UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

POSGRADO EN CIENCIAS FÍSICAS
INSTITUTO DE FÍSICA

POSIBLE ESTRUCTURA TERMODINÁMICA SUBYACENTE A LAS LEYES DE ZIPF Y BENFORD

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS FÍSICAS

PRESENTA:
CARLO ANDRÉS ALTAMIRANO ALLENDE

TUTOR
ALBERTO ROBLEDO NIETO
INSTITUTO DE FÍSICA

COMITÉ TUTOR:
PIERRE DENIS BOYER
OCTAVIO MIRAMONTES VIDAL
INSTITUTO DE FÍSICA

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO
FEBRERO 2019

RESUMEN

Esta investigación demuestra que las leyes de Zipf y Benford, obedecidas por decenas de datos numéricos generados por muchos y diversos tipos de fenómenos naturales y actividades humanas, están relacionadas con la expresión focal de una estructura termodinámica generalizada. Esta estructura se obtiene a partir de un tipo deformado de mecánica estadística que surge cuando el espacio de fase configuracional es visitado de forma incompleta de manera estricta. Específicamente, la restricción es que la fracción accesible de este espacio tiene propiedades fractales. Esta expresión corresponde con una transformación (incompleta) de Legendre entre dos potenciales de entropía (o de Massieu) que cuando se particulariza a los primeros dígitos lleva a una generalización previamente existente de la ley de Benford. La función inversa de esta expresión conduce a la ley de Zipf; pero naturalmente incluye las curvas o colas observadas en datos reales para rangos pequeños y grandes. Notablemente, encontramos que todo el problema es análogo a la transición al caos a través de la intermitencia exhibida por los mapas no lineales de baja dimensión.

ÍNDICE

1. INTRODUCCIÓN

Durante más de medio siglo, la asombrosa ubicuidad de las leyes empíricas de Zipf y Benford ha dejado perplejos a un sinnúmero de observadores, dada su aparente validez universal. Como ya se conoce ampliamente, la ley de Zipf se refiere al comportamiento aproximado en ley de potencia que exhiben diversos conjuntos de datos al ser ordenados en rango dependiendo de su tamaño (en relación con el tamaño de la población, la frecuencia de las palabras, la magnitud de los factores de impacto, etc.) [1]. Este comportamiento indica que la mayoría de los objetos del conjunto de datos tendrán un tamaño menor, mientras que pocos exhibirán un gran tamaño. Uno de los ejemplos más antiguos y más estudiados es la distribución de ingresos, observada por Pareto, quien en 1906 consideró que la frecuencia en su distribución seguía un comportamiento en ley de potencia. Sin embargo, este comportamiento se ha observado en distintos conjuntos de datos como poblaciones de ciudades, palabras en textos, factores de impacto de revistas científicas, etc. Se considera a la ley de Pareto como el análogo de la ley de Zipf de variable continua.

De igual manera, aunque de forma menos conocida, existe un comportamiento similar observado en tablas numéricas que comprenden una amplia gama de fenómenos conocido como Ley de Benford, en honor a Frank Benford, quien en 1938 publicó [2] un compendio de datos a los que les asoció empíricamente una regla logarítmica simple para la frecuencia de aparición del

primer dígito en listas de datos como el mercado bursátil, censos de población, capacidades térmicas de los productos químicos, etc.

El presente trabajo estudia estos dos comportamientos, que pueden ser caracterizados en una primera aproximación por una ley de potencias, y explora una posible relación termodinámica de origen, de forma natural a este comportamiento.

En una investigación inicial [15] se relacionó a las leyes de Zipf y Benford con una estructura sustentada y fundamentada en una mecánica estadística subyacente correspondiente a una generalización de la mecánica estadística clásica de Boltzmann-Gibbs.

Más aún, se ha argumentado que la ley de Benford es un caso especial de la ley de Zipf [3]; de hecho, la relación entre ambas se ha derivado de forma explícita hace algunos años [4] mediante la primera generalización de la ley de Benford a partir de la suposición básica de que la distribución de probabilidad subyacente $P(N)$ de los datos N considerados es invariable en escala y, por lo tanto, tiene la forma de la ley de potencias

$$P(N) \sim N^{-\propto}, \propto \geq 1. \qquad (1)$$

Una simple integración sobre $P(N)$, para obtener la probabilidad relativa para números enteros consecutivos n y $n+1$, conduce, cuando $\alpha = 1$, a

$$\pi(n) = log\left(\frac{n+1}{n}\right), \qquad (2)$$

que es La ley de Benford.

El siguiente paso en el trabajo de Pietronero *et al.* [4] fue la obtención del rango k de $P(N)$, esta vez como una integración sobre $P(N)$ desde $N(k)$, el número de datos para el rango k, hasta un número finito N_{max} que corresponde al primer valor del rango. En el límite $N_{max} \to \infty$ obtienen:

$$N(k) \sim k^{1/(1-\alpha)}, \qquad (3)$$

que es la ley de Zipf con exponente $1/(1-\propto)$, cuando $\propto > 1$. Para muchos conjuntos de datos reales $\alpha \simeq 2$ y la ley estándar de Zipf, en la forma representada por la ecuación (3), es para $\alpha = 2$ [4].

En este trabajo nos dimos a la tarea de ampliar e ir más allá de los resultados reportados por Pietronero y colegas [4]. Nuestro primer paso, simple, es mantener N_{max} finito, pero como veremos más adelante, esta consideración facilita la articulación de una inferencia mayor sobre la naturaleza física de las leyes de Zipf y Benford. Aquí sostenemos que estas leyes están relacionadas con una expresión termodinámica general, aunque para un tipo especial de estructura termodinámica obtenida a partir del camino usual a través de un parámetro de deformación escalar representado por la potencia α. La expresión termodinámica general se ve representada por una transformación (incompleta) de Legendre (similar a una energía libre de Landau o una función de densidad de energía libre) entre dos potenciales termodinámicos, y la expresión que relaciona las funciones de partición correspondientes se convierte entonces en una ley generalizada de Zipf. En este trabajo se identifican estas cantidades, así

como las variables conjugadas implicadas, que son el rango k y el inverso del número total de datos N^{-1}.

