modelar la base de datos completa pero no para analizar la dinámica de jugadores individuales. La ausencia de *burstiness* en el modelo de Cattuto sugiere que el fenómeno de *burstiness* –el cual puede explicar la aparición de efectos de memoria de largo alcance– no es necesario para explicar la presencia de correlaciones de largo alcance.

Con el fin de producir una dinámica de tiempo entre eventos con *bursts* de actividad se modificó el proceso estocástico de Yule-Simon introduciendo un núcleo de memoria de tamaño finito, y a este modelo se lo llamó *Bounded Memory Preferential growth* (crecimiento preferencial con memoria limitada) o modelo BMPG. Se caracterizó el modelo BMPG a través del estudio de diversas propiedades estadísticas de las series de elementos generadas por el el mismo mediante simulaciones numéricas y herramientas estándar de la física estadística.

Se encontró que las distribuciones de tiempos de vida de las distintas clases de elementos en las series siguen una ley de potencia con un exponente robusto ante variaciones de los parámetros del modelo. Se derivó la ecuación maestra que gobierna la probabilidad de tener un número particular de copias de una dada clase de elementos dentro del núcleo a un dado tiempo, y se encontró que esta ecuación es similar a las propuestas para modelos de tiempos de vida de especies en ecologías, los que involucran procesos de nacimiento y muerte de especies.

El modelo BMPG genera también elementos cuyas popularidades están distribuidas de acuerdo a una ley de potencia. Esto es consistente con las distribuciones altamente sesgadas y de cola larga del modelo original de Yule-Simon y de la versión modificada por Cattuto et al., la cual incluye un núcleo de memoria de largo alcance. En particular, el exponente de la distribución en el caso del núcleo de tamaño finito puede ser explicado en términos de un proceso de ramificación con una distribución binomial, donde el número máximo de hijos que puede generar un nodo es igual al tamaño del núcleo. Este enfoque funciona bien para núcleos lo suficientemente grandes y es también capaz de explicar la cola de decaimiento exponencial observado en las distribuciones obtenidas con este modelo en las simulaciones numéricas.

Las correlaciones observadas en las series de elementos producidas por el modelo BMPG son de corto alcance, como se esperaba, a diferencia de las correlaciones de largo alcance observadas en el modelo de Cattuto. Sin embargo, la longitud de correlación del núcleo finito crece casi cuadráticamente con la extensión del núcleo.

Un efecto interesante asociado a núcleos de tamaño finito, el cual no se observa en los modelos de Yule-Simon o Cattuto, es que las distribuciones de tiempos entre

eventos de los elementos de la serie con un nivel de actividad definido, decae como ley de potencia con exponente ≈ 3, y es independiente del nivel de actividad del conjunto de elementos. Esto significa que las secuencias generadas con un conjunto de elementos muestras períodos de alta actividad seguidos por períodos de latencia, dado un rango de los parámetros del modelo. Más aún, las distribuciones de tiempos entre eventos colapsan cuando se reescalan de acuerdo al nivel de actividad medio del conjunto de elementos, y además se encontró que el nivel de actividad aumenta con la popularidad de los elementos de la serie. A partir de un análisis del parámetro de *burstiness* se obtuvo que éste es mayor a cero en un rango de valores de los parámetros del modelo, reforzando la evidencia de presencia de *burstiness* en las series generadas. El mecanismo de está dinámica *bursty* puede ser asociada a la superposición de procesos de Poisson con una distribución particular de tasas. Particularmente, este mecanismo fue usado para explicar la distribución de tiempos entre evento de tormentas solares [92] las que, al igual que en el modelo BMPG, siguen una ley de potencia con exponente 3. En contraste, los resultados de los estudios del modelo de Yule-Simon y Cattuto en los capítulos anteriores muestran que las distribuciones de tiempos entre eventos de las series generadas con estos modelos pueden ser explicadas con un proceso de Poisson con una sola tasa.

Con el fin de explicar la presencia de la dinámica *bursty* en las secuencias generadas por el modelo BMPG, se caracterizó el estado del núcleo de memoria mediante la definición de un parámetro de orden que mide la tasa de ocupación del núcleo por la clase de elementos más popular. Además, como medida de la distribución de las diferentes clases de elementos dentro del núcleo, se calculó la entropía del núcleo. Mediante el estudio del valor medio y las fluctuaciones del parámetro de orden y la entropía, se encontró que el estado del núcleo sufre una transición de un estado ordenado (valor pequeños de p) a un estado desordenado (valores grandes de p), y que existe un punto crítico en $p = p_c = 1/\kappa$ donde las fluctuaciones de ambas cantidades alcanzan un máximo. Particularmente, las fluctuaciones de la entropía están relacionadas a la aparición de *burstiness* en el modelo BMPG ya que implican que la tasa a la cual un elemento es copiado es una cantidad variable. Más aún, como todas las curvas colapsan cuando se grafican en función de $p\kappa$, la magnitud de las fluctuaciones del parámetro de orden y la entropía resultan independientes del tamaño del núcleo.

Ajedrez

A

A.1 Reglas y Aspectos Generales del Juego

El Ajedrez es un juego de mesa de estrategia entre dos jugadores y toma lugar en un tablero con 64 cuadrados en una cuadrícula de 8x8. Cada jugador comienza con 16 piezas: un rey, una dama, dos torres, dos caballos, dos alfiles y ocho peones, cada una de las cuales se mueve de forma diferente. Las piezas son utilizadas para atacar y capturar las piezas del oponente, con el objetivo de realizar "jaque mate" al rey del oponente colocándolo situación de inminente captura. El curso del juego esta dividido en tres etapas: apertura, medio juego y final.

El tablero de ajedrez consiste en ocho filas, denotadas por números del 1 al 8, y ocho columnas, denotadas por letras de "a" a "h" (Figura A.1). Las piezas se dividen convencionalmente en blancas y negras, y los jugadores son referidos como "las blancas" y "las negras".

El jugador blanco siempre mueve primero. Un jugador no puede realizar ningún movimiento el cual deje en situación de jaque a su rey. Si un jugador no tiene movimientos legales posibles el juego concluye, ya sea en jaque mate (el jugador sin posibilidad de movimientos legales pierde) si el rey está en situación de jaque, o en ahogado (empate) si el rey no está en jaque.

A continuación se describen los movimientos de cada pieza:

- Rey: puede moverse un cuadrado en cualquier dirección. Esta pieza también tiene un movimiento especial llamado enroque, una sola vez en el juego cada rey tiene permitido moverse dos espacios a lo largo de la primera fila hacia la torre y luego la torre es colocada en el último cuadrado cruzado por el rey.

- Torre: puede moverse cualquier número de casilleros a lo largo de cualquier fila o columna.

- Alfil: puede moverse cualquier número de casilleros diagonalmente.

- Dama: combina el poder del alfil y la torre.

- Caballo: puede moverse en forma de "L", dos casillero verticalmente y uno horizontalmente, o uno verticalmente y dos horizontalmente. El caballo es la unica pieza que puede saltar sobre otras piezas.

- Peón: puede moverse hacia adelante a lo largo de la misma columna de a un casillero a la vez si el mismo está desocupado, excepto en su primer movimiento donde tiene permitido desplazarse dos cuadrados; o puede moverse a un casillero ocupado por una pieza del oponente si se encuentra diagonalmente a un solo movimiento de distancia. El peón también posee la capacidad de promoción, cuando avanza hasta la octava fila debe ser intercambiado por otra pieza a elección del jugador.

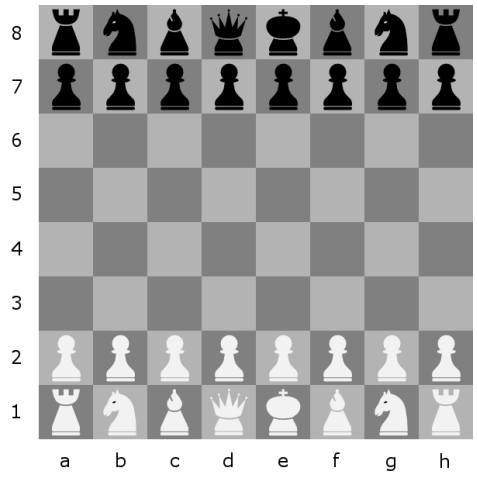

Figura A.1.: Posición inicial de las piezas en el tablero de ajedrez.

Las jugadas y posiciones en el ajedrez son registrados mediante una notación especial, llamada notación algebraica la cual consiste de una letra mayúscula que indica la pieza en movimiento (K o ♔para el rey, Q o ♕para la dama, R o ♖para la torre, B o ♗para el alfil y N ♘para el caballo) más la coordenada de destino de la misma. Por ejemplo, Qg5 (♕g5) significa que la dama realiza un movimiento al casillero g5 (fila 5, columna g). La letra P que indica al peón no se utiliza, por lo tanto e4 simplemente significa que el peón se desplaza hacia el casillero e4. En la situación particular donde dos piezas de la misma especie pueden moverse al mismo casillero se incluye una letra adicional indicando la coordenada de partida, por ejemplo Ngf3 (♘gf3) significa que el caballo de la columna g realiza un movimiento hacia la posición f3.

Si una pieza realiza una captura en uno de sus movimiento, se incluye una "x" antes del casillero de destino, entonces Bxf3 (♗xf3) significa que el alfil captura la pieza ubicada en f3. Cuando un peón realiza una captura se utiliza la designación de la columna de la cual parte en lugar de la inicial de la pieza, y el número de fila es omitido en caso de ser inequívoco, por ejemplo, exd5 significa que el peón en la columna "e" captura

a la pieza localizada en d5. El enroque es indicado por las notaciones especiales, 0-0 para el enroque hacia el flanco del rey y 0-0-0 para el enroque hacia el flanco de la reina. El símbolo "+" indica que el jugador ha colocado al rey del oponente en situación de jaque.

Al finalizar la partida "1-0" indica que las blancas ganaron, "0-1" que las negras ganaron y "1/2-1/2" si la partida concluyó en empate.

