



TESIS DOCTORAL

MODELOS COMBINATORIOS DE SISTEMAS COMPLEJOS: MÉTODOS Y ALGORITMOS

AUTOR

ING. MARIANO GASTÓN BEIRÓ

DIRECTOR DE TESIS

DR. ING. JOSÉ IGNACIO ALVAREZ-HAMELIN – FIUBA

JURADO DE TESIS

DR. PABLO BALENZUELA – FCEN-UBA

DRA. FLAVIA BONOMO – FCEN-UBA

PH.D. ALESSANDRO VESPIGNANI – NU (USA)

DR. MIGUEL VIRASORO – UNGS

LUGAR DE TRABAJO

GRUPO DE REDES COMPLEJAS Y COMUNICACIÓN DE DATOS

DEPARTAMENTO DE ELECTRÓNICA, FACULTAD DE INGENIERÍA – UBA,

EN EL MARCO DE LA BECA DOCTORAL PERUILH FIUBA Y LA BECA TIPO II CONICET.

FACULTAD DE INGENIERÍA – UNIVERSIDAD DE BUENOS AIRES

BUENOS AIRES, NOVIEMBRE DE 2013

Modelos Combinatorios de Sistemas Complejos: Métodos y Algoritmos

Mariano G. Beiró

Índice general

Resumen	1
<i>Overview</i>	3
1. Introducción	5
1.1. Introducción a los Sistemas Complejos	7
1.1.1. Definición y ejemplos	9
1.1.2. Surgimiento y evolución histórica	18
1.1.3. Sistemas Complejos como una ciencia interdisciplinaria	19
1.1.3.1. La Matemática y los sistemas complejos	21
1.1.3.2. La Física y los sistemas complejos	21
1.1.3.3. Las Ciencias de la Computación y los sistemas complejos	21
1.2. Modelos de sistemas complejos	22
1.2.1. Problemas inherentes al modelado de sistemas complejos	27
2. Modelos Combinatorios de Sistemas Complejos	29
2.1. Introducción a los grafos de red	29
2.1.1. Notación y representación de grafos de red	30
2.1.2. Invariantes de los grafos de red	36
2.1.2.1. Conectividad	36
2.1.2.2. Arista-conectividad	37
2.1.2.3. Diámetro	37
2.1.2.4. Coeficiente de agrupamiento	37
2.1.2.5. Distribución de grados y grado medio	39
2.1.2.6. Distribución de grados de los vecinos	39
2.1.2.7. Concordancia por grados de los vértices	40
2.1.3. Medidas de centralidad de los vértices y aristas	41
2.1.3.1. Intermediación	41
2.1.3.2. Cercanía	42

2.1.3.3.	Centralidad de vector propio	42
2.1.3.4.	Índice de capa	43
2.1.3.5.	Índice de denso	45
2.1.4.	Resumen de notación	45
2.2.	Resultados teóricos y experimentales en redes complejas	47
2.3.	Modelos de redes complejas	53
2.3.1.	El modelo Erdős-Rényi	55
2.3.2.	Modelos de Internet	56
2.3.2.1.	El modelo de Waxman	57
2.3.2.2.	El modelo Barabási-Albert	58
2.3.2.3.	El modelo FKP	62
2.3.3.	Generalizaciones del modelo Erdős-Rényi	63
2.3.4.	Modelos de Redes Sociales	65
2.3.4.1.	El modelo Watts-Strogatz	65
2.3.4.2.	El modelo <i>planted l-partition</i>	68
2.3.4.3.	El modelo LFR	68
3.	Descubrimiento de Comunidades en Redes Sociales	73
3.1.	Introducción a la noción de comunidad	74
3.2.	Métodos de descubrimiento de comunidades	77
3.3.	Métricas de comparación	82
3.4.	Análisis del funcional Q (modularidad)	86
3.4.1.	Limitaciones	93
3.5.	El método FGP	93
3.5.1.	Formalización del algoritmo de Lancichinetti <i>et al.</i>	94
3.5.2.	Funciones objetivo	96
3.5.3.	El <i>fitness growth process</i> (FGP)	99
3.5.4.	Extracción de las comunidades	100
3.5.5.	Fundamentación del método en el límite termodinámico	102
3.5.6.	Complejidad computacional	103
3.5.7.	Resultados obtenidos	108
4.	Estudio de la Conectividad en Internet	121
4.1.	Introducción	121
4.2.	Estimación de la conectividad empleando k -núcleos	125
4.2.1.	Formalización del método	125
4.2.1.1.	Un teorema de expansión	125

4.2.1.2.	Arista-conectividad en sentido estricto y en sentido amplio	133
4.2.1.3.	Construcción de conjuntos núcleo-conexos	133
4.2.2.	Resultados obtenidos	135
4.2.2.1.	Árboles de Gomory-Hu	137
4.3.	Visualización de la conectividad en Internet	137
5.	Estudio del Agrupamiento en Redes Complejas	145
5.1.	Introducción	145
5.2.	Cálculo de la descomposición en k -densos	147
5.3.	Visualización de los modelos de agrupamiento	147
6.	Conclusiones	153
A.	Leyes de Potencias	157
A.1.	Propiedades matemáticas de las leyes de potencias continuas	158
A.2.	Ajuste empírico de leyes de potencias continuas	159
A.3.	Propiedad libre de escala	163
A.4.	Leyes de potencias discretas	165
A.4.1.	Ajuste de leyes de potencias discretas con leyes continuas	165
A.5.	Otras distribuciones de cola larga	166
B.	Redes Utilizadas	167
	Bibliografía	179
	Índice alfabético	193

Índice de figuras

1.1. Plegamiento de proteínas	10
1.2. Experimento de mundo pequeño	13
1.3. Red del club de karate de Zachary	14
1.4. Distribución de grados de la Web	15
1.5. Juego de la Vida	16
1.6. Modelo de la pila de arena de Bak <i>et al.</i>	17
1.7. Formalización de los modelos de sistemas complejos de R. Rosen	23
1.8. Modelos basados en agentes	26
2.1. Representación de un grafo	30
2.2. Cortes y cortes por aristas en grafos	35
2.3. Coeficiente de agrupamiento	38
2.4. Intermediación (<i>betweenness</i>)	42
2.5. Cercanía (<i>closeness</i>)	43
2.6. Centralidad de vector propio (<i>eigenvector centrality</i>)	43
2.7. Descomposición en k -núcleos (k -cores)	44
2.8. Descomposición en k -densos (k -densities)	47
2.9. Red de actores	49
2.10. Red de interacciones entre proteínas de <i>S. Cerevisiae</i>	51
2.11. Modelo Erdős-Rényi. Visualización	56
2.12. Modelo Erdős-Rényi	57
2.13. Modelo de Waxman. Visualización	58
2.14. Modelo de Waxman	59
2.15. Modelo Barabási-Albert	61
2.16. Modelo FKP	64
2.17. Modelo de configuración y modelo de grafo aleatorio con grados esperados específicos	65
2.19. Modelo Watts-Strogatz	67
2.18. Modelo Watts-Strogatz. Visualización	67

2.20. Modelo planted l -partition	69
2.21. Modelo LFR	71
3.1. Métodos espectrales de descubrimiento de comunidades. Red de <i>football</i> .	87
3.2. Interpretación de la modularidad como medida signada	89
3.3. Límite de resolución de la modularidad. Ejemplos	90
3.4. El proceso de crecimiento uniforme en la red de football	104
3.5. Método FGP. Comunidades descubiertas en la red de football	105
3.6. Método FGP. Estructuras mantenidas para optimizar el proceso	106
3.7. Resultados de los bancos de prueba BENCH1-4 (Parte I)	111
3.8. Resultados de los bancos de prueba BENCH1-4 (Parte II)	114
3.9. Método FGP. Una comunidad del grafo Web de stanford.edu	116
3.10. Comunidades obtenidas por Louvain en LiveJournal	119
4.1. La noción de distancia contraída	126
4.2. Frontera de Q	127
4.3. Ilustración del Teorema 1	129
4.4. Ilustración del Corolario 1	131
4.5. k -capas y clusters de un grafo	134
4.6. Cálculo de la arista-conectividad con árboles de Gomory-Hu	137
4.7. Arista-conectividad en la red AS-CAIDA 2013	138
4.8. Arista-conectividad en la red AS-DIMES 2011	139
4.9. Descomposición en k -cores y conjunto núcleo-conexo en sentido estricto de la red AS-CAIDA 2011	140
4.10. Descomposición en k -cores y conjunto núcleo-conexo en sentido estricto de la red AS-DIMES 2011	140
4.11. Evolución del núcleo central de Internet según CAIDA entre 2009 y 2013	143
5.1. Procedimiento para la descomposición en k -densos	148
5.2. Descomposición en k -densos del grafo de Sistemas Autónomos de Internet	150
5.3. Descomposición en k -densos del grafo de la red PGP	151
5.4. Descomposición en k -densos del grafo de la red metabólica de E. Coli . .	152
A.1. Leyes de potencias	159
A.2. Estimación de leyes de potencias	163

Índice de cuadros

1.1. Clasificación de los problemas científicos realizada por W. Weaver en 1948	7
1.2. Algunos hechos destacados en el estudio de los sistemas complejos	20
2.1. Resumen de la notación de Teoría de Grafos empleada en este trabajo . .	46
3.1. Estructuras cohesivas utilizadas para estudiar grupos sociales.	76
3.2. Resumen de la notación relativa a estructuras comunitarias (Parte 1) . .	78
3.3. Resumen de la notación relativa a estructuras comunitarias (Parte 2). . .	82
3.4. Comunidad natural de un vértice para $\alpha = 1$	97
3.5. Listado de bancos de prueba y sus parámetros	109
3.6. Listado de redes reales y sus parámetros	110
3.7. Resultados del banco de prueba BENCH5	112
3.8. Resultados del banco de prueba BENCH6	113
3.9. Resultados obtenidos para la red de bandas de jazz	115
3.10. Resultados obtenidos para la porción del grafo Web de <code>stanford.edu</code> . .	117
3.11. Resultados obtenidos para el grafo de la red social LiveJournal	118
4.1. Listado de grafos de exploraciones de Internet	142
4.2. Núcleo-conectividad de los grafos de Internet	142
B.1. Red de <i>football</i>	168
B.2. Red de bandas de <i>jazz</i>	169
B.3. Red de la Web de <code>stanford.edu</code>	170
B.4. Red AS-CAIDA 2009	171
B.5. Red AS-CAIDA 2011	172
B.6. Red AS-CAIDA 2013	173
B.7. Red AS-DIMES 2011	174
B.8. Red LiveJournal	175
B.9. Red de confianza de PGP	176
B.10. Red metabólica de <i>E. Coli</i>	177

Resumen

El objeto de estudio de esta tesis son los sistemas complejos: sistemas integrados por una gran cantidad de elementos que interactúan entre sí y de cuya interacción surge un comportamiento colectivo organizado. El tamaño de estos sistemas hace que sea prácticamente imposible estudiar su evolución a nivel microscópico. Esto obliga a desarrollar metodologías de trabajo en parte distintas a las de otras áreas de la ciencia.

En el área de Sistemas Complejos es de fundamental importancia la construcción de modelos. A través de los modelos se busca reproducir el comportamiento macroscópico de los sistemas complejos e inferir lo que sucede a pequeña escala en término medio, o cuál será el comportamiento macroscópico si el sistema aumenta su tamaño.

La *simulación* de un sistema es la ejecución de un modelo con el fin de reproducir sus comportamiento. Durante la simulación se aplican las reglas de interacción a las variables definidas en el modelo. Para que el modelo sea útil, y considerando que los sistemas están constituidos por una cantidad grande de componentes, es importante que las reglas del modelo sean simples, y escalen lo más eficientemente posible con el tamaño del sistema. Un buen modelo debe entonces encontrar un punto intermedio adecuado entre la minuciosidad de la descripción, la precisión de sus resultados y la escalabilidad.

La variedad de modelos existentes en esta área se corresponde con la imposibilidad de que un único modelo capture en forma íntegra el comportamiento del sistema. En esta tesis estudiamos los modelos combinatorios de sistemas complejos, que son aquellos en que el sistema se modela a través de una red, a la que denominamos *red compleja*. En su forma más general, una red es un conjunto de nodos conectados entre sí a través de aristas, y se representa matemáticamente a través de un grafo.

Nuestro aporte consiste en el desarrollo de métodos y algoritmos en el marco de los modelos combinatorios, para caracterizar ciertas propiedades de los sistemas complejos.

La tesis está organizada de la siguiente forma:

- En el Capítulo 1 introducimos el área de Sistemas Complejos y su evolución histórica, ofrecemos algunos ejemplos de sistemas complejos, y nos introducimos en la problemática del modelado.
- El Capítulo 2 explora el estado del arte en el área de modelado combinatorio, poniendo especial énfasis en aquellos trabajos o líneas de investigación que sirven de antecedente a los aportes de esta tesis. Este capítulo introduce gran parte de la notación utilizada durante todo el trabajo.
- En el Capítulo 3 abordamos una característica presente fundamentalmente en aquellas redes complejas que tienen una componente humana, como las redes sociales:

la estructura comunitaria. Desarrollamos una metodología para obtener comunidades en redes de gran escala. Utilizamos un marco formal para describir nuestro método, y en este marco ofrecemos fundamentos microscópicos de su funcionamiento. Aplicamos diversas métricas y herramientas de visualización para mostrar los resultados en redes reales y en bancos de prueba. Ponemos énfasis también en el estudio de la complejidad computacional y mostramos que el método es escalable.

- En el Capítulo 4 estudiamos la Internet como red de flujo de información, y aportamos un método para garantizar cotas inferiores de conectividad en la red en un tiempo lineal con el tamaño de la misma. El estudio de la conectividad de Internet es relevante porque permite a los proveedores de servicio mejorar la robustez de la red y aumentar la tolerancia a fallos. Nuestro algoritmo permite por ejemplo identificar puntos débiles de la red.
- Finalmente, en el Capítulo 5 desarrollamos una herramienta de visualización para estudiar el fenómeno del agrupamiento (*clustering*) propio de las redes complejas. Estudiamos diversas redes que poseen una estructura modular y jerárquica y mostramos que a partir de la visualización es posible diferenciar a simple vista algunos modelos de agrupamiento.

Overview

The subject of this dissertation are complex systems, which are systems formed by multiple elements interacting between them. From these interactions, an organized collective behavior emerges. The size of these systems makes it almost impossible to study their evolution on the microscopical level, so that typical methodologies in Complex Systems are essentially different from those in other fields of science.

Model building is of major importance in Complex Systems. Models are built in order to reproduce macroscopic behavior of these systems and then infer what happens in a small scale from a statistical point of view, or how the macroscopic behavior will evolve if the system grows.

System *simulation* is the execution of a model in order to reproduce the system's behavior. Throughout a simulation, interaction rules are applied between the variables defined in the model. In order for the model to be useful, and considering that these systems are formed by a great number of components, it is important for the rules to be as simple as possible, and to scale efficiently with the size of the system. Thus, a good model should find a trade-off between refinement, precision of its results and scalability.

The variety of existing models in this field is due to the inability for a single model to capture the full behavior of the system. In this dissertation we study combinatorial models of complex systems, in which the representation of the system is a network, which we call *complex network*. In general terms, networks are formed by nodes and edges connecting them. They are mathematically described by graphs.

Our contribution here is to develop methods and algorithms for combinatorial models, in order to study and characterize some properties of complex systems.

This dissertation is organized as follows:

- In Chapter 1 we introduce the Complex Systems field and some of its historical milestones. We offer some examples of complex systems and we introduce the modeling problem.
- Chapter 2 explores the state of the art in combinatorial modeling. We mainly focus in those results or research lines which are most related with our contributions and serve as precedent for this work. This chapter also introduces most of the notation used throughout the entire work.
- In Chapter 3 we deal with a property which is mainly found in networks with a human component, like social networks: community structure. We develop a methodology for obtaining communities in large-scale networks. We describe the method by using a formal framework in which we also offer microscopical arguments

for its correct behavior. By means of comparison metrics and visualization tools, we show the obtained results in both real networks and benchmarks. We also focus on the computational complexity and show that our method scales efficiently with the size of the networks.

- In Chapter 4 we study the Internet as an information flow network and we contribute with a method that provides lower bounds for network connectivity in linear time. Studying Internet connectivity is quite relevant because it allows service providers to improve the quality of service and increase fault tolerance. Our algorithm is able to identify weak points in the network, for example.
- Finally, in Chapter 5 we develop a visualization tool for studying the clustering phenomenon in complex networks. We analyze several hierarchical and modular networks. We use different types of clustering models on them and, by means of visualization, we show that one of the models better reproduces the original networks, and that it is possible to distinguish the models at a glance.

Capítulo 1

Introducción

“It is merely suggested that some scientists will seek and develop for themselves new kinds of collaborative arrangements; that these groups will have members drawn from essentially all fields of science; and that these new ways of working, effectively instrumented by huge computers, will contribute greatly to the advance which the next half century will surely achieve in handling the complex, but essentially organic, problems of the biological and social sciences.”

WARREN WEAVER, “SCIENCE AND COMPLEXITY”, 1948 [156]

“Complexity is the property of a real world system that is manifest in the inability of any one formalism being adequate to capture all its properties.”

DONALD MIKULECKY, 2001 [108]

Fenómenos como la traslación de la Tierra alrededor de su astro, el Sol, o el choque de dos bolas de billar, pueden ser correctamente modelados y explicados a partir de las leyes de la Mecánica Clásica. En cambio, la evolución de un gas dentro de un recipiente, a pesar de obedecer a esas mismas leyes, es irresoluble desde el punto de vista práctico debido al gran número de variables involucradas. La Física Estadística ofrece las herramientas necesarias para deducir –a partir de las leyes de la Mecánica Clásica– las propiedades macroscópicas del sistema en el estado de equilibrio.

Extrapolar este método del estudio de gases en un recipiente al estudio del comportamiento de las personas en una sociedad no parece en principio factible. No disponemos de leyes físicas elementales, y el comportamiento de las personas se tildaría de *impredecible* y *complejo*. Sin embargo, en una gran cantidad de situaciones se observa un *comportamiento macroscópico organizado*, como de hecho sucede en las movilizaciones masivas, la conformación de modas sociales y la difusión de un rumor. Más allá de preguntarnos si

estos hechos son deducibles a partir de leyes más elementales, nos planteamos si pueden ser entendidos como el resultado de las interacciones entre los individuos.

Esta digresión inicial nos permitirá comprender la clasificación que hiciera en 1948 el matemático Warren Weaver, pionero en prever el estudio de los Sistemas Complejos como una ciencia interdisciplinaria. Weaver clasificó a los problemas en aquellos de **complejidad desorganizada** (*disorganized complexity*) y de **complejidad organizada** (*organized complexity*), en términos de la dificultad de tratar con ellos y encontrar su solución [156].

Los problemas de complejidad desorganizada son aquellos en que las leyes que vinculan a las variables son conocidas, pero el número de variables es muy grande y típicamente el estado inicial o entrada del problema se desconoce. Si ese estado inicial o entrada puede considerarse de alguna manera *aleatorio*, entonces los métodos estadísticos entran en acción y permiten predecir ciertas propiedades globales –macroscópicas– del sistema como un todo. Weaver aclara que el enfoque estadístico no se restringe a problemas físicos sino que puede aplicarse a problemas de índole económico o social. Las fórmulas de Erlang¹ para el dimensionamiento de recursos y el Cálculo Actuarial son también consecuencia de este enfoque.

En los problemas de complejidad organizada también hay un gran número de variables. Estas variables se relacionan de una manera complicada, pero de ninguna forma aleatoria. Consideremos por ejemplo el comportamiento de un grupo de personas en una organización, o la manera en que la constitución genética de un individuo se expresa en sus rasgos característicos. Lejos estamos de conocer leyes que gobiernen ambos problemas. Sin embargo somos capaces de percibir que las variables interactúan de cierta forma, dando como resultado un todo orgánico.

En contraposición a estos problemas encontramos los **problemas simples**, en los cuales el número de variables es pequeño, y la forma en que estas variables interactúan es conocida. Estos problemas ocuparon a la física de los siglos XVIII, XIX y XX, y resultaron en grandes avances tecnológicos promoviendo la Revolución Industrial y más tarde el surgimiento de la llamada Era de la Información.

Por último, y para completar el esquema, existen problemas en los que se conocen las leyes rectoras pero en los que la sensibilidad del sistema a las condiciones iniciales impide predecir el resultado. Estos sistemas se denominan **sistemas caóticos**. En ellos, pequeñas alteraciones en la entrada suelen producir grandes perturbaciones observadas a la salida. Los modelos climáticos y los mercados de valores son ejemplos de estos sistemas.

El siguiente esquema muestra la clasificación que acabamos de delinear:

¹Véase “*Teletraffic Engineering and Network Planning*”, V.B. Iversen, 2010, págs. 108 y 232.

CLASE	CARACTERÍSTICAS ESENCIALES	EJEMPLOS
Simplicidad	- Pequeña cantidad de variables - Leyes de interacción conocidas	- Funcionamiento de un motor de combustión (a partir de variables macroscópicas) - Radiación de una antena
Complejidad desorganizada	- Gran cantidad de variables - Leyes de interacción conocidas - Visión macroscópica - Aleatoriedad	- Modelos matemáticos de poblaciones - Modelos de decaimiento radiactivo
Complejidad organizada	- Gran cantidad de variables - Interacción existente pero no formalizada - Visión orgánica	- Estudio de factores genéticos en la enfermedad - Estudio de las relaciones humanas y formación de grupos sociales
Caos	- Leyes de interacción conocidas - Inestabilidad - Dificultad de predicción	- Turbulencia en flúidos - Climatología

Cuadro 1.1: *Clasificación de los problemas científicos realizada por W. Weaver en 1948 [156].*

El objeto de estudio de la presente tesis –los *sistemas complejos*– se enmarca dentro de los problemas de complejidad organizada. Este primer capítulo se encuentra dividido en dos partes: en la primera presentaremos los sistemas complejos a través de sus propiedades y algunos ejemplos, y construiremos una definición. Luego haremos un breve recuento de la evolución histórica de su estudio. En la segunda parte del capítulo entraremos en el problema del modelado y la simulación.

1.1. Introducción a los Sistemas Complejos

Antes de intentar esbozar una definición de los sistemas complejos introduciremos dos nociones fundamentales relacionadas con ellos, y sobre las cuales existe un gran acuerdo de parte de la comunidad científica:

Los sistemas complejos son emergentes. Están constituidos por un gran número de elementos que interactúan entre sí. Estas interacciones son relativamente simples en su composición. Sin embargo, a partir de la multiplicidad de relaciones individuales, el sistema como un todo orgánico presenta ciertas características que han *emergido*, pero que no estaban presentes en los elementos individuales. A este surgimiento de una estructura o patrón novedoso y coherente lo llamamos *emergencia*.

Los sistemas complejos son auto-organizados (*self-organized*). Presentan en gran escala una estructura ordenada que, nuevamente, surge como resultado colectivo de las interacciones individuales. La organización no está controlada por ningún agente externo al sistema, ni tampoco por un miembro en particular del mismo; es más bien espontánea y descentralizada. Esto la hace robusta y tolerante a fallos. Un ejemplo práctico de este fenómeno en el contexto social es el llamado “comportamiento colectivo” (*collective behavior*) de los grupos sociales. En muchos casos esta auto-organización implica una estructura jerárquica.

Sobre los factores que originan la complejidad se ha hablado mucho, sobre todo desde la biología evolutiva, que ha buscado explicar de diversas formas la emergencia a partir de la selección natural. Desde la perspectiva de la ingeniería se han propuesto teorías en que la auto-organización es el resultado de un diseño optimizado frente a la escasez de recursos².

Por último, queremos mencionar un tema que ha sido y es causa de debate. Hemos dicho que los elementos constituyentes de los sistemas complejos interactúan de una forma que no es simple ni formalizable, pero que de esa interacción surgen propiedades globales que no poseían los elementos individuales. Cabe preguntarse cuál es la esencia de dichas interacciones, ya que la respuesta a esta pregunta podría decirnos algo esencial respecto a los sistemas complejos. Por un parte, el **reduccionismo científico** construido por Descartes y que ha impulsado con gran éxito a las ciencias naturales desde el siglo XVI postula que se puede tener un conocimiento completo de un sistema a partir del conocimiento de cada una de las partes que lo constituyen. Este enfoque, que tiene antecedentes en el atomismo griego, llevó por ejemplo a E. Zermelo a buscar un conjunto axiomático completo para las matemáticas, y a R. Dawkins a reducir la complejidad biológica a la selección natural. Según el reduccionismo, las interacciones serían deducibles a partir del conocimiento cabal de los elementos del sistema.

En contraposición al reduccionismo, el **holismo** o **emergentismo** postula la necesidad de ver al sistema como un todo. La comprensión de cada elemento particular no alcanza para comprender al sistema, y entonces concluimos que lo novedoso es la *interacción*. Es decir, la interacción de las partes es la que da como resultado un todo organizado. En esta perspectiva holística se arraigan, por ejemplo, la teoría psicológica de la Gestalt, la biología relacional de Rashevsky-Rosen³ y la filosofía de Hegel.

Incluso dentro del emergentismo se reconocen dos corrientes [40]: los emergentistas fuertes consideran que la auto-organización global no puede reducirse, ni siquiera en

²Véase el modelo Highly Optimized Tolerance (HOT) en el Ejemplo 4 de la Sección 1.1.1.

³Véase el libro de R. Rosen [136].

principio, a interacciones simples entre los individuos. Los emergentistas débiles en cambio, sostienen que a través de reglas sencillas de interacción entre las partes es posible obtener las características de un sistema complejo, como ser patrones globales y una estructura ordenada y jerárquica. El enfoque emergentista débil apunta a la construcción de modelos simples que permitirían simular sistemas complejos. Ejemplos de ellos son el Juego de la Vida⁴ de Conway [75] y los modelos de sistemas complejos basados en agentes.

La cuestión de si las leyes de interacción en los sistemas complejos pueden formalizarse o no aún sigue generando discusión. Pero, por lo pronto, podemos concluir lo siguiente: es necesario revertir el enfoque analítico –que se interesa por la naturaleza de las interacciones– para pasar a un enfoque sistémico –que, en cambio, se interesa por los efectos– para entender el comportamiento colectivo como el resultado macroscópico de complicadas y desconocidas interacciones individuales.