Razonamos también que este tipo de termodinámica deformada surge de la existencia de un fuerte impedimento para acceder a un espacio de fase configurable, que se materializa en un único subconjunto fractal o multifractal de este espacio disponible para el sistema. Una consecuencia cuantitativa de considerar N_{max} finita es la reproducción de la curva de rango pequeño mostrada por los datos reales antes de que se establezca el comportamiento de ley de potencia. El régimen de la ley de potencia en la expresión teórica persiste hasta el rango infinito $k \to \infty$, indicando una especie de 'límite termodinámico'. Esta característica la ilustramos al compararla con los datos disponibles para las frecuencias de las palabras en inglés en textos escritos en dicha lengua [5]. Para efectos de este trabajo de investigación, nos referimos a la aplicación de este esquema al grado de distribución en las redes libres de escala.

Un desarrollo subsecuente es la identificación de una analogía estricta entre la mencionada expresión termodinámica y la de (todas) las trayectorias en la transición al caos a través de la intermitencia en mapas no lineales de baja dimensión; la llamada bifurcación tangente [6]. Estas trayectorias críticas siguen [7] la forma cerrada exacta del mapa de punto fijo del grupo de renormalización de composición funcional (RG, por sus siglas en inglés) [6,8]. En consecuencia, asociamos la misma estructura mecánico-estadística a la dinámica no lineal en esta transición. Además, examinamos las modificaciones introducidas en la ley

generalizada de Zipf por el correspondiente desplazamiento del mapa de la tangencia al régimen caótico. Éstas consisten en la introducción de un límite superior para el rango *k* y la reproducción de la cola observada para el rango grande en datos reales. Ilustramos nuestro esquema comparándolo con los datos numéricos disponibles para el llamado factor propio, o *eigenfactor*, de las revistas especializadas de física[1], las tasas de producción industrial[2] y las emisiones de carbono[3]. La analogía indica que el valor más común para el índice α debe ser α = 2.

Por último, en esta tesis hacemos uso de la interpretación mecánico-estadística para ampliar nuestro análisis. Suponemos que la transformación de Legendre expresada por estas leyes puede ser finalizada de la manera usual y eliminar la variable $\mathcal{N}^{-1}$ a favor de k.

Para lograr este paso es necesario especificar la función de partición $N_{max}(\mathcal{N}^{-1})$, una característica de los datos disponibles o una prerrogativa del recolector de datos, y evaluar la 'ecuación de estado' $k(\mathcal{N}^{-1})$. Al hacer esto queda claro que la universalidad de las leyes se debe a la forma general de la transformación incompleta de Legendre, mientras que los potenciales

[1] Ver la categoría por temas: Physics, 2007 en http://www.eigenfactor.org/index.php
[2] Ver el CIA World Factbook 2010 en http://www.cia.gov/library/publications/the-world-factbook/rankorder/2089rank.html
[3] Ver el International Energy Annual 2005 en http://www.photius.com/rankings/carbon_footprint_of_countries_per_capita_1980_2005.html

termodinámicos iniciales y transformados son particulares de los datos en cuestión.

Por lo tanto, este trabajo se estructura de la siguiente manera: en el siguiente capítulo se reproducen las expresiones obtenidas de la referencia [4] relevantes para esta investigación. Posteriormente, en el capítulo 3 se describe la estructura mecánico-estadística generalizada observada en estas expresiones. En el capítulo 4 se presentan el paralelismo entre la clasificación de los datos y la dinámica crítica en la bifurcación tangente de mapas no lineales y se describe el efecto de tamaño finito de la primera. En el capítulo 5, se amplía la descripción mecánico-estadística y se extraen conclusiones sobre la aparente universalidad de las leyes empíricas de Zipf y Benford. Se concluye con un resumen y una discusión.

Los resultados de este trabajo ya fueron publicados parcialmente en [9] y en su totalidad en [16].

2. DERIVACIÓN DE LAS LEYES DE ZIPF Y BENFORD

La forma más conocida de ley de Zipf se dedujo inicialmente de manera empírica al encontrarse un cierto orden entre la frecuencia en que aparecen las palabras en textos de la lengua inglesa y el rango que ocupan según esta frecuencia. Al clasificar el conjunto de palabras mediante un ordenamiento rango-frecuencia, es decir, a la palabra más frecuente se le asigna el rango 1, a la segunda más frecuente le corresponde el rango 2, y así sucesivamente dentro de un cierto intervalo de validez, la distribución resultante es de la forma de ley de potencias

$$y \sim x^{-\propto}, \qquad (4)$$

donde y corresponde al número de ocurrencias o frecuencia de la palabra y x corresponde al rango. El exponente $\propto$ es medido de forma empírica y varía en valor según el fenómeno en cuestión. En su caso, Zipf encontró que $\propto = 1$.

Por otro lado, la ley de Benford, también conocida como la “ley del primer dígito”, establece que la distribución en la que aparecen los primeros dígitos en series numéricas en notación decimal, de una gran variedad de fuentes distintas, muestra una marcada asimetría. Esta se traduce en una fuerte disposición de los dígitos pequeños sobre los más grandes. Esto quiere decir que, en general, se encuentra que los primeros tres enteros –1, 2 y 3– tienen una frecuencia de aparición del 60%, mientras que los seis dígitos restantes–4 al 9– aparecen en tan solo un 40% de los casos.

La distribución de probabilidad a la que condujo las observaciones de Benford en 1938 para la ocurrencia del primer dígito es

$$P(n) = \log(\frac{n+1}{n}), \qquad (5)$$

donde $n = 1, 2, 3, 4, 5, 6, 7, 8, 9$ es el primer dígito.

Como se mencionó en la introducción, Pietronero et. al. [4] dedujeron una generalización de la ley de Benford a partir de suponer que la distribución de probabilidad que obedecen los datos numéricos donde se observa dicha ley es invariante ante cambios de escala.

Sea $P(N)$ la distribución de probabilidad asociada al conjunto de datos bajo consideración (por ejemplo, la distribución obtenida a partir de un histograma generado por los datos, un total de $\mathcal{N}$ números, dados por las magnitudes de la población de un conjunto de países). Bajo el supuesto de invariancia de escala, la distribución tiene la forma de una ley de potencia

$$P(N) \sim N^{-\propto}, \qquad (6)$$

con $\propto > 0$.

Dado que para los diferentes números enteros n (del 1 al 9), se tiene la misma probabilidad relativa, los autores concluyeron que la probabilidad $\pi(n)$ de observación del primer dígito n del número N viene dada por

$$\pi(n) = \int_n^{n+1} N^{-\propto} dN = \frac{1}{1-\propto}[(n+1)^{1-\propto} - n^{1-\propto}], \qquad (7)$$

donde $\propto \neq 1$.