A continuación se muestra un ejemplo de una partida completa registrada entre Garry Kasparov (blancas) y Viktor Kortschnoj (negras) jugada en Islandia en el año 2000:

1.	e4	e6
2.	d4	d5
3.	Nc3	Nf6
4.	Bg5	Bb4
5.	e5	h6
6.	Be3	Ne4
7.	Qg4	Kf8
8.	a3	Bxc3+
9.	bxc3	c5
10.	Bd3	h5
11.	Qf4	Qa5
12.	Ne2	Nxc3
13.	0-0	Nxe2+
14.	Bxe2	Nc6
15.	c4	cxd4
16.	Bxd4	Nxd4
17.	Qxd4	Bd7
18.	cxd5	exd5
19.	Bf3	Bc6

1/2-1/2

A.2 La Historia del Ajedrez

El juego del ajedrez ha fascinado a la humanidad por más de 1500 años. El predecesor más similar tuvo origen en el norte de la India con el nombre de *chaturanga*[1]. No obstante sus comienzos son tan remotos que resulta imposible determinar su origen exacto. La primera referencia escrita al juego es en un poema de finales del Siglo VI. La teoría más

[1] El término *chaturanga* significa cuatro secciones y refiere a una formación militar. Llevaba este nombre ya que se jugaba entre cuatro personas y en un tablero de 74 casillas.

probable es que el *chaturanga* se expandiera hacia el este en dirección a China y Japón, y hacia el oeste a Persia, donde pasó a llamarse *shatranj*[2][93].

Es desde Persia donde el ajedrez comienza a evolucionar hasta lograr su forma actual. Esta versión más parecida al juego moderno fue transmitida primero a España y de allí al resto de Europa. La diferencia más grande entre el *shatranj* y el ajedrez actual es la movilidad de las piezas equivalentes a la dama y el alfil, las cuales sólo podían avanzar al igual que los peones, y la no existencia del enroque. Por esta diferencia de movilidad de estas piezas tan claves en el ajedrez actual, las aperturas eran, en comparación, increíblemente lentas.

Hacia el Siglo XII el juego del ajedrez se había expandido prácticamente en todo el continente europeo, y dejó de ser simplemente un entretenimiento para convertirse en un atractivo para el arte y la ciencia.

A finales del Siglo XV, con la finalidad de agilizar las aperturas, la movilidad de algunas piezas cambió, el peón ahora podría avanzar dos posiciones en el primer movimiento, y la dama y el alfil adquirieron las capacidades de movilidad de la actualidad. Debido a que la reina se convirtió en la pieza más poderosa la nueva versión del juego fue apodada en algunos libros de los Siglos XV y XVI "ajedrez de la dama".

Entre los Siglos XVII y XIX, la llegada del movimiento cultural e intelectual europeo que trajo consigo la Ilustración y la emancipación del pensamiento, el ajedrez comienza a desligarse de las doctrinas medievales y se establece como el juego predilecto de la clase intelectual[93], al mismo tiempo que comienza a atraer cada ver más la atención de la clase aristocrática y las cortes reales, a las que fueron invitados los jugadores más prominentes de la época.

A medida que el juego cobraba popularidad se establecieron ciertas reglas que persisten en la actualidad, como la limitación del tiempo de juego, el enroque y las reglas de ahogado (o *stalemate*) en la que el juego termina en empate, hasta entonces variaba dependiendo de la época y zona geográfica: victoria para el jugador en posición de tablas, el mismo solo perdía el turno, o simplemente no estaba permitido, entre otras posibilidades.

[2] El término *shatranj* se deriva de la palabla *chaturanga*. En la cultura popular Persa se escribía algunas veces como *sad* ('cien') + *ranj* ('preocupaciones').

A.3 Sistema de puntuación Elo

A través de la historia hubo numerosos intentos por determinar un sistema que fuera capaz de puntuar las capacidades de los jugadores. En 1970 la Federación Internacional de Ajedrez, FIDE, implementó un sistema de puntuación llamado Elo que utiliza un método estadístico para calcular los niveles relativos de habilidad de los jugadores. Este método fue inventado por el físico y ajedrecista aficionado Árpád Élő, y siendo aplicado también a otras formas de competición como scrabble y juegos de rol de participación masiva por Internet como World of Warcraft.

El problema de la calificación de los jugadores es un problema que cae dentro del área de la estadística del modelado de 'comparación de pares', cuyos datos se obtienen de cualquier resultado que indique preferencia por un objeto sobre otro. En el caso ajedrez los resultados de los partidos no son mas que la consecuencia de la comparación entre dos jugadores para determinar cuál de ellos es el 'preferido' (o si no existe 'preferencia' como en el caso del empate).

A partir del estudio de torneos pasados, Élő observó que la distribución de rendimientos, esto es, la distribución de probabilidades de que un jugador se desempeñe a un cierto nivel, era similar a la de una distribución normal[3]. Una de las ventajas de utilizar la distribución normal para modelar los desempeños de los jugadores es que la diferencia entre las distribuciones de rendimiento de dos jugadores es también una distribución normal, solo que más dispersa[74].

En la actualidad la Federación de Ajedrez Estadounidense (USCF) utiliza la distribución logística en lugar de la normal, a pesar de que al analizar datos de comparación de pares no existe una diferencia significativa si se asume una distribución normal o logística para las diferencias entre los rendimientos de los jugadores[94].

En el sistema de Elo cada jugador posee un puntaje numérico el cual no es calculado de forma absoluta sino que es estimado a partir de victorias, derrotas y empates en enfrentamientos contra otros jugadores. Calculando la diferencia entre los Elos de dos jugadores es posible estimar el resultado esperado del partido. Si un jugador A que posee un Elo R_A se enfrenta a un jugador B con Elo R_B, las puntuaciones esperadas de los jugadores serán,

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}},$$

[3] Desviaciones estándar de las distribuciones obtenidas por Élő: $\sigma = 200$ puntos para la distribución individual de un jugador y $\sigma = \sqrt{2}\,200$ puntos para la distribución de dos jugadores.

donde las puntuaciones que un jugador puede obtener en un partido son 1 si el jugador gana, $\frac{1}{2}$ si el juego termina en empate y 0 si pierde. De esta forma una diferencia de 200 puntos significa que el jugador de mayor Elo posee un puntaje esperado de $0{,}75$, que es la probabilidad de victoria P_v más la mitad de probabilidad de empate P_e, ya que en el sistema de Elo un empate se considera media victoria más media derrota, es decir,

$$E_A = P_v + \frac{P_e}{2}.$$

Una de las contribuciones más importantes de Élő fue la introducción de un algoritmo simple que actualiza las calificaciones de los jugadores en base a los resultados de un torneo. Si el jugador en cuestión supera el puntaje esperado su Elo aumenta, y en caso contrario disminuye. Estas actualizaciones se realizan de manera incremental y existe un límite máximo para los ajustes de los Elos de los jugadores por partido llamado K-factor, el cual depende de la categoría ($K = 16\,Elo$ para maestros y $K = 32\,Elo$ para jugadores menos expertos). Suponiendo que un jugador A posee un puntaje esperado de E_A puntos, pero en la realidad obtuvo S_A, su actualización de Elo será,

$$R'_A = R_A + K \cdot (S_A - E_A).$$

La escala de puntuaciones tiene un límite mínimo en cero, y por más que el máximo no está limitado, sería inaudito que un jugador excediera los $3000\,Elo$. En la actualidad, los jugadores de ajedrez poseen Elos menores a 2900, mientras que, debido al desarrollo de las reglas eurísticas, el puntaje de los motores de ajedrez supera los $3000\,Elo$[10].

El sistema de puntuación de Elo es utilizado por FIDE y USCF para clasificar tanto los torneos como los jugadores en categorías. La FIDE clasifica los torneos considerando el promedio de Elo de los jugadores. Las categorías cambian cada 25 puntos, comenzando con la categoría 1 con Elos de 2251 a 2275, hasta la categoría 22 con Elos superiores a 2776 para los hombres, y las mismas categorías para el caso de las mujeres pero con 200 puntos menos, por lo tanto la correspondiente categoría 1 sería de 2051 hasta 2075. Por otra parte la Federación de Ajedrez Estadounidense clasifica a los jugadores en 14 categorías según su Elo, desde la categoría A (Elos de 100 a 199) hasta la categoría Senior Master (Elo 2400 ó superior) en incrementos de $200\,Elo$.

B

Procesos Auto-similares

Los procesos auto-similares fueron introducidos por Kolmogorov (1941) dentro de un contexto teórico. Sin embargo, los estadistas ignoraban la relevancia de dicho concepto hasta que fue introducido por Mandelbrot. No obstante, la idea de auto-similitud es más antigua. Mandelbrot se refiere, por ejemplo, a las pinturas con flujos turbulentos de Leonardo da Vinci las que exhiben torbellinos coexistentes de todos los tamaños y por lo tanto auto-similitud. Una figura geométrica se dice auto-similar de forma determinista si las mismas estructuras geométricas son observadas independientemente de la distancia a la que se la examine.

Desde el punto de vista estocástico, la auto-similitud está definida en términos de la distribución del proceso. Un proceso estocástico con parámetro temporal continuo t se dice auto-similar con parámetro de auto-similitud H si, para todo factor de estiramiento c positivo, el proceso reescalado con escala temporal ct, $c^{-H}Y_{ct}$, es igual en distribución al proceso original, en otras palabras, la distribución posee invariancia de escala. Por lo tanto, recorridos habituales de la muestra son cualitativamente iguales, independientemente de la distancia a la cual se observe.

Un proceso estocástico Y_t tiene incrementos estacionarios si, para todo $k \geqslant 1$ y tiempos $t_1, ..., t_k$ cualesquiera, la distribución de $(Y_{t_1+c} - Y_{t_1+c-1}, ..., Y_{t_k+c} - Y_{t_1+c-1})$ no depende de $c \in \mathbb{R}$. Dada esta definición es posible obtener un resultado de sumo interés.