1.1.1. Definición y ejemplos

Integrando los conceptos introducidos previamente construimos la siguiente definición:

Definición. *Un sistema complejo es el resultado de la integración de componentes –en general heterogéneos– que interactúan entre sí. De esta interacción emerge un comportamiento colectivo, que no presenta ninguno de los componentes por sí solo. El sistema complejo es una estructura auto-organizada –muchas veces jerárquica– a través de cuyo ordenamiento los componentes colaboran armoniosamente para alcanzar una función o resultado global.*

Nuestra definición de sistema complejo está muy probablemente influida por el concepto de *sistema* de Edgar Morin como “*unidad global organizada de interrelaciones entre elementos, acciones o individuos*” [110]. Para Mario Bunge un sistema es “*un todo complejo cuyas partes o componentes están relacionadas de tal modo que el objeto se comporta en ciertos aspectos como una unidad y no como un mero conjunto de elementos*” [32].

A partir de la similitud entre las definiciones, cabe preguntarse si acaso todos los sistemas son de por sí complejos, o si existen sistemas más complejos que otros. Según Rolando García, por ejemplo, un sistema complejo es “*una totalidad organizada en la cual los elementos no son separables y, por lo tanto, no pueden ser estudiados aisla-*

⁴El Juego de la Vida es un famoso autómatas celular del que emergen patrones interesantes a partir de simples reglas. Al ser equivalente computacionalmente a una máquina de Turing, pone en discusión los límites de computabilidad de los sistemas complejos. Véase el Ejemplo 4 de la Sección 1.1.1.

damente” [74]. Sobre esta cuestión epistemológica remitimos al lector a [135] para más detalles.

Daremos a continuación una serie de ejemplos de sistemas complejos:

Ejemplo 1: Plegamiento de proteínas (*protein folding*)

Las proteínas son polímeros complejos de aminoácidos que las células sintetizan con el fin de que desempeñen diversas funciones biológicas. A través de un proceso denominado plegamiento las proteínas adquieren una estructura espacial tridimensional estable, que les permite realizar la función que deben llevar a cabo. Predecir la estructura tridimensional más estable de una proteína implica hallar el mínimo global de la función de energía libre, lo que es un problema difícil desde el punto de vista computacional.

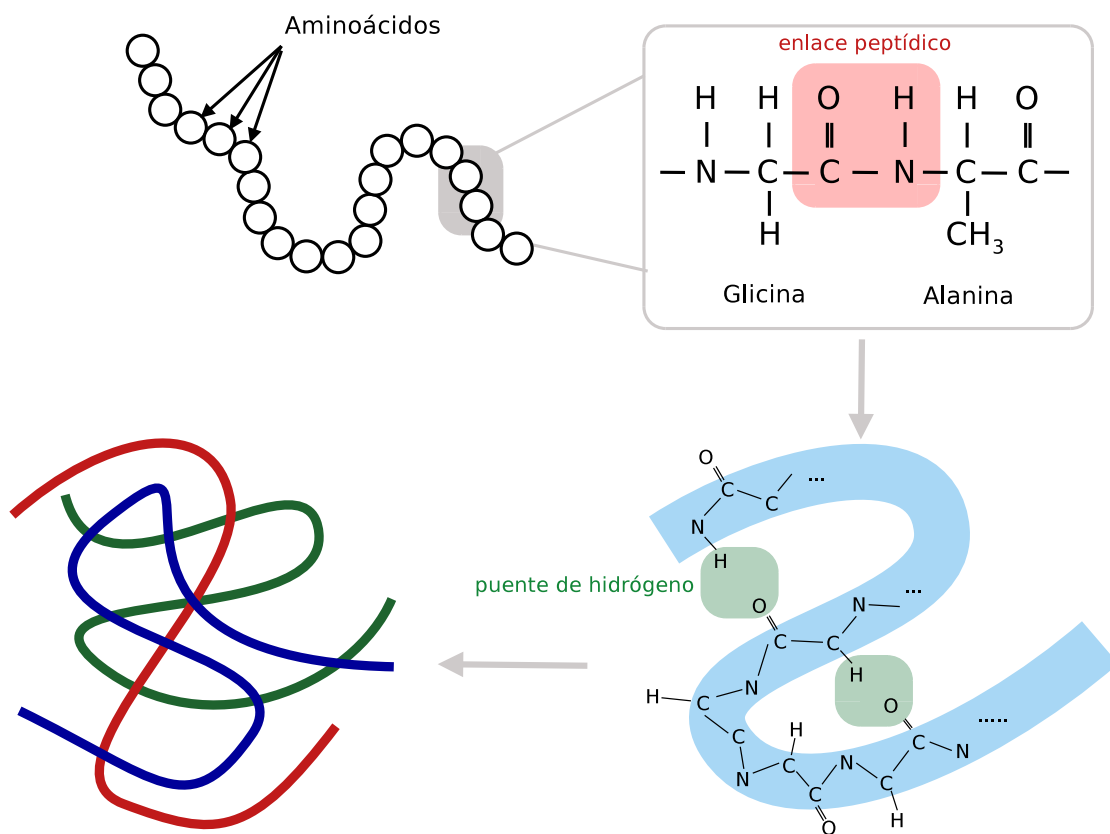


Figura 1.1: *Plegamiento de proteínas*. Las secuencias de aminoácidos que constituyen las proteínas se pliegan espacialmente en forma espontánea a partir de fuerzas iónicas e intermoleculares, desarrollando una estructura tridimensional particular en relación con la función desempeñada.

Aplicando el enfoque de sistemas complejos, nos encontramos con un sistema –la proteína– constituido por una gran cantidad de componentes –los aminoácidos–. El estudio de los aminoácidos en sí no nos permite deducir ni explicar la función desempeñada por la proteína. Sin embargo la proteína como un todo tiene una función global específica,

esta función está vinculada a su estructura, y su estructura es el resultado de la interacción entre los distintos aminoácidos, a través de enlaces covalentes, puentes de hidrógeno y enlaces disulfuro.

El problema computacional de hallar la estructura óptima de una proteína es NP-completo, en parte porque no es posible considerar cada aminoácido y determinar individualmente la posición que debe ocupar; el código de este proceso no está contenido en los aminoácidos sino en la cadena. Esta dificultad computacional contrasta con la simplicidad con que es resuelto el problema en los sistemas biológicos: la evolución temporal del sistema guiada por las leyes de la física conduce indefectible y sorprendentemente a la configuración estable en sólo unos microsegundos [159]. Dicho de otra forma, la naturaleza no explora todo el espacio de soluciones para encontrar la posición final⁵. Este proceso espontáneo es habitual en los sistemas biológicos y se denomina *auto-ensamblaje* (*self-assembly*).

Los métodos computacionales conocidos para resolver el problema de plegamiento de proteínas suelen emplear técnicas de inteligencia artificial y algoritmos de minería de datos (*data-mining*) para buscar la estructura óptima [67].

Ejemplo 2: El comportamiento social

Wilhelm Wundt, considerado el padre de la psicología experimental, postuló hacia el 1900 en su trabajo *Psicología de los Pueblos* la idea de que existe un comportamiento social que no puede ser descripto exclusivamente desde la perspectiva del individuo. Sus conceptos fueron desarrollados en los trabajos de Gustave Le Bon, William McDougall y Sigmund Freud en los años siguientes⁶, y dieron origen a una nueva disciplina conocida como Psicología Social.

A lo largo del siglo XX los psicólogos sociales estudiaron experimentalmente fenómenos como la influencia y la persuasión, la propagación de rumores, la construcción de identidad social, el sentido de pertenencia a un grupo y la cohesión, entre otros. Mencionaremos a continuación tres de ellos:

Experimento de conformidad de Asch. En 1950 Solomon Asch mostró cómo un grupo social puede ejercer presión sobre uno de sus individuos para que cambie su juicio o parecer sobre determinada cuestión. En sus experimentos, Asch planteaba un problema muy sencillo de resolver a un grupo de personas. Los primeros participantes –cómplices del experimentador– respondían en su mayoría incorrectamente, de manera intencional. Luego, el verdadero sujeto del experimento –que desconocía la verdadera naturaleza del

⁵Véase en este sentido la *paradoja de Levinthal* [104].

⁶Véase al respecto la *“Psicología de las masas y análisis del yo”*, S. Freud, 1921.

mismo— mostraba una tendencia a responder incorrectamente, a pesar de conocer la solución correcta.

Seis grados de separación. Stanley Milgram, quien fuera alumno de Asch y recordado por su controversial experimento sobre obediencia a la autoridad de 1963, llevó a cabo en 1967 el denominado *experimento de mundo pequeño* (*small world experiment*) [150]. Este experimento buscaba corroborar una tesis postulada varios años atrás por las ciencias sociales: que en las grandes poblaciones dos personas escogidas al azar están a una distancia promedio entre 5 y 6, medida en la cantidad de intermediarios necesarios para llegar de una a la otra. En este contexto se entiende por intermediario a una persona conocida por la persona anterior en la cadena.

Para corroborar esta hipótesis, Milgram diseñó el siguiente experimento: escogió a un grupo de 296 individuos en Estados Unidos, 196 de ellos habitantes del estado de Nebraska, y 100 habitantes de la ciudad de Boston. Estos 296 individuos oficiaron de iniciadores de un intercambio de correo dirigido a un mismo destinatario: un agente de bolsa de Boston. A cada uno de los 296 individuos se lo proveyó con cierta información básica respecto al destinatario: su nombre, dirección, estudios realizados, trabajo desempeñado, etc., y se le indicó que no intentara contactarlo directamente, sino que enviase su carta a un conocido que considerara como el más adecuado o cercano al destinatario. Este conocido sería el primer intermediario de la secuencia, y debería hacer lo suyo siguiendo las mismas reglas. A través de cadenas de intermediarios, 64 de los 296 individuos lograron hacer llegar sus cartas al destinatario común, y se estimó una distancia promedio de 5.12 intermediarios.

Como conclusión de su experimento Milgram planteó la necesidad de desarrollar modelos teóricos de interconexión en las redes sociales que explicaran este comportamiento de *mundo pequeño*. De entre ellos, destacamos el modelo Watts-Strogatz [154], de gran impacto, que discutiremos más adelante en este trabajo.

La tesis de que todo el mundo está conectado por un promedio de 6 intermediarios, conocida con el nombre de *seis grados de separación*, ha sido validada por recientes resultados experimentales de mayor envergadura [101].

Conflicto y fisión. Entre 1970 y 1972 W. Zachary estudió el comportamiento de los miembros de un club de karate [161]. A raíz de un conflicto entre quienes ejercían el liderazgo —el instructor y el administrador del club— se fueron conformando con el tiempo dos facciones, hasta que finalmente el club se fisiónó y quienes adherían al instructor conformaron una nueva organización. Mientras que los miembros no reconocían conscientemente la existencia de una división política, Zachary observó que efectivamente

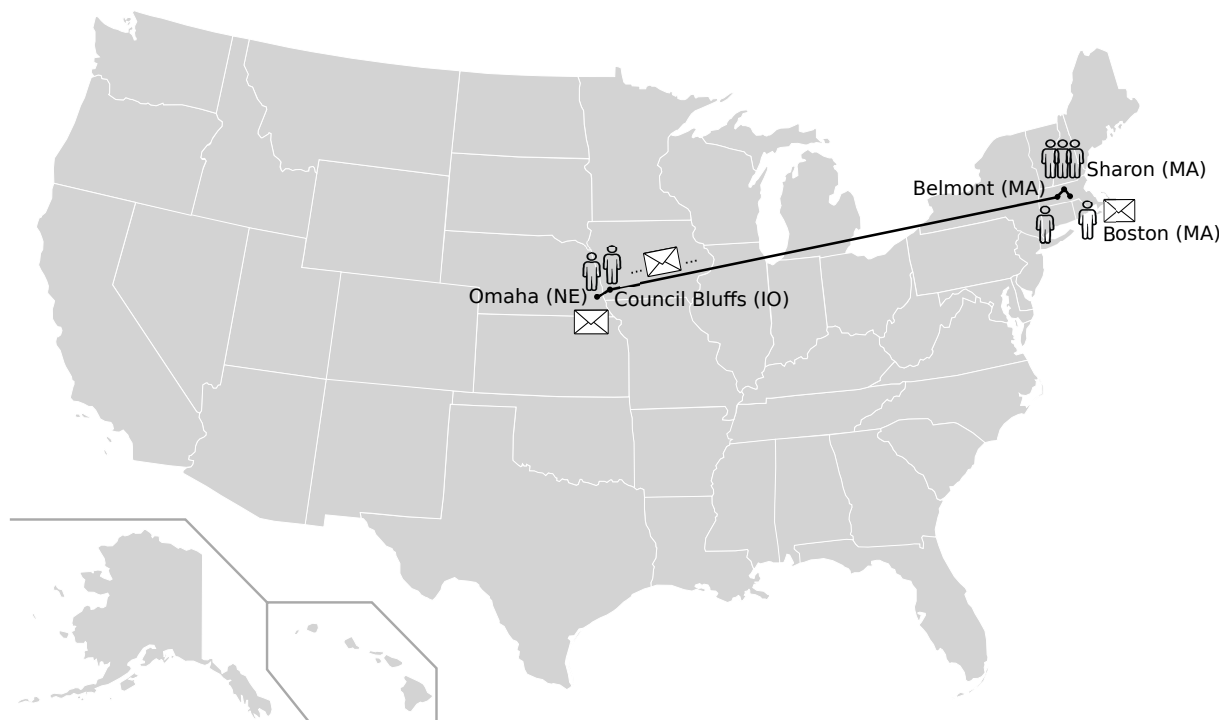


Figura 1.2: *Experimento de mundo pequeño*. 64 cartas llegaron hasta el destinatario en Boston a través de una cadena de intermediarios. Algunas de ellas mostraban un progresivo acercamiento geográfico, mientras otras realizaban un salto directo desde el estado inicial hasta Massachusetts. La distancia promedio fue de 5.2 intermediarios.

emergió una división del club en grupos, sostenida por relaciones de afinidad.

Siguiendo la idea de algunos antropólogos que lo antecedieron, Zachary representó la red social con un grafo, en donde los vértices representan miembros y las aristas reflejan una relación de amistad. A partir de herramientas de la Teoría de Grafos –en particular del teorema de Ford-Fulkerson de flujo máximo y corte mínimo– logró deducir la conformación de los dos grupos, que sería luego validada por la fisión del club.

Ejemplo 3: La World Wide Web

La Web es una red de distribución de información a nivel mundial, no centralizada. Sus unidades de información son los documentos, que se conectan a través de *hipervínculos* (*links*). En 1999, Barabási y Albert realizaron una exploración automatizada de la Web que recolectó datos de alrededor de 300000 documentos conectados por un millón y medio de hipervínculos⁷ [3]. A partir de estos datos analizaron la topología del grafo de la Web –un grafo dirigido en que los vértices son los documentos y las aristas dirigidas representan hipervínculos de un documento a otro–, obteniendo novedosos resultados:

- Observando la distribución de los grados de los vértices, descubrieron que dicha

⁷Los datos de la exploración se encuentran disponibles en el sitio web de Barabási.

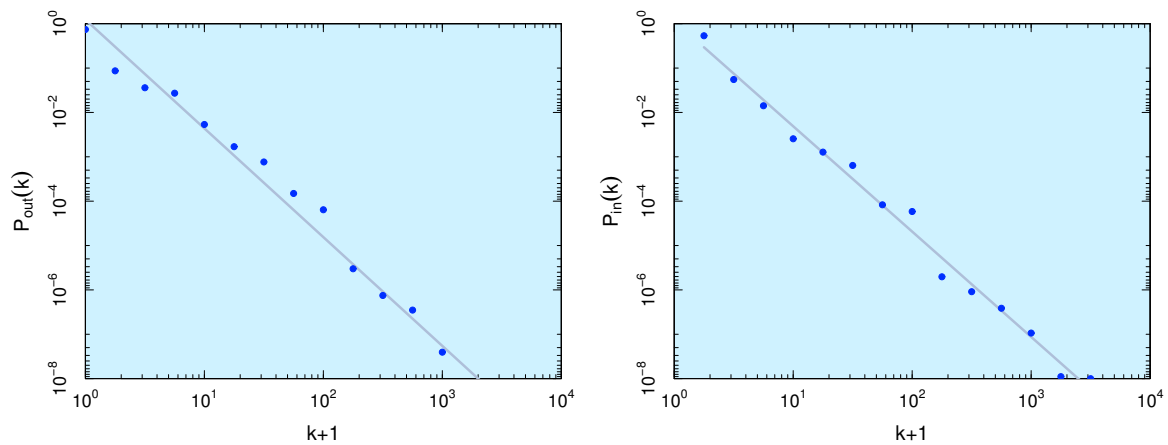


Figura 1.4: *Distribución de grados de la Web*. Barabási descubrió en 1999 que la distribución de la cantidad de hipervínculos que poseen los documentos en la Web sigue una *ley de potencias*. El gráfico muestra la estimación del grado externo (*out-degree*) (*Izq.*) y el grado interno (*in-degree*) (*Der.*) para la exploración realizada por Barabási. El histograma se realizó con un binning logarítmico, y la regresión lineal de los datos en escala log-log se ajusta en gran medida a una ley de potencias.

Ejemplo 4: Los autómatas celulares

Los autómatas celulares se utilizan para modelar sistemas complejos que evolucionan en el tiempo. Fueron propuestos por S. Ulam y J. von Neumann en la década del '40, y adquirieron fama a partir del autómata conocido como Juego de la Vida, desarrollado por J. Conway en 1970.

Un autómata celular es un reticulado cuyos elementos –denominados células– toman un estado de un conjunto finito $\mathbb{K}$. El conjunto de los estados de todas las células en un instante t discreto determinado conforma el estado del autómata en ese instante. El autómata parte de un estado inicial, y evoluciona en el tiempo de manera discreta siguiendo simples reglas. Las reglas expresan el estado de cada célula en el instante $t + 1$ en función de su propio estado y el de sus *vecinas* en el instante t .

El Juego de la Vida. En el Juego de la Vida el reticulado es una cuadrícula bidimensional infinita cuyas células $c_{i,j}$ tienen dos estados posibles: $\mathbb{K} = \{vivo, muerto\}$. Al estado de la célula $c_{i,j}$ en el instante t lo llamaremos $E(c_{i,j}, t)$. El estado en $t + 1$ dependerá del propio estado y de aquel de las células vecinas en t –considerando como vecinas a las 8 células que se encuentran en la periferia de $c_{i,j}$ –. Concretamente, $\mathcal{L}(c_{i,j}, t)$ representará el conjunto de células vecinas *vivas* de $c_{i,j}$ en el instante t , mientras $\mathcal{D}(c_{i,j}, t)$ representará el conjunto de vecinas *muertas* en ese instante. Las reglas de evolución son

entonces:

$$\begin{aligned}
 \text{si } E(c_{i,j}, t) = \text{muerto} \wedge |\mathcal{L}(c_{i,j}, t)| = 3 &\Rightarrow E(c_{i,j}, t+1) = \text{vivo} \\
 \text{si } E(c_{i,j}, t) = \text{vivo} \wedge |\mathcal{D}(c_{i,j}, t)| = 2 &\Rightarrow E(c_{i,j}, t+1) = \text{vivo} \\
 \text{si } E(c_{i,j}, t) = \text{vivo} \wedge |\mathcal{D}(c_{i,j}, t)| = 3 &\Rightarrow E(c_{i,j}, t+1) = \text{vivo} \\
 \text{sino} &\Rightarrow E(c_{i,j}, t+1) = \text{muerto} .
 \end{aligned}$$

En términos coloquiales, diríamos que una célula nace cuando en su vecindad hay 3 células vivas, y se mantiene viva siempre que en su vecindad haya 2 o 3 células vivas, de lo contrario muere.

La Figura 1.5 muestra la evolución del Juego de la Vida para una configuración inicial particular, durante los primeros 5 instantes.

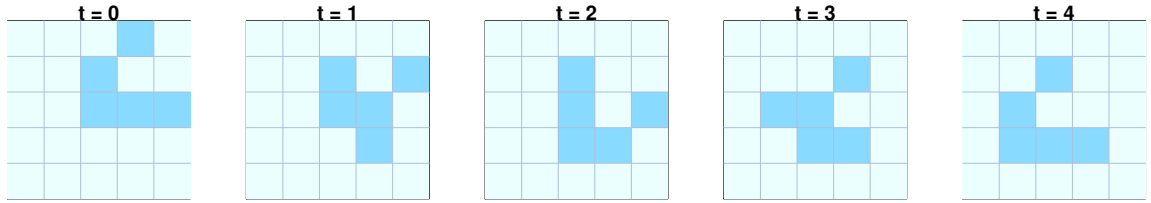


Figura 1.5: *Juego de la Vida*. Evolución durante los primeros 4 instantes a partir de una configuración inicial dada. Los estados se representan a través de los colores oscuro (*vivo*) y claro (*muerto*).

La pila de arena y la *self-organized criticality* (SOC). En 2002 S. Wolfram clasificó a los autómatas celulares en 4 tipos, en función de su comportamiento a largo plazo [158]. Los autómatas del cuarto tipo son los que más nos interesan, ya que presentan características típicas de la complejidad: dependencia de largo alcance y parámetros que siguen distribuciones libres de escala.

El primer autómata celular en el que se encontraron estos dos últimos fenómenos es el modelo de la pila de arena (*sandpile*) propuesto por Bak *et al.* en 1987 [13]. Este modelo considera –en su variante bidimensional– que en cada célula se apilan granos de arena que son arrojados al azar. Cuando sobre una misma célula se acumulan 4 granos, entonces se produce un colapso y los granos se distribuyen entre las 4 células *vecinas* –en este caso se considera vecinas sólo a las células que se encuentran arriba, abajo, a la izquierda y a la derecha–. Al simular este autómata, Bak *et al.* observaron el siguiente comportamiento:

- El colapso de una célula tiene en muchos casos un *efecto dominó o avalancha*,

produciendo el colapso de todo un *cluster de células* (conjunto de células transitivamente vecinas que incluye a la célula original).

- Si se mide el tamaño de los clusters afectados cada vez que hay un colapso, se observa una ley de potencias. Esto quiere decir que en algunas ocasiones –aunque sean pocas estadísticamente– el efecto dominó llega hasta lugares muy lejanos respecto a la célula en que se originó el colapso. Nos encontramos ante un fenómeno muy habitual en los procesos auto-semejantes, denominado *dependencia de largo alcance* (*long-range dependency*).
- Los tiempos de vida de los clusters responden también a una ley de potencias.

Bak *et al.* denominó al comportamiento de este modelo *self-organized criticality* (SOC), porque los estados de equilibrio son estados críticos, en donde una pequeña perturbación produce un fenómeno colectivo libre de escala (la avalancha). El modelo SOC explica el comportamiento de muchos fenómenos reales como los terremotos, las avalanchas y los relámpagos.

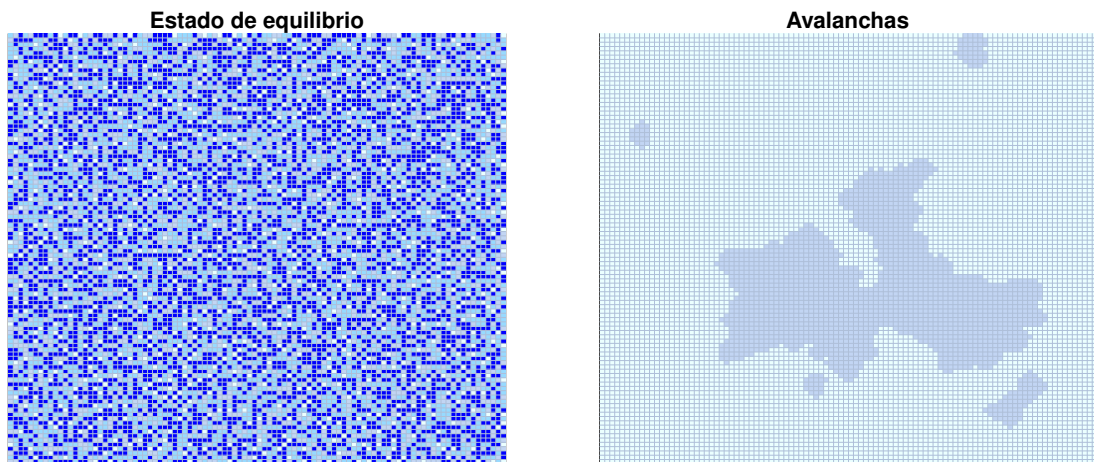


Figura 1.6: *Modelo de la pila de arena de Bak et al.* Para una grilla de 100×100 , se muestra a la izquierda un estado alcanzado luego de arrojar 100000 granos de arena al azar –los colores representan 1 grano (*gris*), 2 granos (*celeste*) o 3 granos (*azul*) acumulados–. A la derecha se muestran, para ese mismo estado, 5 posibles avalanchas. Una avalancha se dispara cuando cae un grano de arena sobre una célula que tiene acumulados 3 granos. Bak observó que la distribución de tamaños de las avalanchas sigue una ley de potencias.

Los autores también hacen una lectura de la evolución de la pila de arena usando *series temporales*, y observan que en ellas la auto-semejanza se refleja como ruido $1/f$ (*pink noise*).

Forest-fires. En 1990 Bak *et al.* propusieron un segundo autómatas celular denominado *forest-fire* [12, 62]. Este autómata simula un bosque en el que dinámicamente nacen árboles y se producen incendios que los destruyen, y presenta, al igual que el primero, el fenómeno de la criticalidad. En particular, Bak *et al.* se interesan aquí por el aspecto energético de la dinámica del sistema, ya que observan que la energía que ingresa al sistema con distribución uniforme en el espacio y el tiempo (modelada por el nacimiento de nuevos árboles) manifiesta una dimensión fractal al disiparse a través del fuego.

Highly Optimized Tolerance (HOT). Observando el comportamiento de los *forest-fires*, Carlson y Doyle cuestionan el mecanismo SOC y proponen un mecanismo de modelado de sistemas complejos denominado *Highly Optimized Tolerance (HOT)* [36]. En su trabajo, los autores sostienen que los sistemas complejos están optimizados (por ejemplo, a través de la selección natural o del diseño)⁹ de manera de ser robustos y eficientes. En este contexto, muestran de qué manera las leyes de potencias surgen como compromisos (*trade-offs*) entre la minimización de los costos y la maximización de la tolerancia a fallos.

En concreto, modifican los modelos de la pila de arena y *forest-fire* incorporando elementos diseñados específicamente para aumentar el *rendimiento* (en términos de la densidad de árboles o la estabilidad de la arena en la pila). En el caso del *forest-fire*, por ejemplo, se incorporan barreras contra incendio cuya disponibilidad es acotada y deben ser distribuidas sobre la grilla de la forma más conveniente. Mientras que en el modelo SOC la complejidad se observaba sólo en el punto crítico –determinado rango de densidades de árboles y tasas de provocación de incendios–, Carlson y Doyle sostienen que con un diseño optimizado la complejidad no depende de los parámetros del modelo.