En el caso $\propto = 1$, se recupera la expresión conocida para la ley de Benford, dada por la ecuación 5. Es decir:

$$\pi(n) = \int_n^{n+1} N^{-1} dN = \int_n^{n+1} d(\log N) = \log\,(1 + n^{-1}),$$

cuando $\propto = 1$.

Este caso, $\propto = 1$, corresponde a una distribución uniforme en el espacio logarítmico; mientras que la ecuación (7) corresponde a una *ley de Benford generalizada*.

Por otro lado, el conjunto de $\mathcal{N}$ números de datos puede ser clasificado y comparado con el ranking de otro conjunto de también $\mathcal{N}$ números extraídos de la distribución básica $P(N) \sim N^{-\propto}$. En este caso, $P(N)$ es la distribución de probabilidad de observación del número *N* dentro del conjunto discreto de datos $\mathcal{N}$ extraídos de $P(N)$.

Si se toma el caso de población de ciudades, por ejemplo, $\mathcal{N}$ estará dado el número de ciudades dentro del conjunto, mientras que *N* es el tamaño de la población de cada ciudad, por lo que, ordenadas todas las ciudades por rango, la ley de Zipf me indica que la probabilidad de encontrar una ciudad con una cierta población $N(k)$ será mayor para números de rango mayores. Es decir, la ciudad más poblada del conjunto $\mathcal{N}$, con una población de N_{max} corresponde al rango $k = 1$, mientras $k_{max} = \mathcal{N}$.

Es decir, el rango k definido por la ley de Zipf se obtiene al realizar una integración sobre la distribución de probabilidad $P(N)$ desde el valor numérico $N(k)$ correspondiente a la frecuencia de aparición que ocupa el rango k hasta un número finito N_{max}, que corresponde al rango $k = 1$, es decir, $N(k = 1) = N_{max}$.

La distribución cumulativa complementaria $\Pi\,(\mathrm{N(k)}, N_{max})$ está determinada por $P(N)$:

$$\Pi(N(k), N_{\max}) = \int_{N(k)}^{N_{max}} P(N')dN'. \qquad (8)$$

Es decir,

$$\mathrm{P(N)} = -\frac{\partial}{\partial N}\Pi(N(k), N_{\max}). \qquad (9)$$

Por definición, la distribución $\Pi\,(\mathrm{N(k)}, N_{max})$ organiza los datos de acuerdo a su tamaño o magnitud; conforme el valor de *N* disminuye de un valor N_{max},la distribución Π crece, tomando valores entre $\Pi\,(N_{max}, N_{max}) = 0$ a $\Pi\,(N_{max}, N_{max}) = 1$. A partir de esto, se puede asociar esta distribución con $k/\mathcal{N}$. De esta manera, obtenemos el rango de la forma

$$\frac{k}{\mathcal{N}} = \int_{N(k)}^{N_{max}} N^{-\alpha} dN = \frac{1}{1-\alpha}[N_{max}^{1-\alpha} - N(k)^{1-\propto}], \qquad (10)$$

con $\propto \neq 1$, donde N_{max} y $N(k)$ corresponden al rango $k = 1$, y al rango no específico $k > 1$, respectivamente. La normalización de $P(N)$ implica que $\Pi(N_{min}, N_{\max}) = 1$.

Al resolver la ecuación (10) para $N(k)$ en el límite cuando $N_{max} \gg 1$ se obtiene la ley de Zipf en la forma

$$k \sim N(k)^{(1-\propto)}; \qquad (11)$$

con lo cual, al invertir esta relación se obtiene

$$N(k) \sim k^{1/(1-\propto)}. \qquad (12)$$

Esta ecuación es la forma comúnmente conocida de la ley de Zipf con exponente $1/(1-\propto)$ cuando $\propto > 1$.

La ecuación (10) introduce una variable de espacio continuo para el rango k en el cual el primer valor del rango es $k = 0$. Esto es una desviación de la representación comúnmente observada del primer rango $k = 1$ y los siguientes rangos dados por números naturales sucesivos.

Este enfoque corresponde a una descripción de variable continua adecuada para conjuntos de datos muy grandes, y para la cual puede obtenerse la restricción a valores enteros del rango mediante el uso de valores adecuados para los límites inferiores de la integración de $N(k)$ en la ecuación (10).

3. LEYES GENERALIZADAS DE BENFORD Y ZIPF COMO RELACIONES TERMODINÁMICAS

Considere la función logarítmica *q*-deformada

$$\log_q(x) \equiv (1-q)^{-1}[x^{1-q} - 1] \qquad (13)$$

con $q \neq 1$ un número real, y su inversa, la función exponencial *q*-deformada

$$\exp_q(x) \equiv [1 + (1-q)x]^{1/(1-q)} \qquad (14)$$

que se reducen, respectivamente, a las funciones ordinarias logarítmica y exponencial cuando $q = 1$.

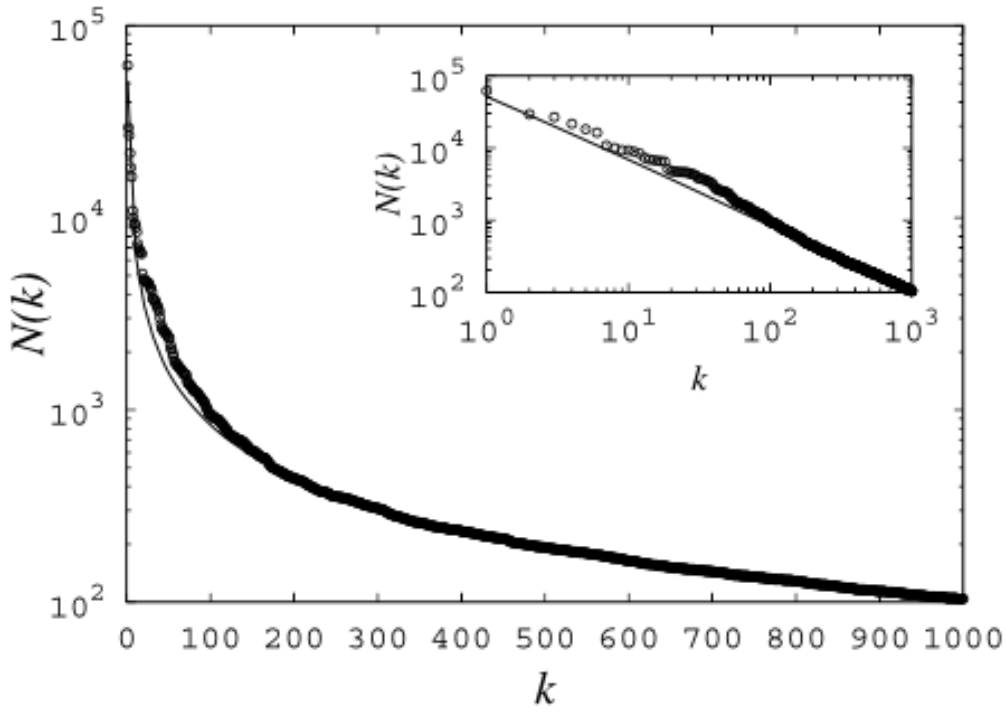

Fig. 1. Estadísticas de orden de rango para la ocurrencia de palabras (círculos vacíos) en el Corpus Nacional Británico [5]. La

ecuación (9) con $\propto = 2.09$ (curva suave) se ajusta a los datos. La línea recta de la inserción se dibuja para fines de visualización.