Suponiendo que Y_t es un proceso estocástico tal que $Y_1 \neq 0$ con probabilidad positiva e Y_t es el límite en distribución de la secuencia de sumas parciales normalizadas

$$\frac{1}{a_n}\sum_{i=1}^{[nt]} X_i = \frac{S_{nt}}{a_n} \rightarrow_d Y_t$$

donde $[nt]$ denota la parte entera de nt, $\rightarrow_d$ significa convergencia en distribución[1], $X_1, X_2, ...$ es una secuencia estacionaria de variables aleatorias, y $a_1, a_2, ...$ es una

[1]Una secuencia de variables aleatorias $X_1, X_2, ...$ se dice converger en distribución a una variable aleatoria X si, $\forall x \in \mathbb{R}$ para el cual F es continua, $\lim_{n\to\infty} F_n(x) = F(x)$, donde F_n y F son las funciones de distribución acumuladas de las variables X_n y X respectivamente.

secuencia de constantes positivas normalizadoras tales que $\log(a_n) \to \infty$. Entonces existe una constante $H > 0$ tal que para todo $u > 0$,

$$\lim_{n\to\infty} \frac{a_{nu}}{a_n} = u^H$$

e Y_t es auto-similar con parámetro de auto-similitud H y tiene incrementos estacionarios. ¿Es decir, independientemente del parámetro de estiramiento u elegido, $\frac{a_{nu}}{a_n}$ se comporta asintóticamente, para $n \to \infty$, como una ley de potencias con el mismo exponente H? Esto significa que, cuando un proceso es el límite de las sumas parciales normalizadas de variables aleatorias, es necesariamente auto-similar. Por lo tanto se puede decir que el rol de los procesos auto-similares dentro de los procesos estocásticos es análogo al rol central de las distribuciones estables dentro de las distribuciones.

B.1 Incrementos Estacionarios en Procesos Auto-similares

Dado un proceso auto-similar Y_t con parámetro de auto-similitud H, la propiedad

$$Y_t =_d t^H Y_1,$$

donde $=_d$ es igualdad en distribuciones, implica el siguiente comportamiento límite de Y_t cuando $t \to \infty$:

1. Si $H < 0$, entonces $Y_t \to_d 0$.

2. Si $H = 0$, entonces $Y_t =_d Y_1$.

3. Si $H > 0$ e $Y_t \neq 0$, entonces $|Y_t| \to_d \infty$.

Análogamente, para $t \to 0$ se tiene:

1. Si $H < 0$ e $Y_t \neq 0$, entonces $|Y_t| \to_d \infty$.

2. Si $H = 0$, entonces $Y_t =_d Y_1$.

3. Si $H > 0$, entonces $Y_t \to_d 0$.

El rango de H puede ser restringido a $H > 0$, ya que si los incrementos del proceso auto-similar son estacionarios, entonces el proceso es matemáticamente patológico

para valores negativos de H. Más específicamente, para $H < 0$, Y_t no es un proceso mensurable.

El aspecto de la función de covarianza $\gamma_y(t,s) = cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)]$ de un proceso auto-similar Y_t con incrementos estacionarios es el resultado de considerar H positivo e $Y_0 = 0$ con probabilidad igual a 1. Asumiendo $E(Y_t) = 0$ a fin de simplificar notación, $s < t$, y denotando por $\sigma^2 = E[(Y_t - Y_{t-1})^2] = E[Y_1^2]$ la varianza del proceso incremental $X_t = Y_t - Y_{t-1}$, entonces,

$$E[(Y_t - Y_s)^2] = E[(Y_{t-s} - Y_0)^2] = \sigma^2(t-s)^{2H}.$$

Por otro lado,

$$E[(Y_t - Y_s)^2] = E[Y_t^2] + E[Y_s^2] - 2E[Y_t Y_s] = \sigma^2 t^{2H} + \sigma^2 s^{2H} - 2\gamma_y(t,s),$$

por lo tanto,

$$\gamma_y(t,s) = \frac{1}{2}\sigma^2[t^{2H} - (t-s)^{2H} + s^{2H}].$$

Las covarianzas de la secuencia de incrementos $X_i = Y_i - Y_{i-1} \quad (i = 1, 2, 3, ...)$ son calculadas de forma similar. Utilizando la auto-similitud se obtiene, para la covarianza entre X_i y $X_{i+k} \quad (k > 0)$,

$$\gamma(k) = \frac{1}{2}\sigma^2[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$$

para $k \geqslant 0$ y $\gamma(k) = \gamma(-k)$ para $k < 0$. Y por lo tanto las correlaciones están dadas por

$$\rho(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$$

para $k \geqslant 0$ y $\rho(k) = \rho(-k)$ para $k < 0$.

El comportamiento asintótico de $\rho(k)$ es analizado mediante la expansión de Taylor: Primero cabe notar que $\rho(k) = \frac{1}{2}k^{2H}g(k^{-1})$ donde $g(x) = (1+x)^{2H} - 2 + (1-x)^{2H}$. Si $0 < H < 1$ y $H \neq 1/2$, entonces el primer término distinto de cero en la expansión de Taylor de $g(x)$, expandido alrededor del origen, es $2H(2H-1)x^2$. Por lo tanto, para $k \to \infty$, $\rho(k)$ es equivalente a $H(2H-1)k^{2H-2}$, es decir,

$$\frac{\rho(k)}{H(2H-1)k^{2H-2}} \to 1$$

para $k \to \infty$. Para $1/2 < H < 1$, esto significa que las correlaciones decaen lentamente de forma que

$$\sum_{-\infty}^{\infty} \rho(k) = \infty,$$

por lo tanto, la Ec. (2.13) es válida, lo que significa que el proceso posee memoria de largo alcance y que el exponente H resulta ser el exponente de Hurst. Para $H = 1/2$, todas las correlaciones para distancias no nulas son cero, y las observaciones X_i resultan no correlacionadas.

C

Fluctuaciones en el modelo de YSM

El proceso de Yule-Simon es un modelo estocástico ideado para explicar el crecimiento de ley de potencia en el número acumulado de ocurrencias individuales de palabras en un texto. El mismo es derivado de una aproximación de campo medio, ya que supone un límite contínuo en el tiempo y en la ocurrencia de palabras.

Este proceso, también conocido como proceso de Simon, funciona de la siguiente manera. En cada paso temporal t se agrega una palabra:

- Probabilidad p: se crea una palabra nueva
- Probabilidad: $\bar{p} = 1 - p$: se copia una palabra ya existente

Sea i el índice de palabras distintas ordenadas ascendentemente por tiempo de creación, $n_i(t)$ el número acumulado de ocurrencias de la palabra i hasta tiempo t y $N(t)$ la longitud de la secuencia a tiempo t. Por definición $N(t) = t$.

De esta forma se puede escribir la probabilidad que una palabra i sea elegida es proporcional al número de ocurrencias de esa palabra:

$$\Pi(i,t) = \frac{s_i(t)}{N(t)} \quad \mapsto \quad \text{Proceso de Yule Simon}$$

Donde s es número acumulado de ocurrencias de palabra. Si $f(s,t)$ número de palabras distintas incluidas en la clase s a tiempo t, la probabilidad de elegir una dada palabra es,

$$\mathcal{P}(s,t) = \frac{sf(s,t)}{N(t)}$$

EL modelo de Barabasi-Albert es un modelo de grafos que describe el crecimiento de la web, donde $p = 1/2$. Hace crecer el grafo agregando nodos uno por uno, resultando en un cierto número de links conectados a los nodos existentes en proporción a su grado.

nodo $\mapsto$ palabra

grado $\mapsto$ número acumulado de ocurrencias de la palabra

En este modelo el grado de un nodo crece como ley de potencia:

$$s_i(t) \propto \left(\frac{t}{t_i}\right)^{1/2}, \tag{3}$$

donde $s_i(t)$ es el grado esperado del nodo i a tiempo t y t_i el tiempo en el que el nodo i ingresó al grafo.

Dentro del marco del modelo de Yule-Simon, el número acumulado de ocurrencias de la palabra i a tiempo t, denotado por $s_i^*(t)$ es,

$$s_i^*(t+\Delta t) = s_i^*(t) + (1-p)\Pi(i,t)\Delta t$$

$$\frac{s_i^*(t+\Delta t) - s_i^*(t)}{\Delta t} = (1-p)\Pi(i,t) \qquad \Pi(i,t) = \frac{s_i^*(t)}{N(t)} = \frac{s_i^*(t)}{t}$$

$$\frac{ds_i^*}{dt} = (1-p)\frac{s_i^*}{t}$$

$$\frac{ds_i^*}{s_i^*} = (1-p)\frac{dt}{t}$$

Entonces, en forma integral:

$$\int \frac{ds_i^*}{s_i^*} = (1-p)\int \frac{dt}{t}.$$

y se obtiene,

$$s_i^*(t) = \left(\frac{t}{t_i}\right)^{1-p} \qquad s_i^*(t_i) = 1. \tag{5}$$

Notar que el modelo de BA es un caso particular de YSM con $p = 1/2$.

La ocurrencia de un palabra individual va a desviarse del valor esperado bajo un período de observación. Hay palabras que ocurren con más frecuencia que el valor esperado y otras con menos frecuencia.

¿Qué forma tiene la distribución de probabilidad de esas fluctuaciones?

Para contestar esta pregunta se definen las siguientes cantidades,

$P(s_i(t) = s)$: probabilidad que el número acumulado de ocurrencias de la palabra i a tiempo t $(s_i(t))$ sea igual a s.

$P(s_i(t) \to s)$: probabilidad que $s_i(t)$ pase de $s-1$ a s justo a tiempo t.

Δt : tiempo transcurrido desde t_i.

$u_i = t_i + \Delta t$: tiempo para medir probabilidades.