En el fondo, Carlson y Doyle consideran que la complejidad de diseño de los sistemas complejos no se manifiesta generalmente en la estructura –salvo en casos específicos como los fractales–. Es decir que no es la estructura la que muestra características auto-asemejantes sino más bien el comportamiento, que es fruto de un diseño planificado y de la optimización.

1.1.2. Surgimiento y evolución histórica

Sería difícil –si no imposible– determinar el momento histórico en que el hombre aplicó por primera vez el enfoque sistémico para resolver un problema. Pero desde el punto de vista de los movimientos científicos del último siglo, reconocemos dos antecedentes importantes: la Escuela Austríaca de Economía y la Cibernética.

⁹Recuérdese la discusión sobre los factores que dan origen a la complejidad en la introducción del capítulo.

Los economistas de la Escuela Austríaca sostenían en los años '30 que en los mercados económicos, a partir del mutuo ajuste de las economías individuales, puede surgir un orden espontáneo beneficioso para la sociedad. Propusieron entonces modelos económicos basados en el libre mercado, la competencia y el *laissez-faire*. Sus mayores exponentes fueron L. von Mises, F. Hayek y C. Menger.

La Cibernética, por su parte, nació con el objetivo de estudiar a los sistemas que se regulan a sí mismos, en particular los seres vivos y las máquinas. Emparentada con la Teoría de Control, estudia a los sistemas desde la óptica de la realimentación (*feedback*). En términos generales los cibernéticos sostienen que la realimentación, en tanto comunicación entre las distintas partes del sistema, genera redundancia, y es esta redundancia la que disminuye la entropía del sistema y lo conduce a su auto-organización. De entre los cibernéticos del siglo XX destacamos a H. von Foerster, N. Wiener y J. von Neumann.

En el Cuadro 1.2 resumiremos algunos hitos históricos en el estudio de los sistemas complejos, desde 1950 hasta la actualidad.

1.1.3. Sistemas Complejos como una ciencia interdisciplinaria

La interdisciplinariedad es un aspecto esencial del trabajo en el área de Sistemas Complejos. W. Weaver, al hablar de los problemas de complejidad en 1948, ya había predicho que esta nueva ciencia requeriría del trabajo en equipo de matemáticos, físicos, ingenieros, psicólogos, entre otros expertos. A través de la especialización, cada área ofrecería sus recursos y técnicas para permitir que el grupo de trabajo tuviera una visión global del problema[156].

Además de las grandes áreas mencionadas por W. Weaver, podemos agregar la Química, la Biología, la Sociología y la Economía, por ejemplo. También un sinnúmero de disciplinas que surgen de la intersección de varias grandes áreas y se orientan a la comprensión de ciertos procesos desde un enfoque sistémico. Ejemplos de ellas son:

- **La Biología Sistémica:** Estudia a los Sistemas Biológicos a partir de sus interacciones, construyendo modelos matemáticos que expliquen, por ejemplo, su evolución y su función.
- **La Economía de Complejidad:** Estudia la auto-organización de la economía a partir de la dinámica de agentes individuales que interactúan entre sí. Utiliza herramientas de la Teoría de Juegos.
- **La Sociología Matemática:** Estudia fenómenos sociales a partir de modelos matemáticos. Permite analizar la estructura social y la formación de redes sociales.

1955	H. Simon propone el mecanismo de <i>acoplamiento preferencial</i> (<i>preferential attachment</i>) para explicar el origen de las leyes de potencias como la Ley de Pareto (1896), la Ley de Gibrat (1931) y la Ley de Zipf (1935).
1967	S. Milgram realiza el experimento de mundo pequeño [150].
1969	T. Schelling (Nobel de Economía, 2005) propone uno de los primeros modelos de sistemas complejos <i>basados en agentes</i> para estudiar la segregación racial.
1970	J. Conway diseña el autómata celular conocido como Juego de la Vida, en el que a partir de simples reglas emergen diversos patrones globales [75].
1975	B. Mandelbrot comienza a desarrollar la teoría fractal.
1984	Se crea el Santa Fe Institute, una referencia a nivel mundial en el estudio de los Sistemas Complejos. J. Holland acuña aquí el concepto de <i>sistemas complejos adaptativos</i> como una evolución de los sistemas basados en agentes. En los sistemas complejos adaptativos, los agentes tienen capacidad adaptativa (pueden aprender y adquirir experiencia).
1985	R. Rosen formaliza la modelización de los sistemas complejos a partir de la Teoría de Categorías.
1987	Bak <i>et al.</i> proponen el concepto de <i>self-organized criticality</i> (SOC) para fundamentar la existencia de distribuciones libres de escala en los sistemas complejos. Según el modelo SOC, los sistemas complejos estarían en el límite entre el orden y el caos. Lo ejemplifican a través del autómata de la <i>pila de arena</i> [13].
1989	Bak <i>et al.</i> introducen el <i>modelo de forest-fire</i> : un autómata celular que presenta la propiedad de <i>self-organized criticality</i> [12].
1993	Leland <i>et al.</i> encuentran que el tráfico de datos en las redes de alta velocidad muestra un comportamiento auto-semejante y dependencia de largo alcance [100].
1998	D. Watts (Santa Fe Institute) y S. Strogatz (Cornell University) proponen el modelo de mundo pequeño que lleva su nombre [154].
1999	A partir del modelo de <i>forest-fire</i> , J. Carlson y J. Doyle diseñan un mecanismo para modelar sistemas complejos del cual emergen leyes de potencias, denominado <i>Highly Optimized Tolerance</i> (HOT) [36].
1999	Barabási y Albert descubren una <i>ley de potencias</i> en la distribución de los enlaces de los documentos web [3].
1999	Faloutsos <i>et al.</i> descubren una <i>ley de potencias</i> en la topología de Internet [66].
1999	El <i>modelo Barabási-Albert</i> (BA), basado en <i>acoplamiento preferencial</i> (<i>preferential attachment</i>) es el primer modelo de grafo que logra capturar las distribuciones libres de escala de la Web e Internet [14].
1999	Fabrikant <i>et al.</i> proponen el <i>modelo FKP</i> : un modelo de grafo con distribución de grados libre de escala [65] inspirado en el mecanismo HOT.

Cuadro 1.2: Algunos hechos destacados en el estudio de los sistemas complejos.

Para el presente trabajo, nos interesan en particular las herramientas aportadas por tres grandes áreas que desarrollaremos a continuación: la Matemática, la Física y las Ciencias de la Computación.

1.1.3.1. La Matemática y los sistemas complejos

A través de la Matemática se formalizan los modelos de sistemas complejos, utilizando por ejemplo la Teoría de Grafos, los Autómatas Celulares, las Ecuaciones Diferenciales y la Teoría de Juegos. Nos interesan en particular los *modelos combinatorios*, basados en la Teoría de Grafos, que serán estudiados en la presente tesis. La representación de un sistema complejo a través de un grafo se denomina *red compleja*.

Es habitual en los sistemas complejos modelar el comportamiento de los elementos del sistema utilizando variables aleatorias, por esta razón son siempre necesarias las herramientas de la Teoría de Probabilidades y la Estadística. En particular, la introducción de componentes probabilísticos en los modelos basados en grafos (*redes complejas*) da lugar a fenómenos estudiados por la Teoría de Percolación y la Teoría de Grafos Aleatorios.

Por último, muchos modelos de sistemas complejos involucran problemas de optimización. En el caso de las *redes complejas* entra en juego la Optimización Combinatoria.

1.1.3.2. La Física y los sistemas complejos

Los sistemas complejos suelen estar constituidos por un gran número de elementos que se encuentran en un estado de equilibrio dinámico –véase por ejemplo el modelo SOC–. Esta característica hace que los métodos de la Física Estadística sean adecuados para predecir el comportamiento macroscópico a partir de interacciones microscópicas que, en muchos casos, suelen modelarse como aleatorias.

Por otra parte, la concepción de los sistemas complejos como sistemas diseñados para optimizar los recursos –recuérdese el modelo HOT– aportó un enfoque *energético* en que el comportamiento del sistema se entiende como una minimización colectiva de la energía. Este enfoque energético se traduce en la búsqueda de un *hamiltoniano global* del sistema. Así, algunos trabajos analizan las interacciones a partir del modelo de Ising o el modelo de Potts de la Mecánica Estadística.

1.1.3.3. Las Ciencias de la Computación y los sistemas complejos

El rol de las Ciencias de la Computación está centrado en la simulación de modelos de sistemas complejos. Los adelantos en la computación de las últimas décadas han permitido realizar simulaciones en gran escala, y a partir de la década de los '90 se hizo posible procesar grandes cantidades de información para observar leyes de potencia –

por ejemplo, en la Web e Internet–, estudiar largas series temporales en los mercados económicos o analizar el genoma humano.

La Computación es también esencial en todos los problemas de optimización combinatoria que se presentan muy frecuentemente en los *modelos combinatorios*, a partir del desarrollo de métodos heurísticos de optimización y del estudio del problema de la complejidad computacional.

Por último, varias ramas de las Ciencias de la Computación se enfocan en el procesamiento de grandes volúmenes de datos para inferir patrones, reglas o características globales. Se trata por ejemplo de la Minería de Datos, el Reconocimiento de Patrones o la Inteligencia Artificial. Es interesante observar la cercanía de estas ramas con el enfoque sistémico de los Sistemas Complejos. La combinación de la Inteligencia Artificial con los modelos basados en agentes dio lugar a los *sistemas multi-agente*.

1.2. Modelos de sistemas complejos

Un modelo es *una representación de un sistema, que permite estudiar y explicar ciertas características del mismo*. En particular, los modelos de sistemas complejos son representaciones simplificadas que capturan sólo algunas de las propiedades del sistema. En muchos casos permiten predecir el comportamiento futuro del mismo y ciertos *patrones globales*, pero no el comportamiento de los agentes individuales [89].

Ya hemos mencionado varios ejemplos de modelos de sistemas complejos: el grafo del club de karate de Zachary, el Juego de la Vida y los *forest-fires*, entre otros. La formalización de los modelos de sistemas complejos se realiza utilizando herramientas de la Matemática.

Desde lo epistemológico, la importancia de los modelos en las ciencias comenzó a ser discutida en 1950 [137] y cuenta con una extensa bibliografía¹⁰. Nos interesa en particular la formalización realizada por R. Rosen en 1985 utilizando la Teoría de Categorías [136]. Rosen definió la relación de modelado como un proceso con 4 etapas (ver Figura 1.7). En el primer paso se *observa* el comportamiento del *sistema natural* en estudio (que se supone regido por leyes causales desconocidas). En un segundo paso se *codifica* el sistema para obtener el llamado *sistema formal*. El tercer paso es la manipulación del sistema formal a través de reglas de *inferencia*, de manera de reproducir o imitar el comportamiento causal del sistema en estudio. Como último paso se *decodifica* el sistema formal para volver al sistema en estudio. Entonces se observa por *comparación* si el sistema formal logró reproducir o no la dinámica causal en el sistema natural. En caso afirmativo, realmente estamos en presencia de un modelo del sistema, que puede ser

¹⁰Como referencia puede consultarse el exhaustivo libro de D. Bailer-Jones [11].

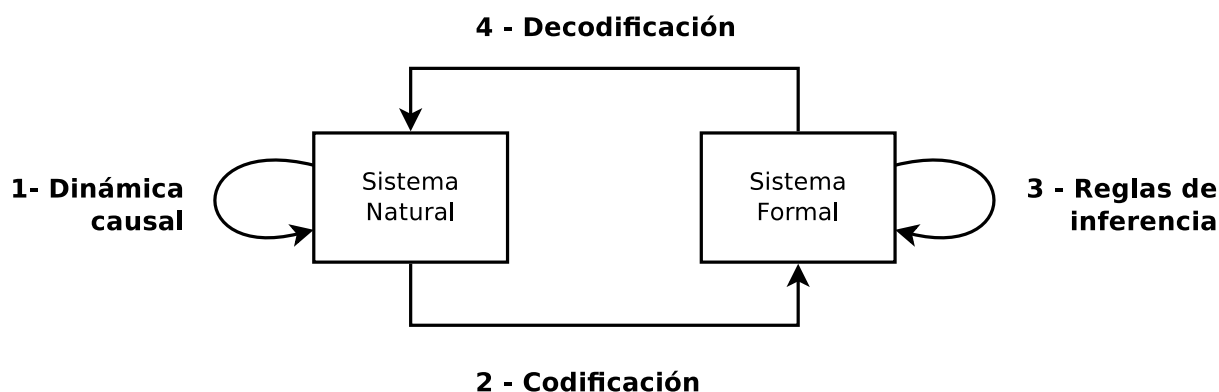


Figura 1.7: *Formalización de los modelos de sistemas complejos de R. Rosen [137]*. El primer paso es la observación del comportamiento del sistema natural en estudio. El segundo paso consiste en la codificación para obtener un sistema formal. En un tercer paso se manipula el sistema formal para definir reglas de inferencia que reproduzcan la dinámica causal del sistema en estudio. El sistema formal es un modelo cuando los pasos $2 + 3 + 4$ logran imitar el comportamiento del sistema natural ($1 = 2 + 3 + 4$).

utilizado para predecir el comportamiento futuro del mismo.

A continuación realizaremos una clasificación no exhaustiva de los modelos matemáticos utilizados en el área de Sistemas Complejos. Dejamos en claro que el tipo de modelo a utilizar depende habitualmente del problema y de las propiedades que se quiere estudiar. Un único modelo no puede capturar todos los aspectos de un sistema complejo, y suelen ser necesarios varios modelos cuando se quieren explorar diversas facetas de un mismo sistema en estudio¹¹.

Modelos en Ecuaciones Diferenciales. En muchos sistemas complejos las variables toman valores continuos, o bien la dimensión del problema permite reemplazar el dominio discreto por un dominio continuo. En estos casos, y en particular cuando se trata con *sistemas dinámicos* –en los que las variables se estudian en función del tiempo– es habitual encontrar modelos planteados en términos de ecuaciones diferenciales.

Los más clásicos de estos modelos son los modelos de evolución de población, como la *ecuación logística* de F. Verhulst (1845) y la *ecuación predador-presa* de Lotka-Volterra (1926). También destacamos los modelos de propagación de epidemias como el *modelo SIR* de Kermack-McKendrick (1927) y todas sus variantes, que han influenciado las políticas sanitarias en el siglo XX. Estos modelos también han sido utilizados desde la década del '60 para el estudio de fenómenos sociales como la propagación de rumores y la distribución de información.

Los modelos mencionados se conocen como *de campo medio* (*mean field*), porque no

¹¹Recuérdese la frase de Mikulecky del comienzo de este capítulo.

consideran la ubicación espacial de los individuos ni la forma en que ellos interactúan entre sí, sino que toman en cuenta únicamente la media estadística del conjunto de todas las interacciones. Los modelos de propagación que utilizan tasas de contagio y los modelos poblacionales con tasas de natalidad son ejemplos de modelos de campo medio. Estos modelos pueden ser tildados de demasiado simplistas o reduccionistas, pero en muchos casos son efectivos para observar un aspecto particular del sistema como, por ejemplo, la cantidad de individuos infectados esperada o la población esperada después de transcurrido cierto tiempo.

También existen modelos en ecuaciones diferenciales que consideran la dinámica en su componente espacial, como es el caso de los *modelos de difusión* y el *movimiento browniano*. La *Dinámica Molecular*, por su parte, modela el movimiento de las partículas de un sistema (átomos o moléculas, por ejemplo), a partir de las fuerzas de interacción entre las mismas y de las leyes básicas de la Mecánica.

Modelos en Ecuaciones de Recurrencia. Son la contrapartida discreta de los modelos en ecuaciones diferenciales. Es el caso del *mapa logístico* de R. May (1976) –que es el análogo discreto de la ecuación logística y tiene un comportamiento caótico– y de la *matriz de Leslie* en ecología poblacional –una ecuación matricial de recurrencia que modela la población de una especie–.

Modelos en Series Temporales. El interés por el análisis de series temporales surgió en 1900 con el análisis de L. Bachelier de los mercados económicos. Bachelier había supuesto una distribución normal e independiente de las variaciones de precios –lo que se conoce como movimiento browniano unidimensional–, pero los datos acumulados en los años siguientes mostraron un claro desvío respecto a su modelo. Mandelbrot observó sus características auto-semejantes recién hacia 1963 y conjeturó que las variaciones de precios seguirían una distribución de Lévy.

Lo cierto es que al analizar series temporales de diversas magnitudes económicas se ha observado frecuentemente un comportamiento libre de escala –manifestado como una ley de potencias en la densidad espectral de la función o ruido $1/f$ – y correlaciones de largo alcance –es decir, con una caída hiperbólica en el tiempo en lugar de exponencial–. El mismo fenómeno se ha visto replicado en los últimos años en las mediciones de tráfico de datos en redes de alta velocidad, en las cuales se suele agregar tráfico proveniente de un número muy grande de usuarios finales [100]. Esto ha acrecentado el interés por estudiar y modelizar matemáticamente estos procesos. Los modelos en series temporales más conocidos para generar correlaciones de largo alcance son el proceso FARIMA (*autoregressive fractionally integrated moving average*) [84] y el *Fractional Gaussian Noise*

(FGN). Ambos son computacionalmente costosos.

La “memoria” de largo alcance de las series temporales suele cuantificarse empleando el *exponente de Hurst*¹². Muchos trabajos vinculan a este exponente con una dimensión fractal, aunque lo cierto es que en principio se trata de dos fenómenos diferentes y no necesariamente correlacionados [79].

Modelos Basados en Agentes. Los modelos basados en agentes consideran a cada uno de los elementos del sistema complejo como un agente, y definen reglas (determinísticas y/o estocásticas) que regulan la interacción entre los agentes. El modelo evoluciona entonces a partir de estas reglas. La modelización basada en agentes es aplicable a una amplia variedad de problemas y, más que una clase de modelos, define toda una concepción desde el punto de vista epistemológico, ya que intenta aportar un enfoque holístico al poner en el centro a la interacción.

Queremos destacar que los modelos de autómatas celulares y los modelos combinatorios –que son el objeto de estudio de esta tesis– son en realidad un caso particular de modelos basados en agentes.

La Figura 1.8 ilustra los modelos basados en agentes con el comportamiento de unas termitas que se organizan de manera descentralizada para acumular madera, extraído del proyecto StarLogo¹³.

Modelos de Autómatas Celulares. Formalmente, un autómata celular está definido por una terna $(G, \mathbb{K}, f)$, en donde:

- G es un grafo cuyos vértices constituyen las células del autómata y cuyas aristas reflejan la relación de *vecindad* entre las mismas.
- $\mathbb{K}$ es un conjunto de estados.
- f es un conjunto de mapeos f_i , uno por cada vértice del grafo, que definen las reglas de transición de los estados de las células en función de su propio estado y de los estados de sus células vecinas.

Los autómatas celulares han demostrado que a partir de muy simples reglas de interacción puede emerger un comportamiento organizado. Esto lo hemos observado ya en autómatas *determinísticos* como la pila de arena¹⁴. Utilizando en cambio autómatas con

¹²H. Hurst estudió en 1965 el régimen de reservas del río Nilo de acuerdo a datos históricos y observó la presencia de correlaciones de largo alcance.

¹³<http://education.mit.edu/starlogo/>, MIT Media Laboratory.

¹⁴Ver Ejemplo 4 de la Sección 1.1

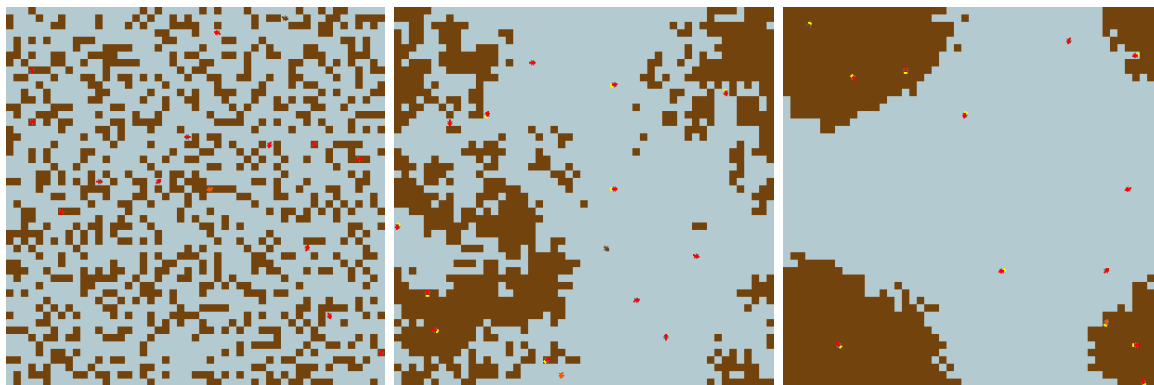


Figura 1.8: *Modelos basados en agentes*. El proyecto StarLogo, ideado por Mitchell Resnick, permite estudiar varios sistemas descentralizados desde la óptica de los modelos basados en agentes. En la imagen vemos el ejemplo de las *termitas*. Un reticulado de 50×50 dispone de astillas de madera (*en marrón*) colocadas aleatoriamente. Un conjunto de 15 termitas se mueven aleatoria e independientemente una de otra, aplicando una simple regla: Cuando encuentran una astilla la toman y continúan caminando. Al encontrar una segunda astilla se proponen encontrar un sitio libre, y en cuanto lo encuentran depositan la primera astilla que habían levantado. (*Izq*) Disposición inicial de las astillas. (*Cen*) Un tiempo después, comienzan a observarse acumulaciones de madera. (*Der*) Finalmente, las termitas logran concentrar la mayoría de las astillas en cuatro pilas.

reglas de transición *estocásticas* –como en el caso de los *forest-fires*– se pueden modelar fenómenos de percolación.

Los autómatas celulares son una forma de implementar la concepción basada en agentes, pasando de un enfoque de campo medio, a un enfoque centrado en la interacción de los agentes. El modelo SIR –que en su forma original es un modelo en ecuaciones diferenciales– tiene su implementación en autómatas celulares. También el modelo de segregación social de Schelling (1969) se implementa a través de un autómata celular.

Es habitual el uso de autómatas celulares en Economía para modelar las interacciones de diversos agentes económicos a partir de la Teoría de Juegos.

Modelos Combinatorios. Los modelos combinatorios modelan a los sistemas complejos con una red de nodos conectados que representan las interacciones entre los elementos del sistema. A esta red asociada al sistema complejo se la denomina *red compleja*. Las redes complejas son muy efectivas para modelar el transporte y flujo de información en los sistemas complejos, de allí que sean la herramienta preferida al abordar el estudio de la Web e Internet. También es muy frecuente su uso para estudiar interacciones de diversa índole entre las personas, que dan lugar a las llamadas *redes sociales*.

La investigación en el área del modelado combinatorio es tan extensa que se designa con el nombre de *Redes Complejas* (*Complex Networks* o *Network Science*) al estudio

de los sistemas complejos desde esta perspectiva.

1.2.1. Problemas inherentes al modelado de sistemas complejos

La modelización de sistemas complejos de acuerdo al método descrito en la Figura 1.7 plantea algunos problemas interesantes que mencionaremos a continuación. El primero de ellos es el concepto de *simulación de un modelo*. La evolución del sistema formal en base a las reglas de inferencia definidas (*Paso 3*) requiere de un procedimiento computacional. Es importante prestar atención a los recursos necesarios para ejecutar este procedimiento (por ejemplo, en términos de tiempo computacional o de cantidad de memoria disponible) y estudiar de qué manera estos recursos escalan con el tamaño del sistema¹⁵. Esta relación es estudiada por la Teoría de la Complejidad Computacional. Hay varios factores que inciden sobre la complejidad computacional de la simulación de un modelo:

- *La sencillez del sistema formal*. Cuanto más sencillo sea el sistema formal en términos de la cantidad de variables y complejidad de las reglas de inferencia, más fácil será su simulación. La sencillez de un modelo compite a veces con la precisión de sus resultados, y entonces se debe buscar una solución de compromiso entre ambos. Aún así, y de acuerdo al principio de parsimonia, entre dos modelos que ofrecen similares resultados se debe preferir siempre el más simple.
- *El procedimiento computacional*. Un mismo modelo puede ser ejecutado más o menos eficientemente de acuerdo al procedimiento computacional utilizado. La optimización de algoritmos y estructuras de datos es un paso importante en el desarrollo de un modelo de simulación.
- *Los criterios de aproximación*. En una gran cantidad de casos el modelo no se simula en forma exacta sino aproximada. Por ejemplo, la resolución de un sistema de ecuaciones diferenciales se realiza habitualmente por métodos aproximados, y requiere definir un nivel de discretización y un criterio de detención; la búsqueda de un máximo en un problema de optimización combinatoria también requiere de criterios de exploración (heurísticas) y de detención. Estas elecciones pueden afectar seriamente a la complejidad computacional. Nuevamente se requiere un compromiso entre la calidad de los resultados y la escalabilidad de la simulación.

¹⁵Recordemos el problema de plegamiento de proteínas en el Ejemplo 1 de la Sección 1.1: mientras que el sistema natural se estabiliza en un tiempo microscópico, la evolución del sistema formal requiere de un tiempo exponencial con la cantidad de aminoácidos.

En síntesis, un buen modelo de simulación debe ser simple, utilizar algoritmos y estructuras de datos eficientes y definir criterios de aproximación adecuados (cuando no se simula en forma exacta).

El segundo problema importante en la modelización es lo que llamamos la *evaluación del modelo*: una vez obtenidos los resultados a partir de la simulación, los mismos deben evaluarse. De acuerdo con la figura 1.7, la evaluación consiste en comparar la dinámica del sistema natural (*Paso 1*) con los resultados que predice el modelo (*Pasos 2+3+4*). Esta comparación no es trivial, ya que en general no se observará una igualdad estricta entre los resultados. Entonces, se hace necesario definir métricas para cuantificar la similaridad entre el modelo y el sistema natural. Aún más, puede ser necesario medir la similaridad entre los resultados de distintos modelos, o entre distintos criterios de aproximación de un mismo modelo. El problema de comparación y de medición es de gran importancia en Sistemas Complejos.

En nuestros aportes a lo largo de esta tesis, pondremos especial énfasis en estas dos cuestiones. En cada modelo propuesto discutiremos la problemática de su simulación y de la complejidad computacional, y estableceremos criterios para evaluar sus resultados y compararlos con lo observado en los sistemas reales.

