UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría en Ciencias (Física)

Instituto de Física

Complejidad y diversidad de palabras en publicaciones científicas

Tesis

QUE PARA OPTAR AL GRADO DE:

Maestría en Ciencias (Física)

PRESENTA:

Fís. José Alberto Ruiz Gayosso

DIRECTOR DE TESIS:

Dr. Alejandro Pérez Riascos

Instituto de Física

COMITÉ TUTOR:

Dr. Marcelo del Castillo Mussot

Instituto de Física

Dr. Pedro Miramontes Vidal

Facultad de Ciencias

CIUDAD DE MÉXICO, MÉXICO, ENERO 2023

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca recibida durante la realización de mi maestría. Fue un apoyo económico que me permitió desarrollar mis estudios de forma satisfactoria.

Expreso también mi agradecimiento a mi tutor, Dr. Alejandro Pérez Riascos, por el tiempo y la dedicación puesta en este trabajo. De la misma forma, agradezco al Dr. Marcelo Del Castillo Mussot y al Dr. Pedro Miramontes Vidal, miembros de mi comité tutor, por sus comentarios y sugerencias que permitieron mejorar este trabajo.

Resumen

En este trabajo se estudia la complejidad y la diversidad de palabras en los títulos de artículos publicados en revistas de la American Physical Society (APS) desde el año 1893 utilizando como herramienta principal la teoría de redes; en particular, la detección de comunidades. En la primera parte se presenta una breve revisión de algunos conceptos teóricos sobre redes, detección de comunidades y el algoritmo de Louvain que constituyen los principales métodos de análisis en esta investigación. Se hace una revisión de los artículos científicos que han utilizado esta base de datos en el contexto de la ciencia de redes y la cienciometría, también se realiza un análisis exploratorio para obtener una visión general de la información contenida en los títulos de artículos publicados por la APS. Con respecto al análisis de estas bases de datos, se describe detalladamente el procesamiento de datos para el filtrado y la obtención de las listas de palabras que constituyen la materia prima de la investigación. Posteriormente, utilizando métodos gráficos y la entropía de Shannon, se analiza la evolución de la diversidad a través de los años de este conjunto de palabras. Finalmente, se define una medida de similaridad con la cual se construyen redes de palabras a partir de las listas generadas mediante la introducción de un parámetro adimensional H que controla la escala a la cual se realiza el análisis; utilizando este parámetro y la detección de comunidades en redes se exploran diferentes escalas que revelan la existencia de patrones en los grupos de palabras definidos por medio de comunidades.

Índice general

Índice de figuras

Índice de tablas

Introducción

El estudio de las redes ha sido de gran importancia para el entendimiento de los sistemas complejos ya que que estos últimos pueden ser representados en términos de redes debido a la simplicidad con la que permiten codificar las complicadas interacciones entre los elementos de un sistemas complejo [1]. Diversos sistemas tienen una estructura que permite su representación como redes y; por lo tanto, se pueden estudiar desde el punto de vista de la teoría de redes la cual ha tenido un gran desarrollo durante las últimas dos décadas. En el caso de la investigación científica es posible definir diferentes tipos de sistemas que admiten este tipo de representación; por ejemplo, redes de colaboración entre investigadores o instituciones o redes de publicaciones científicas. En este contexto se han desarrollado diversos estudios que tienen como objetivo estudiar a la ciencia como un sistema complejo y explorar diversos aspectos relacionados con el desarrollo de la ciencia.

El estudio de la investigación científica a partir de los datos es un área fértil en el presente y de particular importancia para el futuro, debido a esto, es necesario el desarrollo de métodos y técnicas que permitan explorar y aprovechar la información contenida en las extensas bases de datos con las que se dispone en la actualidad. Esto permitirá obtener un mejor entendimiento sobre la evolución de la ciencia y sobre las formas en las que se puede mejorar la investigación científica.

Debido al gran número de publicaciones realizadas en revistas de física durante los siglos XX y XXI existe una amplia cantidad de información contenida en los títulos de artículos pertenecientes a diversas disciplinas de la física. Esta información ha sido poco estudiada desde el punto de vista de la cienciometría; sin embargo, es posible construir redes complejas a partir de las palabras que conforman estos títulos y que representan conceptos representativos de la física. Mediante la aplicación de las técnicas desarrolladas para estudiar estas redes es posible explorar las relaciones entre estas palabras y descubrir patrones ocultos que ofrezcan una visión global de la física como un sistema complejo.

El objetivo general de esta investigación es explorar la base de datos de la American Physical Society desde el punto de vista de la complejidad. Para esto, se plantea realizar una descripción multiescala de la información en esta base en términos de redes y medidas de información. Específicamente, se busca extraer o revelar los patrones contenidos en las redes de palabras que pueden construirse a partir de los títulos de artículos publicados en revistas de la APS y explorar la organización de las palabras en las comunidades conforme se cambia la escala en la que se realiza el análisis.

1. Teoría de redes

1.1. Redes

Una red se define como cualquier sistema que puede representarse mediante una gráfica (también llamada grafo) en la cual sus nodos corresponden a los elementos de dicha red y sus aristas a la presencia de interacciones o algún tipo de relación entre estos elementos [2]. El estudio de las redes ha sido de gran importancia para el entendimiento de los sistemas complejos debido a que estos últimos pueden ser representados en términos de redes debido a la simplicidad con la que estas permiten codificar las complicadas interacciones entre los elementos de un sistemas complejo [1]. Existe una gran variedad de sistemas complejos cuyo estudio, desde el punto de vista de la teoría de redes, ha permitido entender su funcionamiento y revelar características no apreciables a simple vista. Entre estos sistemas se encuentran: sistemas de transporte [3], genomas[4], el Internet[5], la WWW[5, 6], las redes cerebrales[7, 8], las redes tróficas[9], así como los sistemas sociales[10] y económicos[11], entre muchos otros. Una característica común en todos estos sistemas es la existencia de elementos individuales cuya interacción, en muchos casos simple, dota al sistema con propiedades no triviales y que emergen de dicha interacción. Estas propiedades, como la heterogeneidad de los grados y las leyes de potencias o el efecto de mundo pequeño están presentes en un buen número de estos sistemas y, en este sentido, representan características universales de las redes complejas [1].

El estudio de las redes, complejas o no, se ha formalizado a través del aŕea de las matemáticas conocida como teoría de gráficas, las gráficas son objetos matemáticos que se utilizan para describir a las redes y están definidas en términos de un conjunto de vértices y un conjunto de aristas [12, 13]. Formalmente una gráfica G consiste del par (V, E) donde V es un conjunto de nodos o vértices y E un conjunto de aristas de la forma (i, j). En la Figura 1.1 se muestra una representación de una gráfica simple en la cual cada círculo de color rojo corresponde a un elemento de la red y la existencia de una línea que une a cualesquiera dos elementos implica una interacción o relación entre estos elementos. Para estudiar una red se define la matriz de adyacencia $\mathbf{A}$ con elementos A_{ij} dados por:

$$A_{ij} = \begin{cases} 1 & \text{si} \quad (i,j) \in E, \\ 0 & \text{si} \quad (i,j) \notin E. \end{cases} \tag{1.1}$$

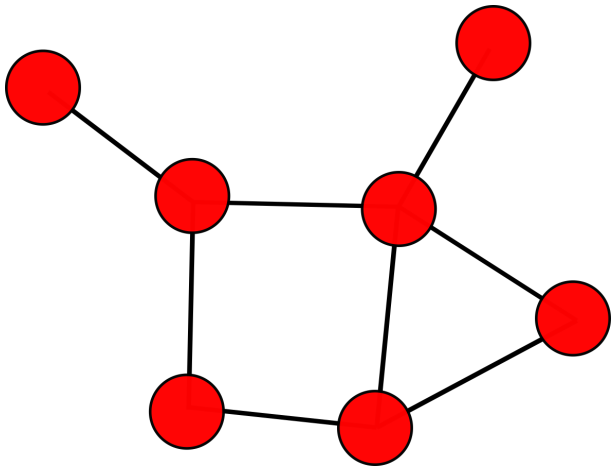

Figura 1.1: Ejemplo de una gráfica (o grafo) simple. En teoría de gráficas, las gráficas o grafos son los objetos matemáticos bajo estudio y permiten modelar relaciones o interacciones entre los elementos de un sistema. Dichos objetos están constituidos por un conjunto de vértices, o nodos, y un conjunto de aristas que representan a los elementos y sus interacciones, respectivamente. La teoría de gráficas es importante en el estudio de las redes complejas pues estas se describen matemáticamente mediante gráficas.

La entrada A_{ij} de la matriz de adyacencia es 1 si existe una conexión o interacción entre los nodos i , j y 0 en caso contrario [12]. Esta interacción puede ocurrir en una sola dirección o en ambas direcciones; en este último caso, la gráfica es dirigida. En casos más generales las conexiones o enlaces entre los nodos pueden ser múltiples o con pesos; por ejemplo, el número de personas que viajan entre dos estaciones de autobuses en una red de transporte o el número de veces que dos palabras aparecen de forma adyacente en un texto en una red de palabras. En estos casos, para considerar la magnitud de las interacciones entre los elementos de la red, se define la matriz de pesos $\mathbf{\Omega}$; en la cual, cada entrada Ω_{ij} especifica la magnitud de la interacción entre los nodos i, j [12].

A partir de la matriz de adyacencia $\mathbf{A}$ se han definido diversas medidas que permiten caracterizar estadísticamente a la red y que pueden generalizarse de forma directa para la matriz de pesos $\mathbf{\Omega}$ [12]. Por ejemplo, una de las principales medidas de caracterización es la importancia de los nodos o aristas. Esta medida se denomina centralidad y expresa la idea de que un nodo o arista es importante o "central" si permite conectar o hacer interactuar a partes distantes de la red. Existen diversas formas de medir la centralidad según las propiedades que se consideren interesantes [12, 13]. La medida más utilizada es el grado de la matriz de adyacencia o de la matriz de pesos. Dada la matriz de adyacencia $\mathbf{A}$ el grado del nodo i de la red se define como:

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{1.2}$$

donde N es igual al número de nodos en la red. La generalización del grado para matrices dirigidas o con pesos puede consultarse en [14].

1.2. Detección de comunidades

Otra parte importante en el estudio de redes es la partición de redes o detección de comunidades. Las comunidades están formadas por grupos de nodos altamente conectados; es decir, conjuntos de nodos cuya principal característica es que presentan un mayor número de conexiones e interacciones entre sí en comparación con el resto de los nodos de la red. Esta partición o división de la red en diferentes grupos es tal que el número de enlaces que conectan a nodos en una misma comunidad es significativamente mayor que el número de enlaces que conectan a diferentes grupos [14]. En aquellas redes que representan a sistemas reales, la detección de comunidades es importante para la identificación de su estructura y organización a grandes escalas [14]; por ejemplo, la detección de comunidades en redes sociales permite identificar grupos de individuos altamente conectados como grupos de amigos y familia.

Existen diversas técnicas para la identificación de comunidades; sin embargo, cada una tiene sus ventajas y desventajas y no existe un algoritmo "universal" que pueda aplicarse en todas la situaciones [13]. Debido a que la mayoría de redes de interés representan sistemas complejos con un número gigante de componentes, el desarrollo de algoritmos para la detección de comunidades es un campo de estudio fértil y relevante en la actualidad. Una gran diversidad de estos algoritmos se han desarrollado para ser aplicados en redes complejas que surgen en diferentes áreas como la biología, ciencias de la computación, economía, ingeniería, política, entre otros [15].

En el problema de detección de comunidades el objetivo es encontrar los grupos que existen de forma intrínseca en la red [16] y por lo tanto el número y tamaño de cada comunidad no son especificados. Por el contrario, el algoritmo debe ser capaz de realizar la identificación de la partición más adecuada considerando únicamente la información contenida en los enlaces que conectan a todos los nodos de la red. Debido a que no existe un consenso respecto al significado de comunidad se han desarrollado una gran variedad de métodos para obtener particiones de una red; sin embargo, un enfoque muy utilizado en la tarea de la obtención de comunidade consiste en realizar una partición de la red mediante la maximización de alguna cantidad que cuantifica la "calidad" de la partición. Por ejemplo, la *modularidad* Q es una medida que compara el número de enlaces dentro de cada una de las comunidades con el número de enlaces dentro de estas mismas comunidades pero en una situación en la cual los nodos son reasignados de forma aleatoria. La modularidad se define como [17]:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \tag{1.3}$$

donde m es el número total de enlaces, δ es la función delta de Kronecker y c_v denota la comunidad que contiene al nodo v. En esta ecuación el factor $\frac{k_v k_w}{2m}$ cuantifica la probabilidad de que los nodos v, w estén conectados si los enlaces son asignados de forma aleatoria. Existen diferentes algoritmos para obtener el número óptimo de comunidades; es decir, el número que maximiza la modularidad.

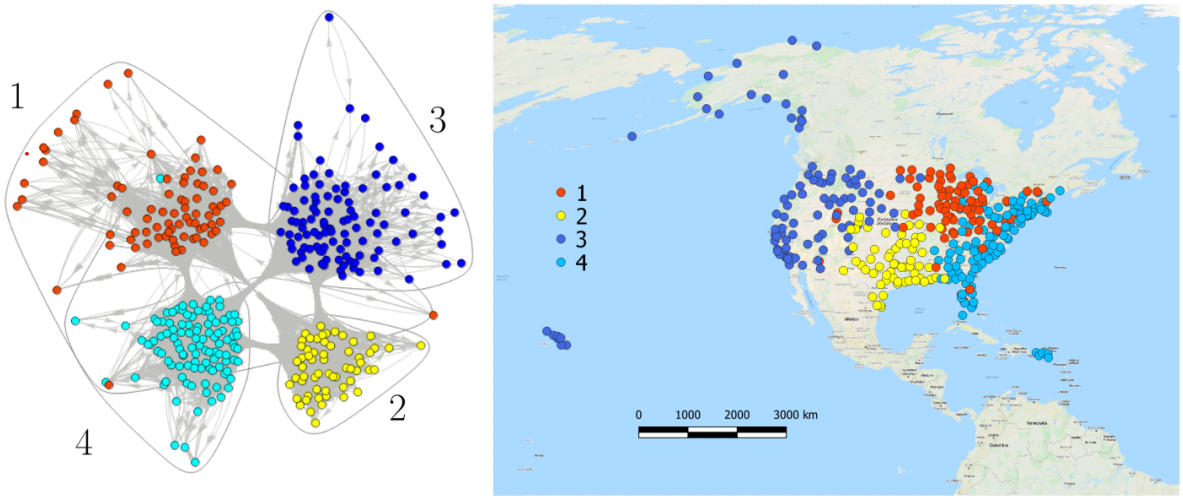

Figura 1.2: Comunidades de la red de aeropuertos de Estados Unidos [18].

En la Figura 1.2 se muestra como ejemplo de una red compleja el sistema de aeropuertos de Estados Unidos y sus respectivas comunidades [18]. En este caso, los aeropuertos corresponden a los nodos de la red y el número de pasajeros entre dos aeropuertos corresponde al peso del enlace entre estos aeropuertos. Para la detección de comunidades se maximiza la modularidad utilizando el algoritmo de Louvain cuya descripción se realiza en la siguiente sección. Este algoritmo arroja como resultado la presencia de cuatro comunidades que al ser graficadas sobre un mapa del territorio estadounidense revela la existencia del mismo número de regiones geográficas (Norte, Sur, Este y Oeste) que separan a este territorio en aŕeas bien definidas mostrando la organización del sistema aeroportuario a una escala de cientos de kilómetros [18]. Nótese que este resultado se obtiene únicamente a partir de la información contenida en la representación del sistema como una red y los patrones de interacción de sus respectivos nodos.

Además de la maximización de la modularidad, existen otros métodos para obtener las comunidades naturales de una red. Uno de estos métodos, está basado en la remoción de enlaces que conectan a nodos en comunidades diferentes, es decir, enlaces que no forman parte de comunidad alguna. Al remover estos enlaces de forma progresiva se tiene como resultado un conjunto de comunidades aisladas. En este método se utiliza el número de caminos más cortos que pasan a través de cada enlace para identificar nodos que yacen entre comunidades. La lógica detrás de este método se basa en que un enlace que conecta a dos comunidades tiene asociado un número mayor de caminos de longitud mínima. También existen otros métodos para la detección de comunidades colectivamente conocidos como de agrupamiento jerárquico (*hierarchical clustering*), en estos algoritmos de tipo aglomerativo se comienza con una partición en la cual cada nodo representa una comunidad; de forma progresiva se crean comunidades con base en una medida de similaridad entre los nodos de la red. Este conjunto de algoritmos corresponde a uno de los más antigüos para la detección de comunidades [14].

1.3. Algoritmo de Louvain

En esta sección se hace una breve descripción del algoritmo de Louvain utilizado para la detección de comunidades en redes con un gran número de nodos y conexiones. La principal ventaja de este algoritmo es su rapidez en la detección de comunidades en redes con un gran número de elementos y en tiempos relativamente cortos; por ejemplo, este algoritmo ha sido utilizado para extraer las comunidades de una red con 118 millones de nodos en tan solo 152 minutos. Puesto que en redes con un número de nodos del orden de los millones se espera la existencia de diferentes niveles de organización (jerarquía), otro aspecto interesante de este algoritmo es que, debido a su implementación, este permite identificar estos niveles de organización de la red obteniendo así una visión completa de la red a diferentes resoluciones [19].