$\boxed{s = 1}$:

$P(s_i(u_i) = 1)$: probabilidad que la palabra i haya ocurrido una sola vez en el intervalo $[t_i : u_i]$.

$$P(s_i(u_i) = 1) = \prod_{t=t_i}^{u_i - 1} \left(p + \bar{p} \frac{t-1}{t} \right)$$
$$= \frac{\Gamma(t_i)\Gamma(u_i - \bar{p})}{\Gamma(u_i)\Gamma(t_i - \bar{p})}$$

$$P(s_i(u_i)) = \prod_{t=t_i}^{u_i - 1} \left[\begin{pmatrix} \text{prob. de crear} \\ \text{una nueva} \\ \text{palabra} \end{pmatrix} + \begin{pmatrix} \text{prob. de copiar} \\ \text{una palabra} \end{pmatrix} \begin{pmatrix} \text{prob. de no} \\ \text{copiar la} \\ \text{palabra } i \end{pmatrix} \right]$$
$$= \prod_{t=t_i}^{u_i - 1} [p + \bar{p}(1 - \Pi(t_i))] \qquad \Pi(\tau, t) = \frac{s_i(t)}{t}$$
$$= \prod_{t=t_i}^{u_i - 1} \left(p + \bar{p} \frac{t-1}{t} \right) \qquad s_i(u_i) = 1$$
$$p + \bar{p} = 1$$
$$= \prod_{t=t_i}^{u_i - 1} \left(1 + \frac{\bar{p}}{t} \right)$$
$$= \frac{\Gamma(t_i)\Gamma(s_i - \bar{p})}{\Gamma(u_i)\Gamma(t_i - \bar{p})}$$

$\boxed{s = 2}$:

$$P(s_i(u_i) \to 2) = P(s_i(u_i - 1) = 1)\frac{\bar{p}}{u_i - 1}$$
$$= \bar{p}\frac{\Gamma(t_i)\Gamma(s_i - 1 - \bar{p})}{\Gamma(u_i)\Gamma(t_i - \bar{p})}$$

$$P(s_i(u_i) \to 2) = \begin{pmatrix} \text{prob. que la palabra} \\ i \text{ haya ocurrido 1 vez} \\ \text{en el intervalo } [t_i + 1 : u_i - 1] \end{pmatrix} \begin{pmatrix} \text{prob. que } i \\ \text{se copie en} \\ \text{el paso } s_i \end{pmatrix}$$
$$= P(s_i(u_i - 1) = 1)\frac{\bar{p}}{u_i - 1}$$
$$= \frac{\Gamma(t_i)\Gamma(u_i - 1 - \bar{p})}{\Gamma(u_i - 1)\Gamma(t_i - \bar{p})}\frac{\bar{p}}{u_i - 1} \qquad (u_i - 1)\Gamma(u_i - 1) = \Gamma(u_i)$$
$$= \bar{p}\frac{\Gamma(t_i)\Gamma(u_i - 1 - \bar{p})}{\Gamma(u_i)\Gamma(t_i - \bar{p})}$$

$$P(s_i(u_i) = 2) = \sum_{v=t_i+1}^{u_i} \left[P(s_i(v) \to 2) \prod_{t=v}^{u_i-1} \left(p + \bar{p}\frac{t-2}{t} \right) \right]$$

La probabilidad de elegir la palabra i por 2° vez en el intervalo $[v : u_i]$ es:

$$\begin{pmatrix} \text{prob. que en el paso } u \\ \text{se copie la palabra } i \end{pmatrix} \begin{pmatrix} \text{prob. que la palabra } i \\ \text{no sea copiada en los } u_i - v \\ \text{pasos siguientes} \end{pmatrix}$$
$$= P(s_i(v) \to 2) \prod_{t=v}^{u_i-1} [p + \bar{p}(1 - \Pi(\tau, t))] \qquad \Pi(\tau, t) = \frac{s_i(t)}{t} = \frac{2}{t}$$
$$= P(s_i(v) \to 2) \prod_{t=v}^{u_i-1} \left[p + \bar{p}\frac{2}{t} \right]$$

El tiempo de la 2° ocurrencia puede ser cualquier tiempo en el intervalo $[t_i + 1 : u_i]$, entonces:

$$P(s_i(u_i) = 2) = \sum_{v=t_i+1}^{u_i} \left[P(s_i(v) \to 2) \prod_{t=v}^{u_i-1} \left(p + \bar{p}\frac{t-2}{t} \right) \right]$$

$$P(s_i(v) \to 2) = \bar{p}\frac{\Gamma(t_i)\Gamma(v-1-\bar{p})}{\Gamma(v)\Gamma(t_i-\bar{p})}$$

$$\prod_{t=v}^{u_i-1}\left(p+\bar{p}-\frac{2\bar{p}}{t}\right) = \prod_{t=v}^{u_i-1}\left(1-\frac{2\bar{p}}{t}\right) = \frac{\Gamma(v)\Gamma(u_i-2\bar{p})}{\Gamma(u_i)\Gamma(v-2\bar{p})}$$

Por lo tanto:

$$P(s_i(u_i)=2) = \bar{p}\frac{\Gamma(t_i)\Gamma(u_i-2\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+1}^{u_i}\frac{\Gamma(v-1-\bar{p})\Gamma(v)}{\Gamma(v)\Gamma(v-2\bar{p})} = \bar{p}\frac{\Gamma(t_i)\Gamma(u_i-2\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+1}^{u_i}\frac{\Gamma(v-1-\bar{p})}{\Gamma(v-2\bar{p})}$$

$\boxed{s=3}$:

$$P(s_i(u_i)\to 3) = P(s_i(u_i-1)=2)\bar{p}\frac{2}{u_i-1} \qquad \Pi(i,u_i-1) = \frac{s_i}{u_i-1} = \frac{2}{u_i-1}$$

$$= 2\bar{p}^2\frac{\Gamma(t_i)\Gamma(u_i-1-2\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+1}^{u_i-1}\frac{\Gamma(v-1-\bar{p})}{\Gamma(v-2\bar{p})}$$

$$P(s_i(u_i)\to 3) = P(s_i(u_i-1)=2)\bar{p}\frac{2}{u_i-1} = \bar{p}\frac{\Gamma(t_i)\Gamma(u_i-1-2\bar{p})}{\Gamma(u_i-1)\Gamma(t_i-\bar{p})}\sum_{v=t_i-1}^{u_i-1}\frac{\Gamma(v-1-\bar{p})}{\Gamma(v-2\bar{p})} = 2\bar{p}^2\frac{\Gamma(t_i)\Gamma(u_i-1-2\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+1}^{u_i-1}\frac{\Gamma(v-1-\bar{p})}{\Gamma(v-2\bar{p})}$$

$$P(s_i(u_i)=3) = \sum_{v=t_i+1}^{u_i}\left[P(s_i(v)\to 3)\prod_{t=v}^{u_i-1}\left(p+\bar{p}\frac{t-3}{t}\right)\right] = 2\bar{p}^2\frac{\Gamma(t_i)\Gamma(u_i-3\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+2}^{v_i}\left[\frac{\Gamma(v-1-2\bar{p})}{\Gamma(v-3\bar{p})}\sum_{w=t_i+1}^{v-1}\frac{\Gamma(w-1-\bar{p})}{\Gamma(w-2\bar{p})}\right]$$

$$\prod_{t=v}^{u_i-1}\left(p-\bar{p}\frac{t-3}{t}\right)=\prod_{t=v}^{u_i-1}\left(1-\frac{3\bar{p}}{t}\right)$$
$$=\frac{\Gamma(v)\Gamma(u_i-3\bar{p})}{\Gamma(u_i)\Gamma(v-3\bar{p})}$$

$$\Rightarrow P(s_i(u_i)=3)=\sum_{v=t_i+2}^{u_i}\left[2\bar{p}^2\frac{\Gamma(v)\Gamma(v-1-2\bar{p})}{\Gamma(v)}\sum_{w=t_i+1}^{v-1}\frac{\Gamma(w-1-\bar{p})}{\Gamma(w-2\bar{p})}\frac{\Gamma(v)\Gamma(u_i-3\bar{p})}{\Gamma(u_i)\Gamma(v-3\bar{p})}\right]$$
$$=2\bar{p}^2\frac{\Gamma(t_i)\Gamma(u_i-3\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+2}^{u_i}\frac{\Gamma(v-1-2\bar{p})}{\Gamma(v-3\bar{p})}\sum_{w=t_i+1}^{v-1}\frac{\Gamma(w-1-\bar{p})}{\Gamma(w-2\bar{p})}$$

$\boxed{s=4}$:

$$P(s_i(u_i)\to 4)=P(s_i(u_i-1)=3)\bar{p}\frac{3}{u_i-1}\qquad\qquad \Pi(i,u_i-1)=\frac{s_i}{u_i-1}=\frac{3}{u_i-1}$$
$$=6\bar{p}^3\frac{\Gamma(t_i)\Gamma(u_i-1-3\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}\sum_{v=t_i+2}^{u_i-1}\left[\frac{\Gamma(v-1-2\bar{p})}{\Gamma(v-3\bar{p})}\sum_{w=t_i+1}^{v-1}\frac{\Gamma(w-1-\bar{p})}{\Gamma(w-2\bar{p})}\right]$$

$$P(s_i(u_i)=4)=\sum_{v=t_i+3}^{u_i}\left[P(s_i(v)\to 4)\prod_{t=v}^{u_i-1}\left(p+\bar{p}\frac{t-4}{t}\right)\right]$$
$$=\underbrace{6\bar{p}^3\frac{\Gamma(t_i)\Gamma(u_i-4\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}}_{(s-1)!\bar{p}^{s-1}\frac{\Gamma(t_i)\Gamma(u_i-s\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}}\overbrace{\sum_{v=t_i+3}^{u_i}}^{\sum_{\phi=t_i+s-1}^{u_i}}\underbrace{\left[\overbrace{\frac{\frac{\Gamma(\phi-1-(s-1)\bar{p})}{\Gamma(\phi-s\bar{p})}}{\Gamma(v-1-3\bar{p})}}^{\frac{\Gamma(\phi-1-(s-1)\bar{p})}{\Gamma(\phi-s\bar{p})}}\overbrace{\sum_{w=t_i+2}^{v-1}}^{\sum_{\psi=t_i+s-2}^{\phi-1}}\overbrace{\left[\frac{\Gamma(w-1-2\bar{p})}{\Gamma(w-3\bar{p})}\sum_{q=t_i+1}^{w-1}\frac{\Gamma(q-1-\bar{p})}{\Gamma(q-2\bar{p})}\right]}^{S_{s-1}(\psi)}\right]}_{S_s(\phi)}$$