En términos de estas funciones, la ecuación (10) y su inverso pueden ser escritos más económicamente como

$$\log_\alpha N(k) = \log_\alpha N_{max} - \mathcal{N}^{-1}k \ , \qquad (15)$$

y

$$N(k) = N_{max}\exp_\propto[-N_{max}^{\propto-1}\mathcal{N}^{-1}k]. \qquad (16)$$

Aquí observamos que la ecuación (16) es una generalización de la ley de Zipf que toma en cuenta, apropiadamente, el comportamiento para k de rango bajo observado en datos reales en donde, como se esperaría, N_{max} es finito. En el caso $\propto = 1$, la expresión anterior adquiere la forma exponencial ordinaria

$$N(k) = N_{max}\exp[-\mathcal{N}^{-1}k].$$

En la Figura 1 comparamos los números de ocurrencias de palabras inglesas en un corpus con $N(k)$ según lo dado por la ecuación (16) donde la reproducción de la curva de rango pequeño mostrada por los datos antes de que se establezca el comportamiento como ley de potencia es evidente.

En la expresión teórica, este régimen persiste hasta el rango infinito $k \to \infty$. Alternativamente, podemos recuperar de la ecuación (16) la ley de potencia $N(k) \sim k^{1/(1-\propto)}$ en el límite cuando $N_{max} \gg 1$ cuando $\propto > 1$. Cuando $\propto =$

2 recuperamos la forma clásica de la ley de Zipf $N(k)\sim k^{-1}$. Observamos que para los listados de datos *N* clasificados en términos de rango, la normalización de su distribución $P(N)$ implica que el rango máximo k_{max} es igual al número de datos $\mathcal{N}$. La normalización de $P(N) = N^{-\propto}$ conduce a $k_{max} = \mathcal{N}$ con ambos $k_{max} \to \infty$ y $\mathcal{N} \to \infty$, pero $\mathcal{N}^{-1}k$ generalmente es finito. La suposición de una forma de ley de potencia pura para $P(N)$ no puede representar un conjunto con un número finito de datos.

Ahora, ¿cómo podemos interpretar o asociar la ecuación (15) a un fenómeno físico? Para responder esta pregunta, hay que poner atención a las cantidades contenidas en ella. Observamos que tanto el $\log_{\propto} N_{max}$ como el $\log_{\propto} N(k)$ están dados por las integrales

$$\log_{\propto} N_{max} = \int_1^{N_{max}} N^{-\propto} dN \qquad (17a)$$

y

$$\log_{\propto} N(k) = \int_1^{N(k)} N^{-\propto} dN \ . \qquad (17b)$$

En el mismo sentido, la distribución $P(N) = N^{-1}$ se puede interpretar como el valor de una distribución microcanónica de N configuraciones en el espacio fase, de la cual, cuando $\propto = 1$, se pueden construir las entropías

$$\widehat{S_1} = \log N_{max} = \int_1^{N_{max}} N^{-1} dN \qquad (18a)$$

y,

$$S_1 = \log N(k) = \int_1^{N(k)} N^{-1} dN, \qquad (18b)$$

donde la probabilidad de *N* configuraciones igualmente probables en el espacio fase es $P(N) = N^{-1}$. Ahora, Si permitimos $\propto > 1$, podemos mantener la misma interpretación,

$$\widehat{S_{\propto}} = \log_{\propto} N_{max} \qquad (19a)$$

y

$$S_{\propto} = \log_{\propto} N(k), \qquad (19b)$$

con $P(N) = N^{-\propto}$ todavía vista como la probabilidad de N configuraciones de espacio fase igualmente probables, y con N_{max} y $N(k)$ jugando los roles de números de configuración totales o funciones de partición. Por lo tanto, la ecuación (15) puede reescribirse como

$$S_{\propto} = \widehat{S_{\propto}} - \mathcal{N}^{-1}k, \qquad (20)$$

y por tanto, puede leerse como la expresión de lo que llamamos una transformación de Legendre *incompleta* del potencial de Massieu $\widehat{S_{\propto}}(\mathcal{N}^{-1})$, que es una función del inverso del número $\mathcal{N}$, a la entropía $S_{\propto}(k)$, que es una función del rango *k*. Las variables conjugadas $\mathcal{N}^{-1}$ y k podrían ser vistas, por ejemplo, como tomando el papel de la temperatura inversa β y la energía u en la descripción de un sistema térmico, respectivamente.

Como sabemos, una transformación de Legendre se realiza en dos pasos, el primero es sumar (restar) el producto de dos variables conjugadas de un

potencial termodinámico; y el segundo es eliminar la variable en el primer potencial a favor de la otra variable para obtener el segundo potencial.

El último paso implica la derivada del primer potencial, ya que la transformación de Legendre se asocia a un valor extremo. Pero vale la pena detener el procedimiento en el primer paso y el uso del potencial generalizado que depende de las dos variables conjugadas, y ahora explicaremos el porqué.

Ejemplos familiares de transformaciones incompletas de Legendre son la energía libre de Landau (la descripción de un imán depende tanto de la magnetización como del campo externo) y las funciones de densidad de energía libre asociadas a muchos problemas térmicos. La ecuación (16), siendo la inversa de la ecuación (15), establece la misma relación, pero en términos de las "funciones de partición" $N(k)$ y $N_{max}(\mathcal{N}^{-1})$. La ausencia de un límite superior para el rango *k* indica una condición a la que nos referimos como límite termodinámico en nuestra interpretación mecánico-estadística de la ecuación (15).