El algoritmo de Louvain consiste de dos fases que se repiten de forma iterativa hasta encontrar la partición óptima en términos de la modularidad Q. Si se considera una red con N nodos o vértices la primera fase comienza con una partición en la cual cada nodo se asigna a una comunidad diferente, es decir, existen tantas comunidades como nodos en la red. Posteriormente, para cada nodo v se considera el conjunto de nodos vecinos w y se calcula el cambio en la modularidad que resulta de mover al nodo i hacia la comunidad de cada uno de sus vecinos. Si para alguno de estos movimientos se obtiene un valor positivo en el cambio ΔQ de la modularidad entonces se escoge el máximo de estos valores y el nodo i se mueve a la comunidad correspondiente. Este paso se repite para cada uno de los nodos de la red hasta que se alcanza un máximo local en la modularidad lo cual ocurre cuando no es posible incrementar la modularidad mediante movimientos individuales de un solo nodo [19].

La eficiencia de este algoritmo radica en la rapidez con la cual se puede calcular el cambio ΔQ de la modularidad que se obtiene al mover un nodo i hacia una comunidad C. Este cambio está dado por [19]:

$$\Delta Q = \left[\frac{\sum_{in}+K_{i,in}}{2m} - \left(\frac{\sum_{tot}+K_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{1.4}$$

donde $\sum_{in}$ es igual a la suma de los pesos de los enlaces contenidos en la comunidad C, $\sum_{tot}$ es la suma de los pesos de los enlaces que inciden sobre todos los nodos en C, k_i es la suma de los pesos de enlaces incidentes sobre i, $K_{i,in}$ la suma de pesos de enlaces desde i hacia nodos dentro de C y m es igual a la suma de los pesos de todos los enlaces en la red. De esta forma, el cálculo del cambio en la modularidad no requiere obtener la modularidad para las dos particiones involucradas (aquella con el nodo i en su comunidad original y aquella en la que el nodo i se ha movido a la comunidad C) que requiere considerar a todos los nodos de la red sino únicamente considerar el nodo i y los nodos dentro de la comunidad C, m asume un valor constante.

En la segunda fase del algoritmo se crea una nueva red a partir de las comunidades obtenidas en la primera fase, cada una de las comunidades pasa a formar un nodo en la

nueva red y los nuevos enlaces entre los nodos asumen un peso que es igual a la suma de los pesos de enlaces que conectan a comunidades diferentes en la red original, los enlaces entre nodos de una misma comunidad en la red original aparecen como bucles en la nueva red. De esta forma la segunda fase únicamente consiste en crear una nueva red a partir de las comunidades de la fase anterior reduciendo en el proceso el número de nodos con respecto a la red original. Una vez creada esta red se puede aplicar nuevamente la primera fase del algoritmo para generar una nueva partición en meta-comunidades y posteriormente la segunda fase para producir una nueva red con un número cada vez más pequeño de nodos. Esto último hace posible la gran eficiencia del algoritmo pues la mayor parte del tiempo de cómputo ocurre durante las primeras iteraciones en las cuales el número de nodos es muy grande [19]. Este procedimiento se repite hasta que no exista un cambio positivo en la modularidad al mover los nodos entre comunidades; es decir, hasta que se alcance un máximo de la modularidad.

Aunque el algoritmo de Louvain es comúnmente utilizado en la optimización de la modularidad, existen otras cantidades paramétricas que pueden optimizarse mediante este algoritmo y que permiten resolver problemas asociados con la modularidad como el llamado límite de resolución [20, 21, 22].

2. Base de datos de la American Physical Society (APS)

2.1. Introducción

En este capítulo se realiza la descripción de la base de datos de la American Physical Society (APS) utilizada en este trabajo. Esta base de datos, que puede solicitarse en el sitio web oficial de la revista [23], contiene información acerca de todos los artículos publicados en revistas pertenecientes a la APS, parte de esta información se remonta hasta el año 1893 y abarca todo el material publicado desde esta fecha hasta el año 2020. Debido a que las revistas de la APS representan una parte significativa de la literatura de la investigación en física, esta base de datos constituye una fuente importante de información que permite estudiar la evolución de la física a lo largo de los siglos XX y XXI.

En la primera parte del capítulo se presenta una revisión de la investigación que se ha realizado utilizando esta base de datos; en particular, en las áreas de la ciencia de redes y la cienciometría. Esta revisión tiene como objetivo presentar el contexto y las diversas formas en las que se ha utilizado esta base de datos y también introducir la motivación para la realización de este trabajo.

En la segunda parte se describe detalladamente el contenido de los datos que se utilizan en este trabajo y que corresponden a los metadatos de todos los artículos publicados en revistas de la APS desde el año 1893 hasta el año 2020. Esta información que incluye, además de otros datos, fecha y título de cada artículo permite estudiar la evolución de los conceptos más importantes de la física a lo largo de los años en términos de su diversidad. También se presentan tablas y figuras con el objetivo de describir de forma general a los datos y de mostrar algunas de las características temporales de cada revista. Además, se realiza un análisis preliminar de las palabras que conforman los títulos de los artículos publicados en cada revista.

2.2. Base de datos de la APS

En esta sección se describe el conjunto de datos de la American Physical Society (APS). Este conjunto de datos se divide en dos partes [24]: la primera consiste de la colección de pares de artículos en revistas de la APS en los cuales uno de los artículos ha sido citado por el otro y permite construir diversos tipos de redes de colaboración científica. La segunda parte consiste de los metadatos correspondientes a cada uno de los artículos publicados en revistas de la APS; por ejemplo, identificador (DOI), revista de publicación, volumen, número, título, número de páginas, autores, afiliación de los autores, números PACS, etc [24].

Este conjunto contiene información de aproximadamente 450,000 artículos publicados en Physical Review (Serie I, Serie II, A, B, C, D y E), Physical Review Letters, Review of Modern Physics y otras revistas de la American Physical Society. Para propósitos de este trabajo únicamente se consideraron los artículos publicados en Physical Review y Physical Review Letters de los cuales se tiene registro desde el año 1893 hasta el año 2020. Los artículos publicados en Review of Modern Physics no se tomaron en cuenta debido a que el número de publicaciones por año es pequeño en comparación con el resto de las revistas y, por lo tanto, no permite obtener resultados estadísticamente significativos. De manera similar, otras revistas no son consideradas en esta investigación debido a que estas fueron creadas en su mayoría en algún momento del siglo XXI por lo que no se dispone de una cantidad suficiente de información. En la tabla 2.1 se presenta, para cada una de las revistas consideradas, los años durante los cuales se publicó o ha sido publicada cada revista y el número total de artículos (incluyendo editoriales y comentarios).

Esta base de datos ha sido utilizada extensamente ya que es posible crear y analizar di-

Tabla 2.1: Revistas de la APS publicadas desde 1893 hasta la fecha. La revista Physical Review ha presentado diversos cambios durante el siglo pasado. En 1958 esta se dividió para dar origen a Physical Review Letters y en 1970 se crearon cuatro revistas para cubrir campos importantes de la física. En 1993, Physical Review A se dividió y se creó Physical Review E. Para cada revista se muestra el nombre, años en los que fue o ha sido publicada, número total de artículos (incluyendo editoriales y comentarios) en la base de datos.

Revista	Años de publicación	Número total de artículos
Physical Review A	1970 - Presente	83632
Physical Review B	1970 - Presente	197564
Physical Review C	1970 - Presente	41895
Physical Review D	1970 - Presente	94833
Physical Review E	1993- Presente	61903
Physical Review Letters	1958 - Presente	129116
Physical Review (Serie I)	1893 - 1912	1469
Physical Review (Serie II)	1913 - 1969	47940
Review of Modern Physics	1929 - Presente	3423

ferentes tipos de redes de colaboración entre autores que han publicado en revistas de la APS. Diversas técnicas y metodologías se han aplicado para analizar la información en este conjunto de datos. Esta información ha permitido estudiar el impacto de la interdisciplinariedad y la diversidad de temas en el éxito o influencia de los artículos en términos del número de citas. Por ejemplo, en [25], los autores analizaron la red de acoplamiento bibliográfico para estudiar el papel que juega la innovación y la interdisciplina en la relevancia de los artículos científicos. El análisis de las propiedades estadísticas de esta red en términos del número absoluto de citas muestra que los esquemas de incentivos actuales promueven, en mayor medida, la producción de artículos que abordan temas convencionales o que han dominado durante un largo periodo de tiempo en comparación con artículos que exploran temas innovadores. En [26], se estudió la correlación entre la diversidad de artículos, medida con el índice de diversidad de Weitzman utilizando como categorías los códigos PACS de cada artículo, y el número de citas. Los resultados obtenidos sugieren que, a largo plazo, los artículos y autores con valores muy bajos o muy altos de diversidad reciben un número de citas pequeño en comparación con aquellos artículos y autores con una diversidad moderada.

También se ha estudiado la relación entre la interdisciplinariedad y el éxito de las carreras científicas. En [27] los autores utilizan códigos PACS para mostrar que una carrera científica exitosa requiere cierto grado de interdisciplinariedad. Adicionalmente, proponen un modelo basado en agentes el cual reproduce los resultados obtenidos con la base de datos de la APS y que incorpora efectos como la aleatoriedad (suerte) y su influencia en el éxito de las carreras científicas. La base de datos de la APS también ha sido utilizada en [28] para analizar características estructurales de la red de cascadas de citas y de la red de cascadas de referencias. Estas cascadas consisten de todos artículos que pueden conectarse mediante relaciones de citación a lo largo de diferentes generaciones. Este tipo de análisis permite entender cómo un artículo en particular puede influenciar la investigación futura en términos de su grado de interdisciplinariedad.

Por otra parte, la base de datos de la APS ha sido utilizada para estudiar las distribuciones de probabilidad del número de citas de artículos publicados por la APS y otras propiedades estadísticas. Por ejemplo, en [29], se analizaron las distribuciones del número de citas de artículos publicados en Physical Review en el periodo 1985-2009. Se encontró que una comparación equitativa de estas distribuciones para diferentes áreas y años requiere de un procedimiento de reescalamiento por el promedio del número de citas. Estos resultados son importantes para la evaluación cuantitativa del desempeño en la investigación. De forma similar, en [30] se analizaron las distribuciones del número de citas en el periodo 1893-2003 y se encontró que una distribución log-normal describe adecuadamente los datos. También se estudiaron otras propiedades estadísticas relacionadas con la edad de los artículos y la evolución temporal de artículos clásicos en términos del número de citas.

El estudio y análisis de propiedades en redes simples y redes multiplex son otros ejemplos en los cuales se ha utilizado esta base de datos. En [31], se estudió la generalización de la paradoja de los amigos en redes de colaboración científica. Esta paradoja establece que en una red de social un individuo tiene, en promedio, un número menor de amigos que sus propios amigos. Se encontró que esta paradoja ocurre a nivel individual (nodos) y nivel

de redes y es válida para varias características como el número de coautores, número de citas y número de publicaciones. En [32] se definió, utilizando métodos de teoría de la información, una red multiplex de colaboración científica y a partir de esta se analizó la relación entre los patrones de colaboración y la organización del conocimiento en las diferentes áreas de la física. La red de institución-citas se analizó en [33] y se propuso un nuevo modelo, IPRank, para cuantificar el impacto de las instituciones y de los artículos en la investigación científica. Este modelo se comparó con los resultados obtenidos de la aplicación del algoritmo PageRank a esta misma red y se encontró que IPRank es un mejor modelo para identificar instituciones y artículos con gran impacto.

Otra área en la que ha analizado la base de datos de la APS es la asignación de crédito en artículos escritos por más de un autor. En [34] se propuso un nuevo modelo para la asignación de crédito en publicaciones con múltiples autores y este fue validado mediante la identificación de artículos cuyos autores han ganado el premio Nobel. De manera similar, en [35], se definió un nuevo método de asignación dinámica de crédito en términos del algoritmo PageRank.

La base de datos de la APS también se ha utilizado para estudiar el fenómeno de hibernación en artículos y autores. En este fenómeno se observa un incremento repentino en el número de citas después de un periodo de tiempo durante el cual el artículo o autor permanecieron en el anonimato. Por ejemplo en [36], se analizó el fenómeno de las llamados *bellas durmientes* (en inglés sleeping beauties) que son artículos cuya importancia es reconocida muchos años después de haber sido publicados. En este trabajo se introdujo una medida no paramétrica para determinar si un artículo puede considerarse en esta categoría o no, incorporando factores como el periodo de hibernación y la intensidad con la que estos artículos emergen. Los resultados obtenidos muestran que este fenómeno no es excepcional y ocurre principalmente en áreas interdisciplinarias. De forma similar, en [37], se analizó este fenómeno para el caso de autores de artículos científicos. Se introdujo un algoritmo basado en "genes sociales" que describen factores asociados a las actividades de cada autor como número de artículos publicados, número de coautores, años en activo, etc. Los resultados mostraron que este algoritmo permite detectar más casos de autores emergentes en comparación con algoritmos basados en caminatas aleatorias en redes bibliográficas.

En este punto es importante mencionar que una de las limitaciones de esta base de datos es que la información corresponde únicamente a artículos y autores que han publicado en alguna de las revistas de la APS. Por lo tanto, las redes de colaboración como la red de citas o de co-citación, construidas exclusivamente con esta base de datos no incluyen información de artículos publicados o autores que han publicado en revistas no pertenecientes a la APS. Esto puede reducir, por ejemplo, el número de citas de un artículo de un carácter interdisciplinario cuyas citas provienen, en mayor medida, de artículos pertenecientes a otras áreas de la ciencia. Esto puede introducir sesgos en favor de artículos poco interdisciplinarios o que abordan temas exclusivos de la física.

Los trabajos en los cuales se ha utilizado y analizado la base de datos de la APS, mencionados en párrafos anteriores, muestran que el estudio y análisis de la ciencia misma es un

Tabla 2.2: Ejemplo extraído del conjunto de datos de PRL. Para este trabajo se consideraron la fecha de publicación, título, volumen y número de autores (no mostrado en la tabla) para cada artículo publicado en revistas de la APS.

date	title	format	volume.number
1958-07-01	Editorial	html+mathml	1
1958-08-01	Investigation of Time-Reversal Invariance in t...	html+mathml	1
1958-08-01	Experimental Limit of the Neutrino Rest Mass	html+mathml	1
1958-08-01	Experimental Evidence for the Influence of Ato...	html+mathml	1
1958-08-01	Cosmic-Ray Increases Produced by Small Solar F...	html+mathml	1
1958-08-01	Frequency of Cesium in Terms of Ephemeris Time	html+mathml	1
1958-08-01	Further Observations on the Nature of the Curr...	html+mathml	1
1958-08-01	Photoproduction of <span class=“aps-inline-for..	html+mathml	1

área importante de trabajo. En efecto, la ciencia de las ciencias en inglés *The Science of Science*(SoS) es un campo que se encarga de analizar aspectos importantes relacionados con la ciencia. En una era digital en la cual la cantidad de datos es cada vez más mayor, la información acerca del financiamiento de la ciencia, la colaboración y productividad científica, datos de citación, y la movilidad de los investigadores permiten entender la estructura y evolución de la ciencia. El objetivo final de la SoS es el desarrollo de herramientas y políticas encaminadas al desarrollo acelerado de la ciencia [38]. La importancia de este tipo de investigación radica en el potencial para proveer soluciones a diferentes problemáticas asociadas con el desarrollo de la ciencia como la selección de candidatos a puestos académicos y de investigación o la asignación de recursos a proyectos específicos.

En particular, el estudio de la ciencia desde el punto de vista de los sistemas complejos y las redes complejas es útil para estudiar la interacción entre diferentes factores que determinan la estructura y evolución de las ciencias y revelar la existencia de propiedades y patrones no evidentes a simple vista que ocurren en diferentes escalas [39]. Lo mencionado anteriormente motivó la realización de este trabajo cuyo objetivo es analizar la evolución y la diversidad de la física en términos de los conceptos (ideas) físicos que han aparecido en revistas de la APS durante los siglos XX Y XXI.

Con el propósito de estudiar la evolución y diversidad de los conceptos de la física a lo largo del tiempo se consideró la información contenida en los títulos de artículos publicados en Physical Review (A,B,C,D,E, serie I y II, Letters). De esta forma, se analizó la base de datos correspondiente a los siguientes metadatos de cada uno de los artículos: fecha de publicación, título, formato, volumen de cada artículo publicado y autores (véase tabla 2.2). Esta información permite extraer las palabras (conceptos de la física) utilizadas en los títulos de artículos publicados en cada uno de los años considerados.