Entonces, la forma exacta de la distribución de probabilidades resulta:

$$P(s_i(u_i)=s)=\begin{cases}\dfrac{\Gamma(t_i)\Gamma(u_i-\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})} & s=1\\ (s-1)!\bar{p}^{s-1}\dfrac{\Gamma(t_i)\Gamma(u_i-n\bar{p})}{\Gamma(u_i)\Gamma t_i\bar{p}}\displaystyle\sum_{\phi=t_i+s-1}^{u_i}S_s(\phi) & s>1\end{cases} \tag{10}$$

$$S_s(\phi)=\begin{cases}\dfrac{\Gamma(\phi-1-\bar{p})}{\Gamma(\phi-2\bar{p})} & s=1\\ \dfrac{\Gamma(\phi-1-(s-1)\bar{p})}{\Gamma(\phi-s\bar{p})}\displaystyle\sum_{\psi=t_i+s-2}^{\phi-1}S_{s-1}(\psi) & s>1\end{cases}$$

Tomando t_i y u_i suficientemente grandes, y utilizando la relación asintótica $\lim_{t\to\infty}\frac{\Gamma(t-a)}{\Gamma(t)}\sim t^{-a}$, se obtiene:

$\boxed{s=1}$:

$$\lim_{\substack{t_i\to\infty\\ u_i}}\frac{\Gamma(t_i)\Gamma(u_i-\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}=\lim_{t_i\to\infty}\frac{\Gamma(t_i)}{\Gamma(t_i-\bar{p})}\lim_{u_i\to\infty}\frac{\Gamma(u_i-\bar{p})}{\Gamma(u_i)}\sim t_i^{\bar{p}}u_i^{-\bar{p}}$$

$\boxed{s>1}$:

$$\lim_{\substack{t_i\to\infty\\ u_i}}\frac{\Gamma(t_i)\Gamma(u_i-s\bar{p})}{\Gamma(u_i)\Gamma(t_i-\bar{p})}=\lim_{t_i\to\infty}\frac{\Gamma(t_i)}{\Gamma(t_i-\bar{p})}\lim_{u_i\to\infty}\frac{\Gamma(u_i-s\bar{p})}{\Gamma(u_i)}\sim t_i^{\bar{p}}u_i^{-s\bar{p}}$$

Entonces,

$$P(s_i(u_i)=s)\sim\begin{cases}t_i^{\bar{p}}u_i^{-\bar{p}} & s=1\\ (s-1)!\bar{p}^{s-1}t_i^{\bar{p}}u_i^{-s\bar{p}}\displaystyle\sum_{\phi=t_i+s-1}^{u_i}S_s(\phi) & s>1\end{cases} \tag{12}$$

Además, si $t_i\to\infty$ y $u_i\to\infty$, entonces $\phi\to\infty$:

$\boxed{s=2}$:

$$\begin{aligned}\lim_{\phi\to\infty}\frac{\Gamma(\phi-1-\bar{p})}{\Gamma(\phi-2\bar{p})} &= \lim_{\phi\to\infty}\frac{\Gamma(\phi-p-\bar{p}-\bar{p})}{\Gamma(\phi-2\bar{p})}\\ &= \lim_{\phi\to\infty}\frac{\Gamma(\phi-p-2\bar{p})}{\Gamma(\phi-2\bar{p})}\\ &\sim \lim_{\phi\to\infty}\frac{\Gamma(\phi-p)}{\Gamma(\phi)} \sim \phi^{-p}\end{aligned}$$

$\boxed{s>2}$:

$$\begin{aligned}\lim_{\phi\to\infty}\frac{\Gamma(\phi-1-(s-1)\bar{p})}{\Gamma(\phi-s\bar{p})} &= \lim_{\phi\to\infty}\frac{\Gamma(\phi-\bar{p}-p-s\bar{p}+\bar{p})}{\Gamma(\phi-s\bar{p})}\\ &= \lim_{\phi\to\infty}\frac{\Gamma(\phi-p)}{\Gamma(\phi)} \sim \phi^{-p}\end{aligned}$$

Por lo tanto,

$$S_s(\phi) \sim \begin{cases}\phi^{-p} & s=1\\ \phi^{-p}\displaystyle\sum_{\psi=t_i+s-2}^{\phi-1} S_{s-1}(\psi) & s>1\end{cases} \tag{13}$$

Tomando $p\to 0$ o $\bar{p}\to 1$ $\phi^{-p}\Rightarrow 1$, $\phi^{-p}=1\Rightarrow S_2(\phi)=1$, entonces:

$$S_s(\phi) = \sum_{\psi=t_i+s-2}^{\phi-1} S_{s-1}(\phi) \qquad s>2$$

$s=3: S_3(\phi)=\sum_{\psi=t_i+1}^{\phi-1} 1 = \phi - t_i - 1$

$s=4: S_4(\phi)=\sum_{\psi=t_i+2}^{\phi-1}(\psi-t_i-1) = \frac{1}{2}(\phi-t_i-2)(\phi-t_i-1)$

$s=5: S_5(\phi)=\sum_{\psi=t_i+3}^{\phi-1}(\psi-t_i-2)(\psi-t_i-1) = \frac{1}{6}(\phi-t_i-3)(\phi-t_i-2)(\phi-t_i-1)$

$$\Rightarrow\ S_s(\phi) = \frac{1}{(s-2)!}\frac{(\phi-t_i-1)!}{(\phi-t_i-(s-1))!}$$

$$\begin{aligned}
\sum_{\phi=t_i+s-1}^{s_i} S_s(\phi) &= \frac{1}{(s-2)!}\sum_{\phi=t_i+s-1}^{u_i}\frac{(\phi-t_i-1)!}{(\phi-t_i-(s-1))!} \\
&= \frac{1}{(s-2)!}\frac{(u_i-t_i)!(-u_i+s+t_i-2)}{(s-1)(u_i-s-t_i+2)!} \\
&= \frac{1}{(s-1)!}\frac{\Delta t!(\Delta t-s+2)}{(\Delta t-s+2)!} \\
&= \frac{1}{(s-1)!}\frac{\Gamma(\Delta t+1)}{\Gamma(\Delta t-s+2)} \\
&= \frac{1}{(s-1)!}\frac{\Gamma(\Delta t+1-s+2+s-2)}{\Gamma(\Delta t-s+2)} \\
&= \frac{1}{(s-1)!}\frac{\Gamma(\Delta t-s+2+(s-1))}{\Gamma(\Delta t-s+2)} \\
&\sim \frac{1}{(s-1)!}(\Delta t-s+2)^{s-1}
\end{aligned}$$

Entonces obtenemos el 'valor específico' para las sumas de las Ec. 10 y 10:

$$\sum_{\phi=t_i+s-1}^{u_i} S_s(\phi) \overset{p\to 0}{=} \frac{(\Delta t-s+2)^{n-1}}{(s-1)!} \tag{14}$$

$s>1$:

$$\begin{aligned}
P(s_i(u_i)=s) &\sim (s-1)!\bar{p}^{s-1}t_i^{\bar{p}}u_i^{-s\bar{p}}\sum_{\phi=t_i+s-1}^{u_i} S_s(\phi) \\
&= (s-1)!t_i^{\bar{p}}u_i^{-s\bar{p}}\frac{(\Delta t-s+2)^{s-1}}{(s-1)!} \qquad \bar{p}=1 \\
&= t_i u_i^{-s}(\Delta t-s+2)
\end{aligned}$$

$$P(s_i(u_i)=s) \overset{p\to 0}{\sim} t_i u_i^{-s}(\Delta t-s+2)^{s-1} \tag{15}$$

Para $p>0$ hay que calcularlo numéricamente.

¿Cuál es la escala de la desviación del número acumulado de apariciones de palabras individuales del valor esperado?

Valor esperado: $s_i^*(t)=(t/t_i)^{1-p}$. El valor esperado aumenta más lentamente para t_i más grades, entonces la desviación de esas palabras deberá ser mas chica que para

palabras con tiempo de la primera aparición, t_i, más chicos, si tuviéramos el mismo Δt.

Normalizando el tamaño de la desviación con la dependencia de t_i utilizando distintos Δt_i:

$$u_i = t_i + \underbrace{\Delta t_i}_{\text{período de observación}} = \overbrace{\lambda}^{\text{constante (para cada palabra)}} t_i$$

Entonces,

$$s_i^* = \left(\frac{\lambda t_i}{t_i}\right)^{1-p} = \lambda^{1-p} \overset{p\to 0}{=} \lambda$$

$\downarrow$

valor de referencia para medir la escala de la desviación para cada palabra

Reemplazando $n = x\lambda^{\bar{p}}$ (x veces el valor de referencia) en (15):

$$\underbrace{P(s_i(u_i) = x\lambda)}_{p\to 0 \Rightarrow \lambda^{1-p}\sim\lambda} \overset{p\to 0}{\sim} t_i(\lambda t_i)^{-x\lambda}[(\lambda-1)t_i - x\lambda + 2)^{x\lambda-1}$$
$$= \lambda^{-1}\left(1 - \frac{1}{\lambda} - \frac{x - 2/\lambda}{t_i}\right),$$

Donde x es la escala de la desviación y $s = x\lambda^{\bar{p}} = x\lambda^{1-p} = x\ s_i^*(u_i)$ es el valor esperado. Por otro lado $s = s_i(u_i)$ es el valor real, y entonces $s_i(u_i) = xs_i^*(u_i)$,

$$x = \frac{s_i(u_i)}{s_i^*(u_i)}.$$

Para la mayoría de las palabras se asume $t_i \gg x \sim 1 \Rightarrow \frac{x-2/\lambda}{t_i} \sim 0$. Por lo tanto se obtiene que la distribución de probabilidad de la fluctuación para todas las palabras resulta:

$$P(s_i(u_i) = x\lambda) \overset{p\to 0}{\sim} \frac{1}{\lambda - 1}\left(1 - \frac{1}{\lambda}\right)^{x\lambda},$$

la cual decae exponencialmente y no depende de t_i.