Para completar la transformación de Legendre de $\widehat{S_\propto}(\mathcal{N}^{-1})$ en $S_\propto(k)$ y eliminar la variable $\mathcal{N}^{-1}$ a favor de *k*, sería necesario optimizar $S_\propto$, es decir, mediante el uso de una `ecuación de estado'

$$k = \frac{d}{d\mathcal{N}^{-1}} \log_\propto N_{max}(\mathcal{N}^{-1}) \ . \qquad (21)$$

Esta ecuación de estado adquiere sentido más adelante.

4. ANALOGÍA CON LA BIFURCACIÓN TANGENTE

Notablemente, existe una analogía estricta entre la ley generalizada de Zipf, las ecuaciones (15) y (16), y la dinámica no lineal para el mapa de punto fijo RG en la bifurcación tangente, como se anotó de manera original en la referencia [8]. En consecuencia, estos dos problemas en aparentemente diferentes comparten la misma interpretación mecánico-estadística indicada anteriormente, y esta equivalencia ofrece una alternativa para avanzar en nuestro análisis. En específico, en la caracterización de los efectos de tamaño finito para la ley generalizada en términos del desplazamiento del mapa fuera de la tangencia.

La analogía puede ser más evidente después de una revisión inmediata del tratamiento RG de la bifurcación tangente que media la transición entre los atractores caóticos y periódicos [6]. El procedimiento común para estudiar la transición al caos a partir de una trayectoria de período *n* comienza con la composición $f^{(n)}(x)$ de un mapa unidimensional $f(x)$ en dicha bifurcación, seguido de una expansión para la vecindad de uno de los *n* puntos de tangencia a la línea con pendiente unitaria [6]. Con total generalidad se obtiene

$$x' = f^{(n)}(x) = x + ux^z + \ldots, \; z > 1, \qquad (22)$$

donde $x^z \equiv \text{sign}(x)|x|^z$. El mapa de RG de punto fijo es la solución $f^*(x)$ de

$$f^*\big(f^*(x)\big) = \lambda^{-1} f^*(\lambda x) \qquad (23)$$

junto con un valor específico para λ que al expandirse alrededor de $x = 0$, reproduce la ecuación (22). Se obtuvo una expresión analítica exacta para $f^*(x)$

en la referencia [8] con el uso de la propiedad de traducción de una variable auxiliar $y = x^{1-z}$. Esta propiedad está escrita como

$$x'^{1-z} = x^{1-z} + (1-z)u \qquad (24)$$

o, de forma equivalente, como

$$x' = x\exp_z(ux^{z-1}). \qquad (25)$$

Directamente se puede corroborar que $x' = f^*(x)$, dado por la ecuación (25), satisface la ecuación (23) con $\lambda = 2^{1/(z-1)}$. La iteración repetida de la ecuación (24) conduce a

$$x_t^{1-z} = x_0^{1-z} + (1-z)ut \qquad (26)$$

o,

$$\log_z x_t \equiv \log_z x_0 + ut\ . \qquad (27)$$

De forma que la dependencia del número de iteración, o tiempo *t*, de todas las trayectorias viene dada por

$$x_t = x_0\exp_z[x_0^{z-1}ut], \qquad (28)$$

donde x_0 son las posiciones iniciales. Las propiedades *q*-deformadas de la bifurcación tangente se discuten con mayor amplitud en la referencia [7]. El paralelo entre las ecuaciones (27) y (28) con las ecuaciones (15) y (16), respectivamente, es claro y evidente, y por lo tanto, podemos concluir con seguridad que el sistema dinámico representado por el mapa de punto fijo $f^*(x)$ opera de acuerdo con la misma propiedad mecánico-estadística descrita en la sección anterior para las leyes generalizadas.

Para enfatizar que la analogía es firme, y va mucho más allá que una simple semejanza casual, entre el ranking de los datos y las secuencias de iteraciones en la bifurcación tangente, mostramos que hay una fuente común detrás de las ecuaciones (15) y (27). Es decir, nos referimos a la restricción de la accesibilidad al espacio de fase previamente mencionada. Esto se ve fácilmente al considerar el reemplazo, válido para un tiempo τ largo, de la diferencia $x_{\tau+1} - x_\tau$ por $dx_\tau/d\tau$ en la ecuación (22), escrito como

$$x_{\tau+1} - x_\tau = u x_\tau^z \ . \qquad (29)$$

La integral de lado izquierdo de la ecuación (29) de la forma diferencia resultante

$$\frac{dx_\tau}{x_\tau^z} = u d\tau \qquad (30)$$

entre x_0 y x_t; y la integral del lado derecho de *0* a *t* lleva inmediatamente a las ecuaciones (26) y (27). La cantidad x_τ^z en la ecuación (30) toma el mismo papel que la distribución de ley de potencias $P(N) \sim N^{-\propto}$.

Observamos que la ausencia de un límite superior para el rango *k* en las ecuaciones (15) y (16) es equivalente a la condición de tangencia en el mapa. En consecuencia, observamos los cambios en $N(k)$ provocados por el desplazamiento del mapa de tangencia correspondiente (ver Fig. 2); es decir, consideramos las trayectorias x_t con las posiciones iniciales x_0 del mapa

$$x' = x \exp_z(u x^{z-1}) + \epsilon, \ \ 0 < \epsilon \ll 1 \qquad (31)$$

con las identificaciones $k = t$, $\mathcal{N}^{-1} = -u$, $N(k) = x_t + x^*$, $N_{max} = x_0 + x^*$ y $\propto = z$, donde la traslación x^* asegura que todos los $N(k) \geq 0$.

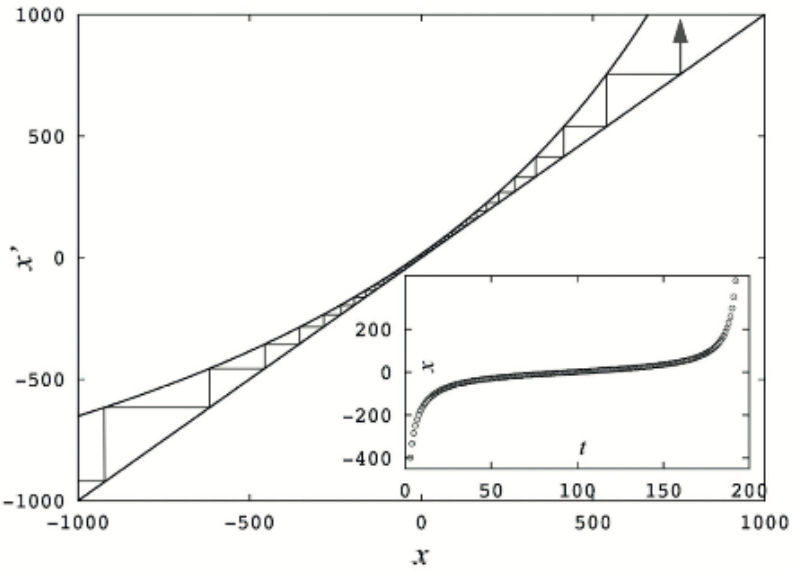

Fig 2. El mapa en la ecuación (31) con una trayectoria. El recuadro muestra la dependencia temporal de la trayectoria.