Esta forma de estudiar la diversidad de los artículos de física difiere de los enfoques utilizados en otros trabajos; por ejemplo, en [25, 26, 27] se estudia la diversidad y multidisciplinariedad utilizando una clasificación con códigos PACS. Sin embargo, una limitación importante de esta clasificación es que comenzó a utilizarse a partir del año 1975 y por lo tanto los artículos publicados en fechas anteriores no pueden clasificarse utilizando esta

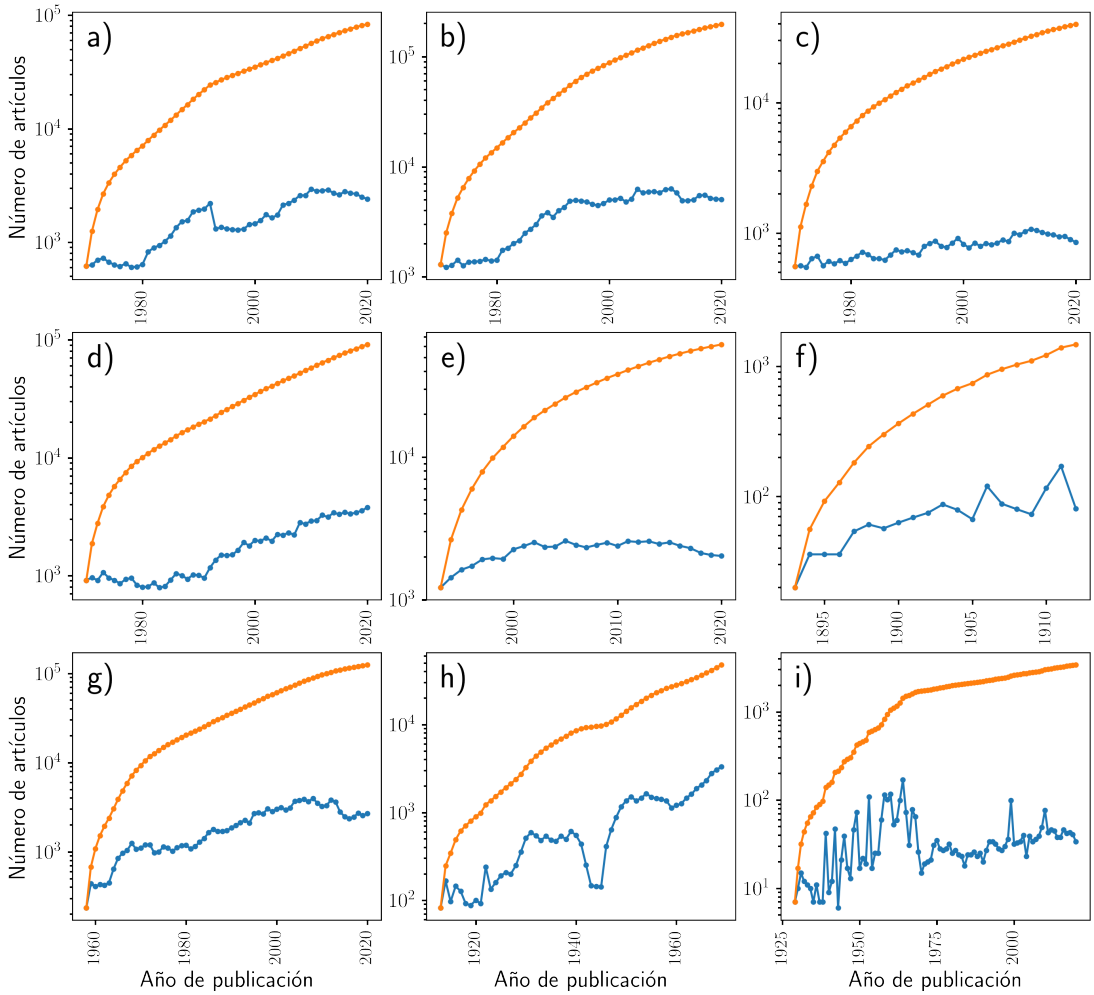

Figura 2.1: Número de artículos (color azul) y número de artículos acumulados (color naranja) en cada año de publicación para: (a) PRA, (b) PRB, (c) PRC, (d) PRD, (e) PRE, (f) PR(serie I), (g) PRL, (h) PR(serie II), (i) RMP. Nótese que el número de artículos en Review of Modern Physics es significativamente menor que para otras revistas, lo mismo ocurre en el caso de Physical Review (Serie I). Los resultados se muestran en escala logarítmica para una mejor visualización

metodología. Adicionalmente, no todos los artículos cuentan con esta clasificación. Esto se traduce en una pérdida importante de información ya que una fracción significativa de los artículos en Physical Review y Physical Review Letters fueron publicados antes de 1975. En efecto, todos los artículos de Physical Review (Serie I y II) fueron publicados antes de 1959 como puede observarse en la Figura 2.1(f) y 2.1(h). Además esta información es necesaria para estudiar de forma completa la evolución temporal de la física durante este periodo. Explorar y analizar esta base de datos utilizando este nuevo enfoque puede ser útil para extraer patrones antes no identificados y obtener nuevos resultados que permitan entender la dinámica de la física durante un periodo de tiempo determinado.

Analizando la información en la tabla de metadatos 2.2 es posible identificar, de forma inmediata, ciertas características de la evolución de la física durante los siglos XX y XXI. Por ejemplo, en la Figura 2.1, se muestra el número de artículos publicados en

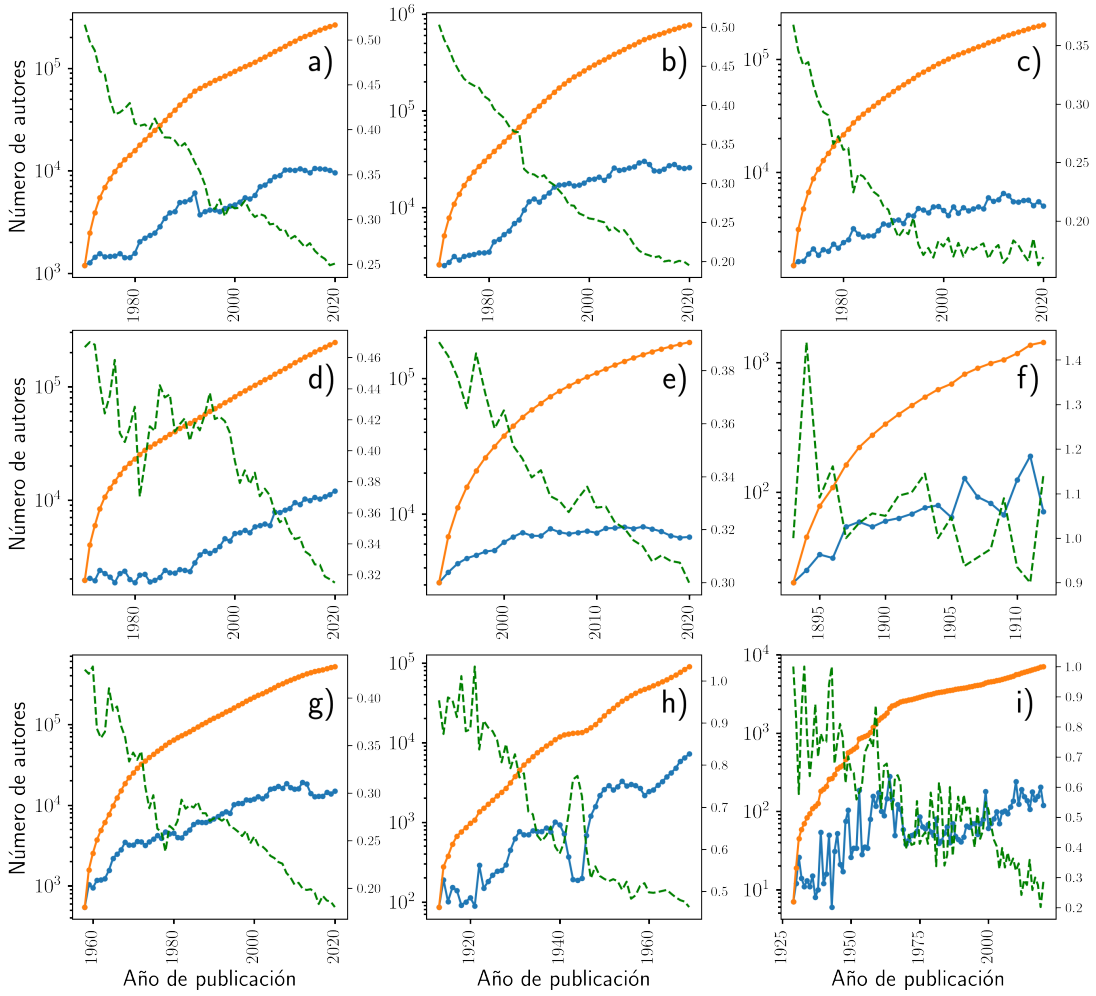

Figura 2.2: Número de autores (color azul) y número de autores acumulados (color naranja) por año para (a) PRA, (b) PRB, (c) PRC, (d) PRD, (e) PRE, (f) PR(serie I), (g) PRL, (h) PR(serie II), (i) RMP. En el eje vertical derecho se muestra, en línea punteada (color verde), el cociente entre el número de artículos y el número de autores

cada año para cada una de las revistas consideradas. Se puede observar un incremento en el número de artículos publicados a lo largo del tiempo; en particular, al analizar el número de artículos acumulados (línea de color naranja) es evidente que este incremento ha permanecido constante, tal y como puede notarse al observar la pendiente de las curvas de las revistas que aún permanecen vigentes. Ciertos aspectos relacionados con la historia de las revistas también son visiblemente evidentes. Por ejemplo, en el caso de PR(Serie II) se observa una disminución importante en el número de artículos durante el periodo de la segunda guerra mundial [2.1(h)]. De forma similar, se puede observar la división de PRA, reflejada en la disminución en el número de artículos, para dar lugar a PRE en el año 1993 [2.1(a)]. Otro aspecto importante de estas gráficas es el número absoluto de artículos publicados en cada revista. En el caso de Review of Modern Physics 2.1, el número de artículos es de un orden de magnitud menor ($\sim 10^4$) en comparación con otras revistas en las que se observa ($\sim 10^5$). Debido a esta diferencia en el número de artículos

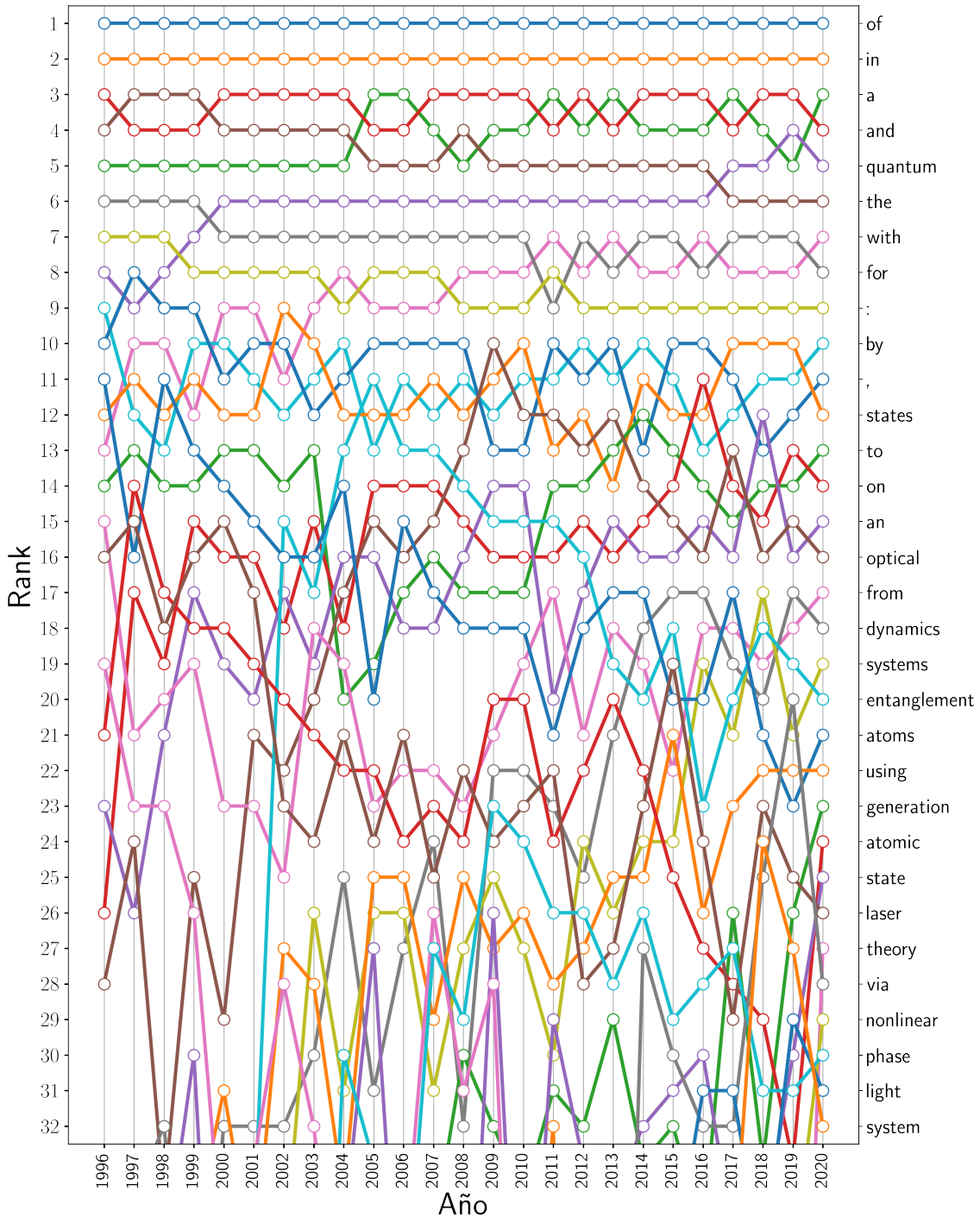

Figura 2.3: Evolución del ranking de palabras en títulos de artículos publicados en Physical Review A. Se muestra la evolución temporal del ranking de las treinta palabras más utilizadas en títulos de artículos publicados en PRA en el año 2020. Cada curva de color representa la dinámica de una palabra distinta cuya posición en el ranking se indica en el eje vertical izquierdo. Nótese que los títulos no han sido filtrados para eliminar palabras o símbolos sin alguna connotación física.

acumulados, no se considera RMP en los siguientes capítulos.

De forma similar, en la Figura 2.2, se muestra el número de autores que publicaron artículos en cada uno de los años correspondientes a cada revista. Estas gráficas muestran algunas de las características de la Figura 2.1, al comparar las gráficas respectivas de cada revista en cada una de las figuras se puede identificar una tendencia similar de las curvas; es decir, el número de artículos publicados y el número de autores crecen con el tiempo pero se observa un comportamiento de estabilización al final de cada periodo. Sin embargo, el crecimiento de ambas cantidades no es exactamente el mismo como puede observarse en el comportamiento de las curvas en color verde en la Figura 2.2, que representan el cociente entre el número de artículos y el número de autores; el comportamiento decreciente indica que la física se mueve hacia una situación en la cual la investigación es cada vez más colaborativa. Por ejemplo, en el caso de Physical Review Letters (Figura 2.2.g), este cociente tenía un valor de 0.5 en el año 1960 mientras que en 2020 este fue menor a 0.20, esto significa pasar de un promedio de dos autores por artículo a un promedio de cinco autores por artículo.

Además de explorar la información general de esta base de datos, es interesante analizar la información contenida en los títulos de artículos publicados en revistas de la APS. Esta información puede analizarse mediante la separación de cada título en palabras o *tokens*.

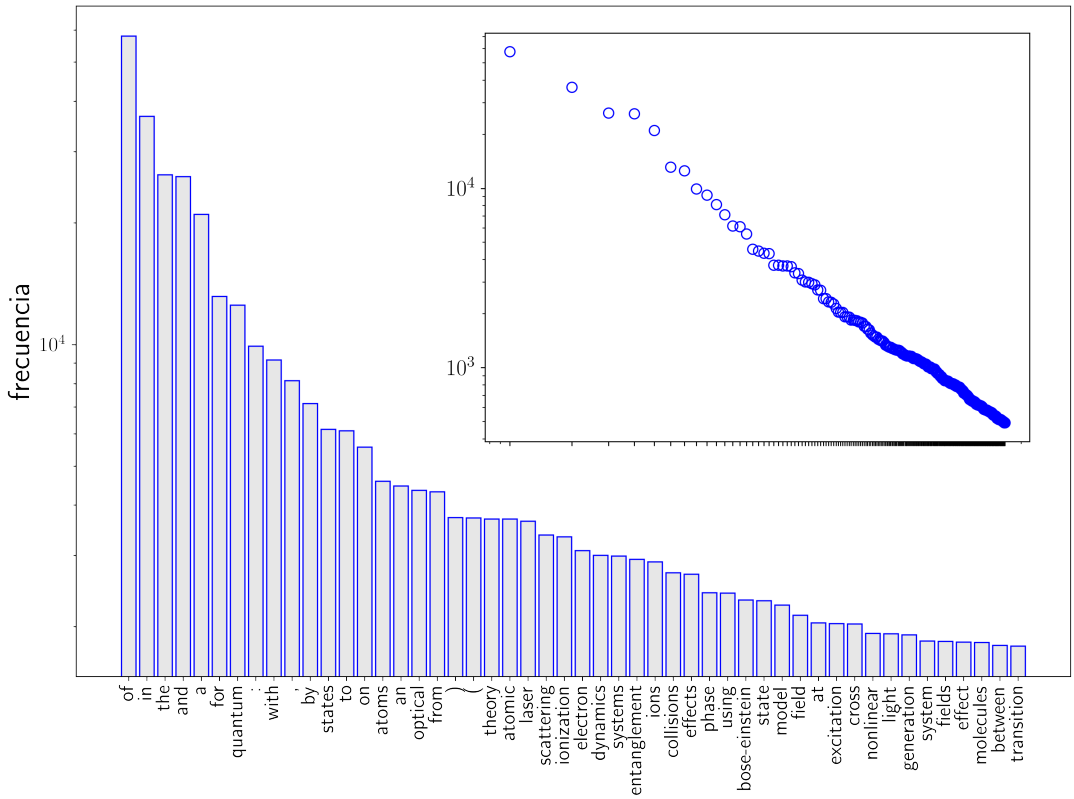

Figura 2.4: Frecuencia de palabras en el periodo 1893-2020. El eje vertical describe la frecuencia de aparición de cada palabra y la altura de cada barra corresponde al número de apariciones de la palabra en todos los títulos de artículos publicados en las revistas consideradas. El recuadro interior muestra la gráfica de frecuencias para las primeras 250 palabras en escala logarítmica.