D

Un proceso de Yule-Simon con memoria

Asumimos un estado inicial con n_0 palabras. A cada paso temporal t hay dos opciones: introducir una palabra nueva (probabilidad p) o copiar una palabra ocurrida anteriormente (probabilidad $\bar{p} = 1-p$). En esta última opción queda por determinar cual de las palabras anteriores será copiada. La probabilidad que la palabra copiada sea la ocurrida a tiempo $t - \Delta t$ es:

$$Q(\tau) = \frac{C(t)}{\kappa_c + \Delta t} \tag{D.1}$$

donde κ_c es una escala de tiempo característica sobre la cual palabras recientemente añadidas tienen probabilidades comparables de ser elegidos y $C(t)$ es una normalización logarítmica dado por la condición de normalización,

$$\begin{aligned} 1 &= \sum_{\Delta t=1}^{t-1} Q(\Delta t) \\ &= \sum_{\Delta t=1}^{t-1} \frac{C(t)}{\kappa_c + \Delta t} \\ \implies C(t) &= \left(\sum_{\Delta t=1}^{t-1} \frac{1}{\kappa_c + \Delta t} \right)^{-1} \end{aligned}$$

Modelo de Simon: La probabilidad de elegir una palabra existente, la cual ya ha ocurrido s veces al tiempo t, es $\bar{p}s\pi(s,t)$, donde $\pi(s,t)$ es la fracción de palabras con frecuencia s a tiempo t.

Preferential Attachment: Si el histograma de frecuencias de palabras que han sido copiadas, donde el peso de la contribución de cada palabra es pesada de acuerdo con el factor $1/\pi(s,t)$, es directamente proporcional a s, entonces hay crecimiento preferencial.

A partir de ahora se va a usar: $C = C(t)$ y $\alpha(t) = \alpha \equiv \bar{p}C(t)$.

$P(\Delta t)$: probabilidad de que la próxima ocurrencia de i sea en $t + \Delta t$.

Palabra $i \longrightarrow$ ocurrió por primera vez a tiempo t_i:

$P(\Delta t) = P(s_i(u_i) \to 2)$: prob. que la próxima ocurrencia de i sea en $t_i + \Delta t = u_i$

$\boxed{\Delta t = 1}$:

$$P(1) = \left\{ \begin{array}{c} \text{Prob. que se repita una} \\ \text{palabra ya existente} \end{array} \right\} \left\{ \begin{array}{l} \text{Prob. que se repita la} \\ \text{palabra anterior} = Q(1) \end{array} \right\}$$

Entonces,

$$\begin{aligned} P(1) &= \bar{p}\frac{C}{\kappa_c + 1} \\ &= \frac{\alpha}{\kappa_c + 1} \end{aligned}$$

$\boxed{\Delta t > 1}$:

$$P(\Delta t) = \left\{ \begin{array}{c} \text{Prob. de no elegir } i \text{ por} \\ \Delta t - 1 \text{ pasos consecutivos} \end{array} \right\} \left\{ \begin{array}{c} \text{Prob. de elegir} \\ i \text{ en } t + \Delta t \end{array} \right\}$$

$$\begin{aligned} P(\Delta t) &= \left\{ \prod_{\tau=1}^{\Delta t-1} [p + \bar{p}(1 - Q(\tau))] \right\} \bar{p} Q(\Delta t) \\ &= \left\{ \prod_{\tau=1}^{\Delta t-1} \left[p + \bar{p}\left(1 - \frac{C(t+\Delta t)}{\kappa_c - \tau}\right)\right] \right\} \bar{p}\frac{C(t+\Delta t)}{\tau_c - \Delta t} \\ &= \left(\frac{\bar{p}C}{\tau_c + \Delta t}\right) \prod_{\tau=1}^{\Delta t-1} \left[p + \bar{p}\left(1 - \frac{C}{\kappa_c + \tau}\right)\right] \qquad \Delta t << t_i \Rightarrow C = cte \end{aligned}$$

El primer factor que multiplica a la productoria es la probabilidad de copiar una palabra anterior ($\bar{p}$) por la probabilidad de copiar la palabra ocurrida Δt pasos atrás ($Q(\Delta t)$). Cada factor de la productoria es la probabilidad de que todas las palabras hasta $\Delta t - 1$ sean nuevas más la probabilidad de copiar palabras anteriormente ocurridas en los $\Delta t - 1$ pasos, pero que no sean la palabra i.

Recordando que $\alpha(t) = \alpha \equiv \bar{p}C(t)$ reescribimos $P(\Delta t)$ como:

$$P(\Delta t) = \left(\frac{\alpha}{\tau_c + \Delta t}\right) \prod_{\tau=1}^{\Delta t-1} \left(p + \bar{p} - \frac{\alpha}{\kappa_c + \tau}\right)$$
$$= \left(\frac{\alpha}{\tau_c + \Delta t}\right) \prod_{\tau=1}^{\Delta t-1} \left(1 - \frac{\alpha}{\kappa_c + \tau}\right)$$

Veamos la productoria:

$$\prod_{\tau=1}^{\Delta t-1} \left(1 - \frac{\alpha}{\kappa_c + \tau}\right) = \frac{(-\alpha - \kappa_c + 1)_{\Delta t-1}}{(\kappa_c + 1)_{\Delta t-1}}$$

Donde $(x)_n \equiv \frac{\Gamma(x+n)}{\Gamma(x)}$ es el símbolo de Pochhammer (rising factorial).

$$\Rightarrow \prod_{\tau=1}^{\Delta t-1} \left(1 - \frac{\alpha}{\kappa_c + \tau}\right) = \frac{\Gamma(-\alpha + \kappa_c + 1 + \Delta t - 1)}{\Gamma(-\alpha + \kappa_c + 1)} \frac{\Gamma(\kappa_c + 1)}{\Gamma(\kappa_c + 1 + \Delta t - 1)}$$
$$= \frac{\Gamma(-\alpha + \kappa_c + \Delta t)}{\Gamma(-\alpha + \kappa_c + 1)} \frac{\Gamma(\kappa_c + 1)}{\Gamma(\kappa_c + \Delta t)}$$

Entonces,

$$P(\Delta t) = \bar{p} Q(\Delta t) \prod_{\tau=1}^{\Delta t-1} \left(1 - \frac{\alpha}{\kappa_c + \tau}\right)$$
$$= \bar{p} \frac{C}{\kappa_c + \Delta t} \frac{\Gamma(-\alpha + \kappa_c + \Delta t)\Gamma(\kappa_c + 1)}{\Gamma(-\alpha + \kappa_c + 1)\Gamma(\kappa_c + \Delta t)}$$

Usando que $(x + a)\Gamma(x + a) = \Gamma(x + a + 1)$:

$$P(\Delta t) = \alpha \frac{\Gamma(-\alpha + \kappa_c + \Delta t)\Gamma(\kappa_c + 1)}{\Gamma(-\alpha + \kappa_c + 1)\Gamma(\kappa_c + \Delta t + 1)}$$
$$= \alpha \frac{\Gamma(\kappa_c + \Delta t + 1 - (\alpha - 1))}{\Gamma(\kappa_c + \Delta t + 1)} \frac{\Gamma(\kappa_c + 1)}{\Gamma(-\alpha + \kappa_c + 1)}$$

Tomando $\lim_{t_i \to \infty}$ y $\lim_{u_i \to \infty}$, y usando que $\lim_{t \to \infty} \frac{\Gamma(t-a)}{\Gamma(t)} \sim t^{-a}$:

$$P(\Delta t) = \alpha(\Delta t + \kappa_c + 1)^{-\alpha-1}(\kappa_c + 1)^{\alpha}$$

Finalmente tomado $u_i = t_i + \Delta t >> \Delta t >> 1$:

$$P(\Delta t) \simeq \alpha(1+\kappa_c)^{\alpha}(\kappa_c + \Delta t)^{-\alpha-1}. \tag{D.2}$$

La dependencia temporal de $P(\Delta t)$ es a través de α, entonces la distribución de probabilidad de intervalos Δt es no estacionaria.

Por simplicidad se toma $\kappa_c = 0$.

Tiempo característico de retorno: $\langle \Delta t \rangle$:

$$\begin{aligned}
\langle \Delta t \rangle &= \sum_{\Delta t=1}^{t} P(\Delta t)\Delta t \\
&\simeq \sum_{\Delta t=1}^{t} \alpha(1+\kappa_c)^{\alpha}(\kappa_c+\Delta t)^{-\alpha-1}\Delta t \\
&= \sum_{\Delta t=1}^{t} \alpha \Delta t^{1-\alpha} \\
&\simeq \frac{\alpha}{1-\alpha} t^{1-\alpha}
\end{aligned}$$

Frecuencia s_i de la palabra i:

$$\frac{ds_i}{dt} = \bar{p}\Pi_i \tag{D.3}$$

Donde Π_i es la probabilidad de elegir una aparición previa de la palabra i.

La probabilidad de elegir una palabra ocurrida Δt paso atrás es: $\frac{C}{\kappa_c + \Delta t}$ con $\kappa_c = 0$. Queremos la probabilidad de que a tiempo t se copie la palabra ocurrida en t_i: $\Delta t = t - t_i$. Entonces,

$$\Pi_i = C \sum_{j=1}^{s_i} \frac{1}{t - t_j^{(i)}} \tag{D.4}$$

donde $t_j^{(i)}$ con $j = 1, ..., s_i$ son los tiempos donde apareció la palabra i y s_i la cantidad de veces que i se repitió.