Capítulo 2

Modelos Combinatorios de Sistemas Complejos

Los grafos son la herramienta utilizada para representar modelos combinatorios. Por ello comenzaremos el capítulo con una breve introducción a la Teoría de Grafos y una reseña de la notación matemática empleada a lo largo de este trabajo.

A continuación haremos una síntesis de algunos de los resultados teóricos y experimentales más importantes obtenidos en el área de Redes Complejas. Esto permitirá entender cómo interactúan los procesos de construcción de modelos y de observación de redes reales.

Finalmente exploraremos varios de los modelos combinatorios más utilizados para estudiar redes complejas. Algunos de ellos –como el modelo Barabási-Albert– intentan explicar el surgimiento de leyes de potencias en la Web o Internet; otros –como el modelo Watts-Strogatz– se interesan particularmente por el fenómeno de mundo pequeño (*small world*). Cada modelo aborda una o más problemáticas particulares del sistema modelado e intenta reproducirlas de la mejor manera posible. En general la propuesta de un modelo genera discusiones en la comunidad científica, a las que sigue un proceso de validación y ajuste que –como se describió en el primer capítulo– en muchos casos lo refuerza y en otros lo descarta o bien lo reemplaza por otro modelo superador. En los casos en que sea pertinente, comentaremos como se ha dado esta dinámica y la evolución histórica que ha tenido el modelo.

2.1. Introducción a los grafos de red

Los grafos de red nos permiten representar matemáticamente las interacciones entre los elementos de un sistema complejo. Cada elemento estará representado por un *vértice* del grafo, mientras que la interacción entre un par de elementos estará representada por

una *arista*. Un grafo puede entonces ser visualizado como un conjunto de puntos unidos por segmentos, como ilustra la Figura 2.1.

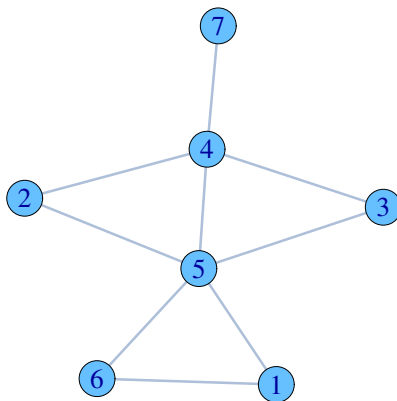


Figura 2.1: *Representación de un grafo*. Representación visual de un grafo G con 7 vértices y 9 aristas.

Las variaciones sobre este esquema general son numerosas: es muy habitual utilizar *grafos dirigidos*, en donde la arista es un par ordenado. También es común asociar valores numéricos a los vértices o aristas, obteniendo lo que se llama un *grafo pesado*. Por último, si bien poco común, es posible que las interacciones involucren a más de dos elementos, o a una cantidad variable de ellos, en cuyo caso la representación se denominará *hipergrafo*.

El bagaje de herramientas que ofrece la Teoría de Grafos es muy amplio. Sugerimos como bibliografía los libros de West [157] y Bollobás [26]. La notación que introducimos a continuación está basada en el libro de West.

2.1.1. Notación y representación de grafos de red

Un grafo G es una terna determinada por los siguientes tres elementos:

- Un conjunto de vértices, $V(G)$.
- Un conjunto de aristas, $E(G)$.
- Una relación que asocia a cada arista con un par de vértices a los que se llama sus *extremos*.

Orden y tamaño del grafo. La *cantidad de vértices y aristas de un grafo G* la llamaremos respectivamente $n(G) = |V(G)|$ (*orden del grafo*) y $e(G) = |E(G)|$ (*tamaño del grafo*)¹.

¹Dado un conjunto A , la notación $|A|$ indicará el cardinal del conjunto.

Tipos de grafos. Un grafo es *simple* cuando no posee bucles (aristas cuyos extremos coinciden en el mismo vértice) ni aristas repetidas. Cuando posee aristas repetidas se lo denomina *multigrafo*.

Por otra parte, cuando las aristas son *pares ordenados* de vértices, se lo denomina *grafo dirigido* o *digrafo*. De lo contrario, el grafo es *no dirigido*.

Cuando los vértices y/o aristas del grafos poseen un valor numérico asociado —denominado *peso*— se dice que el grafo es *pesado*. De lo contrario, el grafo es *no pesado*.

En esta sección se considerarán únicamente grafos simples no pesados, ya sean no dirigidos o dirigidos. A lo largo de todo el trabajo se tomará la misma consideración, salvo casos en que se aclare explícitamente.

Relación de adyacencia. En grafos no dirigidos, si los extremos de una arista e son u y v diremos que $e = uv = vu$. Diremos que dos vértices u y v son *adyacentes* o *vecinos* cuando $uv \in E(G)$. A la relación de adyacencia la notaremos $u \leftrightarrow v$. Si $u \leftrightarrow v$, también diremos que $u \rightarrow v$ y que $v \rightarrow u$.

En cambio, en grafos dirigidos, la arista es un par ordenado y lo notaremos $e = (u, v)$. Diremos que $u \rightarrow v$, que u es la cabeza del vértice e y que v es su cola.

En ambos casos, si $u \rightarrow v$ diremos que v es vecino de u , que u precede a v o que v sucede a u , diremos también que la arista correspondiente va de u a v , que sale de u y que es incidente en v .

Matriz de adyacencia. Habitualmente enumeraremos los vértices de un grafo de manera consecutiva, como $v_1, v_2, \dots, v_{n(G)}$. A partir de esta enumeración, un grafo G es representado de manera unívoca por su *matriz de adyacencia* $A(G)$, una matriz de tamaño $n(G) \times n(G)$ definida como:

$$A(G) = (a_{ij}) = (\mathbf{1}\{v_i \rightarrow v_j\}) \ .$$

Para el ejemplo de la Figura 2.1 la matriz de adyacencia es

$$A(G) = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \ .$$

La matriz de adyacencia suele ser una matriz rala. Para grafos no dirigidos es una matriz simétrica, porque $(v_i \rightarrow v_j) \Rightarrow (v_j \rightarrow v_i)$. En grafos dirigidos, en cambio, es en general no simétrica.

Grados y vecindades en grafos no dirigidos. El *grado de un vértice*, $d(v)$, se define como la cantidad de aristas incidentes en él. Es decir:

$$d(v) = |\{e \in E : e \text{ es incidente en } v\}| .$$

El grado también puede calcularse a partir de la matriz de adyacencia como

$$d(v_k) = \sum_{i \neq k} a_{ik} .$$

Los grafos no dirigidos verifican la *fórmula de la suma de grados*:

$$\sum_{v \in V(G)} d(v) = 2e(G) .$$

La *vecindad* de un vértice v , $\mathcal{N}(v)$, es el conjunto de vecinos de v :

$$\mathcal{N}(v) = \{u : v \rightarrow u\} .$$

En grafos *simples* el cardinal de la vecindad $\mathcal{N}(v)$ coincide con el grado del vértice v .

Grados en grafos dirigidos. En grafos dirigidos se define el grado interno de un vértice v , $d^-(v)$ como la cantidad de aristas para las cuales v es su cabeza, y el grado externo, $d^+(v)$ como la cantidad de aristas para las cuales v es su cola.

$$d^-(v) = |\{e = (x, y) \in E : x = v\}| \quad d^+(v) = |\{e = (x, y) \in E : y = v\}| .$$

Los grafos dirigidos verifican la *fórmula de la suma de grados para grafos dirigidos*:

$$\sum_{v \in V(G)} d^-(v) = \sum_{v \in V(G)} d^+(v) = e(G) .$$

Caminos y distancias. En grafos no dirigidos dos aristas son *adyacentes* entre sí cuando comparten un vértice. En grafos dirigidos una arista e_1 es adyacente a una arista e_2 cuando la cola de e_1 coincide con la cabeza de e_2 .

Un *camino* entre dos vértices u, v es una secuencia de aristas $(e_1, e_2, \dots, e_n)$ tal que cada arista es adyacente a la siguiente en la secuencia, e_1 sale de u y e_n es incidente en

u y v se denominan *extremos del camino*. La *longitud* de un camino es la cantidad de aristas que lo constituyen. Para todo vértice u existe un camino de u a sí mismo de longitud cero, es decir, que no contiene aristas.

Un *camino* es un *ciclo* cuando su longitud es no nula y sus dos extremos coinciden en el mismo vértice.

Dos vértices u, v están *conectados* cuando existe un camino entre ellos.

Dos caminos son *arista-disjuntos* cuando no comparten aristas.

Dos caminos son *vértice-disjuntos* cuando no comparten vértices a excepción de sus extremos.

La máxima cantidad de caminos vértice-disjuntos de a pares entre u y v la denominamos $\lambda(u, v)$.

La máxima cantidad de caminos arista-disjuntos de a pares entre u y v la denominamos $\lambda'(u, v)$.

Propiedad: Todo conjunto de caminos entre u y v vértice-disjuntos de a pares es también un conjunto de caminos arista-disjuntos de a pares. Luego, $\lambda'(u, v) \geq \lambda(u, v)$.

La *distancia* entre dos vértices u y v que están conectados es la menor de las longitudes de los caminos entre ellos, y la representamos $d(u, v)$. Todo camino entre u, v que realiza esta distancia es un *camino mínimo entre u, v* . Cuando dos vértices u, v no están conectados se define $d(u, v) = \infty$.

Propiedad: La matriz de adyacencia es útil en el cálculo de la distancia entre vértices. Dos vértices v_i y v_j distintos están a distancia d si y sólo si para todo entero positivo $k < d : [A(G)^k]_{ij} = 0$ y $[A(G)^d]_{ij} \neq 0$. El elemento $[A(G)^l]_{ij}$ indica la cantidad de caminos distintos de longitud l entre v_i y v_j .

A través de una búsqueda en anchura (BFS, por *breadth first search*) se puede hallar un camino mínimo entre dos vértices u, v en un tiempo de $O(e(G))^2$.

Subgrafos. Un grafo H es un subgrafo de G si y sólo si $V(H) \subset V(G)$ y $E(H) \subset E(G)$ y las aristas en $E(H)$ tienen la misma asignación de extremos en H que en G . Cuando $V(H) = V(G)$, H se puede obtener por eliminación sucesiva de las aristas en $M = E(G) \setminus E(H)$. En este caso diremos que $H = G - M$.

El *subgrafo de G inducido por el conjunto de vértices T* , con $T \subset V(G)$, es aquel grafo que se obtiene por eliminación de los vértices en $\bar{T} = V(G) \setminus T$ y de todas las aristas incidentes en algún vértice en $\bar{T}$. Lo denotamos $G[T]$ o $G - \bar{T}$.

²Para grafos pesados en general –con pesos no negativos en las aristas– el *algoritmo de Dijkstra* encuentra un camino mínimo en $O(e(G) + n(G) \log n(G))$

Componentes conexas. En grafos no dirigidos, la relación “estar conectado” entre vértices es una relación de equivalencia. Esto permite definir clases de equivalencia $C_1, C_2, \dots, C_{c(G)}$ que constituyen una partición del conjunto de vértices $V(G)$. Los subgrafos $G[C_i]$ inducidos por esta relación de equivalencia se denominan *componentes conexas de G* . Como no existen aristas entre vértices que pertenezcan a distintas clases de equivalencia, la unión de las componentes conexas de G es igual a todo el grafo G . La cantidad de componentes conexas de G la denotamos $c(G)$.

Decimos que un grafo es *conexo* cuando tiene una única componente conexa, es decir, cuando para todo par de vértices $u, v \in V(G)$, u y v están conectados. En caso contrario, el grafo es *no conexo*.

Las clases de equivalencia presentan respecto a la conexión una propiedad de maximalidad: todo subgrafo conexo de G está incluido en alguna componente conexa de G . Las componentes conexas de G son *subgrafos conexos maximales* con respecto a esta propiedad.

En grafos dirigidos, al hablar de conectividad estaremos aludiendo a la *conectividad fuerte*: diremos que dos vértices u y v de un grafo dirigido están fuertemente conectados cuando existe un camino de u a v y también un camino de v a u . Cuando hagamos referencia a las componentes conexas de un grafo dirigido, se asumirá implícitamente que aludimos a las componentes fuertemente conexas.

Cortes. Dados $S, T \subset V(G)$, denotamos $[S, T]$ al conjunto de aristas que salen de un vértice en S y son incidentes en un vértice en T ³:

$$[S, T] = \{e : e \text{ sale de } x \text{ y es incidente en } y, x \in S \wedge y \in T\} .$$

Un *corte por aristas* (*edge-cut*) es un conjunto de aristas de la forma $[S, \bar{S}]$, con $S \neq \emptyset$ y $\bar{S} \neq \emptyset$.

La *capacidad de un corte por aristas* es la cantidad de aristas que posee, y la denotamos $|[S, \bar{S}]|$.

Dado un grafo conexo G , todo corte por aristas es un *conjunto separador de G* , en el sentido que $G - [S, \bar{S}]$ es no conexo.

Un *corte- (u, v) por aristas* (*(u, v) -edge-cut*) es un corte por aristas que deja a u y v en dos componentes conexas distintas de $G - [S, \bar{S}]$.

Un *corte- (u, v) por vértices* o simplemente *corte- (u, v)* (*(u, v) -cut*) S es un conjunto de vértices $S \subset V(G) - \{u, v\}$ tal que $G - S$ tiene a u y v en componentes conexas

³En particular, si S y T tienen intersección no nula y los extremos de una arista pertenecen ambos a la intersección, entonces la arista se contará dos veces en el corte.

distintas.

El *tamaño de un corte* S es la cantidad de vértices del conjunto S .

El mínimo de los tamaños de los cortes- (u, v) se denomina $\kappa(u, v)$, y puede ser calculado empleando el algoritmo de Ford-Fulkerson [69].

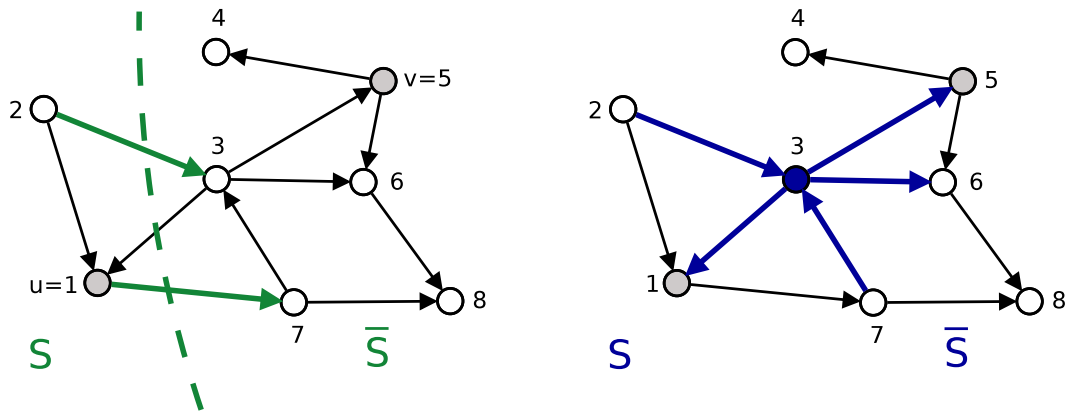


Figura 2.2: *Cortes y cortes por aristas en grafos.* (Izq.) Un corte- $(1, 5)$ por aristas en un grafo dirigido, en donde $S = \{1, 2\}$. Es un corte- $(1, 5)$ por aristas porque $1 \in S$ y $5 \in \bar{S}$. La capacidad de este corte por aristas es 2. No es un corte- $(1, 5)$ por aristas *mínimo*, ya que existen cortes- $(1, 5)$ por aristas de capacidad 1. (Der.) Un corte- $(1, 5)$ en el mismo grafo. Aquí $S = \{3\}$, y el tamaño de este corte es 1. Es un corte- $(1, 5)$ porque la eliminación del vértice 3 deja a 1 y 5 en componentes distintas.

Arista-conectividad y conectividad entre vértices. La mínima cantidad de aristas que deben ser eliminadas para dejar a u y v en componentes conexas distintas la denominamos *arista-conectividad entre u y v* , y la denotamos $\kappa'(u, v)$.

Teorema de Menger (aristas): ([157], pág. 168) La mínima cantidad de aristas que deben ser eliminadas para dejar a u y v en componentes conexas distintas es equivalente a la cantidad de caminos arista-disjuntos de u a v :

$$\kappa'(u, v) = \lambda'(u, v) .$$

La mínima cantidad de vértices que deben ser eliminados para dejar a u y v en componentes distintas la denominamos *conectividad entre u y v* , y la denotamos $\kappa(u, v)$. Es equivalente al mínimo de los tamaños de los cortes- (u, v) :

$$\kappa(u, v) = \min\{|S|, S \text{ es un corte-}(u, v)\} .$$

Teorema de Menger (vértices): ([157], pág. 167) La máxima cantidad de caminos vértice-disjuntos de u a v es igual al mínimo de los tamaños de los cortes-

(u, v) :

$$\lambda(u, v) = \min\{|S|, S \text{ es un corte-}(u, v)\} .$$

De la definición de $\kappa(u, v)$ y del Teorema de Menger se desprende que la conectividad entre u y v equivale a la máxima cantidad de caminos disjuntos de pares entre u y v :

$$\kappa(u, v) = \lambda(u, v) .$$

Al trabajar con varios grafos en simultáneo, indicaremos eventualmente en cada parámetro el nombre del grafo como subíndice, de manera de saber en qué grafo se aplica. Así, al escribir $d_G(v)$ estaremos diciendo “el grado de v en el grafo G ”. En los casos en que consideramos que no existe ambigüedad optaremos por omitir la referencia al grafo.

2.1.2. Invariantes de los grafos de red

Un *invariante* de un grafo es una función del grafo que sólo depende de la estructura abstracta del mismo, es decir que no varía con las enumeraciones de sus vértices (isomorfismos) o con la forma en que el grafo se representa. Algunas invariantes de los grafos son: el orden, el tamaño, la conectividad, la arista-conectividad, el diámetro, la cromaticidad, la arboricidad, el polinomio característico, la concordancia (*assortativity*) y el coeficiente de agrupamiento (*clustering coefficient*) global. A continuación desarrollaremos algunos de ellos. En la sección siguiente, “Medidas de centralidad de los vértices y aristas”, veremos que algunas de dichas medidas también dan lugar a invariantes globales.

2.1.2.1. Conectividad

La *conectividad* de un grafo es el mínimo cardinal de un conjunto de vertices $S \subset V$ tal que $G - S$ es desconexo o tiene sólo un vértice. En otras palabras, es la mínima cantidad de vértices a ser removidos para obtener un grafo no conexo o bien con un solo vértice⁴. La conectividad de un grafo G se denota $\kappa(G)$. Equivalentemente:

$$\kappa(G) = \min_{u, v \in V(G)} \kappa(u, v) = \min_{u, v \in V(G)} \lambda(u, v) = \min\{|S|, S \text{ es un corte}\} .$$

Un grafo G es *k-conexo* cuando su conectividad es al menos k .

⁴La posibilidad de llegar a un grafo con un solo vértice se agrega a los efectos de que la conectividad se relacione con la mínima cantidad de caminos vértice-disjuntos entre todo par de vértices.

2.1.2.2. Arista-conectividad

La *arista-conectividad* de un grafo G es el mínimo cardinal de un conjunto de aristas $F \subset E(G)$ tal que $G - F$ es desconexo. La arista-conectividad de un grafo G se denota $\kappa'(G)$. Equivalentemente:

$$\kappa'(G) = \min_{u,v \in V(G)} \kappa'(u, v) .$$

En vistas del Teorema de Menger para aristas, se sigue que:

$$\kappa'(G) = \min_{u,v \in V(G)} \lambda'(u, v) .$$

Como consecuencia del *teorema de flujo máximo y corte mínimo de Ford-Fulkerson* ([157], pág. 180), el mínimo de las capacidades de los cortes por aristas de u, v es igual a la máxima cantidad de caminos arista-disjuntos de u a v :

$$\min\{|[S, \bar{S}]|, [S, \bar{S}] \text{ es un corte por aristas de } u, v\} = \lambda'(u, v) .$$

A partir de los dos últimos resultados, se deduce que la arista-conectividad de un grafo G equivale al mínimo de las capacidades de todos sus cortes por aristas:

$$\kappa'(G) = \min_{S \subset V(G), S \neq \emptyset} \{|[S, \bar{S}]|\} .$$

Un grafo G es *k-arista-conexo* cuando su arista-conectividad es al menos k .

2.1.2.3. Diámetro

El *diámetro* de un grafo G es el máximo de las distancias entre todos sus pares de vértices:

$$\text{diam}(G) = \max_{u,v \in V(G)} d(u, v) .$$

El diámetro de un grafo es infinito si y sólo si el grafo es no conexo.

2.1.2.4. Coeficiente de agrupamiento

En grafos no dirigidos⁵ el *coeficiente de agrupamiento* (*clustering coefficient*) asociado a un vértice es una medida de la densidad de aristas entre sus vecinos [154]. Dado un vértice u de grado $d(u) \geq 2$, la máxima cantidad de aristas entre sus vecinos es $\frac{1}{2}d(u)(d(u) - 1)$. El coeficiente de agrupamiento se define entonces –para vértices de grado mayor a 1– como la proporción de aristas que unen a dos vecinos del vértice, con

⁵Existen extensiones del coeficiente de agrupamiento para grafos pesados [16].

respecto a esa cantidad máxima:

$$cc(u) = \frac{2 \sum_{\{v,w\} \subset \mathcal{N}(u)} \mathbf{1}\{vw \in E(G)\}}{d(u)(d(u) - 1)} .$$

Es frecuente también analizar la distribución del coeficiente de agrupamiento de los vértices, en función del grado.

El *coeficiente de agrupamiento global de un grafo* es un invariante, y se calcula como la cantidad de triángulos ordenados del grafo en relación con la cantidad de tripletas⁶. Un triángulo ordenado es una terna ordenada (u, v, w) tal que $u \rightarrow v, v \rightarrow w, w \rightarrow u$, mientras que una tripleta (*triplet*) es una terna ordenada (u, v, w) tal que $u \rightarrow v, v \rightarrow w$:

$$cc(G) = \frac{\sum \triangle}{\sum \triangleleft} = \frac{\sum_{u,v,w} \mathbf{1}\{u \rightarrow v, v \rightarrow w, w \rightarrow u\}}{\sum_{u,v,w} \mathbf{1}\{u \rightarrow v, v \rightarrow w\}} .$$

El coeficiente de agrupamiento global así definido es equivalente al denominado índice de transitividad (*transitivity ratio*), que cuantifica la transitividad de las adyacencias. Se encuentra entre 0 y 1.

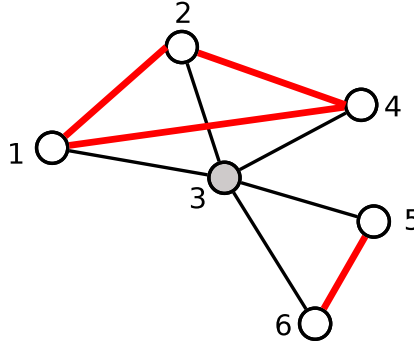


Figura 2.3: *Coeficiente de agrupamiento*. Observamos la vecindad del vértice 3 de un grafo, formada por $\mathcal{N}(3) = \{1, 2, 4, 5, 6\}$. Las aristas entre vecinos de 3 se muestran en rojo. El coeficiente de agrupamiento de 3 es entonces $cc(3) = \frac{2 \cdot 4}{d(3) \cdot (d(3) - 1)} = 0,4$.

⁶Algunos autores, en cambio, denominan coeficiente de agrupamiento de un grafo al promedio de los coeficientes de agrupamiento de los vértices:

$$\frac{1}{n(G) - |\{u \in V(G), d(u) = 1\}|} \sum_{u \in V(G), d(u) > 1} cc(u) .$$

Nosotros sin embargo optaremos por la definición que hemos dado anteriormente, y a este último lo llamaremos simplemente *coeficiente de agrupamiento promedio*, $\overline{cc}(G)$. Nuestra definición coincide de todas formas con un promedio *ponderado* de los coeficientes de agrupamiento de los vértices, en el que cada vértice u se pondera con un peso $\frac{d(u)(d(u)-1)}{2}$.

2.1.2.5. Distribución de grados y grado medio

La *secuencia de grados* de los vértices, en orden creciente o decreciente, es también un invariante del grafo. A partir de los grados $d(v)$ de los vértices, se puede definir la función de probabilidad del grado de los vértices, $p_v(k)$, como⁷:

$$p_v(k) = \mathbb{P}_v[d(v) = k] = \frac{\sum_{v \in V(G)} \mathbf{1}\{d(v) = k\}}{n(G)}, k \in \mathbb{Z}^+.$$

El valor de $p_v(k)$ representa la probabilidad de que al tomar un vértice al azar del grafo –con distribución uniforme sobre el conjunto $V(G)$ –, el mismo tenga grado k .

El valor medio de esta distribución, $\mathbb{E}_v[d(v)]$, se denomina *grado medio del grafo*. Notaremos a la varianza de la distribución de grados como $\sigma_v^2(d(v))$. Para simplificar la notación también utilizaremos $\bar{d} = \mathbb{E}_v[d(v)]$ y $\sigma^2(d) = \sigma_v^2(d(v))$.

Al máximo (mínimo) de entre los grados de todos los vértices lo denominamos *grado máximo (mínimo)*, $d_{\max}(G)$ ($d_{\min}(G)$). Tener una distribución de grados $p_v(k)$, un grado medio $\bar{d}$, varianza $\sigma^2(d)$ o un cierto grado máximo (mínimo) son invariantes del grafo.

2.1.2.6. Distribución de grados de los vecinos

También nos interesará la *distribución de grados de los vecinos de vértices de grado k* , que se define a partir del subconjunto de vértices de grado k de un grafo de la siguiente forma⁸:

$$p_{uv}(k'|k) = \mathbb{P}_{uv}[d(v) = k' | d(u) = k] = \frac{1}{p_v(k)n(G)} \sum_{u \in V(G), d(u)=k} \frac{\sum_{uv \in E(G)} \mathbf{1}\{d(v) = k'\}}{k}.$$

Es equivalente a la distribución de probabilidad que surge de tomar un vértice al azar de entre los vértices de grado k , y luego elegir uno de sus k vecinos al azar (es decir, con distribución uniforme) y observar su grado.