De esta forma, es posible estudiar la estadística de las palabras que aparecen en cada uno de los años de publicación o la evolución temporal del ranking de la frecuencia con la que cada palabra se utilizó. En la Figura 2.3 se presenta la evolución temporal (en el periodo 1996-2020) del ranking de las treinta palabras más utilizadas en títulos de artículos publicados en Physical Review A en el año 2020. Nótese que hasta este momento no se ha aplicado ningún filtro para eliminar palabras sin un significado asociado con la física. En efecto, la palabra *of* es la palabra más utilizada durante ese periodo y por lo tanto siempre se mantiene en la misma posición dentro del ranking. El primer concepto físico (*Quantum*) aparece en la posición cinco y el segundo concepto (*states*) en la posición doce, después de la posición dieciséis casi todas las palabras en el ranking corresponden a conceptos físicos. Al observar las curvas asociadas a estos conceptos se pueden identificar comportamientos no triviales en los movimientos dentro del ranking: aquellas palabras en los puestos más altos tienden a permanecer en un rango de posiciones estrecho, mientras que las palabras en posiciones más bajas presentan un comportamiento más amplio e irregular. Esta es una propiedad general de diversos sistemas en los cuales se puede definir un ranking [40].

Cuando se analiza la frecuencia de aparición de las palabras en los títulos de artículos publicados en cada una de las revistas durante el periodo 1893-2020 se encuentra que la distribución obedece un comportamiento similar al descrito por la ley de Zipf tal y como puede notarse en la Figura 2.4. Este tipo de comportamiento es común en otros sistemas como textos de lenguaje natural, tamaños de poblaciones de ciudades, distribuciones de riqueza, etc. Esto muestra que la información contenida en los títulos de artículos está caracterizada por propiedades no triviales.

3. Evolución de la diversidad de la investigación científica

3.1. Introducción

En el capítulo anterior se realizó la descripción completa del contenido de la base de datos de la APS que se utilizó en la realización de este trabajo. En el capítulo presente se analiza el contenido de la base de datos para obtener información acerca de la evolución de la diversidad de los conceptos de la física a lo largo de los siglos XX y XXI. Lo anterior se realiza mediante técnicas que ofrecen una herramienta cuantitativa para entender cómo ha evolucionado esta diversidad.

En la primera parte se explica cómo se realiza el procesamiento y filtrado de los datos para obtener un conjunto de palabras, a partir de los títulos publicados en revistas de la APS, que corresponden a los conceptos o ideas más importantes de la física. También se analizan las distribuciones de probabilidad del número de apariciones de estas palabras de forma global y de forma individual (para cada revista). En este último caso también se consideran las distribuciones de probabilidad anuales en todo el intervalo de tiempo durante el cual se ha publicado la revista.

En la segunda parte del capítulo se analiza la diversidad de las listas de palabras generadas en el paso anterior. Se define, para cada revista, una matriz que contiene información acerca de la similaridad que existe entre dos años diferentes en términos de las listas de palabras correspondientes. Estas matrices permiten observar periodos de tiempo en los cuales los conceptos utilizados en la investigación han sido similares. También se analiza la evolución temporal de la diversidad de cada revista en términos de la entropía de Shannon.

3.2. Procesamiento y filtrado de datos

En esta sección se hace una descripción detallada del procesamiento de datos para extraer la información que permite estudiar la evolución y diversidad de los conceptos de la física

durante los siglos XX y XXI. El objetivo es obtener, a partir de los títulos de artículos publicados en revistas de la APS las palabras o conceptos, relacionados con la física, utilizados en cada uno de los años de publicación. De esta forma, para cada una de las revistas consideradas PRA, PRB, PRC, PRD, PRE, PR Y PRL se obtiene una lista de palabras y su respectivo número de apariciones en cada año. Con esta información es posible analizar cómo los diferentes conceptos de la física han surgido, evolucionado y, en casos especiales, desaparecido con el paso del tiempo.

Con el fin de obtener las listas de palabras se considera, para cada una de las revistas, la tabla de datos con la siguiente información: fecha de publicación, título, formato del título y número de volumen. La Tabla 2.2 muestra los primeros ocho registros de esta base de datos para la revista Physical Review Letters cuya publicación comenzó en el año 1958. Para generar las listas de conceptos asociados a la física se requirió realizar un procesamiento del conjunto de títulos reportados en estas tablas. En este trabajo un concepto físico se define como cualquier palabra que haya aparecido en el título de algún artículo registrado en la base de datos después de haberse realizado el procesamiento de datos y que tenga una frecuencia de aparición mayor a un cierto valor umbral. El procesamiento de datos requirió llevar a cabo una serie de pasos. Primero, se separó el título en palabras individuales, *tokens*; este paso generó, para cada título, una lista de tokens. Posteriormente estas listas fueron filtradas para eliminar palabras sin contenido (*stopwords*), adjetivos, adverbios, verbos, y signos de puntuación. También se eliminaron símbolos matemáticos, unidades de medida (ej. m, s, eV), números, símbolos asociados a partículas y se aplicó un filtro para extraer la raíz etimológica de las palabras resultantes. Al final de este proceso se obtiene una lista de títulos en la que cada uno de estos tiene una estructura que consiste de una colección de palabras conectadas con la física. Para estudiar la evolución temporal de la diversidad de los conceptos físicos se generó para cada revista y cada año de publicación una de estas listas.

Por otra parte, el filtrado de los títulos no garantiza la eliminación completa de palabras sin alguna connotación física por lo que es necesario un filtro adicional con respecto al número total de apariciones de estas palabras. Lo anterior se justifica al inspeccionar la lista de palabras de cada revista que resulta después del procesamiento y observar que las palabras no asociadas a la física aparecen en las últimas posiciones al ser ordenadas de acuerdo a la frecuencia de aparición. En la Tabla 3.1 se muestra el top 10 de palabras con mayor número de apariciones en cada una de las revistas consideradas. Inmediatamente se pueden identificar algunas palabras distintivas de la física como *quantum*, *states*, *spin*, etc. Además, obsérvese que cada lista es diferente y aunque algunas palabras se repiten a lo largo de las revistas su posición en el ranking es distinta. Nótese también que palabras como *of*, *in* o *the* no aparecen en estas lista como sí es el caso de la Figura 2.3, antes del procesamiento de los datos.

Es importante mencionar que el filtrado de los títulos de artículos no es ideal y al final de estos procedimientos aun permanecen palabras sin algún significado físico y que introducen ruido en los análisis de estas listas. Lo anterior ocurre debido a que las herramientas de procesamiento de lenguaje natural están orientadas, en mayor medida, a procesar textos no científicos en los cuales términos técnicos o conceptos asociados a la ciencia son

Tabla 3.1: Top diez de palabras en el año 2020. Para cada revista se muestra la lista de diez palabras con mayor número de apariciones en ese año.

PRA	PRB	PRC	PRD	PRE	PRL	PR
quantum	quantum	nucleus	model	model	quantum	scatter
state	spin	collision	hole	system	state	theory
system	phase	reaction	dark	network	phase	effect
atom	effect	model	matter	quantum	spin	state
field	state	state	theory	particle	transition	electron
entanglement	transition	neutron	gravity	phase	effect	model
effect	model	energy	gravitational	transition	system	study
use	system	measurement	field	lattice	measurement	field
gas	couple	decay	decay	flow	interaction	reaction
model	field	structure	neutrino	plasma	model	decay

poco usuales. Sin embargo, puede demostrarse, utilizando filtros más restrictivos, que los resultados no cambian de forma significativa al eliminar este tipo de palabras.

Para estudiar algunas de las propiedades de estas listas, un primer análisis estadístico consiste en la construcción de las distribuciones de probabilidad de las frecuencias de aparición de las palabras en estas listas. Para cada palabra se calculó el número total de apariciones en títulos en un periodo de tiempo T. Este periodo de tiempo puede ser un año o el intervalo de tiempo total durante el cual se ha publicado una revista. Si se obtiene este número para cada palabra se puede calcular la probabilidad de encontrar una palabra con un número de apariciones igual a cierto valor f, donde f denota la frecuencia de aparición.

En la Figura 3.1 se muestra la distribución de probabilidad $\rho(f)$ para cada una de las revistas consideradas en el caso en el que T corresponde al periodo completo de publicación. De forma inmediata se puede observar que aunque las revistas han sido publicadas en diferentes periodos de tiempo; es decir, de forma asíncrona, los resultados muestran un tipo de universalidad que surge en el comportamiento estadístico de las frecuencias de aparición de las palabras. En otras palabras, en cada revista se observa el mismo comportamiento. Otro detalle que llama la atención en esta gráfica es la aparente existencia de dos regímenes en la distribución de frecuencias: aquellas palabras con una frecuencia mayor o igual a 1 y menor o igual a ~ 20 pertenecen al primer régimen mientras que aquellas con una frecuencia mayor a ~ 20 pertenecen al segundo régimen. Sin embargo, esta observación es de carácter cualitativo y verificar esta hipótesis requiere de un análisis estadístico riguroso en términos de un ajuste a una ley de potencias en ambos regímenes.

Si se realiza el mismo análisis pero para un periodo de tiempo $T = 1$ año se observa el mismo comportamiento. En la Figura 3.2 se muestra para cada revista la distribución de probabilidad $\rho(f)$ en cada uno de los años de publicación (cada color denota un año diferente). De forma similar a la Figura 3.1, es posible apreciar, la aparente existencia de al menos dos regímenes divididos por $f \approx 20$. Nótese que debido a que el conteo de

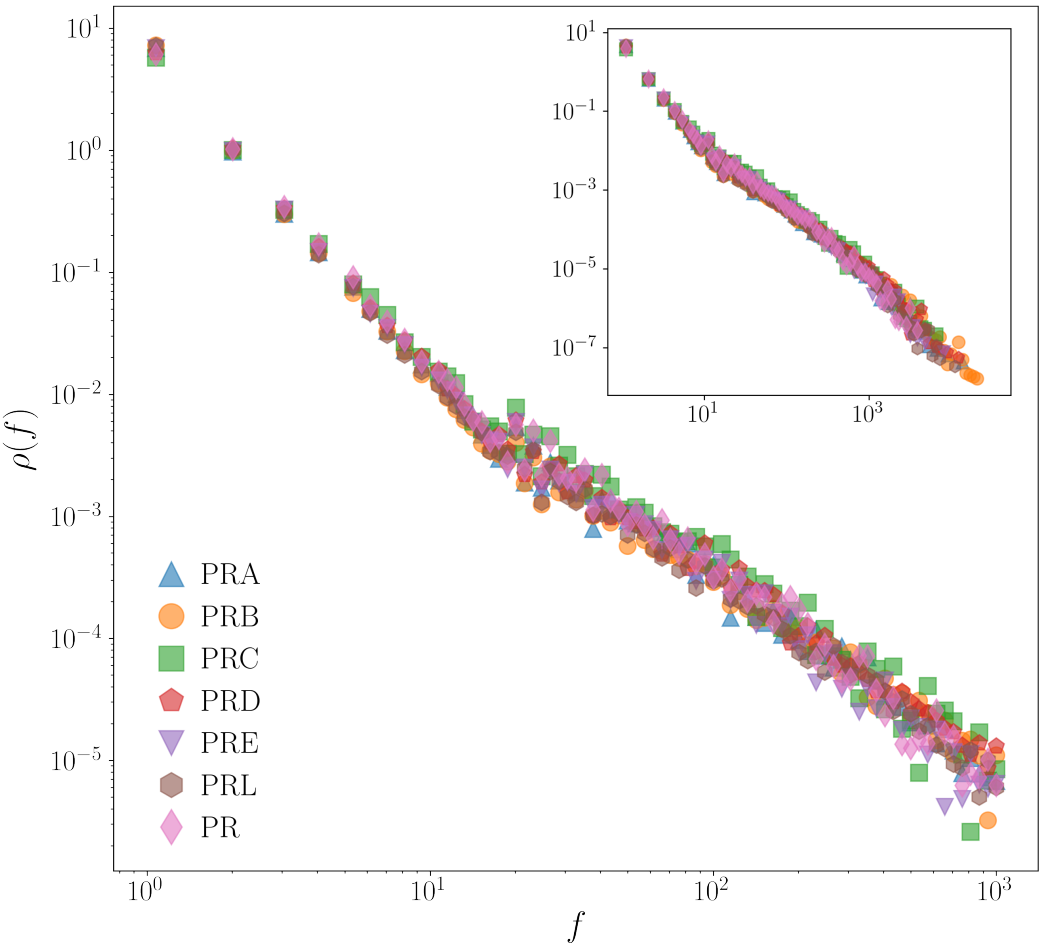

Figura 3.1: Distribución de probabilidad de frecuencias en escala logarítmica. Se muestra la densidad de probabilidad de la frecuencia de aparición de todas palabras que han formado parte de algún título en revistas de la APS. Las palabras corresponden a los títulos de artículos publicados en el periodo 1893-2020. El recuadro interior muestra la misma distribución en el intervalo extendido 10^0-10^4.

palabras es anual la frecuencia máxima no rebasa el valor $f = 10^3$. Algo interesante a notar en este caso es que la evolución temporal de la distribución parece ser constante; es decir, la distribución anual no cambia en el tiempo. Esto es notable tomando en cuenta que las palabras y sus respectivas frecuencias de aparición no son las mismas en cada año.

Lo expuesto en los párrafos anteriores muestra la existencia de distribuciones de cola pesada que describen el comportamiento estadístico de las frecuencias de aparición de palabras en dos escalas diferentes de tiempo. Este resultado podría parecer trivial pues este tipo de comportamiento es similar al observado en lenguajes humanos para los cuales se satisface la ley de Zipf que describe la relación rango-frecuencia en términos de leyes de potencias. Sin embargo, lo que resulta interesante es que estas distribuciones son similares entre revistas y que permanecen constantes en el tiempo como puede observarse en las Figuras 3.1 y 3.2. Una idea que podría explicar este fenómeno es que las listas de palabras

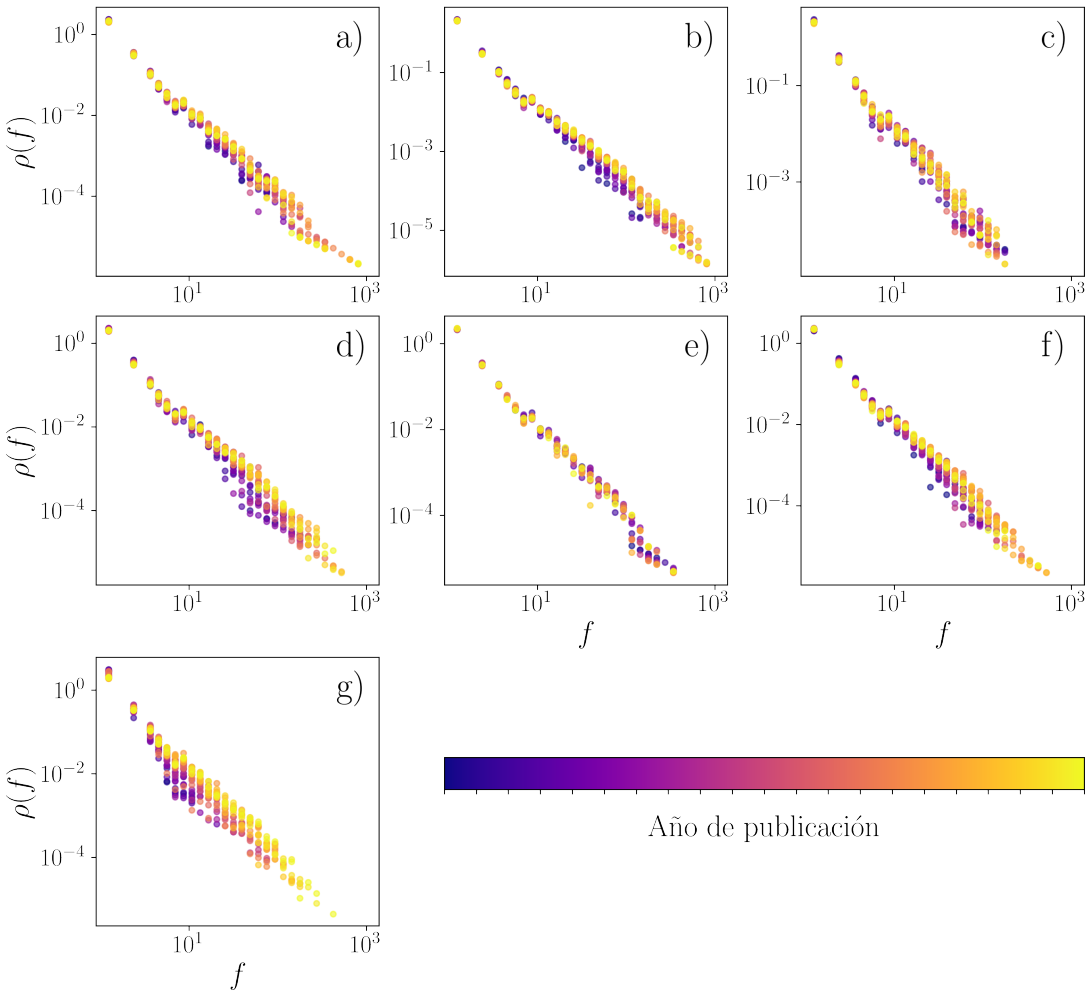

Figura 3.2: Distribuciones de probabilidad de frecuencia por revista en escala logarítmica. Para cada revista y año de publicación se muestra la densidad de probabilidad de la frecuencia de aparición de las palabras en títulos de artículos de **a)** PRA, **b)** PRB, **c)** PRC, **d)** PRD, **e)** PRE, **f)** PR(serie I), **g)** PRL, **h)** PR(serie II), **i)** RMP. Los colores en la barra codifican el año considerado.

y sus respectivas frecuencias son las mismas para toda la colección de revistas, aunque se puede verificar que este no es el caso. Hay que notar que incluso si estas listas son las mismas, el rango de cada una de estas palabras difiere a lo largo de las diferentes revistas como puede notarse en la Tabla 3.1 en la cual se observan palabras cuya posición en el ranking difiere entre una revista y otra.