Campo medio: suponiendo que la sumatoria puede escribirse como,

$$s_i \left\langle \frac{1}{t - t_j} \right\rangle_i$$

donde el último factor es el valor medio de $(t - t_j^{(i)})^{-1}$ sobre los tiempos de aparición $t_j^{(i)}$.

Entonces,

$$\Pi_i \simeq C s_i \left\langle \frac{1}{t - t_j} \right\rangle_j \tag{D.5}$$

Suposición:$\langle\rangle_j$ es dominado por la contribución de la aparición más reciente a tiempo t_{s_i}: $\langle (t - t_j)^{-1} \rangle_j \simeq (t - t_{s_i})^{-1}$, entonces tenemos que,

$$\begin{aligned} \left\langle \frac{1}{t - t_j} \right\rangle_j &\simeq \frac{1}{t - t_{s_i}} \\ &\simeq \frac{1}{\langle \Delta t \rangle} \\ &= \frac{1 - \alpha}{\alpha} \frac{1}{t^{1-\alpha}} \end{aligned}$$

De los $\Delta t = t - t_j$, $t - t_{s_i}$ es el más chico $\Rightarrow (t - t_{s_i})^{-1}$ es el dominante de los Δt^{-1}. Vemos que $\left\langle \frac{1}{t - t_j} \right\rangle_j$ tiene una dependencia de Ley de potencias ($\alpha \gtrsim 0$) en t y una dependencia temporal más lenta (logarítmica) a través de α.

Esta expresión captura correctamente la dependencia temporal si se introduce un factor constante Ω:

$$\left\langle \frac{1}{t - t_j} \right\rangle_j \simeq \frac{1 - \alpha}{\alpha} \frac{1}{t^{1-\alpha}} \tag{D.6}$$

Ω: La necesidad viene de las suposiciones simplificantes, especialmente de la aproximación de campo medio.

Introduciendo las Ec. (D.6) y (D.5) en la Ec. (D.3):

$$
\begin{aligned}
\frac{ds_i^*}{dt} &= \bar{p}\Pi_i \\
&\simeq \bar{p}\, C\, s_i^* \left\langle \frac{1}{t-t_j} \right\rangle_j \\
&\simeq \bar{p}\, C\, s_i^* \, \frac{1}{\Omega} \, \frac{1-\alpha}{\alpha} \, \frac{1}{t^{1-\alpha}} \\
&= \alpha\, s_i^* \, \frac{1}{\Omega} \, \frac{1-\alpha}{\alpha} \, \frac{1}{t^{1-\alpha}}
\end{aligned}
$$

$$
\frac{ds_i^*}{dt} \simeq \frac{s_i^*}{\Omega}\,(1-\alpha)\,t^{\alpha-1} \tag{D.7}
$$

Integrando la Ec.(D.7):

- Despreciando la dependencia temporal lenta de α.
- Límites de integración:
  - Desde el tiempo t_i, cuando la palabra i aparece por primera vez (correspondiente a $s = 1$).
  - A tiempo final t, cuando la palabra i ha alcanzado frecuencia s_i^*.

$$
\begin{aligned}
\frac{ds_i^*}{dt} &= \frac{s_i^*}{\Omega}\,(1-\alpha)\,t^{\alpha-1} \\
\frac{ds_i^*}{s_i^*} &= \frac{(1-\alpha)}{\Omega}\,t^{\alpha-1}\,dt \\
\int_1^{s_i^*} \frac{ds_i^{*\prime}}{s_i^{*\prime}} &= \frac{(1-\alpha)}{\Omega}\int_{t_i}^{t} t'^{\alpha-1}\,dt' \\
\ln s_i^{*\prime}\Big|_1^{s_i^*} &= \frac{1-\alpha}{\alpha\Omega}\,t'^{\alpha}\Big|_{t_i}^{t} \\
\ln s_i^* &= \frac{1-\alpha}{\alpha\Omega}(t^\alpha - t_i^\alpha)
\end{aligned}
$$

Definiendo: $K = \frac{1-\alpha}{\alpha\Omega}$ y $A = e^{Kt^\alpha}$.

$$
\begin{aligned}
s_i^* &= e^{\frac{1-\alpha}{\alpha\Omega}t^\alpha}\; e^{-\frac{1-\alpha}{\alpha\Omega}t_i^\alpha} \\
&= A\, e^{-Kt_i^\alpha}
\end{aligned}
$$

Despejando t_i:

$$\ln\left(\frac{s_i^*}{A}\right) = -K\, t_i^\alpha$$
$$t_i^\alpha = \frac{\ln\left(s_i^*/A\right)}{K}$$
$$t_i = \left[\frac{\ln(s_i/A)}{K}\right]^{1/\alpha}$$

Queremos ver la probabilidad de que la palabra i tenga un número de apariciones s_i^* menor a un cierto s, es decir, queremos calcular $P[s_i^*(t) < s]$.

$$s_i^* < s$$
$$-K\, t_i^\alpha < \ln(s/A)$$
$$t_i > \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}$$
$$\implies P[s_i^*(t) < s] = P\left(t_i > \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right)$$

Las palabras se agregan a intervalos temporales iguales, entonces los t_i tienen una densidad de probabilidad constante,

$$P(t_i) = \frac{p}{n_0 + pt}$$

donde t_i es el tiempo de la primera aparición de la palabra i, y como no se introducen palabras nuevas en cada paso temporal se debe multiplicar por p.

$$
\begin{aligned}
P\left(t_i < \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right) &= \int_0^{\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}} P(t_i)dt_i \\
&= \int_0^{\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}} \frac{p}{n_0+pt}dt_i \\
&= \frac{p}{n_0+pt}\int_0^{\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}} dt_i \\
&= \frac{p}{n_0+pt}\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha} \\
\Rightarrow P\left(t_i > \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right) &= 1 - P\left(t_i < \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right) \\
&= 1 - \frac{p}{n_0+pt}\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}
\end{aligned}
$$

Por último,

$$
\begin{aligned}
P(s) &= \frac{\partial P}{\partial s}[s_i^* < s] \\
&= \frac{\partial}{\partial s}P\left(t_i > \left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right) \\
&= \frac{\partial}{\partial s}\left\{1 - \frac{p}{n_0+pt}\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha}\right\} \\
&= -\frac{1}{\alpha}\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha-1}\frac{s}{A}\left(\frac{-1}{s^2}\right)\frac{p}{n_0+pt}
\end{aligned}
$$

Entonces la Distribución de densidad de probabilidad de frecuencia de palabras es:

$$
\boxed{P(s) = \frac{p}{(n_0+pt)(K\alpha)s}\left[\frac{\ln(A/s)}{K}\right]^{1/\alpha-1}} \tag{D.8}
$$

Bibliografía

[1] Duncan J Watts. A twenty-first century science. *Nature*, 445(7127):489–489, 2007.

[2] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[3] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

[4] C. Sire and S. Redner. Understanding baseball team standings and streaks. *The European Physical Journal B*, 67(3):473–481, 2009.

[5] Y. de Saá Guerra, J.M. Martín González, S. Sarmiento Montesdeoca, D. Rodríguez Ruiz, A. García-Rodríguez, and J.M. García-Manso. A model for competitiveness level analysis in sports competitions: Application to basketball. *Physica A: Statistical Mechanics and its Applications*, 391(10):2997 – 3004, 2012.

[6] Alexander M. Petersen, Woo-Sung Jung, and H. Eugene Stanley. On the distribution of career longevity and the evolution of home-run prowess in professional baseball. *EPL (Europhysics Letters)*, 83(5):50010, 2008.

[7] E. Bittner, A. Nubaumer, W. Janke, and M. Weigel. Football fever: goal distributions and non-gaussian statistics. *The European Physical Journal B*, 67(3):459–471, 2009.

[8] E. Ben-Naim, S. Redner, and F. Vazquez. Scaling in tournaments. *EPL (Europhysics Letters)*, 77(3):30005, 2007.

[9] A. Heuer and O. Rubner. Fitness, chance, and myths: an objective view on soccer results. *The European Physical Journal B*, 67(3):445–458, 2009.

[10] Haroldo V. Ribeiro, Satyam Mukherjee, and Xiao Han T. Zeng. Anomalous diffusion and long-range correlations in the score evolution of the game of cricket. *Phys. Rev. E*, 86:022102, Aug 2012.

[11] Li-Gong Xu, Ming-Xia Li, and Wei-Xing Zhou. Weiqi games as a tree: Zipf's law of openings and beyond. *EPL (Europhysics Letters)*, 110(5):58004, 2015.

[12] Frédéric Prost. On the impact of information technologies on society: an historical perspective through the game of chess. In Andrei Voronkov, editor, *Turing-100*, volume 10 of *EPiC Series*, pages 268–277. EasyChair, 2012.

[13] Bernd Blasius and Ralf Tönjes. Zipf's Law in the Popularity Distribution of Chess Openings. *Phys. Rev. Lett.*, 103:218701, Nov 2009.

[14] Haroldo V. Ribeiro, Renio S. Mendes, Ervin K. Lenzi, Marcelo del Castillo-Mussot, and Luís A. N. Amaral. Move-by-move dynamics of the advantage in chess matches reveals population-level learning of the game. *PLoS ONE*, 8(1):e54165, 01 2013.

[15] Mariano Sigman, Pablo Etchemendy, Diego Fernandez Slezak, and Guillermo A. Cecchi. Response time distributions in rapid chess: A large-scale decision making experiment. *Frontiers in Neuroscience*, 4:1, October 2010.

[16] Philippe Chassy and Fernand Gobet. Measuring chess experts' single-use sequence knowledge: An archival study of departure from 'Theoretical' openings. *PLoS ONE*, 6(11), November 2011. PMID: 22110590 PMCID: PMC3217924.