El *grado medio de los vecinos de los vértices de grado k* lo denominamos $k_{nn}(k)$ y se puede calcular de la siguiente manera [126]:

$$k_{nn}(k) = \sum_{k' \in \mathbb{Z}^+} k' \cdot p_{uv}(k'|k).$$

⁷El subíndice v hace referencia a los elementos del espacio muestral, que en este caso es el conjunto de los vértices, $V(G)$.

⁸Las aristas de $E(G)$ se toman con distribución uniforme. Si el grafo es no dirigido, al extraer una arista uv del conjunto $E(G)$ se la debe ordenar como (u, v) o (v, u) con distribución uniforme. La probabilidad conjunta $p_{uv}(k, k')$ representa entonces la probabilidad de que al tomar una arista (u, v) , los grados de sus extremos sean k y k' respectivamente. En este sentido, $p_{uv}(k'|k)$ se puede interpretar como la probabilidad condicional de $d(v)$ dado $d(u)$.

2.1.2.7. Concordancia por grados de los vértices

La concordancia por grados (*degree assortativity*) es la medida de correlación entre los grados de vértices adyacentes [112]. En grafos no dirigidos se define a partir de valores esperados y desvíos que toman como espacio muestral al conjunto de las aristas del grafo⁹:

$$a(G) = \frac{\mathbb{E}_{uv}[d(u)d(v)] - \mathbb{E}_{uv}[d(u)] \cdot \mathbb{E}_{uv}[d(v)]}{\sigma_{uv}[d(u)] \cdot \sigma_{uv}[d(v)]}.$$

En términos del k_{nn} , la concordancia se puede expresar también como [35]:

$$a(G) = \frac{\bar{d} \sum_{k \in \mathbb{Z}^+} [k^2 p(k) k_{nn}(k)] - \bar{d}^2}{\bar{d} \bar{d}^3 - \bar{d}^2}.$$

Al ser una correlación, la concordancia por grados tiene la siguiente propiedad: si se considera a los grados de los extremos de una arista uv tomada al azar, $d(u)$ y $d(v)$, como variables aleatorias, la concordancia por grados puede interpretarse como la pendiente de la recta de regresión entre estas dos variables¹⁰.

- Una concordancia por grados positiva implica una correlación alta entre los grados de vértices adyacentes: vértices de grado elevado se conectan típicamente a otros vértices de grado elevado, mientras que vértices de grado pequeño son adyacentes a vértices de grado pequeño.
- Una concordancia por grados negativa es muestra también de una alta correlación, aunque en este caso en sentido contrario: vértices de grado pequeño tienen preferencia por vértices de grado elevado, y viceversa.
- En cambio, una concordancia por grados cercana a 0 es signo de una pobre correlación entre los grados de vértices vecinos.

El concepto de concordancia (*assortativity*) no se restringe a los grados sino que puede aplicarse también para comparar atributos categóricos de vértices adyacentes en el grafo¹¹. En esta variante, la concordancia permite estudiar los denominados *patrones de mezclado* (*mixing patterns*), de gran relevancia en las redes sociales. Dado un conjunto de categorías $\mathcal{K} = (K_1, K_2, \dots, K_{|\mathcal{K}|})$ y una función $f_{\mathcal{K}} : V(G) \rightarrow \mathcal{K}$ que asigna categorías

⁹Existen extensiones de la concordancia por grados para grafos dirigidos y grafos pesados [16].

¹⁰En términos generales la correlación entre dos variables aleatorias X e Y se corresponde con la pendiente de la recta de regresión entre las variables normalizadas $X' = \frac{X - \mu_X}{\sigma(X)}$ e $Y' = \frac{Y - \mu_Y}{\sigma(Y)}$. En este caso particular las variables están idénticamente distribuidas y la normalización no es necesaria.

¹¹Sin embargo, las dos medidas de concordancia no son equivalentes matemáticamente. En el caso de valores escalares como los grados, empleamos como medida la correlación de Pearson. Para atributos categóricos, en cambio, utilizamos aquí la medida de acuerdo de Cohen.

a los vértices, se define la *concordancia por $\mathcal{K}$* como: [114]¹²

$$a(G) = \frac{\text{Tr}(\mathbf{e}) - \|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|},$$

en donde $\mathbf{e}$ es una matriz cuyas componentes e_{ij} representan la probabilidad de que en una arista (u, v) tomada al azar sean $f(u) = K_i$ y $f(v) = K_j$.

En términos generales hablaremos de un *comportamiento concordante (assortative)* cuando para determinado atributo de los vértices la concordancia es positiva, y de *comportamiento discordante (disassortative)* cuando la concordancia es negativa.

2.1.3. Medidas de centralidad de los vértices y aristas

Las medidas de centralidad intentan cuantificar la relevancia de los vértices o aristas de un grafo. En general esta relevancia se relaciona con la cercanía del vértice o arista a los demás vértices o aristas, o con la utilización del mismo para establecer caminos para llegar a otros vértices o aristas, o bien con los efectos que tendría su posible eliminación en el grafo. En particular, el grado de un vértice es una medida de centralidad, y posiblemente una de las más simples: puede pensarse que un vértice al que se conectan muchos otros vértices es un vértice importante en un grafo, sin embargo esto no siempre es cierto.

Las medidas de centralidad existentes son numerosas. A continuación desarrollaremos sólo aquellas que nos serán de utilidad: la intermediación (*betweenness*), la cercanía (*closeness*), la centralidad de vector propio (*eigenvector centrality*), el índice de capa (*shell index o coreness*) y el índice de denso (*dense index*). Algunas medidas presentan variaciones y diversas normalizaciones posibles. Aquí daremos para cada una la definición que a nuestro entender es la más simple y adecuada para el objetivo de nuestro trabajo. En las 3 primeras se requerirá que el grafo sea conexo.

2.1.3.1. Intermediación

La *intermediación (betweenness)* fue propuesta por L. Freeman en 1977 [72] y es una de las medidas de centralidad más clásicas. Se obtiene a partir de la cantidad de caminos mínimos en el grafo que pasan por un vértice determinado, y su definición para un grafo conexo es:

$$c_B(v_i) = \sum_{\{v_j, v_k\} \subset V(G), jk \neq i} \frac{L(v_j, v_k | v_i)}{L(v_j, v_k)},$$

¹²Esta definición de concordancia (*assortativity*) de Newman [114] coincide con la medida de acuerdo (*agreement*) propuesta por Cohen [47, 23].

en donde $L(v_j, v_k|v_i)$ es la cantidad de caminos mínimos entre v_j y v_k y pasan por v_i , y $L(v_j, v_k)$ es la cantidad de caminos mínimos entre v_j y v_k . La intermediación cuantifica la utilización de un vértice en los caminos mínimos que conectan a los demás vértices.

En 2002 Girvan *et al.* propusieron un concepto análogo de intermediación para aristas (*arista-intermediación* o *edge-betweenness*) [76].

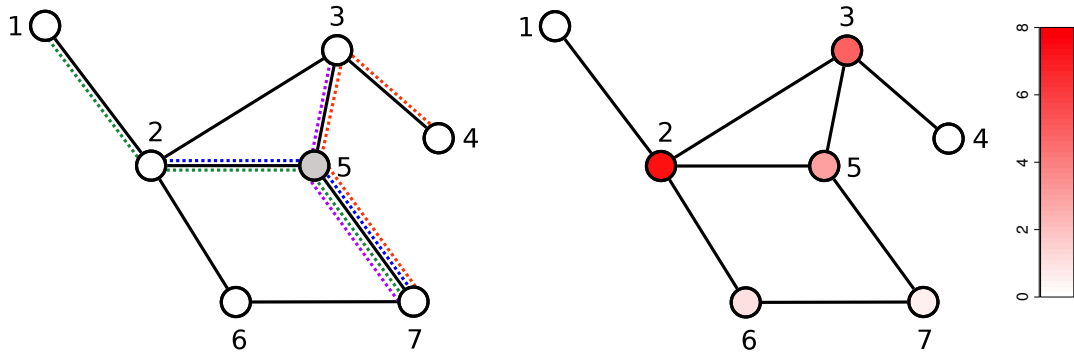


Figura 2.4: *Intermediación (betweenness)*. (Izq.) Cuatro caminos mínimos que atraviesan el vértice 5. Dos de ellos –los caminos de 1 a 7 y de 2 a 7– tienen un camino mínimo alternativo, y por lo tanto su peso en el cálculo de la intermediación es de $1/2$. La intermediación del vértice 5 es entonces $c_B(5) = 3$. (Der.) Los vértices del mismo grafo, coloreados según su intermediación.

2.1.3.2. Cercanía

Dado un grafo conexo, la *cercanía (closeness)* de un vértice se define como la inversa de su distancia promedio a los demás vértices del grafo [73]:

$$c_C(v_i) = \frac{n(G) - 1}{\sum_{v_j \in V(G), j \neq i} d(v_i, v_j)} .$$

Una desventaja de la cercanía respecto a la intermediación es que tiende a concentrarse en un rango relativamente pequeño de valores al aplicarlo a todos los vértices [119].

2.1.3.3. Centralidad de vector propio

Esta medida de centralidad se basa en la descomposición espectral de la matriz de adyacencia de un grafo conexo. Como todos los coeficientes de la matriz de adyacencia $A(G)$ son no negativos y la matriz es irreducible, de acuerdo al teorema de Perron-Frobenius el radio espectral de $A(G)$ es un valor propio con un único vector propio asociado cuyas componentes son todas positivas [144]; a este vector propio lo notaremos $v^1(G)$. Definiremos entonces a la *centralidad de vector propio (eigenvector centrality)* del

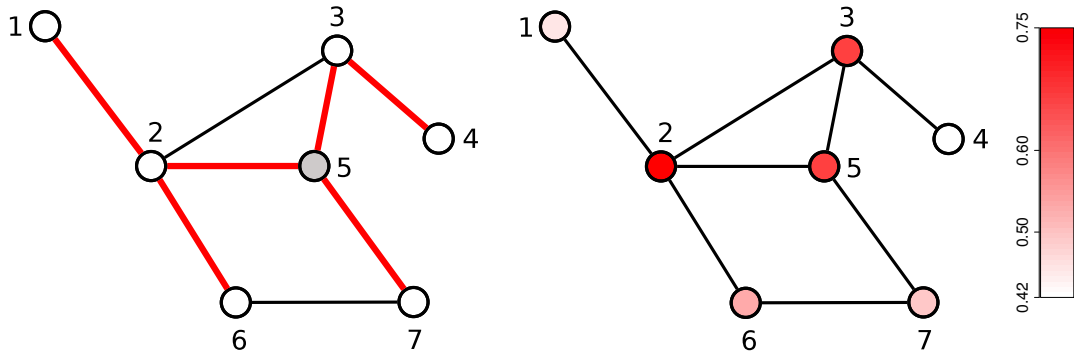


Figura 2.5: *Cercanía (closeness)*. (Izq.) En rojo se muestran las aristas que forman parte de caminos mínimos desde el vértice 5 hacia otros vértices. La distancia promedio de 5 a los demás vértices del grafo es $3/2$, y por lo tanto su cercanía es $c_C(5) = 2/3$. (Der.) Los vértices del mismo grafo, coloreados según su cercanía.

vértice v_i como la componente i -ésima del vector $v^1(G)$ dividida por la norma infinito del vector:

$$c_E(v_i) = \frac{v_i^1(G)}{\max_j \{v_j^1(G)\}} \quad .$$

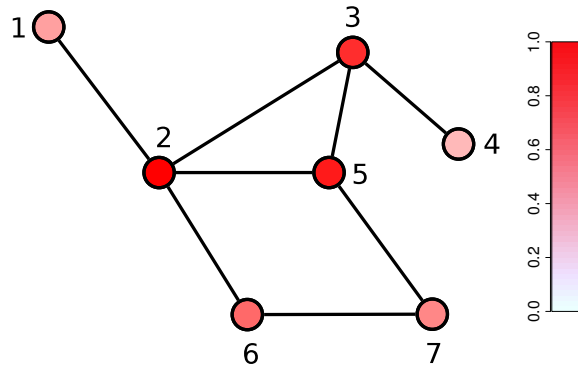


Figura 2.6: *Centralidad de vector propio (eigenvector centrality)*. Los vértices de un grafo, coloreados según su centralidad de vector propio.

Los distintos vectores propios de la matriz de adyacencias están vinculados con la dinámica de caminos aleatorios (*random walks*) y procesos de difusión sobre el grafo [144]. En particular, aquellos vectores propios asociados a los valores propios de mayor valor son los que más incidencia tienen en esta dinámica. De aquí que logren capturar en cierta forma la relevancia de los vértices en el grafo.

2.1.3.4. Índice de capa

La medida de centralidad que denominamos índice de capa (*shell index o coreness*) es aquella basada en la descomposición en k -núcleos (*k-cores*) de un grafo, que presen-

taremos a continuación.

La descomposición en k -núcleos fue introducida por Seidman en 1983 [142] y organiza a los vértices de un grafo en una estructura de capas anidadas, llamadas *núcleos* (*cores*), en que las capas más céntricas –de mayor k – contienen vértices con mayor cantidad de conexiones entre ellos respecto a las capas más periféricas. En efecto, un k -núcleo se define como un subgrafo inducido maximal tal que cada uno de sus vértices tiene al menos k vecinos en el subgrafo inducido. Es decir:

$$\mathcal{C}_k(G) = G[S] \Leftrightarrow \{\forall v \in V(G[S]) : d_{G[S]}(v) \geq k\} \wedge S \text{ es maximal con esta propiedad} ,$$

en donde recordamos que el grado de v se mide en el subgrafo de G inducido por S .

Diremos que un vértice v tiene *índice de capa* $c_K(v) = k$ cuando pertenece al k -núcleo pero no pertenece al $(k + 1)$ -núcleo.

El máximo valor de k en un grafo G para el cual el k -núcleo de G no es vacío es un invariante, y se denomina *núcleo-profundidad* (*core number*). Lo notaremos $k_{\max}(G)$.

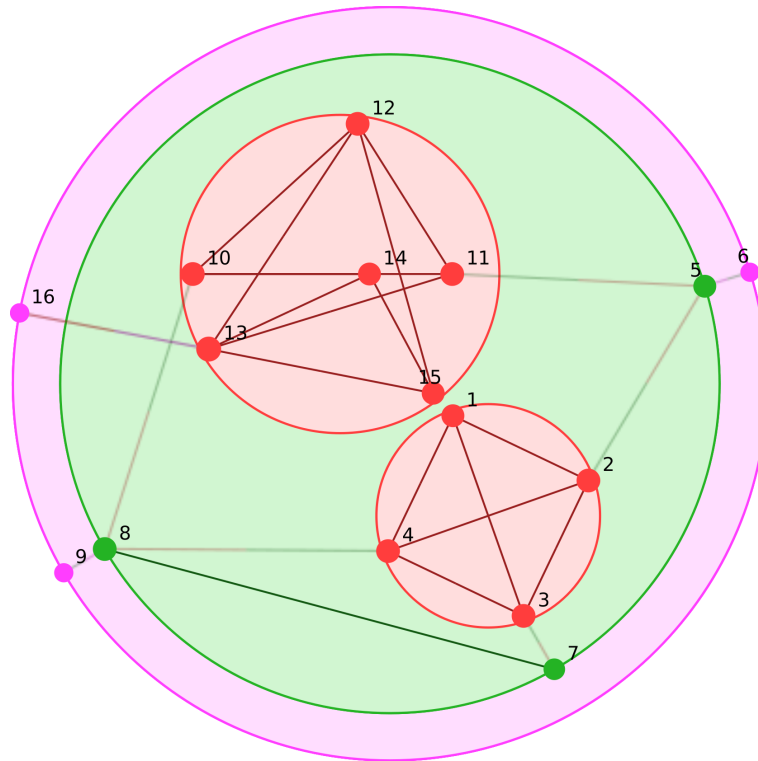


Figura 2.7: *Descomposición en k -núcleos* (*k-cores*). Descomposición en k -núcleos de un grafo con 16 vértices. Los vértices en rojo tienen índice de capa 3, es decir que tienen al menos 3 conexiones entre ellos. Los vértices 5, 7 y 8 (en verde) tienen índice de capa 2 –observar que si bien el vértice 8 tiene 3 conexiones, al eliminar al vértice 7, le quedarán sólo dos, y por eso no puede acceder al 3-núcleo–. Los vértices en rosa tienen índice de capa 1. Obsérvese que el 3-núcleo es no conexo, mientras que el 2-núcleo y el 1-núcleo tienen una única componente conexa.

Los diversos k -núcleos de un grafo pueden obtenerse eliminando recursivamente del mismo los vértices de grado menor a k . Basado en este procedimiento, el algoritmo presentado por Batagelj y Zaversnik [18] encuentra la descomposición en k -núcleos de un grafo conexo en un tiempo de $O(e(G))$.

2.1.3.5. Índice de denso

La descomposición en k -densos (k -densos) de un grafo es el análogo de la descomposición en k -núcleos, pero centrándose en las aristas en lugar de en los vértices. Si la descomposición en k -núcleos observaba el grado del vértice en el grafo inducido, aquí observaremos la *multiplicidad* de la arista, $m(e)$, que se define como la cantidad de vértices que son vecinos simultáneos de ambos extremos de la arista en el subgrafo. Por último, como el k -denso se obtiene a partir de un conjunto de aristas, es un subgrafo del grafo original, pero no necesariamente es un subgrafo inducido. Definiremos entonces al k -denso de un grafo G , $\mathcal{D}_k(G)$, para $k \geq 2$ de la siguiente manera [141]:

$$\begin{aligned} E(\mathcal{D}_k(G)) = S &\Leftrightarrow \{ \forall e \in S : m_{G-\bar{S}}(e) \geq k-2 \} \wedge S \text{ es maximal con esta propiedad} \\ V(\mathcal{D}_k(G)) &= \{ u \in V(G) / \exists v \in V(G) : uv \in E(\mathcal{D}_k(G)) \} . \end{aligned}$$

Es decir, primero construimos el subconjunto maximal de aristas de multiplicidad al menos $k-2$ entre ellos, $E(\mathcal{D}_k(G))$, y luego definimos al conjunto de vértices como formado por aquellos vértices en los que incide alguna arista de $E(\mathcal{D}_k(G))$.

La descomposición en k -densos de un grafo puede obtenerse eliminando recursivamente del mismo las aristas de multiplicidad menor a $k-2$, para valores de k crecientes a partir de $k=2$.

Si una arista e pertenece a cierto k -denso pero no pertenece al $(k+1)$ -denso diremos que tiene *índice de denso* (*dense-index*) k , o que $c_D(e) = k$.

El máximo de los *índices de denso* de los vértices de un grafo es un invariante al que denominamos *denso-profundidad* (*dense number*), $k_{\max}^{\text{dense}}(G)$.

2.1.4. Resumen de notación

$n(G)$	orden de G
$e(G)$	tamaño de G
$V(G)$	conjunto de vértices de G
$E(G)$	conjunto de aristas de G
$A(G)$	matriz de adyacencia de G
a_{ij}	elemento i, j de la matriz de adyacencia

$d(v)$	grado de v
$\mathcal{N}(v)$	vecindad de v
$d^-(v)$	grado interno del vértice v (grafos dirigidos)
$d^+(v)$	grado externo del vértice v (grafos dirigidos)
$\lambda(u, v)$	máxima cantidad de caminos vértice-disjuntos de u a v
$\lambda'(u, v)$	máxima cantidad de caminos arista-disjuntos de u a v
$d(u, v)$	distancia entre u y v
$G[T]$	subgrafo de G inducido por $T \subset V(G)$
$c(G)$	cantidad de componentes de G
$[S, S]$	corte por aristas
$ [S, S] $	capacidad de un corte por aristas
$\kappa(u, v)$	corte mínimo entre u y v
$\kappa'(u, v)$	arista-conectividad entre u y v
$\kappa(G)$	conectividad de G
$\kappa'(G)$	arista-conectividad de G
$\text{diam}(G)$	diámetro de G
$cc(v)$	coeficiente de agrupamiento (<i>clustering</i>) de v
$cc(G)$	coeficiente de agrupamiento (<i>clustering</i>) global de G
$\overline{cc}(G)$	coeficiente de agrupamiento (<i>clustering</i>) promedio de G
$p_v(k)$	distribución de grados
$\overline{d}, \overline{d^k}$	grado medio, k -ésimo momento de la distribución de grados
$\sigma^2(d)$	varianza de la distribución de grados
$d_{\text{máx}}$	grado máximo
$p_{uv}(k' k)$	distribución de grados de los vecinos de vértices de grado k
$k_{nn}(k)$	grado medio de los vecinos de vértices de grado k
$a(G)$	concordancia por grados (<i>degree assortativity</i>) de G
$c_B(v)$	intermediación (<i>betweenness</i>) de v
$c_C(v)$	cercanía (<i>closeness</i>) de v
$c_E(v)$	centralidad de vector propio (<i>eigenvector centrality</i>) de v
$c_K(v)$	índice de capa (<i>shell index</i>) de v
$\mathcal{C}_k(G)$	k -núcleo (<i>k-core</i>) de G
$k_{\text{máx}}(G)$	núcleo-profundidad (<i>core number</i>) de G
$c_D(e)$	índice de denso (<i>dense index</i>) de e
$\mathcal{D}_k(G)$	k -denso (<i>k-dense</i>) de G
$k_{\text{máx}}^{\text{dense}}(G)$	denso-profundidad (<i>dense number</i>) de G

Cuadro 2.1: Resumen de la notación de Teoría de Grafos empleada en este trabajo. Tomamos como referencia el libro de West [157].

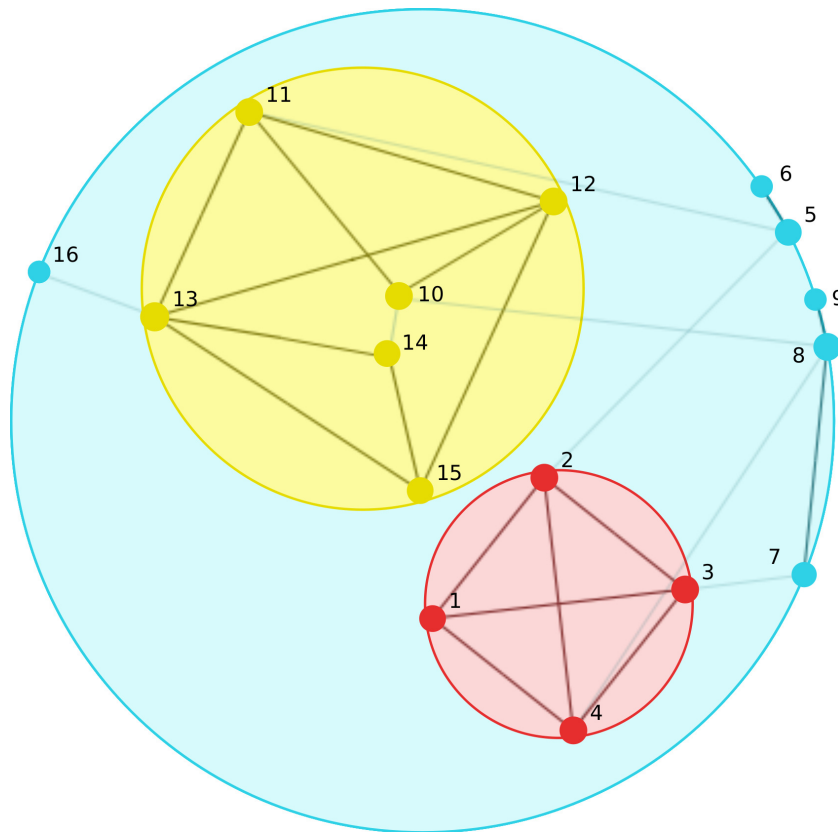


Figura 2.8: *Descomposición en k -densos (k -densos)*. Descomposición en k -densos del mismo grafo que el de la figura anterior. Las aristas en rojo tienen índice de denso 4, es decir que cada una de ellas participa de al menos dos triángulos dentro del 4-denso. Las aristas en amarillo pertenecen al 3-denso, y todas participan de al menos un triángulo. Las aristas en celeste tienen índice de denso 2. Los vértices toman el color de aquella de sus aristas incidentes con mayor índice de denso, de acuerdo con la definición. Obsérvese que la arista $10 \leftrightarrow 14$ tiene índice de denso 2 porque, si bien conecta a vértices del 3-denso, no participa de ningún triángulo.

2.2. Resultados teóricos y experimentales en redes complejas

En esta sección realizaremos una síntesis de algunos de los resultados teóricos y experimentales más importantes obtenidos en el área de Redes Complejas, dejando para la próxima sección lo relativo al desarrollo de modelos. Ejemplificaremos la discusión con algunas *redes sociales*, *tecnológicas* y *biológicas* –en particular las redes metabólicas y de interacciones entre proteínas–. Dejamos al margen otras redes que no desarrollaremos en el presente trabajo pero sobre las cuales se ha avanzado mucho también en el área, como

es el caso de las redes semánticas y de tipos particulares de redes biológicas: las redes ecológicas y las redes neuronales. Para ampliar los resultados sugerimos consultar [115, 35, 58].

Comenzamos nuestra revisión en el año 1999, en que se descubrió que los grafos de diversas redes complejas se ajustan a *leyes de potencias* (*power-laws*)^{13,14}, es decir que varios de sus atributos siguen leyes de la forma $f(x) \propto x^{-\alpha}$. Se destacan en particular:

- El trabajo de los hermanos Faloutsos [66], que observaron una ley de potencias en la distribución de grados de Internet. A partir de diversas *exploraciones* que recolectaron información sobre cerca de 4000 *routers* de Internet y sus conexiones en ese entonces, mostraron que la cantidad de conexiones que tiene un *router* tomado al azar se ajusta muy bien a una ley de potencias, con exponente α entre 2,0 y 2,5 dependiendo de la exploración. Mostraron también que la ley de potencias en la distribución de grados origina por sí misma leyes de potencias en la distribución de las distancias entre pares de *routers* de la red, y en la distribución de las distancias desde un *router* determinado hacia el resto de la red.
- Los trabajos de Barabási y Albert [3, 14] que encontraron leyes de potencias en:
 - Una porción del grafo de la Web, con 325729 vértices representando *documentos web*, conectados entre sí por *hipervínculos* (*links*) que permiten saltar de un documento a otro. Como los hipervínculos son direccionales, la Web se modela más convenientemente como un grafo dirigido. Así lo hicieron Albert y Barabási, y mostraron que el grado interno d^- y el grado externo d^+ siguen sendas leyes de potencias con exponentes 2,1 y 2,45 respectivamente.
 - Una red de 212250 actores, en la que las aristas entre actores representan su trabajo conjunto en algún *film*. Aquí encontraron una ley de potencias con exponente 2,3 en la distribución de la cantidad de actores que coparticipó en algún *film* con un actor dado.
 - La red de distribución eléctrica de los Estados Unidos, integrada por 4941 estaciones y subestaciones, conectadas por líneas de alta tensión. La cantidad de líneas conectadas a un determinado nodo de la red se ajusta a una ley de potencias con exponente 4.