3.3. Diversidad de palabras en títulos de artículos científicos

En la sección anterior se describió un tipo de universalidad estadística observada en las distribuciones de frecuencia de palabras. Esta universalidad muestra la existencia de un

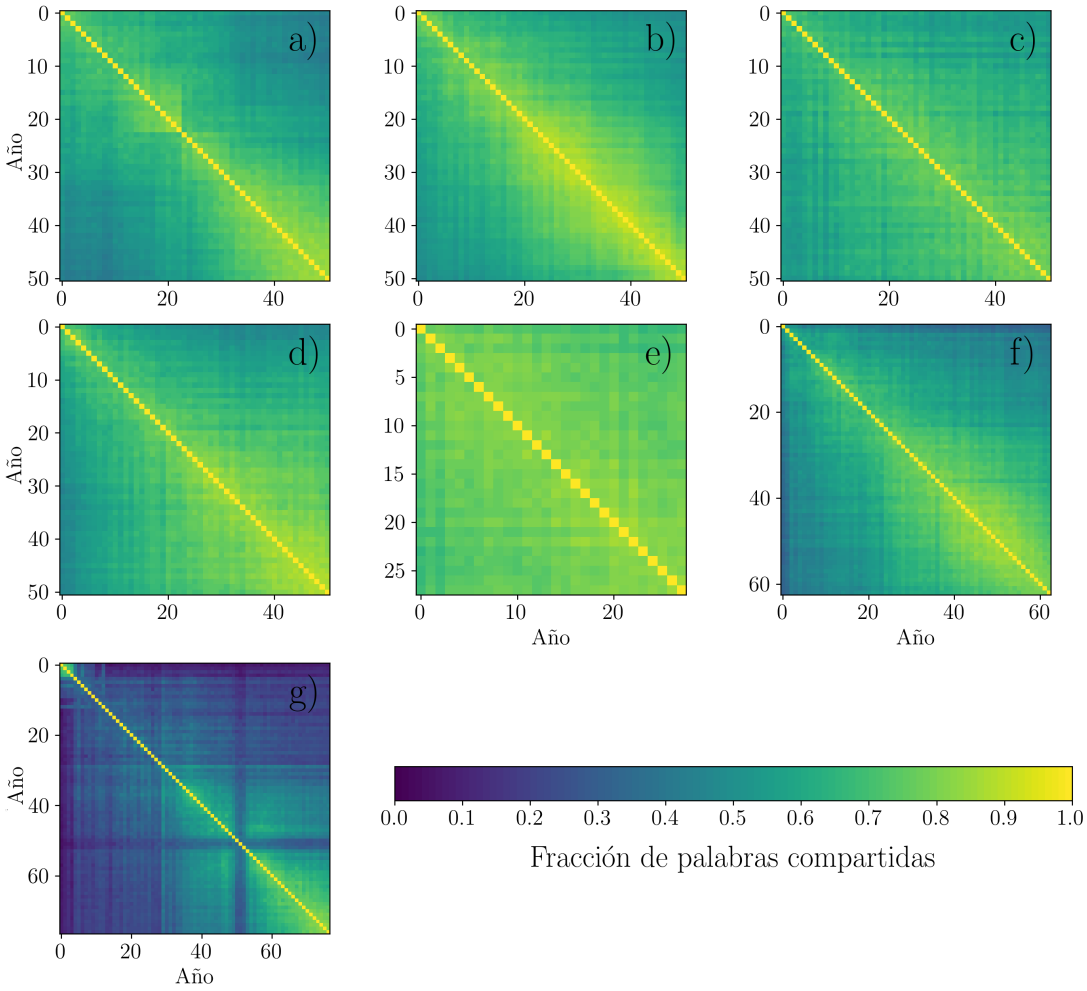

Figura 3.3: Matrices de similaridad para $n = 200$. Para las revistas **a)** PRA, **b)** PRB, **c)** PRC, **d)** PRD, **e)** PRE, **f)** PR(serie I), **g)** PRL, **h)** PR(serie II), **i)** RMP se presenta la matriz de fracción de palabras compartidas. Cada entrada S_{ij} corresponde a la fracción de palabras que comparten las listas de 200 palabras más comunes en los años i, j, estos valores son codificados en la barra de color.

comportamiento común entre todas las revistas analizadas sin importar el periodo de publicación o el campo de estudio de cada revista. Sin embargo, ese tipo de análisis no ofrece información completa acerca de la evolución de la diversidad a lo largo de los años; por ejemplo, la forma en que cada palabra se mueve dentro del ranking no se ve reflejada en estas distribuciones. Esto implica una pérdida de información que es relevante para entender la dinámica de la diversidad de las palabras que representan conceptos utilizados en la base de datos de la APS.

Para entender la dinámica de evolución de estos conceptos se puede analizar las listas de palabras utilizando diversas herramientas. Una primera manera de visualizar la dinámica consiste en comparar el top n de conceptos de cada lista de palabras para cada par de años. Para un valor fijo de n se obtienen los conjuntos de palabras o conceptos $W_i = \{w_k^i\}_{k=1}^n$, $W_j = \{w_k^j\}_{k=1}^n$ de los años i, j respectivamente, w_k^i denota la palabra en la posición k dentro del ranking en el año i. Posteriormente, se calcula la intersección entre estos

conjuntos $I_{ij} = W_i \cap W_j$, se calcula la cardinalidad de este conjunto y se divide entre el número n:

$$S_{ij} \equiv \frac{|I_{ij}|}{n}. \tag{3.1}$$

Este número, que asume valores entre 0 y 1, representa una medida de similaridad entre las listas de palabras de diferentes años. Un valor cercano a 1 implica que los años son similares en cuanto a los conceptos físicos utilizados mientras que un valor cercano a 0 implica diferencias entre las listas de palabras.

Es importante destacar que S_{ij} no respeta la posición de las palabras, si el valor de n es pequeño (10, por ejemplo) entonces I_{ij} también asume un valor pequeño pues el top de palabras cambia significativamente de un año a otro; por otra parte, si el valor de n es grande (1000, por ejemplo) entonces I_{ij} también asume un valor grande pues las listas anuales contienen muchas palabras en común. Para un valor intermedio $n = 200$ se tiene un equilibrio entre estos dos efectos y es posible observar las diferencias o similitudes entre las listas de conceptos de diferentes años. Si se considera cada par de años i, j entonces es posible definir una matriz S cuyas entradas corresponden a los valores S_{ij}. En la Figura 3.3 se muestra esta matriz para una de las revistas consideradas y para el caso $n = 200$. En la mayoría de los casos es posible observar un comportamiento no trivial en el cual los ranking de palabras anuales se organizan en bloques, esto es más evidente en el caso de PRL para la cual se observa varios bloques alrededor de la diagonal. Por ejemplo, el intervalo de años $30 - 50$ exhibe una alta similaridad en los rankings de palabras. En el caso de PRE el comportamiento es más uniforme como puede observarse en la poca diversidad de colores de la matriz correspondiente.

El análisis anterior muestra la existencia de una dinámica temporal en los rankings de palabras. Otra forma de estudiar esta dinámica es por medio de la diversidad. Existen varias maneras de analizar la diversidad de un sistema, por ejemplo, en casos en los cuales es posible definir un rango de acuerdo a alguna característica de los elementos que lo conforman. En [41], se definió la diversidad de rango para estudiar la dinámica de las palabras en lenguajes humanos. A partir del ranking anual de palabras obtenido de Google Ngram se calculó la diversidad del rango k como el número de palabras que han ocupado la posición k dentro del ranking en un periodo de tiempo T.

El sistema considerado en este trabajo consiste de la lista de palabras o conceptos utilizados en títulos de artículos publicados en revistas de la APS y el rango se define en términos de la frecuencia con la que se ha utilizado cada palabra. Al calcular esta diversidad de rango para los rankings de conceptos se obtiene un segundo tipo de universalidad. Esto se muestra en la Figura 3.4 en la cual se puede observar la diversidad para cada una de las revistas y las primeras quinientas posiciones del ranking. En todos los casos el comportamiento es similar; los primeros rangos presentan una menor diversidad y conforme se avanza hacia rangos menores estos presentan una mayor diversidad. De forma interesante, puede observarse como las revistas PRA, PRB, PRC y PRD tienen el mismo comportamiento en términos de la diversidad de rango. Este resultado no es trivial pues como ya se ha mencionado anteriormente, el ranking de palabras difiere entre una revista y otra. Nótese también que a partir de un rango límite la diversidad alcanza un

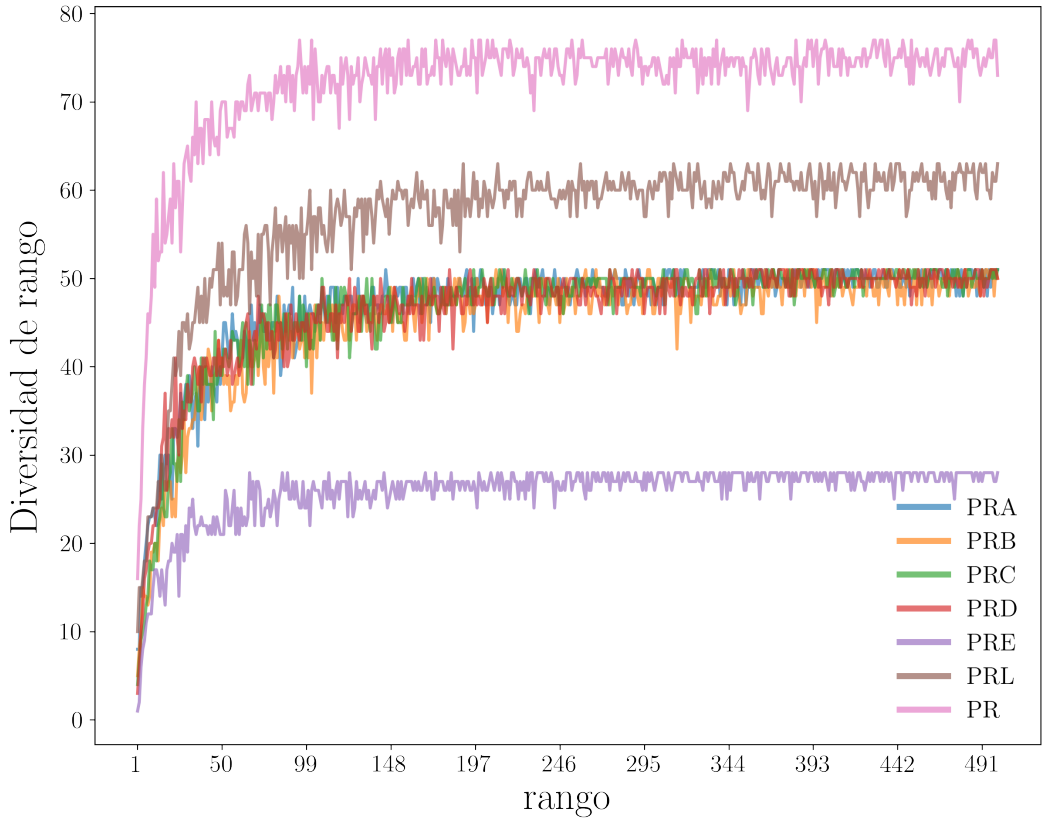

Figura 3.4: Diversidad de rango. Para cada revista se muestra la diversidad de rango definida como el número de palabras que han ocupado una posición dentro del ranking en el periodo de tiempo considerado.

valor estable y todos los rangos menores a este valor límite tienen la misma diversidad. Este comportamiento implica que las palabras más utilizadas durante todo el periodo de publicación de una revista tienden a permanecer en las primeras posiciones durante mucho tiempo; sin embargo, aquellas con una frecuencia de uso menor tienden a moverse continuamente de posición dentro del ranking. Esto es consistente con lo observado en otros sistemas en los cuales es posible definir un ranking; por ejemplo, deportes, países, instituciones, palabras, personas, etc [40].

Otra forma de medir la diversidad en un sistema es a partir de la entropía de información, también llamada entropía de Shannon. Para un sistema conformado por N tipos o especies de elementos, la entropía H se define como:

$$H(\{p_i\}) = -\sum_{i=1}^{N} p_i \log(p_i) \tag{3.2}$$

donde p_i es la frecuencia relativa de la especie i en el sistema [42]. De esta forma, si existe una sola especie o tipo en el sistema entonces $H = 0$ y no existe diversidad. Por otra parte, la diversidad del sistema es máxima cuando la frecuencia relativa de cada una de las N

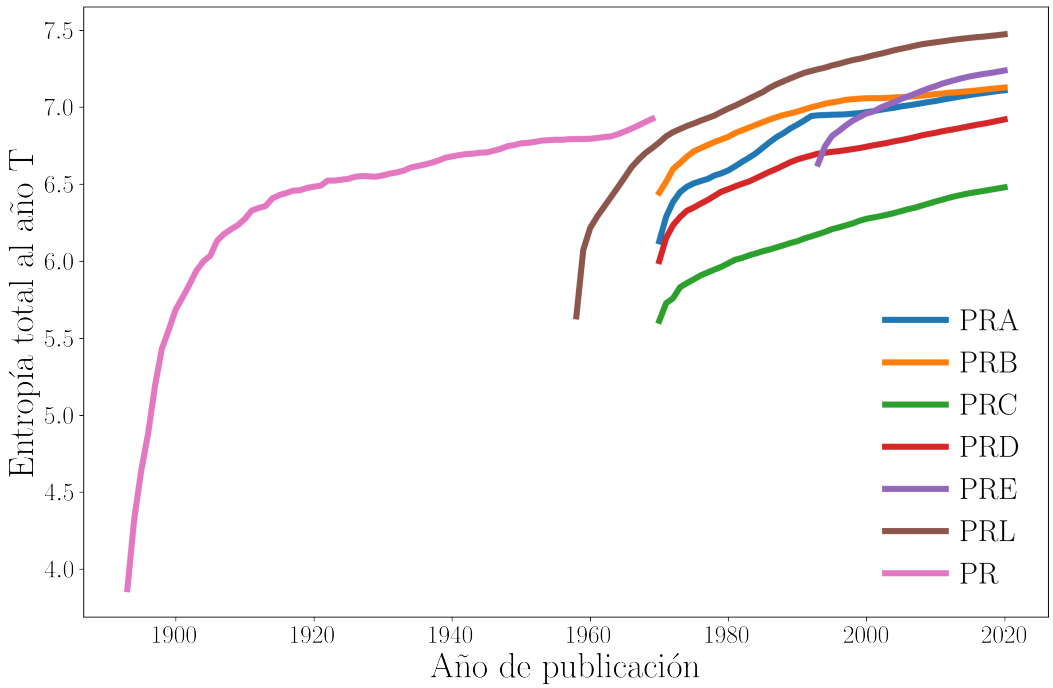

Figura 3.5: Evolución de la entropía total. Para cada revista se muestra la entropía total como función del año de publicación.

diferentes especies es idéntica; es decir, $p_i = \frac{1}{N}$ y por lo tanto $H = \log(N)$. En el contexto de este trabajo, la entropía se entiende como una medida asociada a la diversidad de las palabras que se han utilizado en cada año para hablar o expresar ideas asociadas con la física.