En las Figuras 3 a 5 ilustramos la capacidad de este enfoque para reproducir datos cuantitativamente reales para clasificar los factores propios (o *eigen*factores) (una medida del valor total) de las revistas especializadas de física, de las tasas de crecimiento de la producción industrial por país y de las emisiones de dióxido de carbono per cápita por país o región, respectivamente. En la ruta de intermitencia fuera del caos es relevante determinar la duración de los llamados episodios laminares [6], es decir, el tiempo promedio que pasan los trayectos por el "cuello de botella" formado en la región donde el mapa está más cerca de la línea de pendiente unitaria. Naturalmente, la duración de los episodios laminares diverge en la bifurcación tangente cuando desaparece el exponente de Lyapunov para la separación de trayectorias.

Curiosamente, es esta propiedad de la dinámica no lineal la que se traduce en las propiedades de tamaño finito ($k_{max} < \infty$) de la función de rango de ocurrencia $N(k)$, que hemos obtenido sin averiguar los detalles de la salida de la distribución básica $P(N)$ de la ley de potencias $N^{-\propto}$ pura. Otro resultado importante que se desprende de la analogía entre la dinámica no lineal y la ley de rango es que el valor más común para el grado de no-linealidad en la tangencia es $z = 2$, obtenido cuando el mapa es analítico en $x = 0$ con una segunda derivación distinta de cero, lo cual implica $\alpha = 2$, cerca de los valores observados para la mayoría de los conjuntos de datos reales.

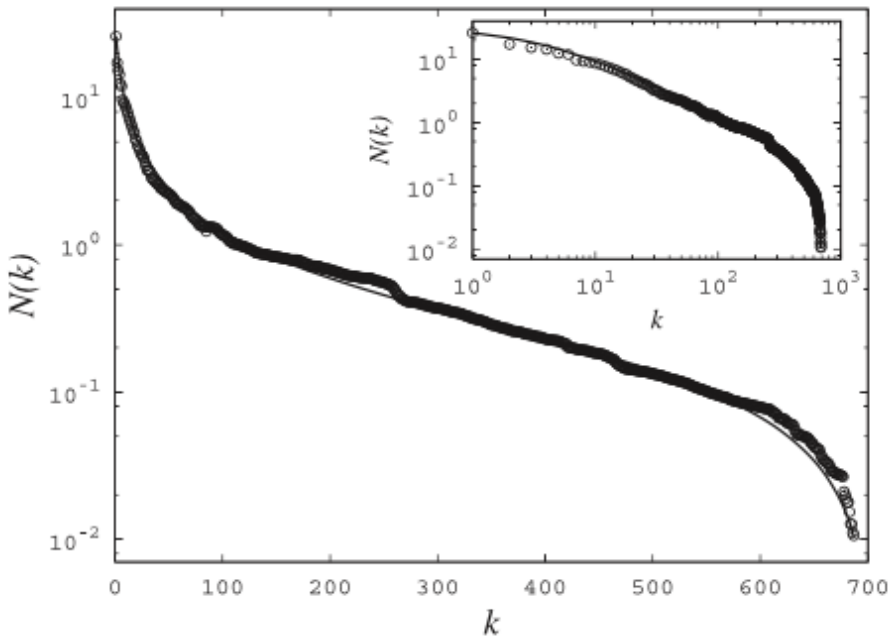

Fig 3. Ordenamiento de rango-frecuencia para el *eigenfactor* de revistas especializadas de Física (círculos vacíos) obtenidos de[1]. Se ajusta la ecuación (31) a los datos con los parámetros $\propto= 2.01$ y $\varepsilon =$

-0.00064 (curva suave). El recuadro muestra la misma gráfica en escala log-log.

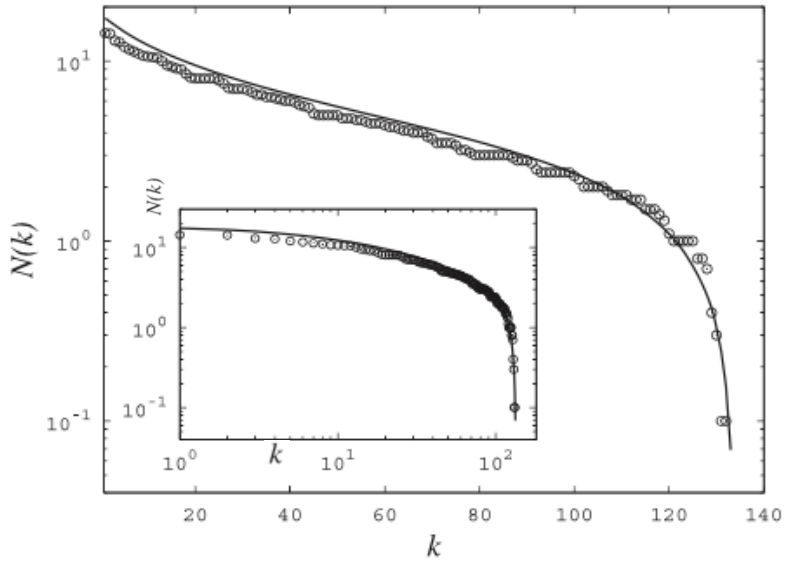

Fig 4. Ordenamiento de rango-frecuencia para las tasas de crecimiento de la producción industrial por país (círculos vacíos) tomados de[2]. Se ajusta la ecuación (31) a los datos con los parámetros $\propto= 2.13$ y $\varepsilon = -0.058$ (curva suave). El recuadro muestra la misma gráfica en escala log-log.