¹³Si bien la discusión sobre distribuciones libres de escala comenzó a tomar forma en esta época, destacamos el trabajo pionero de Price, que ya en 1976 había observado una ley de potencias en una red de colaboraciones científicas [129].

¹⁴Para una formalización sobre las leyes de potencias puede consultarse el Apéndice A del presente trabajo.

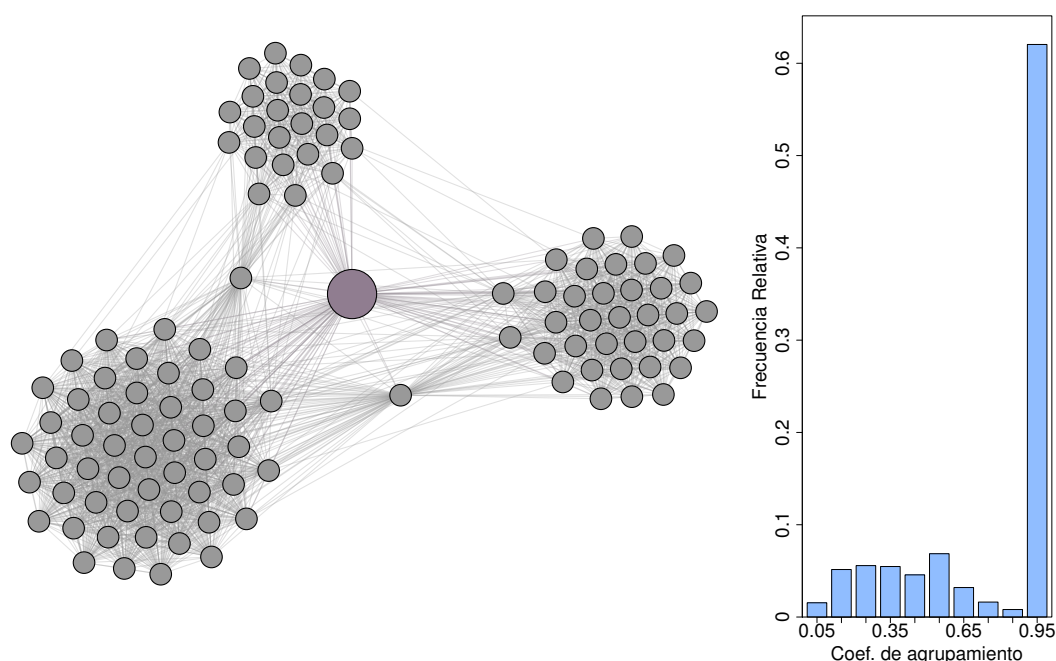


Figura 2.9: *Red de actores*. (Izq.) Visualización de la vecindad de un actor en la red de actores, generada con el software Gephi. El actor al que hacemos referencia se indica en violeta en el centro de la red. (Der.) Frecuencias relativas de los coeficientes de agrupamiento de los vértices, agrupados con un *binning* lineal. El coeficiente de agrupamiento global de la red es de 0,78.

Por otra parte, en [3] Albert y Barabási también mostraron que la distancia promedio entre documentos en la Web (es decir, la cantidad media de clics necesarios para llegar de un documento a otro) en 1999 era de tan sólo 18,59, ajustándose linealmente al logaritmo de la cantidad de documentos. Esto renovó el interés por las redes *de mundo pequeño* (*small-world*) que había estudiado Milgram en los '60 en su conocido experimento. En el mismo año Watts y Strogatz también observaron la propiedad de mundo pequeño en la red de actores y en una red de interacciones entre proteínas¹⁵.

Estos dos fenómenos –el comportamiento *libre de escala* de la distribución de grados y la propiedad de mundo pequeño– se han encontrado en numerosísimas redes complejas, y tienen importantes consecuencias en su dinámica:

- En el 2000 Jeons *et al.* [90] analizaron la estructura de redes de interacciones entre proteínas y, además de encontrar leyes de potencias, observaron una estructura formada por *hubs*, es decir, vértices de elevado grado, que conectan a los vértices

¹⁵En los organismos vivos tienen lugar procesos bioquímicos que cumplen determinada función o responden a cierta necesidad del organismo. Cada uno de estos procesos se rige por la presencia de ciertas proteínas. De esta forma, decimos que dos proteínas interactúan cuando participan de un mismo proceso bioquímico.

de grado pequeño. Llegaron a la conclusión de que estas redes tienen una respuesta robusta frente a la eliminación aleatoria de algunos de sus nodos –manifestándose, por ejemplo, en la estabilidad del diámetro, la distancia media, o la conectividad– mientras que lo que podría llamarse un “ataque planeado” o intencional a uno o varios hubs podría poner en serio peligro la conectividad de la red, o aumentar considerablemente el diámetro. Este tipo de comportamiento de las redes libres de escala, al que Doyle *et al.* denominaron *robust-yet-fragile*¹⁶ [61], también fue encontrado en la Web e Internet [4, 48].

- En 2001 Pastor-Satorras y Vespignani estudiaron la difusión de información y la propagación de epidemias¹⁷ en redes libres de escala, y observaron que dichos fenómenos se ven beneficiados por un diseño que tiende a optimizar el flujo de información [127]. Con un análisis termodinámico mostraron que la propagación de infecciones no tiene un punto crítico, lo que permite por ejemplo que un virus prolifere, independientemente de que su tasa de contagio sea baja. Estos resultados también se aplican a la propagación de rumores y de información en redes sociales.

Las redes libres de escala han sido objeto de importantes estudios teóricos, y con frecuencia se aplicó el potencial de la Mecánica Estadística para estudiar sus propiedades en el límite termodinámico [2, 59]. Estos resultados sirvieron como realimentación para el ajuste de los modelos que se desarrollaban a la par.

Ha despertado gran interés el estudio de la topología de Internet. La evolución permanente de la red mundial y varias cuestiones técnicas y de seguridad dificultan la obtención de una instantánea completa, por eso se desarrollaron varios proyectos para obtener exploraciones fieles de Internet, como CAIDA [34], DIMES [56] y RouteViews [151]. Mencionamos los siguientes resultados:

- Alvarez-Hamelin *et al.* estudiaron la descomposición en k -núcleos del grafo de Internet tanto a nivel de *routers* como a nivel de sistemas autónomos y observaron una ley de potencias en la distribución de tamaños de los k -núcleos [7]. También mostraron que el grado de los vértices y el índice de capa se hallan correlacionados: los *routers* más centrales de la red desde el punto de vista de los k -núcleos suelen tener grado elevado [8].
- Pastor-Satorras *et al.* encontraron un comportamiento discordante por grados de los vértices [126], ajustando el $k_{nn}(k)$ a una ley de potencias con $\alpha \approx 0,5$. En otras palabras, esto muestra que los nodos centrales de la red tienen preferencia

¹⁶Literalmente, *robustas pero frágiles*.

¹⁷Estos fenómenos son análogos a los procesos de difusión en sistemas físicos.

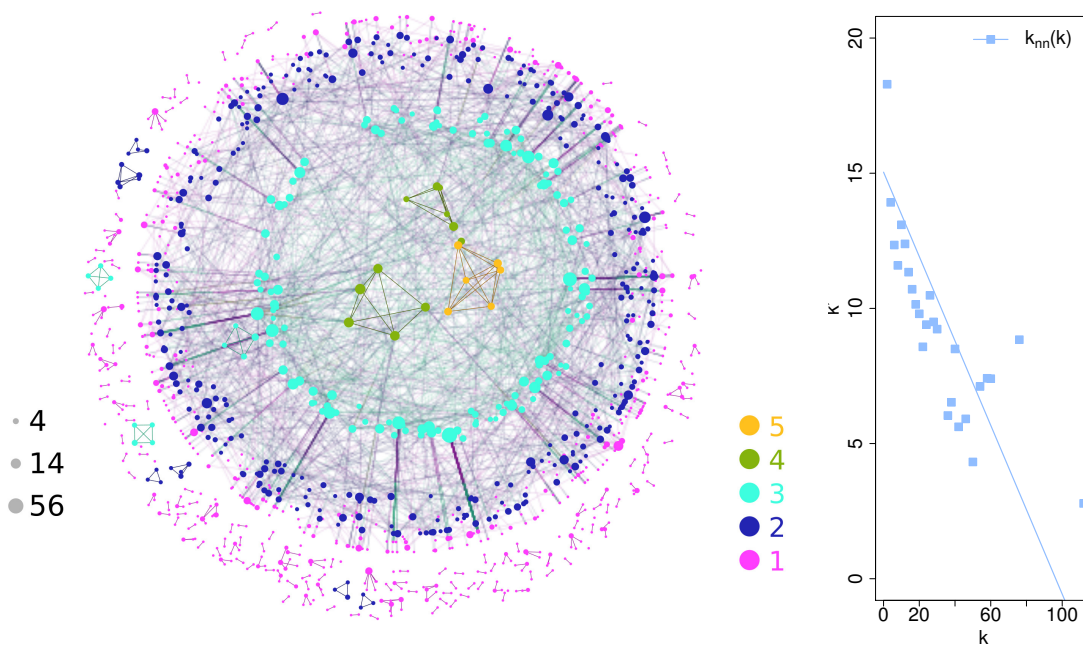


Figura 2.10: *Red de interacciones entre proteínas de S. Cerevisiae*. Visualización de la red de interacciones entre proteínas de la bacteria *S. Cerevisiae* generada con el software LaNet-vi. Las distintas capas de la visualización corresponden a los k -núcleos del grafo. La escala de la izquierda representa el grado y la de la derecha identifica el índice de capa. Se observa que el índice de capa está correlacionado con el grado de los vértices en la red. A la derecha, el $k_{nn}(k)$ en función de k muestra un comportamiento discordante, característico de estas redes en que algunas proteínas actúan como *hubs*, interactuando con muchas otras que tienen pocas interacciones [90]. La concordancia por grados de la red es -0.156 [114].

—en término medio— por establecer conexiones con nodos periféricos, y viceversa: los nodos periféricos prefieren conectarse a nodos centrales —lo que va en acuerdo con la hipótesis de *acoplamiento preferencial* de Barabási—.

- Los k -núcleos se han vinculado también con la conectividad. Luczak había probado en 1991 que en los grafos Erdős-Rényi los k -núcleos son k -conexos con alta probabilidad [107]. Estudios efectuados sobre Internet también han mostrado que los k -núcleos del grafo de Internet son k -conexos [37, 7]. En el Capítulo 4 de este trabajo estudiaremos la k -arista-conectividad de los k -núcleos de los grafos de Internet a nivel de Sistemas Autónomos.

En el área de redes sociales se han investigado los *patrones de mezclado*, es decir, las correlaciones entre determinados atributos de los miembros (edad, sexo, profesión, grado en el grafo de la red, etc.) y sus conexiones. Se evidencia frecuentemente un comportamiento concordante por grados: las personas muy populares —con muchas conexiones—

en determinada red tienden a conectarse con otras personas también muy populares. El fenómeno se ha estudiado en las redes de colaboraciones científicas, de actores y de intercambio de correo electrónico [114], entre otras.

También se ha vuelto de gran relevancia el estudio y descubrimiento de *estructura comunitaria* en redes sociales. Con este término se designa a la organización de los nodos de la red en grupos de afinidad. Dentro de estos grupos los nodos se conectan con mayor densidad que hacia afuera de ellos. El descubrimiento de estructura comunitaria permite capturar la formación de grupos de amistad, de afinidad política o ideológica o de trabajo conjunto y extraer información valiosa de la red. Discutiremos este tema en el Capítulo 3 de nuestro trabajo.

El abordaje de la Web como una red social de intercambio de información permitió aplicar herramientas de redes complejas a la búsqueda de documentos. El poderoso motor del buscador Google, denominado *PageRank*, utiliza una variante de la centralidad de vector propio para clasificar a los documentos web de acuerdo a las conexiones que establecen con otros documentos [122]. *PageRank* calcula periódicamente el vector propio asociado al mayor valor propio de la matriz de adyacencias de toda la Web: una matriz rala con millones de filas y columnas.

La vinculación entre la presencia de distribuciones libres de escala y los procesos auto-semejantes o fractales es bastante controversial. Song *et al.* desarrollaron un marco para analizar la estructura de las redes complejas en busca de auto-semejanza, que verificaron en varias redes [148]. Desde este enfoque, las distribuciones libres de escala serían sólo una manifestación de la naturaleza auto-semejante de los sistemas. Otros trabajos han vinculado la auto-semejanza con la concordancia por grados, sosteniendo que las redes fractales son discordantes, mientras que las no fractales tendrían un comportamiento concordante [160]. Johnson *et al.* [91] por su parte mostraron que la discordancia por grados es el comportamiento esperado en los sistemas que evolucionan guiados por la maximización de la entropía. El comportamiento concordante se limitaría a aquellos sistemas con un fuerte componente humano en las interacciones, como es el caso de las redes sociales. También respecto al uso de la correlación como medida de la concordancia por grados, un reciente trabajo de Hofstad mostraría que la misma no es una medida adecuada en redes con distribuciones de grados libres de escala y de gran tamaño [105].

Por último, Ravasz y Barábasi entre otros han estudiado la *estructura jerárquica* de las redes complejas y sostienen que la misma permite explicar la convivencia, dentro de una misma red, de coeficientes de agrupamiento elevados y leyes de potencias [132]. La organización jerárquica también ha sido muy discutida en el contexto del descubrimiento de comunidades.

2.3. Modelos de redes complejas

Los modelos de redes complejas pretenden reproducir algunos de los patrones que se observan en estas redes, con el fin de predecir su comportamiento o evolución futura. En general son modelos probabilísticos –no determinísticos– que se formalizan desde la Teoría de Grafos Aleatorios. Por este motivo creemos conveniente comenzar esta sección introduciendo el concepto de *grafo aleatorio*. Seguido a esto haremos un recorrido histórico por los modelos de redes complejas más conocidos, y en las subsecciones siguientes desarrollaremos algunos de ellos.

Consideraremos una familia de *grafos aleatorios* parametrizada con n , G_n ¹⁸, donde para cada n , G_n es un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ en donde Ω es un conjunto de grafos de n vértices, cada uno de los cuales tiene una probabilidad determinada de ser extraído. Una instancia del grafo aleatorio es entonces una muestra tomada de este espacio de probabilidad, mientras que los diversos invariantes del grafo aleatorio pueden pensarse como variables aleatorias en el mismo espacio. En este marco, los resultados de la Teoría de Grafos Aleatorios se expresan usualmente en términos de:

1. *Distribuciones de probabilidades de los invariantes.* Puede calcularse la distribución de probabilidades de invariantes como el diámetro, el grado de los vértices, o el coeficiente de agrupamiento, de un grafo aleatorio G_n . Diremos que un invariante $f(G_n)$ del grafo *converge asintóticamente* a cierto $h(n)$ si:

$$\lim_{n \rightarrow \infty} P[(1 - \epsilon)h(n) < f(G_n) < (1 + \epsilon)h(n)] = 1, \quad \forall \epsilon > 0 .$$

2. *Propiedades esperadas con alta probabilidad.* Diremos que G_n tiene una propiedad $\mathcal{P}$ con alta probabilidad (*with high probability*) cuando la probabilidad de que G_n tenga dicha propiedad tiende a 1 cuando $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} P[G_n \in \mathcal{P}] = 1 .$$

Cuando mencionemos que determinado modelo de grafo aleatorio tiene una propiedad $\mathcal{P}$ siempre deberá interpretarse que tiene dicha propiedad con alta probabilidad.

Para ampliar el estudio de grafos aleatorios recomendamos consultar [27, 28].

El concepto de grafo aleatorio fue introducido por P. Erdős y A. Rényi en 1959 en el

¹⁸En términos más generales un grafo aleatorio $G_{p_1, p_2, \dots, p_s}$ puede tener varios parámetros $p_1, p_2, \dots, p_s$, de los cuales uno de ellos suele ser el tamaño, $n(G)$. En nuestra definición mencionamos sólo este parámetro porque nos resultará esencial para introducir la noción de *alta probabilidad*.

modelo que lleva su nombre [64]¹⁹ y que genera grafos con distribuciones de grados poissonianas y con coeficiente de agrupamiento nulo –recuérdese que hablamos en términos de alta probabilidad–.

Durante la década del '70 este modelo inicial fue expandido en busca de obtener grafos aleatorios con otras distribuciones de grados. Surgieron entonces los grafos aleatorios con grados esperados específicos (*random graphs with given expected degrees*) [43], y el modelo de configuración (*configuration model*) o de grafos aleatorios con distribuciones de grados específicas (*random graphs with specified degree distributions*) [21]. A partir de estos modelos se hizo factible obtener grafos con distribuciones libres de escala; sin embargo, ninguno de ellos pretendía ni lograba explicar cómo surgían estas distribuciones a partir de reglas más simples.

En la década del '80, con el interés de modelar la topología de Internet, surgieron diversos *generadores* de topología, como el modelo de Waxman (1988) [155] que introdujo una variable geográfica, y los modelos jerárquicos de Doar (1996) [57] y Zegura (1997) [162]. Hacia fines de los '90, frente a los resultados de las exploraciones de Internet y la Web que mostraban un comportamiento libre de escala, Barabási y Albert propusieron un modelo basado en el *acoplamiento preferencial* que logró reproducir una ley de potencias en la distribución de grados [14]. Fabrikant *et al.* (2002) [65] también obtuvieron grafos con distribuciones libres de escala, pero a partir de un proceso de optimización con limitación de recursos.

En el ámbito de las redes sociales hubo particular interés en reproducir el fenómeno de mundo pequeño (*small-world*). El modelo más destacado es el propuesto por Watts y Strogatz en 1998 [154] que, partiendo de una topología en anillo, construye una red de mundo pequeño con bajas distancias promedio y coeficientes de agrupamiento elevados, aunque con distribuciones de grados aún poissonianas. También Kleinberg (2000) [92] reprodujo el fenómeno de mundo pequeño con un modelo que parte de un retículo (*lattice*) y agrega algunas conexiones de largo alcance.

La concordancia por grados (*degree assortativity*) parece ser una propiedad difícil de modelar; la mayoría de los modelos generan redes con concordancia por grados nula. Algunas excepciones son el modelo de Bianconi y Barabási, que genera redes con comportamiento concordante [22] y se ha utilizado para modelar la Web, y el modelo de Catanzaro *et al.* [39] que es capaz de producir redes con comportamiento discordante.

Por último mencionaremos algunos modelos vinculados con la organización jerárqui-

¹⁹Téngase en cuenta que para muchos autores la noción de *grafo aleatorio* era sinónimo de grafo Erdős-Rényi, en particular algunas décadas atrás. De allí el uso de la nomenclatura *grafo aleatorio generalizado* o *grafo aleatorio con distribuciones de grados específicas*, ya que estos modelos se entendían como una extensión del modelo de grafo aleatorio original. Hoy en día el concepto de grafo aleatorio es mucho más rico, tal como muestra la definición que hemos dado.

ca y la estructura comunitaria. El modelo Community Guided Attachment (CGA) de Leskovec *et al.* (2005) [102] estudia la emergencia de leyes de potencias en el contexto de una estructura jerárquica.

Los modelos que generan una estructura comunitaria, en cambio, no explican el surgimiento de dicha estructura sino que más bien la reproducen, y se emplean habitualmente como bancos de prueba (*benchmarks*) de los distintos algoritmos de descubrimiento de comunidades. Entre ellos destacamos el modelo *relaxed caveman* [153], el *planted l -partition* [51], el modelo jerárquico de Clauset-Moore-Newman (CMN) [44] y el de Lancichinetti-Fortunato-Radicchi (LFR) [97]. Todos ellos son variantes de los grafos aleatorios generalizados y del modelo de configuración que incorporan información sobre la estructura jerárquica y/o comunitaria.

2.3.1. El modelo Erdős-Rényi

El más sencillo de los modelos de grafos aleatorios fue propuesto por Erdős y Rényi hacia 1960 [64]. Este modelo considera un grafo con n vértices, en el que la probabilidad de que dos vértices distintos tomados al azar estén conectados es fija e igual a p .

Los grafos aleatorios Erdős-Rényi (ER) G_{np} presentan las siguientes propiedades:

- El tamaño del grafo sigue una distribución binomial:

$$\mathbb{P}[e(G_{np}) = M] = \binom{N}{M} p^M (1-p)^{N-M}, \quad 0 \leq M \leq N$$

en donde $N = \binom{n}{2}$

- La esperanza del tamaño del grafo es $\mathbb{E}[e(G_{np})] = Np$.
- El grado de los vértices sigue una distribución binomial:

$$\mathbb{P}[d_{G_{np}}(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

- La esperanza del grado de un vértice es $\mathbb{E}[d_{G_{np}}(v)] = (n-1)p$.
- El coeficiente de agrupamiento esperado de un vértice es $\mathbb{E}[cc_{G_{np}}(v)] = p$.
- La concordancia por grados es asintótica a 0 para $n \rightarrow \infty$.
- El diámetro es asintótico a $\ln n / \ln(pn)$ para $n \rightarrow \infty$ [42].
- El grafo es conexo con alta probabilidad.

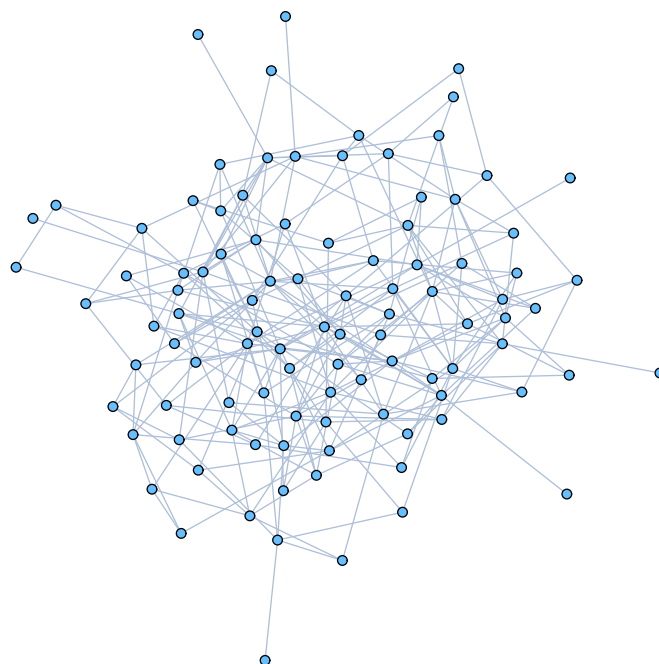


Figura 2.11: *Modelo Erdős-Rényi. Visualización.* Instancia de grafo generado con un modelo Erdős-Rényi con 100 vértices y grado medio esperado 5.

-
- La arista-conectividad es asintótica a $(n - 1)p$ para $n \rightarrow \infty$.

Es habitual estudiar el comportamiento de los grafos Erdős-Rényi cuando $n \rightarrow \infty$ manteniendo np constante, de manera de conservar el grado medio de los vértices. Bajo esta restricción, con $n \rightarrow \infty$ se cumple que:

- La distribución de grados de los vértices converge a una Poisson con media np .
- El coeficiente de agrupamiento de los vértices y global es asintótico a 0.
- El grafo es no conexo (el diámetro es infinito).

Los grafos Erdős-Rényi no son adecuados para modelar redes complejas por tener una distribución de grados con caída exponencial –en lugar de una cola larga– y un coeficiente de agrupamiento pequeño, y por la ausencia de correlaciones que se manifiesta, por ejemplo, en una concordancia por grados cercana a cero.

2.3.2. Modelos de Internet

Describiremos a continuación 3 modelos que se han utilizado para estudiar la topología de Internet: el modelo de Waxman, el modelo Barabási-Albert y el modelo FKP.

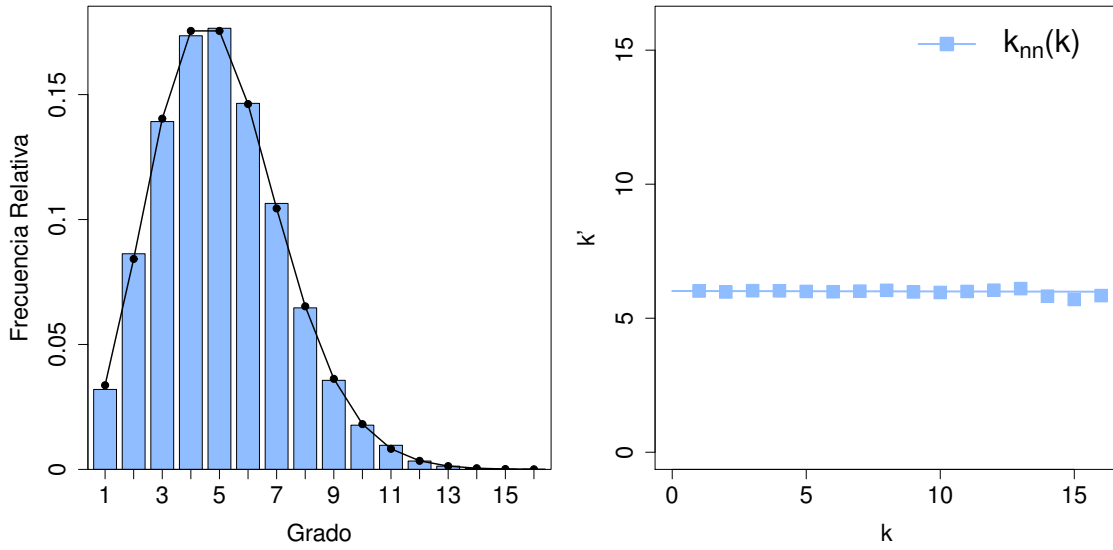


Figura 2.12: *Modelo Erdős-Rényi*. Instancia de grafo generado con un modelo Erdős-Rényi con $p = 0,00025$ y $n = 20000$. El grado medio es $\bar{d} = 5,00$ y el grado máximo es $d_{\text{máx}} = 16$. (Izq.) Frecuencia relativa de los grados de los vértices del grafo, comparada con una distribución binomial de igual media. (Der). Correlación entre los grados de vértices adyacentes. Los puntos representan el valor medio del grado de los vecinos, k_{nn} , en función del grado. La pendiente de la recta de regresión –la concordancia por grados del grafo– es nula. El coeficiente de agrupamiento global de la red también es cero.