Sin embargo, para comparar la diversidad entre un año y otro es conveniente considerar la entropía H_{tot} total al año T. Esta se define como la entropía de las frecuencias relativas de especies o tipos que han existido en el sistema hasta el tiempo T, esta medida permite considerar la contribución de cada especie o tipo en el sistema a la diversidad aun cuando ha dejado de estar presente. De acuerdo a esta definición es claro que la entropía total es una función no decreciente del tiempo. Si se consideran las listas de palabras o conceptos anuales para cada revista y se calculan las frecuencias relativas correspondientes es posible calcular la entropía total como función del tiempo, esto permite analizar la dinámica de la diversidad a lo largo de los años de publicación.

En la Figura 3.5 se muestra la entropía total de cada una de las siete revistas consideradas. Puesto que cada revista ha sido publicada durante periodos de tiempo diferentes es posible observar cómo la división de ciertas revistas ha dado lugar a un cambio en la diversidad. Por ejemplo, en el caso de Physical Review, esta se dividió (y desapareció) para dar origen a Physical Review A, B, C y D. En la figura se observa como la diversidad de PR alcanzó un valor máximo inmediatamente antes de que estas cuatro revistas fueran creadas. En los primeros años de estas revistas su diversidad comienza por debajo del valor máximo de

la diversidad de PR ya que cada una se especializa en un conjunto limitado de temas de la física y, por lo tanto, el número de palabras o conceptos utilizados en títulos de artículos es menor. De forma similar, en el año 1993, la revista PRA se dividió para dar lugar a PRE pero en este caso PRA continuó siendo publicada. La disminución en los temas publicados por PRA se ve reflejado en el valor constante de la diversidad (curva azul) durante un periodo corto de tiempo; es decir, inmediatamente después de la creación de PRE, la diversidad de PRA dejó de crecer debido a la separación en dos campos distintos. Nótese que todas las curvas son funciones no decrecientes del tiempo debido a que los conceptos que han dejado de aparecer en títulos siguen contribuyendo a la diversidad total de la revista.

Un aspecto importante a notar es que, para todas las revistas, la diversidad en los primeros años de publicación crece de forma acelerada pero después de un periodo de tiempo esta diversidad reduce su tasa de crecimiento. Esto es particularmente evidente en el caso de PR y PRL, siendo esta última una revista que aborda temas de todas las áreas de la física. Estos resultados muestran que a pesar de la gran variedad de áreas en la física, el número de conceptos o ideas ha permanecido relativamente constante en los últimos años; es decir, gran parte de la investigación actual cubierta por las revistas analizadas consiste en mayor medida de investigación en temas bien establecidos. Esto es consistente con lo descrito en el Capítulo 2 respecto a la interdisciplinariedad e innovación de la ciencia.

4. Ciencia de redes: redes de palabras

4.1. Introducción

En el capítulo anterior se analizó el conjunto de palabras de las revistas publicadas por la APS mediante técnicas estadísticas que ofrecen información acerca de la diversidad en los textos científicos; en particular, la física. En este capítulo se analizan las palabras asociadas a la física utilizando como herramienta principal la teoría de redes. El objetivo es identificar, a partir de los datos, grupos de palabras que permitan definir una partición de la física en diferentes áreas y a diferentes escalas. En términos de la teoría de redes, cada palabra representa un nodo de la red y es posible definir diferentes tipos de enlaces o conexiones entre las palabras que conforman los títulos de artículos de acuerdo al tipo de análisis que se quiera realizar. En este caso el tipo de relación que se analiza es de similitud.

En la primera parte del capítulo, sección 4.2, se describe cómo se establecen las conexiones entre los nodos (palabras) de la red, estas conexiones se definen en términos de una medida de similitud; para esto, se asocia a cada palabra un vector que contiene información acerca de su relación con otras palabras. La similitud entre dos palabras se define como el producto punto (normalizado) entre los dos vectores correspondientes a estas palabras. Las redes de similitud resultantes proveen información sobre la distancia entre cualquier par de palabras: una distancia cercana a cero implica una alta similitud entre las palabras mientras que una distancia cercana a 1 corresponde a pares de palabras poco similares.

En la segunda parte, sección 4.3, se estudia la formación de comunidades de palabras a diversas escalas. En esta sección, el objetivo es encontrar áreas de la física definidas a partir de la información contenida en los datos (títulos de artículos publicados en revistas de la APS). Debido a que el análisis se realiza a diferentes escalas, se define una medida de información, basada en la entropía de Shannon, para cuantificar la escala que ofrece mayor información con respecto a las comunidades obtenidas.

4.2. Matrices de similitud y comunidades

En esta sección se describe el procedimiento para la construcción de las redes de similitud que constituyen el objeto principal de estudio de este capítulo. Para esto, se consideran las listas de títulos (filtradas) descritas en los capítulos anteriores y a partir de estas se genera un conjunto de vectores con información acerca de la relación entre palabras.

Para definir las relaciones o conexiones entre las palabras de la red, se asocia a cada palabra w, un vector $v_w \in R^N$ donde N es el número total de palabras consideradas obtenidas a partir de los títulos. Considerando una lista de títulos se extrae el conjunto de palabras que aparecen en cada uno de los títulos y se ordenan de acuerdo a su frecuencia de aparición. Esto permite asociar un índice numérico a cada palabra de acuerdo al número de veces que esta aparece en los títulos de artículos publicados por la APS. Para una palabra j, se define la entrada i-ésima del vector v_j como el número total de veces en las que la palabra j aparece de forma simultánea, en un mismo título, junto a la palabra i. De esta forma, cada palabra j es asociada con un punto (definido por un vector v_j) en un espacio Euclidiano en el cual puede definirse una distancia o medida de similitud.

La similitud entre dos palabras se obtiene calculando el producto punto entre los vectores asociados y dividiendo entre las normas de los vectores. El número resultante corresponde al coseno entre ambos vectores y se denomina similitud coseno [2]:

$$\chi(i,j) = \frac{v_i \cdot v_j}{|v_i||v_j|} = \cos\theta_{ij} \tag{4.1}$$

donde θ_{ij} es el ángulo entre los vectores v_i, v_j. Nótese que la similitud así definida asume valores entre 0 y 1 siendo 1 la similitud máxima entre palabras.

El cálculo del coseno entre los vectores correspondientes a cada par de palabras posible define una matriz de similitud $\boldsymbol{\chi}$ cuya entrada χ_{ij} corresponde a la distancia entre las palabras i, j. A partir de esta matriz es posible construir diferentes redes de similitud mediante la introducción de un parámetro adimensional H que permite controlar la escala en la cual se realiza el análisis. El uso de este parámetro para explorar diferentes escalas de una red se hizo por primera vez en [43] en el estudio del Sistema Metrobús de la Ciudad de México. Por consistencia con el análisis realizado en este trabajo se define una nueva matriz $\boldsymbol{\xi}$ con elementos ξ_{ij} y dada por $\boldsymbol{\xi} = \mathbf{1} - \boldsymbol{\chi}$ donde $\mathbf{1}$ denota la matriz cuyas entradas son todas iguales a 1. De esta forma, dos palabras i, j son completamente similares si $\xi_{ij} = 0$. Para construir las redes de similitud que representan las diferentes escalas del sistema se toma un valor fijo del parámetro H en el intervalo $[0, 1]$ y se define la matriz de adyacencia correspondiente $\mathbf{A}^H$ de la siguiente manera:

$$A_{ij}^H = \begin{cases} 1 & \text{si} \quad \xi_{ij} \leq H \\ 0 & \text{si} \quad \xi_{ij} > H \end{cases} \tag{4.2}$$

En la Figura 4.1 se muestra la densidad de probabilidad ρ y la función de distribución

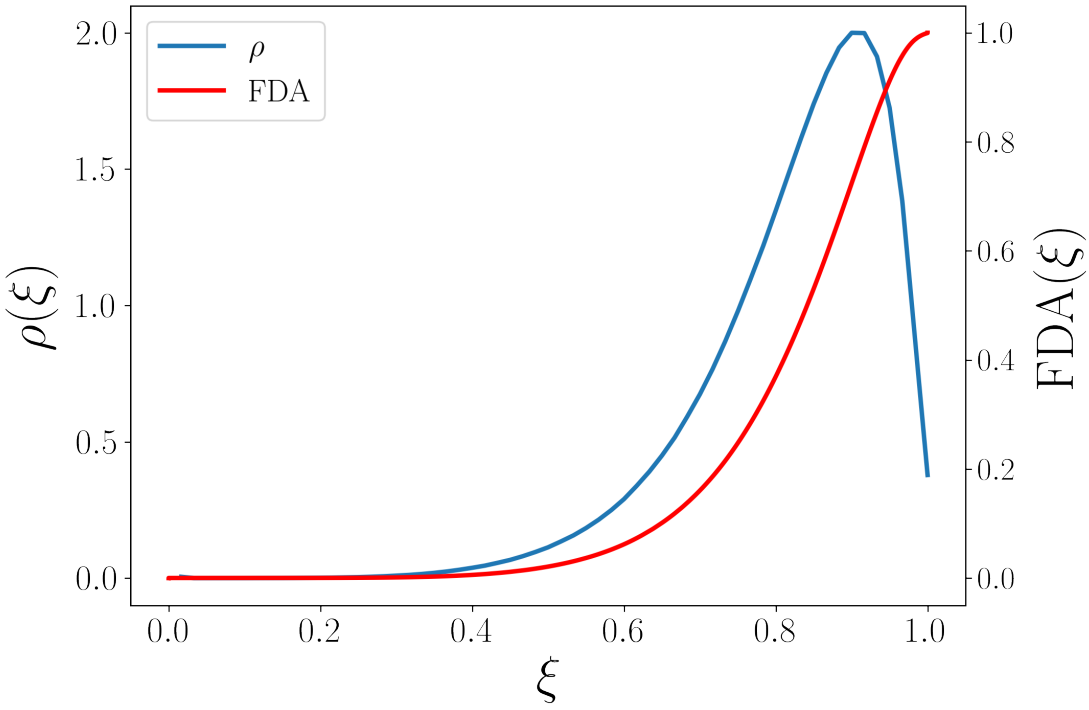

Figura 4.1: Distribución de probabilidad de similitud entre palabras. En color azul se muestra la densidad de probabilidad de la distancia o similitud de todas palabras que han formado parte de algún título en revistas de la APS. En color rojo se muestra la función de distribución acumulada de la distancia o similitud. Las palabras corresponden a los títulos de artículos publicados en revistas de la APS en el periodo 1893-2020.

acumulada (FDA) de las entradas de la matriz $\boldsymbol{\xi}$. La figura muestra que la mayoría de las entradas de esta matriz se concentran en el intervalo $[0.5, 1.0]$ el cual contiene aproximadamente el 98 % del total. Considerando la definición de la matriz de adyacencia lo anterior implica que, para valores de $H < 0.5$, las redes correspondientes (definidas por $\mathbf{A}^H$) tienen un número muy pequeño de conexiones y, al aumentar su valor por encima de 0.5, las redes resultantes muestran un incremento rápido en el número de conexiones entre nodos. Esto sugiere la existencia de varios tipos de comportamiento estructural de las redes de palabras definidas por la ecuación (4.2) dependiendo del valor del parámetro H. Al variar el valor de este parámetro es posible explorar las redes de palabras y extraer información sobre grupos similares de palabras a diferentes escalas. De acuerdo a cómo se han definido las redes de similitud, se espera que valores por debajo de 0.5 permitan observar grupos pequeños de palabras que aparecen de forma frecuente en un título; por ejemplo, (Bose, Einstein), (phase, transition), etc. Por otra parte, para valores de H por encima de 0.5, se espera la aparición de grupos más grandes que corresponden a áreas o subáreas de estudio en la física.

Para verificar las hipótesis del párrafo anterior se pueden analizar las redes de similitud para diferentes valores de H en el intervalo $[0, 1]$. Primero se analiza la evolución del tamaño del componente gigante conforme se incrementa el valor del parámetro H. En

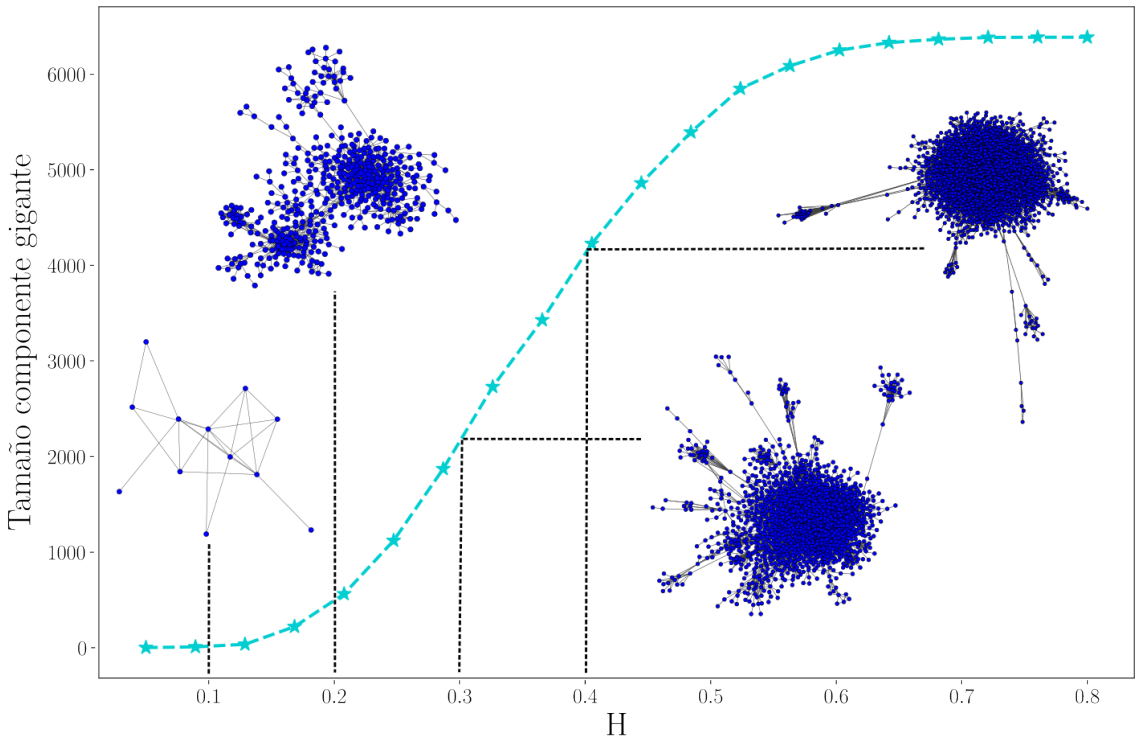

Figura 4.2: Tamaño de la componente gigante de la red de similitud. Se muestra el número de nodos de la componente gigante de la red similitud como función del parámetro H. Las gráficas corresponden a las redes de similitud con parámetro $H = 0.10, 0.20, 0.30, 0.40$. A partir del valor $H = 0.20$ se observa un incremento rápido en el tamaño de la componente gigante y cuando $H \approx 0.60$ la red de similitud es conexa.

la Figura 4.2 se muestra el tamaño de la componente gigante de la red de similitud como función del parámetro H. Para los valores $H = 0.10, 0.20, 0.30$ y 0.40 se muestra la gráfica correspondiente a la componente gigante de cada red de similitud. Nótese que al incrementar el valor del parámetro H la densidad de conexiones incrementa rápidamente. Cuando $0.6 <\approx H$ la red es conexa; de esta manera, la totalidad de nodos forman parte de la componente gigante. Por otra parte, desde $H = 0.2$ se observa un incremento rápido en el tamaño de la componente gigante. Para valores por debajo de 0.2, el número de conexiones es demasiado pequeño y; por lo tanto, el tamaño G de la componente más grande de la red es tal que $G \sim O(10^2)$.

De la discusión en el párrafo anterior se puede inferir que el comportamiento interesante de las redes de similitud, en términos de las comunidades de palabras, ocurre para valores intermedios del parámetro H: valores muy pequeños establecen pocas conexiones entre los nodos y no pueden definirse grupos similares de palabras; por el contrario, valores muy grandes generan redes en las que la mayoría de nodos están conectados y, por lo tanto, se pierde información respecto a la similitud entre palabras. Por ejemplo, para $H = 0.2$, el número de enlaces en la red es igual a 2433. Puesto que el número total de palabras es aproximadamente 6000 esto implica una fracción igual a 0.00006 del número total de enlaces posibles ($\frac{6000\times5999}{2}$). En el caso de H=0.60, el número de enlaces en la red es igual

Tabla 4.1: Evolución de las comunidades correspondiente a la palabra "system". Para cada valor de $H = 0.13, 0.15, 0.20, 0.25, 0.35, 0.50$ se muestran las diez palabras con mayor frecuencia de aparición contenidas en la comunidad de "system".