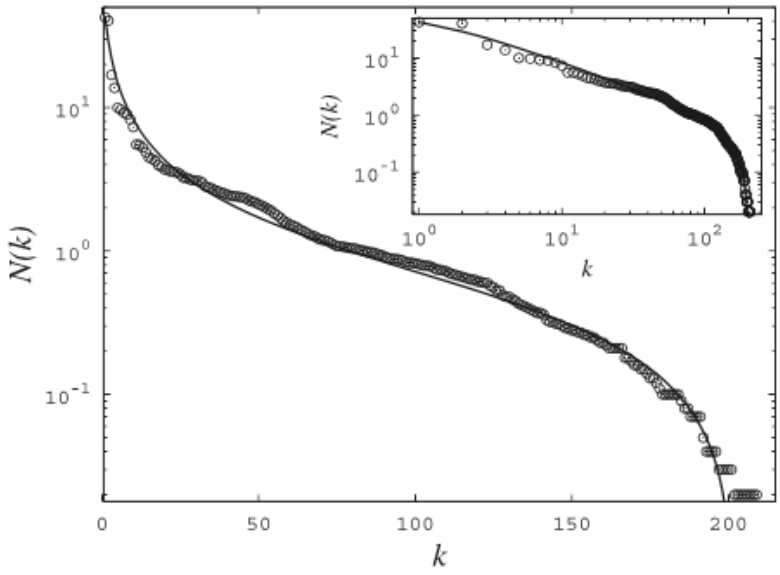

Fig 5. Ordenamiento de rango-frecuencia para las emisiones de dióxido de carbono per capita (círculos vacíos) tomados de[3]. Se ajusta la ecuación (31) a los datos con los parámetros $\propto = 2.06$ y $\varepsilon = -0.0055$ (curva suave). El recuadro muestra la misma gráfica en escala log-log.

.

.

.

5. UNIVERSALIDAD Y SINGULARIDAD DE LA CLASIFICACIÓN DE LOS DATOS

La aproximación del descenso más pronunciado es central para la mecánica estadística (y en un contexto más general, para la teoría de las grandes desviaciones [10]). Esta propiedad facilita la evaluación en el límite termodinámico de una función de partición para un ensamble particular en términos de la función de partición de otro. Termodinámicamente, esta aproximación se relaciona con la transformación de Legendre entre las energías libres o potenciales de Massieu correspondientes, donde se elimina una variable en favor de su conjugado [11]. Como se ha recordado a lo largo de este trabajo, el procedimiento consiste en dos pasos: la suma (o sustracción) del producto de las variables conjugadas a (o desde) el primer potencial para definir el segundo, seguido por el uso de la derivada del primer potencial, o ecuación de estado, para eliminar la variable no deseada. Esto corresponde a la optimización implicada en el método de descenso más empinado. Para fines ilustrativos, asumiremos aquí que el atajo de descenso más pronunciado que subyace en el segundo paso de la transformación de Legendre también es significativo para $\alpha > 1$.

Para realizar el segundo paso de la transformación de Legendre indicado por la ecuación (20) necesitamos una forma explícita para la función $N_{max}\ (\mathcal{N}^{-1})$. Es evidente que la forma de esta función no es única y está

determinada por el conjunto particular de datos $\mathcal{N}$. Para efectos ilustrativos, consideramos un conjunto finito de datos $\mathcal{N}$ extraídos de $P(N) = N^{-\propto}$, aunque una ley de potencia pura no es la distribución correcta en este caso. Sin embargo, si la ecuación del mapa equivalente (31) está muy cerca de la tangencia $\varepsilon \ll 1$ y los datos para el rango máximo, $N_{min} = N(k_{max})$, se eligen de tal manera que su imagen en el mapa está a la izquierda y cerca del punto medio del cuello de botella, entonces $P(N)$ está estrechamente aproximada por la ley de potencias $N^{-\propto}$.

Bajo esta aproximación, la normalización de $P(N)$ sólo produce $k_{max} \sim \mathcal{N}$. Supongamos que los datos disponibles, o la elección del recopilador de datos, fija el valor específico de k_{max} y los límites inferior y superior de la ecuación (2) en N_{min} y N_{max}, respectivamente. Por lo tanto, tenemos

$$k_{max} = \mathcal{N} \int_{N_{min}}^{N_{max}} N^{-\propto} dN = \frac{\mathcal{N}}{1-\propto} [N_{max}^{1-\propto} - N_{min}^{1-\propto}] \ , \qquad (25)$$

con $\propto \neq 1$, o

$$N_{max} = N_{min} \exp_{\propto} [N_{min}^{1-\propto} \mathcal{N}^{-1} k_{max}] \ . \qquad (26)$$

Por ejemplo, un conjunto de datos sobre la población de las ciudades puede estar representado por $k_{max} = 50$ (cincuenta tamaños de ciudad representativos), $N_{min} = 1$ (una ciudad con la población más grande) y $N_{max} = 100$ (cien ciudades con la población más pequeña considerada). La ecuación (26) es la expresión necesaria para $N_{max}(\mathcal{N}^{-1})$ que debe utilizarse en la

`condición de mayor pendiente' o en la `ecuación de estado' (14). El resultado se debe seguir inmediatamente, se trata de $k = k_{max}$.

Como en la termodinámica ordinaria, observamos que la universalidad de las leyes descritas por las ecuaciones (15) y (16) se debe a la forma general de la transformación incompleta de Legendre, mientras que las formas específicas adoptadas por los potenciales dados por las ecuaciones (19a) y (19b):

$$\widehat{S_{\propto}} = \log_{\propto} N_{max}(\mathcal{N}^{-1})$$

y

$$S_{\propto} = \log_{\propto} N(k),$$

son particulares al sistema o a la situación en cuestión.

6. CONCLUSIONES

A lo largo de este trabajo de investigación, se ha ofrecido una interpretación o comprensión termodinámica, o mecánico-estadística, novedosa de las leyes generalizadas de Benford y Zipf. Las expresiones para estas leyes, ecuaciones (3) y (7) (o alternativamente (16)) fueron derivadas por los autores de la referencia [4] bajo la suposición básica de que los conjuntos de datos obedecidos por estas leyes se reproducen estadísticamente bien cuando se extraen de una distribución de leyes de potencia $P(N) \sim N^{-\propto}$.

Señalamos aquí que la desviación de la unidad del exponente $\propto$ implica un acceso restringido al espacio de fase para las configuraciones de datos que cuando se enumeran producen los números N. Esta restricción implica un subconjunto accesible de este espacio con una propiedad invariante de escala; es decir, un conjunto fractal, tal y como implica la ley de potencias $N^{-\propto}$. Este punto de vista se hace evidente cuando se considera que $P(N)$ representa la distribución de probabilidad de N configuraciones igualmente probables en el espacio de fase para los datos y, en consecuencia, sugiere la definición de las entropías generalizadas en las ecuaciones (19a) y (19b).