2.3.2.1. El modelo de Waxman

A partir de observaciones efectuadas sobre Internet, Waxman formuló dos hipótesis respecto a cómo estarían conectados los *routers* en la red. En su trabajo de 1988 Waxman sostuvo que [155]:

1. Los *routers* de Internet se encuentran distribuidos geográficamente, y que esta distribución incide en la forma en que están conectados.
2. Como resultado de un proceso de optimización de recursos, la existencia de conexiones es más probable entre *routers* cercanos que entre *routers* distantes.

A partir de estas hipótesis, y tomando como base el modelo Erdős-Rényi, Waxman realizó una modificación para que la probabilidad de conexión de los nodos fuera dependiente de la distancia entre los mismos. En el modelo de Waxman se distribuyen n nodos en forma aleatoria sobre un cuadrado de lado L . Luego, cada par de vértices (v_i, v_j) se conecta con una probabilidad p_{ij} exponencial en la distancia euclídea entre los vértices, que representamos como $d(v_i, v_j)$:

$$p_{ij} = \beta e^{\frac{-d(v_i, v_j)}{\alpha L}}, \quad 0 < \alpha, \beta \leq 1 .$$

La constante β del modelo permite variar el grado medio mientras que α , al ajustar la caída de la exponencial, regula la probabilidad de que existan conexiones de largo alcance entre los nodos.

Este modelo fue el primero que intentó reproducir la topología de Internet. Sin embargo presenta limitaciones similares a las de su predecesor, el modelo Erdős-Rényi: las distribuciones de grados obtenidas siguen teniendo una caída exponencial.

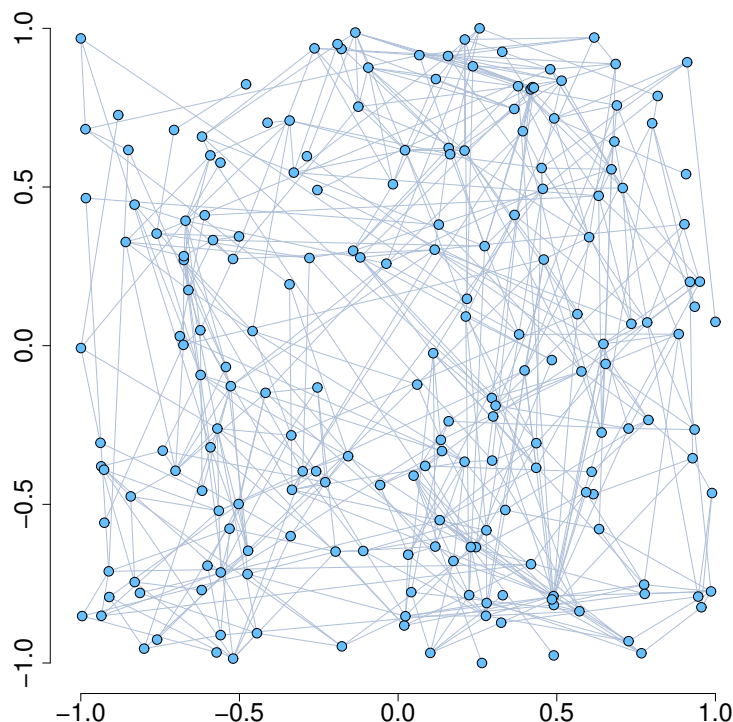


Figura 2.13: *Modelo de Waxman. Visualización.* Instancia de grafo generado con un modelo de Waxman con $\alpha = 0,22$ y $\beta = 0,30$, con $n = 200$ vértices y 529 aristas. El grado medio es $\bar{d} = 5,29$.

2.3.2.2. El modelo Barabási-Albert

El descubrimiento de Faloutsos *et. al* (1999) [66] de que diversas distribuciones de grados en Internet siguen una ley de potencias fue el puntapié inicial de una serie de modelos que intentaron no sólo reproducir este fenómeno, sino también explicarlo. El primero de ellos fue el modelo Barabási-Albert (BA) (1999) [14].

En su trabajo, Barabási y Albert descubrieron la presencia de leyes de escala en las distribuciones de los grados de muchas redes complejas, y la ejemplificaron con la Web, una red de cocitaciones científicas y la red de distribución de energía de los Estados Unidos, entre otras. Asimismo, mostraron que los modelos de redes conocidos hasta ese momento, como el Erdős-Rényi y el Watts-Strogatz, no lograban capturar esta naturaleza

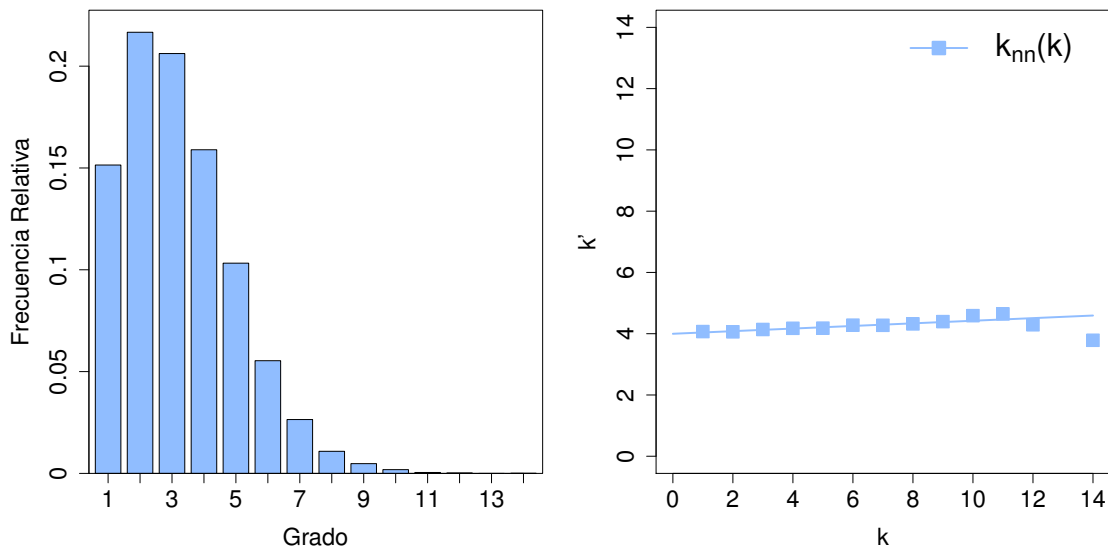


Figura 2.14: *Modelo de Waxman*. Instancia de grafo generado con un modelo de Waxman con $\alpha = 0,15$ y $\beta = 0,0008$, con $n = 20000$ vértices. El grado medio es $\bar{d} = 6$ y el grado máximo es $d_{\text{máx}} = 14$. El coeficiente de agrupamiento global es prácticamente nulo. La concordancia por grados del grafo es 0,043. La distancia media es 8,59 y el diámetro es 19.

libre de escala. Propusieron entonces un nuevo modelo con el objetivo de obtener leyes de potencia en la distribuciones de grados, partiendo de dos hipótesis: el crecimiento de la red y el mecanismo de *acoplamiento preferencial* (*preferential attachment*).

Crecimiento. La primera de las hipótesis es que las redes tienen una evolución dinámica, y que con el tiempo se incorporan nuevos vértices. De esta forma, la topología de la red en un instante determinado es el resultado de la incorporación sucesiva de vértices, uno tras otro. Los autores argumentaron que los modelos previos fallaban en este sentido porque presuponían una cantidad inicial y fija de vértices.

Acoplamiento preferencial. Esta hipótesis sostiene que cuando un nuevo vértice se incorpora a la red, el mismo tiene preferencia por conectarse a vértices de grado elevado. Barabási y Albert ejemplificaron esta hipótesis con la Web. En la Web existen relativamente pocos sitios muy conocidos, y cuando se crean nuevas páginas web es habitual que las mismas incorporen hipervínculos a algunos de estos sitios populares. Con este mecanismo los autores lograron incorporar una idea que estaba presente desde hacía tiempo, en el trabajo de Price sobre las redes de colaboraciones científicas [129], y en el principio social conocido como *the-richer-get-richer*²⁰.

²⁰Literalmente, "los ricos se vuelven más ricos".

El modelo BA (Barabási-Albert) parte de una red formada por una cantidad m_0 de vértices conectados, que llamaremos G_0 , y evoluciona en forma discreta en el tiempo incorporando un nuevo vértice en cada instante. En el instante t , dado el grafo $G_{t-1} = (V_{t-1}, E_{t-1})$ se incorpora un vértice v_t para obtener un grafo G_t . Al ser incorporado, el nuevo vértice se conecta con una cantidad $m \leq m_0$ de vértices en V_{t-1} , cada uno de los cuales se elige con una probabilidad $p(v_j)$ proporcional a su grado:

$$p(v(j)) = \frac{d_{G_{t-1}}(j)}{\sum_{k \leq t-1} d_{G_{t-1}}(v_k)} \quad , j \leq t-1 \quad .$$

A partir de esta simple regla y luego de un tiempo suficiente, la distribución de grados alcanza un estado estacionario libre de escala. Este comportamiento fue mostrado empíricamente por Barabási y Albert y luego probado formalmente por enfoques de campo medio basados en *ecuaciones cinéticas (rate equations)* [15, 93].

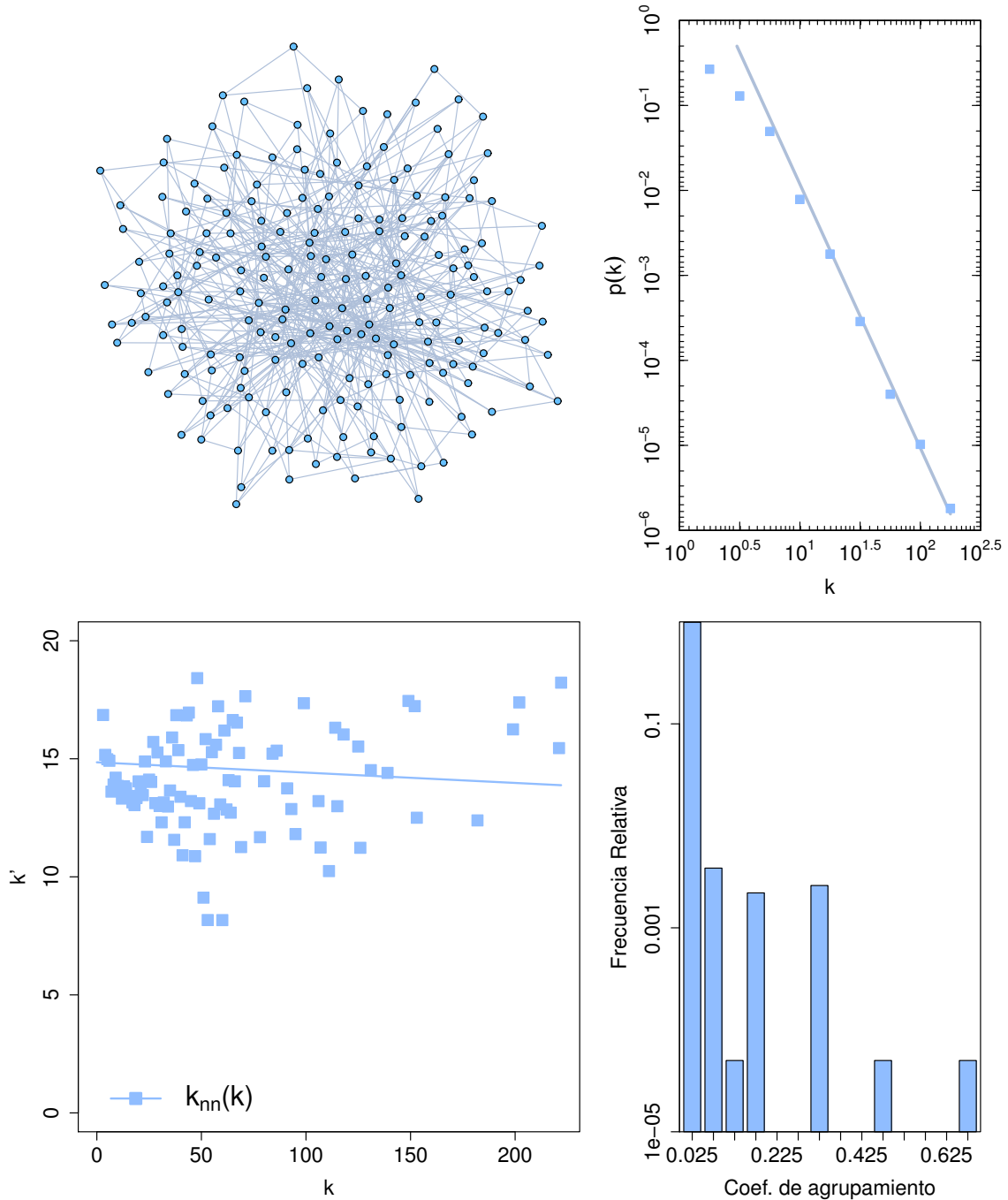


Figura 2.15: *Modelo Barabási-Albert*. Instancia de grafo generado con un modelo Barabási-Albert (BA) con $m = 3$ y $n = 20000$ vértices. El grado medio es $\bar{d} = 6$ y el grado máximo es $d_{\text{máx}} = 222$. Arriba a la izquierda, una visualización del grafo después de haber incorporado los primeros 200 vértices. Arriba a la derecha, el histograma logarítmico de la distribución de grados de los vértices del grafo, ajustada a una ley de potencias con exponente $\alpha = 3,10$ por el método de máxima verosimilitud. Abajo a la derecha, un histograma de los coeficientes de agrupamiento de los vértices, agrupados con un *binning* lineal. Abajo a la izquierda, la correlación entre los grados de vértices adyacentes. Los puntos representan el valor medio del grado de los vecinos, k_{nn} , en función del grado. La pendiente de la recta de regresión –la concordancia por grados del grafo– es $-0,004$. El coeficiente de agrupamiento de la red es prácticamente nulo. La distancia media es 4,71 y el diámetro es 7.

La descripción del modelo en [14] presenta algunas imprecisiones, como fue observado por Bollobás *et al.* [30]. En particular, no se especifica la forma en que están conectados los m_0 vértices iniciales, y al escoger las m conexiones en cada caso no se especifica la distribución conjunta de las m sino solamente la distribución marginal de cada una. Sin embargo, las propiedades libres de escala del modelo no parecen depender de estas elecciones.

Los grafos de red generados con el modelo BA presentan las siguientes propiedades en el estado estacionario ($n \rightarrow \infty$):

- El grado medio $\bar{d}$ es asintótico a $2m$.
- El coeficiente de agrupamiento global es asintótico a $\frac{m-1}{8n(G)} \ln(n(G))^2$ [28].
- La distribución de grados converge a una ley de potencias con exponente $\alpha = 3$.
- Las distancias medias son las propias de redes de mundo pequeño (es decir, son inferiores a $\ln(n(G))$ con alta probabilidad) [49].
- El diámetro es asintótico a $\frac{\ln(n(G))}{\ln \ln(n(G))}$ para $m \geq 2$ [29].
- La concordancia por grados es asintótica a 0.
- El grafo es conexo.

Aunque el modelo BA original tiende a generar leyes de potencias con exponente $\alpha = 3$, con una sencilla variación es posible obtener cualquier ley de potencias con exponente ≥ 2 [60].

En conclusión, hemos logrado imitar las leyes de escala presentes en muchas redes complejas pero no así el fenómeno de mundo pequeño: si bien el modelo BA logra generar redes de diámetro pequeño, tiende a desarrollar un muy pequeño coeficiente de agrupamiento.

2.3.2.3. El modelo FKP

En modelo de Fabrikant *et al.* (FKP) [65] se destaca por haber implementado el mecanismo de Highly Optimized Tolerance (HOT) propuesto por Carlson y Doyle [36] en 1999. Recordemos que el mecanismo HOT sugería que las leyes de potencias en los sistemas complejos surgen como resultado de la optimización de recursos. Siguiendo esta idea, Fabrikant *et al.* propusieron un modelo evolutivo en el que los vértices se agregan dinámicamente y se ubican aleatoriamente en una posición geográfica –de igual modo que en el modelo de Waxman–. Sin embargo, las conexiones que establecen los vértices

incorporados no se determinarán a través de una probabilidad p , sino que se establecerá una conexión con un único vértice previamente existente. De hecho, al incorporar al vértice v_i se evalúa la siguiente función *costo* contra cada uno de los vértices existentes:

$$\Psi(v_i, v_j) = \alpha(n(G))d(v_i, v_j) + \phi(v_j), \quad j \leq i - 1,$$

en donde:

- $\alpha(n(G))$ es función de la cantidad final de vértices, y establece el peso relativo de cada uno de los dos términos.
- $d(v_i, v_j)$ representa la distancia euclídea entre los vértices v_i y v_j .
- $\phi(v_j)$ es una función de *lejanía* –es decir, es inversamente proporcional a alguna medida de centralidad sobre el vértice v_j –, como por ejemplo la inversa de la intermediación o de la cercanía.

Se determina el vértice v_j con $j \leq i - 1$ que minimiza esta función, y se agrega una conexión entre v_i y v_j .

La minimización del funcional $\Psi(v_i, v_j)$ establece un compromiso (*trade-off*) entre dos factores: el costo económico de establecer el enlace –medido por la distancia euclídea– y la utilidad del mismo, representada por la centralidad del vértice en la red. El modelo FKP logra reproducir una ley de potencias en la distribución de grados, pero presenta el inconveniente de que genera un grafo cuya núcleo-profundidad es 1 (su máximo k -núcleo no vacío es el 1-núcleo), con estructura de árbol, y su coeficiente de agrupamiento global es cero. Una extensión del método, propuesta por Alvarez-Hamelin y Schabanel, soluciona esta última limitación [9].

Las propiedades de los grafos generados con el modelo FKP son:

- Para $4 \leq \alpha(n(G)) < \sqrt{n(G)}$ la distribución de grados es asintótica a una ley de potencias con exponente mayor a 1 cuando $n \rightarrow \infty$ (los autores lo prueban tomando como medida de centralidad de los vértices su distancia a un vértice fijo).
- El coeficiente de agrupamiento global es siempre 0.
- El grado medio de los vértices es asintótico a 2.

2.3.3. Generalizaciones del modelo Erdős-Rényi

El modelo Erdős-Rényi original genera grafos de red con distribuciones de grados poissonianas en que los grados de los vértices suelen tener muy poca dispersión. Se suele

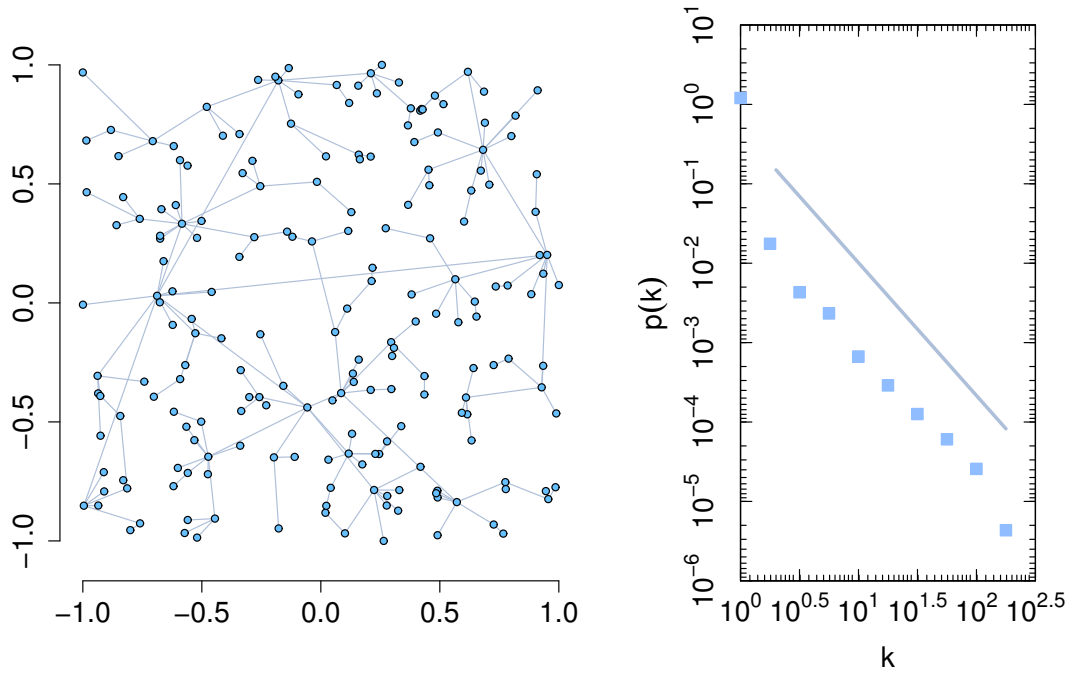


Figura 2.16: *Modelo FKP*. Instancia de grafo generado con un modelo FKP con $\alpha = 25$ y $n = 20000$ vértices. Se utilizó la cercanía como medida de centralidad. El grado medio es $\bar{d} = 2$, el grado máximo es $d_{\text{máx}} = 229$ y el coeficiente de agrupamiento global es nulo. La distancia media es 6,70 y el diámetro del grafo es 12. A la izquierda se muestra una representación del grafo en donde la posición de los vértices en el plano se corresponde con su distribución geográfica, luego de haber conectado los primeros 200 vértices. A la derecha se muestra un histograma logarítmico de la distribución de grados, ajustada a una ley de potencias para $k \geq 2$, con exponente $\alpha = 1,67$, por el método de máxima verosimilitud.

decir en estos casos que el grafo es *homogéneo*. Frente al interés de adaptar el modelo de grafo ER para obtener grafos *heterogéneos* se han propuesto muchas variaciones que intentan generalizar la idea del modelo a otras distribuciones de grados, en particular a las distribuciones libres de escala. De entre estos modelos, desarrollaremos a continuación el modelo de configuración (*configuration model*) y el modelo de grafo aleatorio con grados esperados específicos.

En el *modelo de configuración* [21] se garantiza una secuencia de grados específica. A partir del grado especificado para cada vértice, $d(v_i)$, se arma un conjunto de *stubs* –que pueden pensarse como los extremos de los vértices– de manera que haya $d(v_i)$ *stubs* para cada vértice v_i . De este conjunto de $2e(G)$ *stubs* se escoge un par al azar y se conecta a los dos vértices correspondientes²¹. El proceso se repite, sin reposición de los *stubs* extraídos, hasta que no quede ninguno. Necesariamente, al final del proceso, cada vértice tendrá el

²¹Cada *stub* se elige con distribución uniforme de entre los que quedan. El *configuration model* puede generar grafos con bucles e incluso varias aristas entre un mismo par de vértices.

grado especificado. El modelo presenta la propiedad de muestrear en forma equiprobable el espacio de todos los grafos no isomorfos con la distribución de grados fijada.

En el modelo de grafo aleatorio con grados esperados específicos [43], en cambio, cada par de vértices v_i y v_j se conecta con una probabilidad $p_{ij} = \frac{D_i D_j}{\sum_i D_i}$, de manera que el grado esperado del vértice i es $E[d(v_i)] = D_i$.

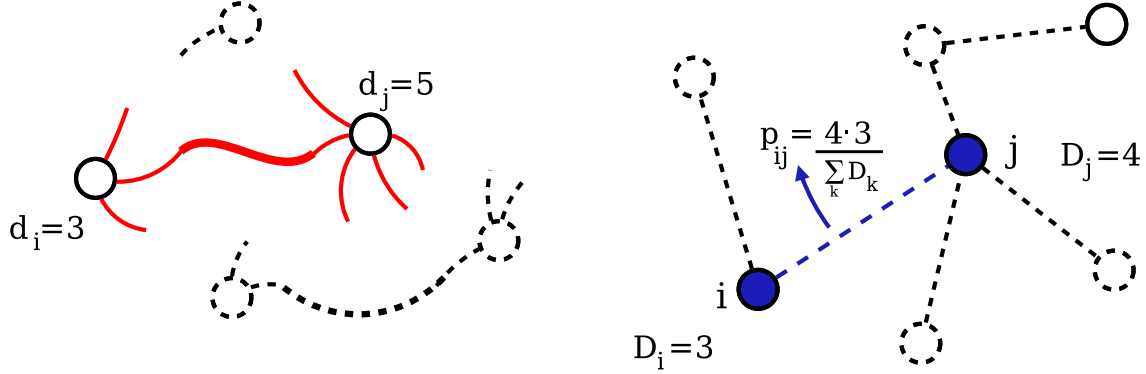


Figura 2.17: *Modelo de configuración y modelo de grafo aleatorio con grados esperados específicos.* En el modelo de configuración (*Izq.*) se conecta a cada vértice una cantidad de *stubs* igual al grado que se le asignó. Se escoge a los *stubs* de a pares en forma aleatoria y se los conecta hasta haber utilizado todos. En el modelo de grafo aleatorio con grados esperados específicos (*Der.*) la probabilidad de conexión de dos vértices v_i y v_j es $p_{ij} = \frac{D_i D_j}{\sum_i D_i}$, en donde D_i es el grado esperado del vértice i .

En ambos modelos al introducir distribuciones de grados libres de escala se reproduce en parte el fenómeno de mundo pequeño: la distancia promedio esperada, para valores de α comprendidos entre 2 y 3, tiende asintóticamente a $\frac{2 \log(\log(n(G)))}{\log(\alpha-2)^{-1}}$ para $n \rightarrow \infty$, mientras que el diámetro es del orden de $n(G)$. Sin embargo, ninguno de los dos logra obtener un coeficiente de agrupamiento apreciable [134, 43].

2.3.4. Modelos de Redes Sociales

Desarrollaremos a continuación las características del modelo Watts-Strogatz, que logró reproducir el fenómeno de *mundo pequeño*, y algunos de los modelos utilizados para generar una estructura comunitaria: el modelo *planted l -partition* y el Lancichinetti-Fortunato-Radicchi (LFR).

2.3.4.1. El modelo Watts-Strogatz

Muchas redes complejas, en particular las redes sociales, presentan el fenómeno de mundo pequeño, que se caracteriza por una distancia promedio pequeña entre dos nodos cualesquiera y un elevado coeficiente de agrupamiento.