H	0.13		0.15		0.20		0.25		0.35		0.50	
	Pos	Palabras	Pos	Palabras	Pos	Palabras	Pos	Palabras	Pos	Palabras	Pos	Palabras
	4	effect	7	field	3	state	1	quantum	1	quantum	3	state
	15	system	15	system	4	effect	2	model	2	model	5	spin
	64	approach	41	use	5	spin	7	field	3	state	15	system
	77	behavior	63	analysis	11	structure	8	theory	5	spin	49	transport
	91	induce	64	approach	13	study	15	system	7	field	66	photon
	138	limit	77	behavior	15	system	20	lattice	8	theory	88	tunnel
	144	application	79	fluctuation	17	surface	30	density	15	system	107	dot
			83	method	18	interaction	31	particle	16	wave	122	hall
			90	disorder	19	property	42	calculation	18	interaction	125	bind
			138	limit	20	lattice	46	function	20	lattice	150	control

a 1261330 y la fracción incrementa hasta un valor de 0.035. Sin embargo al incrementar H hasta un valor de 0.80 esta fracción aumenta su valor hasta 0.20. Por lo tanto, aun cuando existe un rango de valores $[0.6, 1.0]$ en el cual la componente gigante contiene a la mayoría de los nodos de la red, solo una parte de este intervalo es de interés para el análisis de las comunidades.

Para un valor fijo de H, se obtiene una red de similitud que contiene información acerca de la similitud entre palabras, donde H delimita el valor en el cual dos palabras se consideran similares o no. Mediante el incremento en el valor de este parámetro, diferentes grupos de palabras similares pueden unirse y dar lugar a grupos más grandes que indican áreas de estudio de la física. Para entender esta fusión de grupos o comunidades de palabras similares en agregados más grandes, conforme se incrementa el valor de H, se obtienen las comunidades de la red de similitud utilizando el algoritmo de Louvain [19] (mediante la maximización de la modularidad). De esta forma, a cada escala (definida por el parámetro H) corresponde un conjunto de comunidades que representan grupos de palabras altamente relacionados. Para ejemplificar la evolución de estas comunidades al cambiar el valor del parámetro H se considera la palabra "system", la cual ocupa el lugar número 15 en el número de apariciones en títulos de artículos publicados por la APS. Esta palabra es central en la física y se puede suponer que está conectada con un gran número de palabras de la física.

En la Tabla 4.1 se muestra, para los valores de $H = 0.13, 0.15, 0.20, 0.25, 0.35, 0.50$, las diez palabras con mayor frecuencia de aparición en los títulos de artículos y que pertenecen a la misma comunidad que la palabra "system". Cuando $H = 0.10$ la comunidad consiste de una palabra; pero, al aumentar el valor del parámetro nuevas palabras se unen a la comunidad de la palabra "system". Como puede notarse en la Tabla 4.1, las palabras dentro de la comunidad no son las mismas, algunas de estas permanecen dentro de la comunidad para todos los valores de H pero algunas salen para formar parte de otra comunidad, este es el caso de la palabra "quantum", la cual está dentro de la comunidad cuando $H = 0.25$ y $H = 0.35$ pero está ausente al pasar al valor $H = 0.50$. Está situación puede explicarse considerando el hecho de que, al cambiar la escala, se originan nuevas co-

nexiones hacia la palabra "quantum" desde otro grupo de palabras altamente conectadas. De esta forma "quantum" pasa a formar parte de otra comunidad. Este comportamiento también se observa para otras palabras como "spin" o "energy" y da lugar a una dinámica interesante y no trivial conforme se cambia la escala de estudio. En el caso de "quantum", la palabra no necesariamente permanece en una misma comunidad a la cual se le unen nodos al incrementar el parámetro H sino que puede "moverse" hacia otras comunidades si existe una relación más fuerte (más conexiones) con las palabras en estas comunidades.

4.3. Análisis multiescala de comunidades

En la sección anterior se encontró que conforme se incrementa el valor del parámetro H, aparecen nuevas comunidades con un número cada vez mayor de elementos; sin embargo, se requiere encontrar una forma de identificar la escala (dada por el parámetro H) que ofrezca la mayor cantidad de información respecto a la formación de grupos de palabras similares (comunidades). Es posible distinguir, de forma intuitiva, dos casos extremos en los cuales la información que ofrecen las comunidades formadas en la red es de poca utilidad. El primero ocurre cuando cada comunidad consiste de un único nodo y corresponde a valores pequeños del parámetro H, el segundo ocurre cuando existe una única comunidad que contiene a todos los nodos de la red y corresponde a valores grandes de H. En ambos casos existe una alta incertidumbre respecto a la diferencia o similitud de cada nodo (palabra) con el resto de elementos de la red. En el primer caso, todo par de nodos es disimilar pues cada uno de estos se encuentra en una comunidad diferente y no es posible agrupar el conjunto total de nodos de forma no trivial; en el segundo caso, todos los nodos son similares pues están en la misma comunidad y por lo tanto, tampoco existe una distinción no trivial. Entre estos dos extremos se encuentra el caso en el cual existen comunidades no triviales; es decir, diferentes grupos de nodos con más de un elemento que permiten hacer distinciones entre los nodos que los conforman. En este caso, si se seleccionan dos nodos pertenecientes a comunidades diferentes entonces es posible hacer una distinción de acuerdo a las características de los elementos en cada comunidad. Por ejemplo, en el caso de la red de palabras, las comunidades pueden representar diferentes áreas o subáreas de la física.

Para formalizar esta idea, sobre la información que contienen las comunidades de la red, se propone una medida de información basada en la entropía de Shannon. Un requisito que debe satisfacer esta medida es que debe ser mínima en los casos extremos mencionados en el párrafo anterior. De la discusión en la sección anterior, nótese que para cada palabra, el tamaño de la comunidad que la contiene crece de forma monótona como función de H. Eventualmente la comunidad contiene a todos los nodos o palabras y todos los enlaces de la red. En el primer caso extremo existen tantas comunidades como nodos en la red, los enlaces ocurren entre comunidades (enlaces intercomunidad) y no existen enlaces dentro de las comunidades (enlaces intracomunidad). En el segundo caso extremo, existe una sola comunidad la cual contiene a todos los nodos y enlaces de la red; es decir, todos los enlaces son intracomunidad. Debido a que estos casos extremos ofrecen poca información,

la intuición sugiere que la mayor cantidad de información contenida en las comunidades debe ocurrir para un caso intermedio en el cual el número de enlaces intercomunidad y el número de enlaces intracomunidad son distintos de cero. En este caso intermedio puede ocurrir que varias de las palabras en la lista total no formen parte de la componente gigante y sean nodos aislados de la red, esto debido a que no están conectadas a otras palabras por medio de enlaces.

En este contexto, que una escala sea informativa significa que la partición de la red en comunidades contiene información acerca de la diferencia y similitud entre los nodos de la red. En el primer caso, la partición dice que todos los nodos son diferentes y, en el segundo caso, que todos los nodos son iguales; por lo tanto, la medida de información debe estar constituida por dos componentes: un componente que mide la similitud y otro que mide las diferencias entre los diferentes grupos de nodos definidos por la partición. Cuando cada palabra pertenece a una comunidad que consiste de un único nodo, la partición establece que todos los nodos son diferentes y no contiene información respecto a la similitud; por el contrario, en una partición en la que existe al menos una comunidad con más de un nodo incrementa su contenido de información respecto a la similitud entre los nodos que contiene pues ahora al menos dos nodos son parte de una misma comunidad. En el otro extremo (una comunidad que contiene a todos los enlaces de la red), no hay información acerca de la diferencia entre nodos. Si se toma esta única comunidad y se divide en al menos dos partes, esta nueva partición ofrece más información pues hace explícita la diferencia entre dos grupos de nodos.

Tomando en cuenta esta discusión es evidente que esta medida de información debe formularse en términos de los enlaces que contiene cada comunidad. Por lo tanto, se propone la siguiente medida de información (normalizada) en términos de la entropía de Shannon 3.2:

$$I(\{C_i\}) = 1 - \frac{E(\{C_i\}) + \langle E(C_i)\rangle}{\log(M)} = 1 - \frac{-\sum_{i=1}^{K} p_i \log(p_i) + \langle -\sum_{j=1}^{|C_i|} q_j^i \log(q_j^i)\rangle}{\log(M)} \quad (4.3)$$

En la expresión (4.3), el término $E(\{C_i\}) = -\sum_{i=1}^{K} p_i \log(p_i)$ define la entropia del conjunto de comunidades $\{C_i\}$ y corresponde al componente de la medida de información que mide las diferencias entre grupos de nodos, p_i es igual al número relativo de enlaces en la comunidad i; $E(C_i) = -\sum_{j=1}^{|C_i|} q_j^i \log(q_j^i)$ define la entropía de la comunidad i y q_j^i es el peso asignado al enlace j de esta comunidad, este termino mide la información dentro de cada comunidad. Puesto que la redes consideradas no tienen pesos asociados se tiene que $q_j^i = 1$ y el segundo término se reduce a $\langle \log(|C_i|)\rangle$ donde $\langle\rangle$ denota el promedio sobre el conjunto de comunidades, K es igual al número de comunidades y M es el número total de enlaces en la red. El factor $\log(M)$ permite normalizar la información de tal forma que $I <= 1$. Nótese que $I = 0$ en los dos casos extremos mencionados.

En la Figura 4.3 se muestra el valor de la información I para diferentes valores de H en el intervalo $[0.0, 0.8]$. La información es máxima cuando $H \approx 0.48$; sin embargo, todos los valores del parámetro $H \in [0.35, 0.65]$ asumen valores similares en I y, por lo tanto, este intervalo contiene la mayor cantidad de información en términos de las comunidades de la

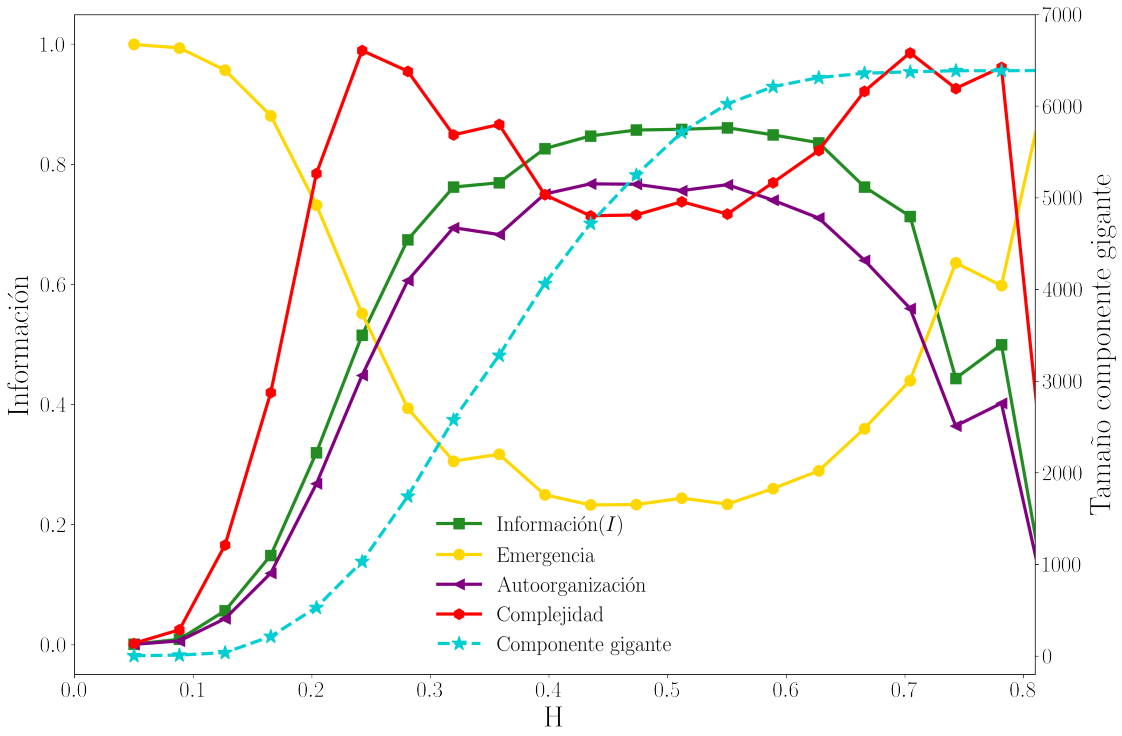

Figura 4.3: Medida de información. Se muestra el valor de la información I para diferentes valores de H en el intervalo $[0.0, 0.8]$, la emergencia E, la autoorganización S, la complejidad C y el tamaño de la componente gigante. Nótese que en el intervalo $[0.35,\ 0.65]$ I asume los valores más altos y por lo tanto este rango contiene la mayor cantidad de información en términos de las particiones de las redes de similitud.

red de similitud. Esto es consistente con lo mencionado en la sección anterior; es decir, que los valores intermedios de H son aquellos para los cuales se observa un comportamiento con comunidades no triviales en las redes de similitud. En este intervalo el número de enlaces es suficientemente grande para revelar relaciones significativas entre las palabras de la red pero suficientemente pequeño para generar redes que no se encuentran completamente conectadas.

Es posible explorar otras medidas para analizar la escala o escalas en la cuales las comunidades son más informativas. Diversas medidas de complejidad han sido introducidas para estudiar la información que se genera en diversos procesos. Por ejemplo, en [44] se propone una medida de complejidad definida como el producto entre la emergencia y la autoorganización de un sistema, de acuerdo a esta noción, la complejidad representa un balance entre cambio y orden. Por una parte, la emergencia cuantifica la cantidad de información producida cuando ocurre un cambio de escala en la descripción de un sistema. Para un sistema discreto la emergencia se calcula como [45]:

$$E = -k \sum_{i=1}^{N} p_i \log_2(p_i) \tag{4.4}$$

donde p_i representa la probabilidad del elemento (o estado) i del sistema. La cantidad $k = \frac{1}{\log_2(b)}$ corresponde a un factor de normalización donde b es igual al tamaño del alfabeto del sistema [45].

Por otra parte, la autoorganización S representa la cantidad de orden en un sistema y se define como el cambio (positivo o negativo) en la información, esta cantidad corresponde al complemento de la emergencia; es decir [45]:

$$S = 1 - E \tag{4.5}$$

Finalmente, la complejidad se calcula como el producto de la emergencia E y la autoorganización S [44, 45]:

$$C = 4 \cdot E \cdot S \tag{4.6}$$

De esta ecuación es fácil notar que la complejidad representa un balance entre diversidad y orden. En el contexto de este trabajo es de interés calcular la complejidad de la partición que define a las comunidades. De esta forma, para calcular la emergencia E de la partición, se define p_i como el número total de enlaces intracomunidad contenidos en la comunidad i divido por el número total de enlaces intracomunidad de toda la red. La emergencia E, autoorganizaación S, y complejidad C se presentan en la Figura 4.3 junto a la medida información I definida en la ecuación (4.3). Se pueden notar diferentes aspectos interesantes de estas curvas. Primero, obsérvese que la medida de información definida en la ecuación (4.3) es muy similar a la autoorganización, aparentemente siendo la única diferencia un factor multiplicativo constante. Los resultados de estos cálculos indican que la medida de información I cuantifica, en gran medida, la cantidad de orden presente en las comunidades. En el intervalo en el cual I y S son más altas, la complejidad asume valores relativamente altos.

Por otra parte, la complejidad es máxima ($C = 1$) para $H \approx 0.24$ y $H \approx 0.7$; esto implica un balance exacto ($E = 0.5$, $S = 0.5$) entre emergencia y autoorganización. De acuerdo a [45], una situación de máxima complejidad ocurre cuando uno o pocos estados del sistema concentran la mayor parte de la probabilidad; en el caso particular de las comunidades, esto se traduce en una o pocas comunidades que concentran la mayor parte de los enlaces intracomunidad. Debido a que I cuantifica la cantidad de orden en la partición, esto implica que para los puntos de máxima complejidad, las comunidades presentan un balance máximo entre diversidad y orden. Para valores de H entre 0.24 y 0.7 la complejidad disminuye debido a que las comunidades correspondientes a estos valores aumentan su organización creando un desbalance entre S (o I) y E.