Es importante aclarar que la estructura mecánico-estadística considerada aquí y obtenida a partir del parámetro habitual de deformación escalar (representado por la potencia $\propto$) no se ajusta a la conocida como estadística no-extensiva [12,13]. Aunque aquí definimos entropías o potenciales de Massieu

con el uso de la función *q*-logarítmica y hacemos uso de su inversa, la *q*-exponencial, no requerimos ni implicamos la optimización de ninguna de estas cantidades mediante el uso de las restricciones empleadas en el formalismo de la mecánica estadística no-extensiva ni implicamos el uso de las llamadas distribuciones *escort* [13].

La clasificación de los datos reales muestra habitualmente desviaciones del régimen de la ley de potencia de Zipf tanto para los rangos pequeños como para los grandes, lo que puede ser observado claramente en las gráficas de escala semi-logarítmica. Como hemos mostrado en la Figura 1, la ley generalizada de Zipf dada por la ecuación (16) es capaz de reproducir con precisión la desviación de bajo rango, pero no la de gran rango como el régimen de ley de potencia en esta ecuación se extiende a $k \to \infty$.

Un límite superior para k sugiere efectos de tamaño finito inherentes a los datos reales. Hemos capturado la naturaleza del límite superior de k demostrando primero una analogía precisa entre la expresión para las leyes de orden d rango, las ecuaciones (15) y (16), y aquellas para la dinámica en la transición al caos vía intermitencia (la bifurcación tangente) en mapas no lineales de baja dimensión.

Se considera que los efectos de tamaño finito en la clasificación de los datos corresponden al desplazamiento de la tangencia en el mapa, de modo que la posición del límite superior para el rango k viene dada por la duración de los episodios laminares de trayectorias caóticas cercanas a la transición al

comportamiento regular. Curiosamente, la interpretación mecánico-estadística propuesta para la ley generalizada de Zipf se extiende a la dinámica crítica de la transición al caos a través de la intermitencia.

Mientras que, en la práctica, los datos para la clasificación de los datos para todos los k son reproducidos cuantitativamente por nuestro formalismo, como se ilustra, respectivamente, en las Figuras 3 a 5 para tres ejemplos específicos: eigenfactor de revistas especializadas de física, tasas de producción industrial, y emisiones de carbono. De acuerdo con las determinaciones empíricas, la analogía implica que el valor más general para el índice $\propto$ es $\propto = 2$.

Como es bien sabido, una estructura mecánico-estadística (compartida por la teoría de las grandes desviaciones [10]) se construye alrededor de la aproximación de mayor pendiente descendente y se expresa como la propiedad de transformación de Legendre que vincula diferentes potenciales termodinámicos. Esto implica una condición de optimización o ecuación de estado que relaciona variables conjugadas. Sólo con fines ilustrativos hemos asumido que esta estructura se extiende a la versión deformada (con un parámetro escalar) que hemos considerado aquí.

Para replicar las circunstancias normalmente encontradas en termodinámica hemos presentado como ejemplo la forma particular tomada por la función $N_{max}(\mathcal{N}^{-1})$ cuando los datos en cuestión están limitados por los números N_{min} y N_{max} , y fija el rango más grande en k_{max}. Luego, se determinó la ecuación de estado y se eliminó la variable $\mathcal{N}^{-1}$ a favor de k, para obtener el

valor de'equilibrio' para $N(k)$. Este ejercicio sugiere que la universalidad de las leyes se debe a la forma general de la transformación incompleta de Legendre, mientras que las expresiones de los potenciales iniciales y transformados son específicas para el diseño de la muestra de datos en consideración.

La interpretación termodinámica que hemos propuesto puede explicar la presencia constante de estas leyes fenomenológicas en una amplia gama de observaciones, incluyendo situaciones muy disímiles. Finalmente, comentamos que nuestros argumentos también se aplican al tema de las redes libres de escala [14]. Dado que la distribución de grados $p(k)$, la distribución del número k de enlaces que conectan un nodo con otros nodos, describe esencialmente el ranking de nodos según el número de enlaces que poseen, podemos tratar los conjuntos de datos desde donde se obtiene fenomenológicamente esta distribución de forma similar a los conjuntos de datos que conducen a la ley de Zipf. Para redes aleatorias, $p(k)$ decae exponencialmene ($\propto = 1$), pero para redes libre de escala, el comportamiento es una ley de potencias aproximada ($\propto > 1$).

REFERENCIAS

1. G.K. Zipf, *Human Behavior and the Principle of Least-Effort* (Addison-Wesley, 1949)
2. F. Benford, Proceedings of the American Philosophical Society **78**, 551 (1938)
3. See J.G. van der Galien (2003) in http://en.wikipedia. org/wiki/Zipfs_law
4. L. Pietronero, E. Tosatti, V. Tosatti, A. Vespignani, Physica A **293**, 297 (2001)
5. G. Leech, P. Rayson, A. Wilson, *Word Frequencies in Written and Spoken English: based on the British National Corpus* (Longman, London, 2001)
6. H.G. Schuster, *Deterministic Chaos. An Introduction*, 2nd revised edn. (VCH, Weinheim, 1988)
7. F. Baldovin, A. Robledo, Europhys. Lett. **60**, 518 (2002)
8. B. Hu, J. Rudnick, Phys. Rev. Lett. **48**, 1645 (1982)
9. C. Altamirano, A. Robledo, *in Complex Sciences*, LNICST (Springer-Verlag, 2009), Vol. 5, p. 2232
10. H. Touchette, Phys. Rep. **478**, 1 (2009)
11. H.B. Callen, *Thermodynamics and an Introduction to Themostatistics*, 2nd edn. (John Wiley & Sons, New York, 1985)
12. C. Tsallis, J. Stat. Phys. **52**, 479 (1988)
13. C. Tsallis, R.S. Mendes, A.R. Plastino, Physica A **261**, 534 (1998)
14. R. Albert, A. Barabási, Rev. Mod. Phys. **74**, 47 (2002)

15. Altamirano, C. Leyes de potencias bajo la Mecánica Estadística no Extensiva: Ley de Zipf y Ley de Benford. Tesis para obtener el título de Físico, Facultad de Ciencias, UNAM. (2008).
16. C. Altamirano, A. Robledo. **Eur. Phys. J. B** 81, 345–351 (2011)