[17] J. I. Perotti, H.H. Jo, A. L. Schaigorodsky, and O. V. Billoni. Innovation and nested preferential growth in chess playing behavior. *EPL (Europhysics Letters)*, 104(4):48005, 2013.

[18] Ana L. Schaigorodsky, Juan I. Perotti, and Orlando V. Billoni. Memory and long-range correlations in chess games. *Physica A: Statistical Mechanics and its Applications*, 394(0):304 – 311, 2014.

[19] Ana L. Schaigorodsky, Juan I. Perotti, and Orlando V. Billoni. A study of memory effects in a chess database. *PLOS ONE*, 11(12):1–18, 12 2016.

[20] Zipf, George K. Human behavior and the principle of least effort. Cambridge, (Mass.): Addison-Wesley, 1949, pp. 573. *Journal of Clinical Psychology*, 6(3):306–306, 1950.

[21] George Kingsley Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Univ. Microfilms International, 1984.

[22] K Okuyama, M Takayasu, and H Takayasu. Zipf's law in income distribution of companies. *Physica A: Statistical Mechanics and its Applications*, 269(1):125 – 131, 1999.

[23] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, Oct 2014.

[24] Edward T. Lu and Russell J. Hamilton. Avalanches and the distribution of solar flares. *The Astrophysical Journal*, 380:L89, October 1991.

[25] G. Udny Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):21–87, 1925.

[26] B. Gutenberg and C. F. Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185, 1944.

[27] G. Neukum and B.A.(1) Ivanov. Crater size distributions and impact probabilities on earth from lunar, terrestrial planeta, and asteroid cratering data, 1994. LIDO-Berichtsjahr=1994,.

[28] Damián H Zanette and Susanna C Manrubia. Vertical transmission of culture and the distribution of family names. *Physica A: Statistical Mechanics and its Applications*, 295(1):1 – 8, 2001. Proceedings of the IUPAP International Conference on New Trends in the Fractal Aspects of Complex Systems.

[29] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[30] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

[31] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[32] Marcelo A. Montemurro and Pedro A. Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461, 2002.

[33] Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, 4(11):e7678, 11 2009.

[34] Eduardo G Altmann and Martin Gerlach. *Statistical Laws in Linguistics*, pages 7–26. Springer International Publishing, 2016.

[35] Eduardo G. Altmann and Giampaolo Cristadoro. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

[36] Álvaro Corral. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys. Rev. Lett.*, 92:108501, Mar 2004.

[37] M. S. Wheatland, P. A. Sturrock, and J. M. McTiernan. The waiting-time distribution of solar flare hard x-ray bursts. *The Astrophysical Journal*, 509(1):448, 1998.

[38] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, May 2013.

[39] Francesc Font-Clos, Gemma Boleda, and Álvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New Journal of Physics*, 15(9):093033, 2013.

[40] C. Cattuto, V. Loreto, and V. D. P. Servedio. A yule-simon process with memory. *EPL (Europhysics Letters)*, 76(2):208, 2006.

[41] Jan Beran. *Statistics for Long-Memory Processes*. Chapman-Hall, 1st edition, 1994.

[42] Richard G. Clegg. A practical guide to measuring the Hurst parameter. *21st UK Performance Engineering Workshop, School of Computing Science Technical Report Series, CSTR-916, University of Newcastle*, pages 43–55, 2006.

[43] Jan W Kantelhardt, Eva Koscielny-Bunde, Henio H.A Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3):441–454, 2001.

[44] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of dna nucleotides. *Phys. Rev. E*, 49:1685–1689, Feb 1994.

[45] Y. Shao, G. Gu, Z. Jiang, W. Zhou, and D. Sornette. Comparing the performance of FA, DFA and DMA using different synthetic long-range correlated time series. *Scientific Reports*, 2:835, Nov 2012/online.

[46] Per Bak, Kim Christensen, Leon Danon, and Tim Scanlon. Unified scaling law for earthquakes. *Phys. Rev. Lett.*, 88:178501, Apr 2002.

[47] Diego Rybski, Sergey V. Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A. Makse. Communication activity in a social network: relation between long-term correlations and inter-event clustering. *Scientific Reports*, 2:560, August 2012.

[48] Justin Keat, Pamela Reinagel, R.Clay Reid, and Markus Meister. Predicting every spike: A model for the responses of visual neurons. *Neuron*, 30(3):803 – 817, 2001.

[49] Albert-Laszlo Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[50] Hang-Hyun Jo, Juan I Perotti, Kimmo Kaski, and Janos Kertesz. Correlated bursts and the role of memory range. *arXiv preprint arXiv:1505.02758*, 2015.

[51] K.-I. Goh and A.-L. Barabasi. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

[52] Armin Bunde, Jan F. Eichner, Jan W. Kantelhardt, and Shlomo Havlin. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett.*, 94:048701, Jan 2005.

[53] Eun-Kyeong Kim and Hang-Hyun Jo. Measuring burstiness for finite event sequences. *Phys. Rev. E*, 94:032311, Sep 2016.

[54] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, May 2005.

[55] Francesc Font-Clos Álvaro Corral. *Self-Organized Criticality Systems (Chapter 5)*. Dr.Markus J. Aschwanden (Ed.), 2014.

[56] Sergei Maslov. Viewpoint: Power laws in chess. *Physics*, 2:97, Nov 2009.

[57] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003.

[58] Claude E. Shannon. Programming a computer for playing chess. *Philosophical Magazine*, 41:256–275, 1950.

[59] `http://scid.sourceforge.net/`.

[60] Didier Sornette. Multiplicative processes and power laws. *Physical Review E*, 57(4):4811, 1998.

[61] Didier Sornette and Rama Cont. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *Journal de Physique I*, 7(3):431–444, 1997.

[62] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1).

[63] Konstantin Klemm, Víctor M Eguíluz, and Maxi San Miguel. Scaling in the structure of directory trees in a computer cluster. *Physical review letters*, 95(12):128701, 2005.

[64] F Tria, V Loreto, V D P Servedio, and SH Strogatz. The dynamics of correlated novelties. *Scientific Reports*, 4:5890, July 2014.

[65] Alexander Gelbukh and Grigori Sidorov. *Zipf and Heaps Laws' Coefficients Depend on Language*, pages 332–335. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[66] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLOS ONE*, 5(12):1–11, 12 2010.

[67] Yasuhiro Hashimoto. Growth fluctuation in preferential attachment dynamics. *Phys. Rev. E*, 93:042130, Apr 2016.

[68] `http://scidvspc.sourceforge.net/`.

[69] Derek de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[70] Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

[71] Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

[72] Damián Zanette and Marcelo Montemurro. Dynamics of text generation with realistic zipf's distribution. *Journal of quantitative Linguistics*, 12(1):29–40, 2005.

[73] Jozef Barunik and Ladislav Kristoufek. On hurst exponent estimation under heavy-tailed distributions. *Physica A: Statistical Mechanics and its Applications*, 389(18):3844–3855, 2010.

[74] Mark E. Glickman. Chess rating systems. *American Chess Journal*, 3:59–102, 1995.

[75] R. M. Bryce and K. B. Sprague. Revisiting detrended fluctuation analysis. *Scientific Reports*, 2(315):1–6, 2012.

[76] Ana V. Coronado and Pedro Carpena. Size effects on correlation measures. *Journal of Biological Physics*, 31(1):121–133, Jan 2005.

[77] Hernán A. Makse, Shlomo Havlin, Moshe Schwartz, and H. Eugene Stanley. Method for generating long-range correlations for large systems. *Phys. Rev. E*, 53:5445–5449, May 1996.

[78] Ana L. Schaigorodsky, Juan I. Perotti, Nahuel Almeira, and Orlando V. Billoni. Short-ranged memory model with preferential growth. *Phys. Rev. E*, 97:022132, Feb 2018.

[79] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[80] Petr Lansky, Federico Polito, and Laura Sacerdote. Generalized nonlinear yule models. *Journal of Statistical Physics*, 165(3):661–679, Nov 2016.

[81] Alessandro Garavaglia, Remco van der Hofstad, and Gerhard Woeginger. The dynamics of power laws: Fitness and aging in preferential attachment trees. *Journal of Statistical Physics*, 168(6):1137–1179, Sep 2017.

[82] M.V. Simkin and V.P. Roychowdhury. Re-inventing willis. *Physics Reports*, 502(1):1 – 35, 2011.

[83] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, May 2001.

[84] S. Lehmann, A. D. Jackson, and B. Lautrup. Life, death and preferential attachment. *EPL (Europhysics Letters)*, 69(2):298, 2005.

[85] R Lambiotte. Activity ageing in growing networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(02):P02020, 2007.

[86] Matú š Medo, Giulio Cimini, and Stanislao Gualdi. Temporal effects in the growth of networks. *Phys. Rev. Lett.*, 107:238701, Dec 2011.

[87] S. Pigolotti, A. Flammini, M. Marsili, and A. Maritan. Species lifetime distribution for simple models of ecologies. *Proceedings of the National Academy of Sciences*, 102(44):15747–15751, 2005.

[88] Barbara Drossel. Biological evolution and statistical physics. *Advances in Physics*, 50(2):209–295, 2001.

[89] K. Sneppen, P. Bak, H. Flyvbjerg, and M H Jensen. Evolution as a self-organized critical phenomenon. *Proceedings of the National Academy of Sciences*, 92(11):5209–5213, 1995.

[90] Michael Drmota. Combinatorics and asymptotics on trees. *Cubo Journal*, 6:2004, 2004.

[91] Jeffrey D. Scargle. Studies in astronomical time series analysis. v. bayesian blocks, a new method to analyze structure in photon counting data. *The Astrophysical Journal*, 504(1):405, 1998.

[92] M. S. Wheatland. The origin of the solar flare waiting-time distribution. *The Astrophysical Journal*, (536):L109–L112, 2000.

[93] C. Padrón Sancho. Juego de Reyes. *Historia y Vida*, vol. 494, May 2009.

[94] Hal Stern. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103 – 117, 1992.