Para analizar los resultados obtenidos a partir de esta medida de información se considera el valor de H para el cual I asume su valor máximo y se genera la red de similitud correspondiente. Al aplicar el algoritmo de Louvain se obtiene un conjunto de 976 comunidades, muchas de las cuales contiene un solo nodo. En la Figura 4.4 se grafican los tamaños de las comunidades más grandes de la partición (aquellas con más de 4 nodos): Únicamente se observa una comunidad con más de 1000, 8 comunidades con menos de 1000 y más de 100 nodos, 7 comunidades con menos de 100 y más de 10 nodos y el resto de comunidades

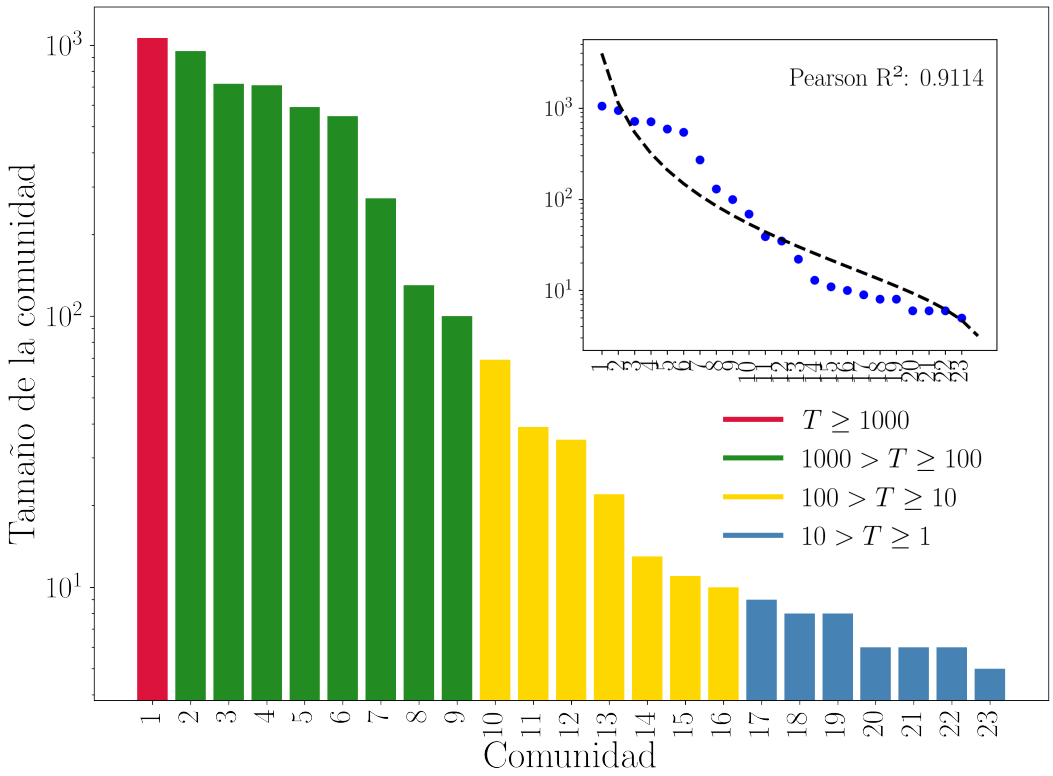

Figura 4.4: Número de palabras en comunidades con más de 4 elementos. Se muestra el tamaño de las 23 comunidades más grandes en la red. El color de cada barra codifica el tamaño de dichas comunidades en escala logarítmica. Nótese que la mayor parte de las palabras se encuentra contenida en las primeras 9 comunidades. El recuadro interior muestra el ajuste de una distribución beta discreta generalizada.

con menos de 10 nodos; por lo tanto, la mayor parte de los nodos se encuentran contenidos en comunidades con un tamaño mayor a 100. En la Tabla 4.2 se presentan las 10 palabras con mayores frecuencias de aparición en las 9 comunidades más grandes de la red. Se puede notar como las palabras dentro de una misma comunidad se encuentran relacionadas. Por ejemplo, en la comunidad 2 se observan palabras como "quantum", "field", "theory", "wave", "particle" que pueden relacionarse con el área de la teoría cuántica de campos. Por otra parte, en la comunidad 3 se observan las palabras "electron", "energy", "scatter", "measurement", "atom", "ion", "reaction", "collision" que están asociadas con teoría de dispersion o en la comunidad 6 se encuentran palabras como "matter", "quark", "boson", "meson", "dark", "qcd", "higgs" que pueden encontrarse en la física de altas energías. Nótese que palabras como "energy" son comunes en todas la áreas de la física por lo que las comunidades encontradas en este análisis no deben interpretarse como fronteras bien marcadas entre las diferentes disciplinas de la física.

Por otra parte, es interesante notar que la gŕafica en la Figura 4.3 parece estar descrita por una distribución beta discreta generalizada, distribución de dos parámetros que permite describir el comportamiento de una gran diversidad de fenómenos que surgen en las artes y en las ciencias sociales y naturales. Esta distribución está dada por $f(r) = A\frac{(N-1+r)^b}{r^a}$

Tabla 4.2: Comunidades más grandes de la red de similitud correspondiente a $H = 0.48$. Para cada comunidad se muestran las diez palabras más frecuentes en las listas de títulos publicados por revistas de la APS. En cada columna P denota Palabra, T denota el tamaño de cada comunidad, y F denota la frecuencia de aparición.

C	T	P1	F	P2	F	P3	F	P4	F	P5
1	1062	structure	29047	study	26543	surface	22235	property	18210	temperature
2	951	quantum	51432	model	45696	field	33569	theory	31667	wave
3	720	electron	35565	energy	30400	scatter	26372	measurement	15777	resonance
4	712	state	43193	spin	35726	system	26265	transport	10358	photon
5	591	effect	39909	phase	30087	transition	28683	order	16666	couple
6	547	production	9594	matter	7958	quark	7548	boson	6235	meson
7	272	gravitational	4941	factor	4256	cosmic	3871	binary	3449	propagation
8	130	reversal	1492	walk	1416	synchronization	1355	front	1172	delay
9	100	break	4140	granular	2360	shear	2339	turbulent	1116	viscosity

F	P6	F	P7	F	P8	F	P9	F	P10	F
16698	crystal	14429	spectrum	12882	ray	12194	liquid	11994	film	11451
22681	interaction	20819	lattice	17348	charge	16051	density	13492	particle	13423
15210	ion	14213	atom	12874	reaction	12267	collision	11415	excitation	10966
9061	tunnel	7271	dot	6252	hall	5887	bind	5824	control	5370
16268	behavior	8063	fluctuation	8009	disorder	7116	induce	7036	polarization	6731
6047	dark	5600	qcd	4857	search	4785	higgs	3899	vector	3002
2701	star	2509	rotate	2225	background	2059	gamma	1863	detector	1652
161	cone	831	real	811	waals	798	passage	604	fourier	414
1097	convection	1089	taylor	572	breaking	510	intermittency	448	vesicle	317

donde r es el valor del rango, N es el rango máximo y A es una constante de normalización [46]. Nótese que esta distribución se reduce a una ley de potencias en el caso cuando b = 0. Al realizar el ajuste de esta distribución por medio de una regresión lineal múltiple se obtiene como resultado un coeficiente de correlación de Pearson igual a $R^2 = 0.9114$ y valores de los parámetros $a = 1.77$, $b = 0.47$. Este ajuste se muestra en el recuadro interior de la Figura 4.4.

De esta manera, los métodos explorados en este capítulo permiten analizar las palabras en títulos de artículos publicados en revistas de la APS utilizando el formalismo de la teoría de redes. Partiendo de las listas de títulos se generaron redes de similitud definidas en términos de un parámetro adimensional H que define la escala a la que se realiza el análisis. Utilizando el algoritmo de Louvain y mediante la variación de este parámetro se mostró que es posible explorar las comunidades de nodos de estas redes de similitud y encontrar particiones que revelan las relaciones entre las palabras de la física a diferentes escalas. Se encontró que algunas de estas palabras pueden moverse de una comunidad a otra conforme se cambia la escala definida por el parámetro H lo cual muestra que cada palabra puede estar asociada con diferentes grupos de palabras.

Este tipo de análisis es interesante pues muestra la existencia de una dinámica no trivial en las relaciones entre palabras determinadas por la comunidades de las redes de similitud y permite explorar cómo las nodos forman grupos de palabras diferentes dependiendo la escala. Otro aspecto importante es que la información de estas relaciones entre grupos de palabras proviene únicamente de los datos en los títulos de artículos y de la descripción de esta información en términos de la teoría de redes.

Conclusiones

En este trabajo se analizó la base de datos [24] que contiene listas con los títulos de artículos publicados en revistas de la American Physical Society (APS) desde el año 1893. Aunque esta base de datos ha sido analizada en el contexto de las redes de colaboración científica y el impacto de la interdisciplinariedad, la información contenida en los títulos de artículos no ha sido explotada ni estudiada en algún trabajo previo. Debido a que el número de artículos y por lo tanto, el número de títulos, es muy grande ($\sim$ 450,000 artículos) la información contenida en este segmento de la base de datos puede revelar la existencia de patrones en los datos previamente obviados.

Para extraer la información de esta base de datos se utilizaron diversos métodos que permitieron explorar varios aspectos de esta base de datos. Debido a la naturaleza de la base de datos se implementó un filtrado de los datos utilizando técnicas de procesamiento de lenguaje natural y a partir de la entropía de Shannon como medida de diversidad de un sistema se encontró que las revistas de la APS muestran una diversidad constante en los últimos años, resultado que es consistente con trabajos previos que han explorado esta base de datos. También se introdujo un método gráfico basado en matrices para la identificación de periodos de tiempo durante los cuales la investigación científica en física ha permanecido estable; es decir, sin cambios en los temas de investigación.

En la parte principal de este trabajo se construyeron redes de palabras con $N \approx 6300$ nodos a partir del procesamiento de las listas de títulos de artículos. La representación de palabras como una red permitió analizar la base de datos a diferentes escalas, determinadas por un parámetro adimensional H, en términos de comunidades de palabras. Al analizar las comunidades de estas redes para diferentes valores de H, en el intervalo [0.0, 1.0], se encontró la existencia de una estructura no trivial en la cual las palabras son parte de diferentes grupos dependiendo de la similaridad con otras palabras. La variación de este parámetro permitió explorar estos grupos a diferentes escalas y encontrar áreas o subáreas de la investigación que se corresponden, en algunos casos, con áreas de la física bien establecidas. Debido a que la modularidad no cuantifica la escala de H que contiene más información, se introdujo una medida de información I que asume valores entre 0 y 1, basada en la entropía de Shannon, para extraer la escala más informativa partiendo de dos casos extremos: aquel en el cual todos los nodos se encuentran contenidos en una sola comunidad y el otro en el cual cada nodo corresponde a una comunidad. Se encontró que existe un rango de valores $[0.35, 0.65]$ de H para el cual la información de las comunidades es significativa; es decir, en este intervalo la medida de información propuesta es muy alta

$(0.80 \leq I \leq 0.85)$ y es posible observar particiones que contienen grupos de palabras altamente conectadas. Como caso particular se analizó el valor $H = 0.48$ para el cual $I = 0.85$, se encontró un total de 976 comunidades entre las cuales se pueden distinguir grupos de palabras que pueden asociarse con aŕeas de la física como teoría cuántica de campos o física de altas energías.

Trabajo futuro

En esta sección se presentan algunas ideas sobre las posibles direcciones que puede tomar este trabajo en el futuro. En primer lugar, sería deseable mejorar el filtrado de los datos para obtener listas de palabras cuyo significado sea estrictamente físico, esto permitiría crear redes de palabras más robustas y resultados más significativos respecto a la partición en comunidades. En segundo lugar, como se demostró en el capítulo 3, se ha observado la existencia de "universalidades" a largo de las revistas de la APS; en este sentido, se podría estudiar si este tipo de patrones están presentes en revistas pertenecientes a otras áreas de la investigación científica. Respecto a las redes de palabras, se mostró la existencia de una dinámica no trivial en el movimiento de palabras entre comunidades conforme se varía la escala (determinada por el parámetro H); sin embargo, este aspecto no se exploró detalladamente para cada una de las palabras consideradas. En un trabajo futuro sería interesante analizar y caracterizar esta dinámica de las palabras y estudiar si es posible realizar una clasificación de acuerdo al tipo de dinámica que presenta cada palabra. Por otra parte, sería importante buscar formalizar la medida de información I a través de argumentos más rigurosos y explorar si esta cantidad puede ser aplicada a otro tipo de sistemas en los cuales también se pueden definir diferentes escalas.

Bibliografía

[1] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[2] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[3] AP Riascos and José L Mateos. Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. *Scientific reports*, 10(1):1–14, 2020.

[4] Ricard V. Solé and Romualdo Pastor-Satorras. *Complex networks in genomics and proteomics*, chapter 7, pages 145–167. John Wiley Sons, Ltd, 2002.

[5] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Jan 2003.

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[7] Olaf Sporns, Dante R. Chialvo, Marcus Kaiser, and Claus C. Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, 2004.

[8] Victor M. Eguiluz, Dante R. Chialvo, Guillermo A. Cecchi, Marwan Baliki, and A. Vania Apkarian. Scale-free brain functional networks. *Physical Review Letters*, 94(1):018102, 2005.

[9] Jose M. Montoya and Ricard V. Solé. Small world patterns in food webs. *Journal of Theoretical Biology*, 214(3):405–412, 2002.

[10] Albert-László Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.

[11] Frank Schweitzer, Giorgio Fagiolo, Didier Sornette, Fernando Vega-Redondo, Alessandro Vespignani, and Douglas R. White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009.

[12] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[13] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, Feb 2011.

[14] Mark Newman. *Networks*. Oxford university press, 2018.

[15] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb 2010.

[16] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, Nov 2004.

[17] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Dec 2004.

[18] J. A. Ruiz Gayosso. Movilidad humana en sistemas de transporte aéreo, 2021. Tesis Licenciatura en Física, Universidad Nacional Autónoma de México.

[19] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008.

[20] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul 2006.

[21] V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*, 84:016114, Jul 2011.

[22] V. A. Traag, G. Krings, and P. Van Dooren. Significant scales in community structure. *Scientific Reports*, 3(1), Oct 2013.

[23] American physical society journals. `https://journals.aps.org/`.

[24] Descripción de la base de datos de la aps. `https://journals.aps.org/datasets`.

[25] Stefan Thurner, Wenyuan Liu, Peter Klimek, and Siew Ann Cheong. The role of mainstreamness and interdisciplinarity for the relevance of scientific papers. *PLOS ONE*, 15(4):1–14, 04 2020.

[26] Murali Krishna Enduri, I. Vinod Reddy, and Shivakumar Jolad. Does diversity of papers affect their citations? evidence from american physical society journals. In *Proceedings of the 2015 11th International Conference on Signal-Image Technology amp; Internet-Based Systems (SITIS)*, page 505–511, USA, 2015. IEEE Computer Society.

[27] Alessandro Pluchino, Giulio Burgio, Andrea Rapisarda, Alessio Emanuele Biondo, Alfredo Pulvirenti, Alfredo Ferro, and Toni Giorgino. Exploring the role of interdisciplinarity in physics: Success, talent and luck. *PLOS ONE*, 14(6):1–15, 06 2019.

[28] Jiawei Xu, Chao Min, Win-Bin Huang, and Yi Bu. Interdisciplinarity vs. unidisciplinarity: A structural comparison of multi-generation citations and references. In *Proceedings of the 18th international conference on scientometrics and informetrics (ISSI 2021)*, Aug 2021.

[29] Filippo Radicchi and Claudio Castellano. Rescaling citations of publications in physics. *Phys. Rev. E*, 83(4):046116, Apr 2011.

[30] Sidney Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58(6):49–54, 2005.

[31] Young-Ho Eom and Hang-Hyun Jo. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific Reports*, 4(1), Apr 2014.

[32] Jacopo Iacovacci, Zhihao Wu, and Ginestra Bianconi. Mesoscopic structures reveal the network between the layers of multiplex data sets. *Phys. Rev. E*, 92:042806, Oct 2015.

[33] Xiaomei Bai, Fuli Zhang, Jin Ni, Lei Shi, and Ivan Lee. Measure the impact of institution and paper via institution-citation network. *IEEE Access*, 8:17548–17555, 2020.

[34] Hua-Wei Shen and Albert-László Barabási. Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, 111(34):12325–12330, 2014.

[35] Jiang-Pan Wang, Qiang Guo, Lei Zhou, and Jian-Guo Liu. Dynamic credit allocation for researchers. *Physica A: Statistical Mechanics and its Applications*, 520:208–216, Apr 2019.

[36] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[37] Zhaolong Ning, Yuqing Liu, and Xiangjie Kong. Social gene — a new method to find rising stars. In *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6, 2017.

[38] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359(6379):eaao0185, 2018.

[39] An Zeng, Zhesi Shen, Jianlin Zhou, Jinshan Wu, Ying Fan, Yougui Wang, and H. Eugene Stanley. The science of science: From the perspective of complex systems. *Physics Reports*, 714-715:1–73, 2017.

[40] Gerardo Iñiguez, Carlos Pineda, Carlos Gershenson, and Albert-László Barabási. Dynamics of ranking. *Nature Communications*, 13(1), Mar 2022.

[41] Germinal Cocho, Jorge Flores, Carlos Gershenson, Carlos Pineda, and Sergio Sánchez. Rank diversity of languages: Generic behavior in computational linguistics. *PLOS ONE*, 10(4):1–12, Apr 2015.

[42] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[43] J. U. Martínez-González and A. P. Riascos. Activity of vehicles in the bus rapid transit system Metrobús in Mexico City. *Scientific Reports*, 12:1–11, Jan 2022.

[44] Carlos Gershenson and Nelson Fernández. Complexity and information: Measuring emergence, self-organization, and homeostasis at multiple scales. *Complexity*, 18(2):29–44, 2012.

[45] Guillermo Santamaría-Bonfil, Carlos Gershenson, and Nelson Fernández. A package for measuring emergence, self-organization, and complexity based on Shannon entropy. *Frontiers in Robotics and AI*, 4, 2017.

[46] Gustavo Martínez-Mekler, Roberto Alvarez Martínez, Manuel Beltrán del Río, Ricardo Mansilla, Pedro Miramontes, and Germinal Cocho. Universality of rank-ordering distributions in the arts and sciences. *Plos One*, 4(3):e4791, 2009.