Watts y Strogatz se plantearon entonces el problema de reproducir esta situación en un modelo de red con grado medio fijo [154]. En el modelo Erdős-Rényi hemos visto que esto no es posible, dado que manteniendo el producto np constante, con n suficientemente grande se obtiene una red disconexa y con coeficiente de agrupamiento tendiendo a cero. Los autores compararon esta situación con el caso de determinados reticulados, en los que el coeficiente de agrupamiento es elevado pero la distancia entre dos nodos puede ser muy grande. Buscando un punto intermedio, propusieron un modelo que parte de una estructura de tipo anillo en la que los nodos se conectan con sus $2 \cdot k$ vecinos más cercanos en el anillo –de manera de obtener un alto coeficiente de agrupamiento– y luego realiza un proceso de *recableado* en que, con cierta probabilidad p , las aristas uv se eliminan y se establece una arista uw con algún nodo w al azar. El procedimiento de recableado no modifica la cantidad de aristas del grafo, por lo cual se conserva el grado medio. El incremento de la probabilidad p disminuye la distancia promedio pero también el coeficiente de agrupamiento global. Sin embargo, para una amplia variedad de valores de p (en el rango $n^{-1} \ll p \ll 1$) el modelo logra obtener grafos con baja distancia promedio y coeficiente de agrupamiento elevado.

Los grafos aleatorios obtenidos con el modelo Watts-Strogatz presentan las siguientes propiedades [17]:

- El tamaño del grafo es kn .
- Para $n \rightarrow \infty$ y $p \rightarrow 1$, la distribución de grados de los vértices converge a una distribución de Poisson de media k .
- En la región $n^{-1} \ll p \ll 1$ el coeficiente de agrupamiento esperado de un vértice es $\frac{3(k-1)}{2(2k-1)}$.
- En la región $n^{-1} \ll p \ll 1$ la distancia esperada entre vértices es $\ln n / \ln k$.

Si bien la distribución de los grados en el modelo Watts-Strogatz sigue siendo una distribución de Poisson, el gran mérito de este modelo es haber sido el primero en construir redes de mundo pequeño.

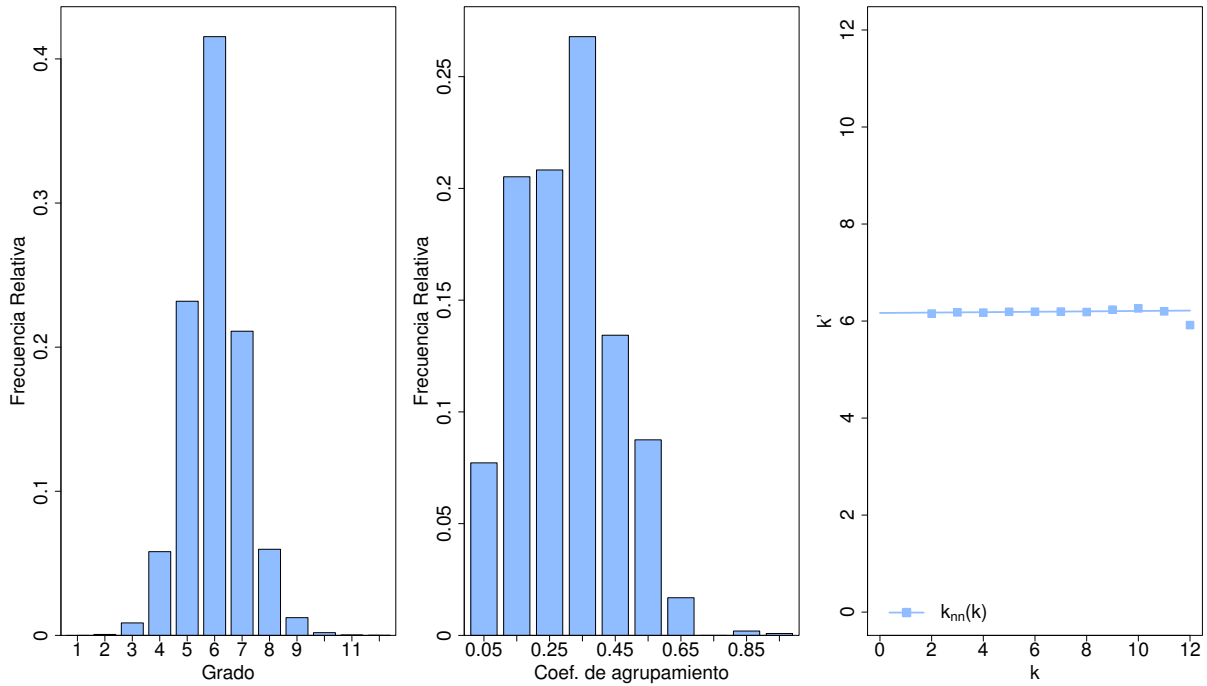


Figura 2.19: *Modelo Watts-Strogatz*. Instancia de grafo generado con un modelo Watts-Strogatz con $p = 0,1$, $k = 3$ y $n = 20000$ vértices. El grado medio es $\bar{d} = 6$ y el grado máximo es $d_{\text{máx}} = 12$. (Izq.) Distribución de grados de los vértices del grafo. (Cen.) Frecuencias relativas de los coeficientes de agrupamiento de los vértices, agrupados con un *binning* lineal. (Der.) Correlación entre los grados de vértices adyacentes. Los puntos representan el valor medio del grado de los vecinos, k_{nn} , en función del grado. La pendiente de la recta de regresión –la concordancia por grados del grafo– es 0,004. El coeficiente de agrupamiento global de la red es 0,302. La distancia media es 7,58 y el diámetro es 12.

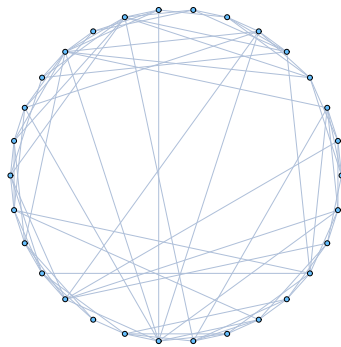


Figura 2.18: *Modelo Watts-Strogatz*. Visualización. Instancia de grafo generado con un modelo Watts-Strogatz con $p = 0,2$, $k = 3$ y $n = 30$ nodos. El grafo tiene 90 aristas y grado medio $\bar{d} = 6$.

2.3.4.2. El modelo *planted l-partition*

El modelo *planted l-partition* fue propuesto por Condon y Karp en 2001 [51] en el contexto de la minería de datos, para ser utilizado como banco de prueba en la tarea de *clustering*.

Este modelo construye un grafo G_n con n vértices agrupados en l comunidades, todas de igual tamaño, que forman una partición del conjunto de vértices. A partir de este agrupamiento inicial, se considera cada par de vértices (u, v) y se los conecta con una probabilidad p_i si pertenecen a la misma comunidad, y con una probabilidad $p_o < p_i$ si pertenecen a comunidades distintas. De esta forma, los vértices quedan más densamente conectados dentro de sus comunidades que hacia afuera.

Los grafos obtenidos bajo este modelo tienen distribuciones de grados homogéneas, con un grado esperado de $E[d] = p_i \left(\frac{n}{l} - 1\right) + p_o \frac{n(l-1)}{l}$, y muy poca dispersión.

El banco de prueba de Girvan-Newman (GN) [76], con $n = 128$ y $l = 4$, es un caso particular del modelo *planted l-partition*, en donde las probabilidades p_i y p_o se eligen de manera que el grado esperado de los nodos sea $E[d(v)] = 16$, lo que conduce a la relación

$$31p_i + 96p_o = 16, \quad p_o < p_i \quad .$$

2.3.4.3. El modelo LFR

Este modelo propuesto por Lancichinetti, Fortunato y Radicchi en 2008 [97] genera grafos con distribuciones heterogéneas tanto en los grados de los vértices como en los tamaños de las comunidades. Se ajusta a través de los siguientes parámetros²²:

- n , el tamaño del grafo, $n(G)$.
- γ , el exponente de la ley de potencias para la distribución de grados de los vértices.
- $\bar{d}$, el grado medio de la ley de potencias.
- $d_{\text{máx}}$, el grado máximo de los vértices.
- β , el exponente de la ley de potencias para la distribución de tamaños de las comunidades²³.
- $s_{\text{mín}}$, el tamaño mínimo de las comunidades.

²²Adicionalmente a los parámetros que mencionamos, existen dos parámetros que permiten definir un *solapamiento (overlapping)* entre las comunidades, que no desarrollaremos aquí.

²³Los autores definen el tamaño de una comunidad como la suma de los grados de sus vértices.

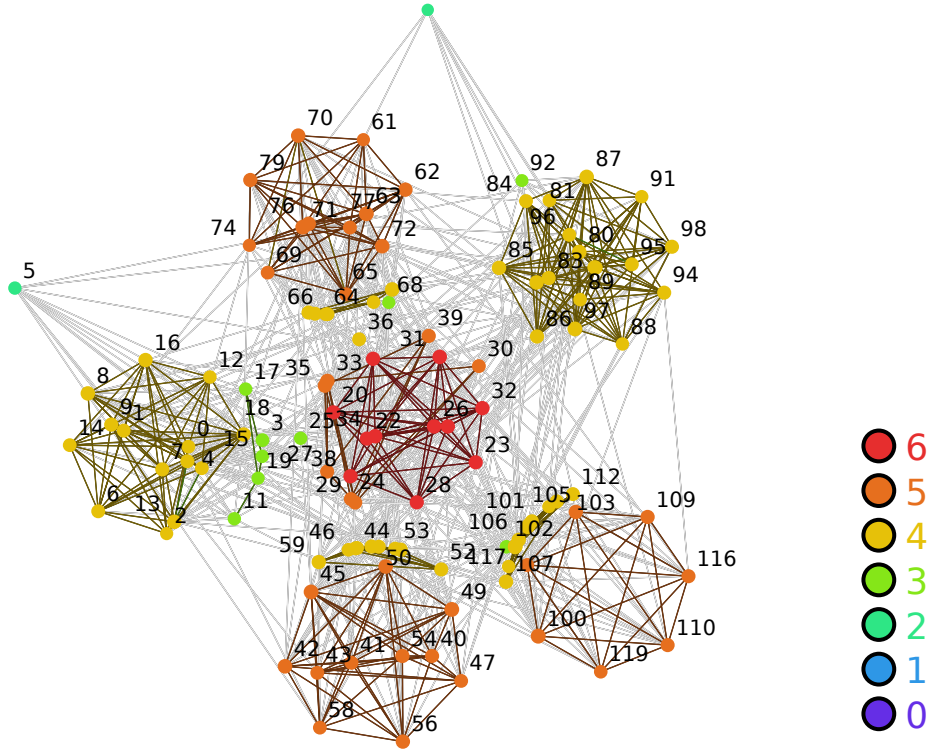


Figura 2.20: *Modelo planted l -partition*. Instancia de grafo generado con el modelo *planted l -partition*, con 120 vértices organizados en 6 comunidades. Las probabilidades de conexión dentro y fuera de las comunidades son 0,4 y 0,02 respectivamente. La visualización se realizó con el software LaNet-vi empleando la descomposición en k -densos del grafo. La escala de la derecha representa el índice de denso de los vértices. Los vértices que pertenecen a la misma comunidad tienen números contiguos, de manera que un vértice v_i pertenece a la comunidad $\lceil \frac{i}{20} \rceil$

- $s_{\text{máx}}$, el tamaño máximo de las comunidades.
- μ , el denominado parámetro de mezclado (*mixing parameter*), que especifica la proporción deseada de conexiones que los vértices tienen hacia afuera de su comunidad, respecto a su grado.
- C , el coeficiente de agrupamiento global deseado.

La construcción del grafo se realiza siguiendo los siguientes pasos:

1. Se asigna a cada vértice un grado tomado de una ley de potencias truncada ($d \leq d_{\text{máx}}$) con exponente γ y esperanza $\bar{d}$.
2. Se realizan las conexiones de la misma forma que en el modelo de configuración.

3. Se definen los tamaños de las comunidades a partir de una ley de potencias truncada ($s \leq s_{\text{máx}}$) con exponente β y tamaño mínimo $s_{\text{mín}}$.
4. Se asigna a cada vértice una comunidad al azar, bajo la restricción de que luego de su inclusión la comunidad no exceda el tamaño que se le había asignado. A través de iteraciones sucesivas se refina el procedimiento hasta que todos los vértices tengan una comunidad.
5. Se realiza un proceso de recableado (*rewiring*) para ajustar el μ de cada vértice lo máximo posible al μ especificado.
6. Finalmente se realiza un segundo recableado para ajustar el coeficiente de agrupamiento promedio al valor deseado.

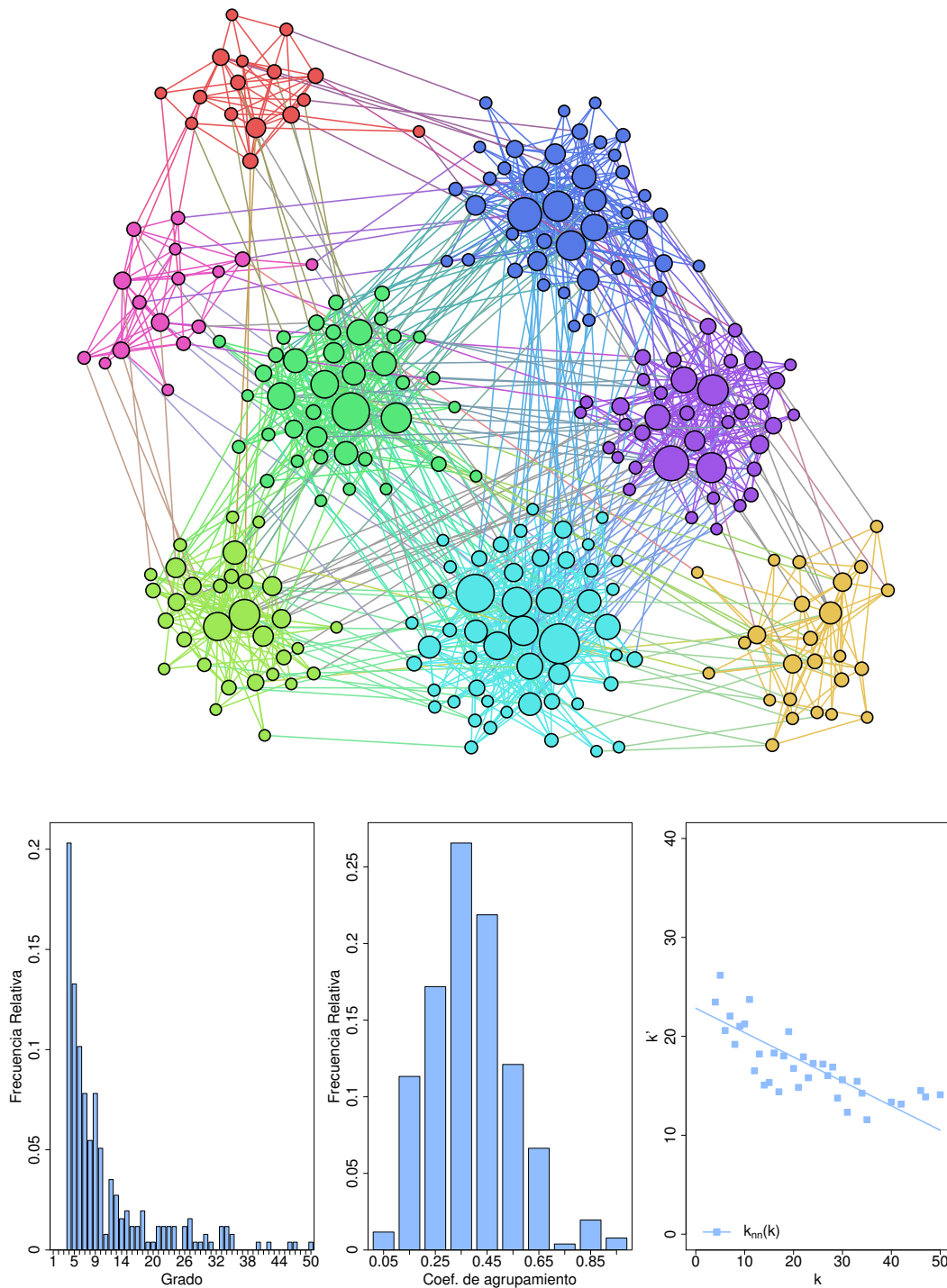


Figura 2.21: *Modelo LFR*. Instancia de grafo generado con el modelo Lancichinetti-Fortunato-Radicchi (LFR) con los siguientes parámetros: $n = 256$, $\bar{d} = 10$, $d_{\text{máx}} = 50$, $\gamma = 2,0$, $\beta = 1,0$, $\mu = 0,2$, $s_{\text{mín}} = 10$, $s_{\text{máx}} = 50$, $C = 0,4$. La instancia generada tiene un grado medio de 10,84, un μ promedio de 0,199 y un coeficiente de agrupamiento promedio de 0,41. La visualización se realizó con el software Gephi. El color de los vértices indica su comunidad de pertenencia y su tamaño es proporcional al grado. Abajo se muestran la distribución de grados, un histograma del coeficiente de agrupamiento de los vértices, y el k_{nn} en función del grado.

Capítulo 3

Descubrimiento de Comunidades en Redes Sociales

La *estructura comunitaria* surge de la organización de los elementos que conforman una red en grupos de pertenencia, a los que llamamos *comunidades*. Esta organización es característica de varios tipos de redes complejas, en particular de las siguientes:

- *Redes sociales*. El descubrimiento de comunidades permite estudiar las asociaciones entre personas de distinta índole o con distintos fines: relaciones de amistad, de trabajo, de familia o de afinidad, por ejemplo. Con el surgimiento de las redes virtuales a través de Internet se han reducido las barreras geográficas y entonces se ha facilitado la formación de *comunidades virtuales*, en las que las personas se asocian muchas veces por su afinidad cultural, política o ideológica. El hecho de que estas comunidades se sustenten en la tecnología de las redes de información tiene algunas consecuencias importantes. Por una parte, ofrece un gran volumen de datos para el estudio científico, requiriendo de métodos eficientes de procesamiento. Por otro lado tiene un gran potencial económico: la información sobre la *vida virtual* de una persona o grupo tiene valor comercial, permite a las empresas descubrir potenciales clientes y ofrecer sus servicios de forma eficiente. Por último, plantea un serio debate sobre la seguridad de la información y la privacidad en el mundo virtual.
- *Redes de colaboraciones científicas* Los científicos trabajan en comunidades de acuerdo con sus áreas de investigación. En algunas áreas cada investigador colabora con muchos otros, resultando en grandes comunidades. En otras áreas, en cambio, los grupos de trabajo son más *cerrados* y las comunidades más pequeñas [111].
- *La Web*. Los sitios en la Web se organizan en comunidades alrededor de determinados temas. Estas comunidades surgen espontáneamente a partir de los enlaces

entre los sitios [68].

- *Redes metabólicas.* Este tipo de redes biológicas representan la evolución de los compuestos (en este contexto, metabolitos) en los ciclos o procesos químicos. En cada proceso ocurren una serie de reacciones en que ciertos metabolitos reaccionan para producir otros. La red formada por todos estos procesos posee comunidades que representan la organización de los metabolitos en *módulos*. Cada módulo está correlacionado con uno o más ciclos o procesos [86].
- *Redes de interacciones entre proteínas.* Las proteínas en los organismos vivos se asocian dentro de las células para participar de procesos vitales que tienen una función específica en el organismo. El descubrimiento de comunidades es entonces una potencial herramienta para inferir la funcionalidad de estas sustancias a partir de la estructura que emerge de las interacciones [41].
- *Redes tróficas.* En los ecosistemas, el descubrimiento de comunidades permite estudiar las relaciones tróficas entre las especies. La noción de comunidad está asociada en este caso al concepto de *compartimento ecológico* [94].

En términos generales, el potencial del descubrimiento de comunidades está vinculado con la capacidad de inferir relaciones entre los elementos de la red, de predecir su comportamiento o decisiones futuras y de entender de qué manera las comunidades surgen y evolucionan.

Este capítulo se estructura de la siguiente manera: en la primera sección discutimos el concepto de comunidad y sus diversas interpretaciones; en 3.2 hacemos un recorrido histórico por los métodos de descubrimiento de comunidades que consideramos más relevantes; luego discutimos la necesidad de definir *métricas de comparación* adecuadas. En las secciones 3.4 y 3.5 presentamos nuestro aporte al descubrimiento de comunidades en redes complejas, plasmado en los artículos [33, 20].

3.1. Introducción a la noción de comunidad

Reconocemos como antecedente del estudio de comunidades en redes complejas al problema de minería de datos conocido con el nombre de *agrupamiento de datos* (*data clustering*). En el problema de agrupamiento de datos se debe agrupar a los elementos de un conjunto de datos en una determinada cantidad de *clusters* en función de sus propiedades –generalmente modeladas como coordenadas en un espacio n -dimensional–. Típicamente en este problema está definido el concepto de *distancia* entre los elementos del conjunto, y la asignación de elementos a los *clusters* busca de alguna manera que

estos sean compactos, es decir, que la distancia entre elementos de un mismo *cluster* sea pequeña.

En el problema de descubrimiento de comunidades, en cambio, existen dos diferencias importantes:

1. La existencia de comunidades puede ser incierta, de manera que se espera que el método de descubrimiento indique si existen comunidades, además de cuántas y cuáles son.
2. La pertenencia de los vértices de la red a las comunidades está determinada principalmente por las conexiones que existen en la red. No es necesaria, en principio, una noción de distancia.

Sin embargo algunos métodos de descubrimiento de comunidades se basan en una definición de distancia y aplican métodos tradicionales de agrupamiento de datos, en particular los denominados de *agrupamiento jerárquico* (*hierarchical clustering*).

Otro antecedente importante del concepto de comunidad lo constituyen los estudios sobre *cohesión* en grupos sociales. La *cohesión* (fortaleza de los vínculos) de un grupo puede determinar uniformidad de opinión o influencia entre las personas que pertenecen a él. Desde la década del '40 en adelante los sociólogos han definido conceptos como los *cliques* [106], los *n-cliques* [1], los *k-plexos* (*k-plexes*) [143], los *n-clanes* (*n-clans*) [109], los *n-clubes* (*n-clubs*) [109] y los *conjuntos LS* (*LS sets*) [98] en el marco del estudio de la cohesión social (véase la Figura 3.1).

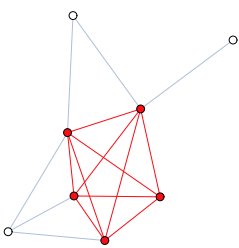
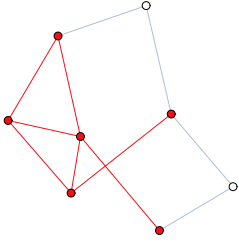
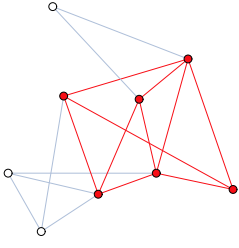
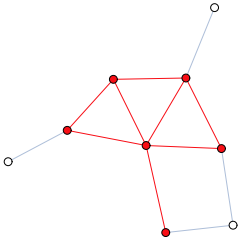
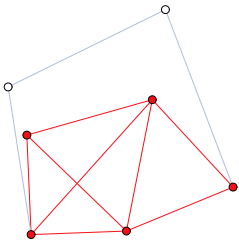
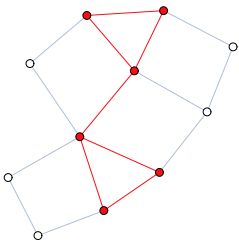
En el campo de las redes complejas la noción de comunidad comenzó a tomar forma con los trabajos de Flake *et al.* (2000) [68] y Newman y Girvan (2001) [111]:

- Flake *et al.* [68] propusieron el concepto de *comunidad web* como un conjunto de vértices $C \subset V(G)$ tal que cada uno de ellos tiene más vecinos dentro del conjunto que fuera de él. Esto se expresa, según la notación que introducimos en el Cuadro 3.2, como:

$$\forall v \in C : d_C^{in}(v) > d_C^{out}(v) .$$

- Newman analizó en [111] el concepto de comunidad en el contexto de una red de colaboraciones científicas. Allí observó que la existencia de comunidades estaba vinculada con un alto coeficiente de agrupamiento: dos científicos que habían colaborado cada uno con un tercero tenían altas probabilidades de haber trabajado en forma conjunta entre ellos.

A partir de entonces el interés por estudiar la estructura comunitaria de las redes complejas ha ido en aumento.

Estructura	Definición	Ejemplo
clique de orden k	subgrafo completo maximal de k vértices.	
n -clique	conjunto maximal de vértices a distancia menor o igual a n tomados de a pares. <i>Ejemplo: 2-clique.</i>	
k -plexo	subgrafo maximal de n vértices en que cada vértice se conecta con al menos $n - k$ vértices del subgrafo. <i>Ejemplo: 3-plexo.</i>	
n -clan	n -clique cuyo subgrafo inducido tiene diámetro menor o igual a n . <i>Ejemplo: 2-clan.</i>	
n -club	subgrafo maximal de diámetro menor o igual a n . <i>Ejemplo: 2-club.</i>	
conjunto LS	conjunto de vértices en que cada vértice tiene más conexiones con otros vértices del conjunto que hacia fuera.	

Cuadro 3.1: Estructuras cohesivas utilizadas para estudiar grupos sociales.

3.2. Métodos de descubrimiento de comunidades

A continuación explicaremos como han evolucionado los métodos de descubrimiento de comunidades desde sus comienzos, y mencionaremos algunos de los más conocidos. Veremos que en algunos de ellos las comunidades surgen de un planteo sobre todo el grafo; a estos métodos se los denomina *globales*. En los métodos *locales*, en cambio, las comunidades surgen localmente y no se ven afectadas por lo que sucede en partes lejanas del grafo. Veremos que en el estudio de las comunidades hubo una evolución de los métodos globales a los locales, y hoy en día se prefieren estos últimos. Para un desarrollo más extenso del estado del arte sugerimos el exhaustivo *survey* de Fortunato [70].

Ya hemos mencionado el trabajo seminal de Newman sobre la red de colaboraciones científicas. Un año después de aquel, Newman desarrolló un método de agrupamiento jerárquico divisivo para el descubrimiento de comunidades basado en la *arista-intermediación* (2002) [76] (véase nuestra Subsección 2.1.3.1). El método se basa en la idea de que las aristas que conectan vértices internos a una misma comunidad deberían tener una arista-intermediación menor que las que conectan vértices en comunidades distintas, porque estas últimas tienen más posibilidades de formar parte de caminos mínimos entre dichos vértices. El algoritmo propuesto calcula la arista-intermediación de todas las aristas y las elimina una a una comenzando por las que tienen mayor valor. A medida que el grafo se desconecta se construye un *dendrograma* en el que las ramas representan las componentes conexas. El algoritmo actualiza la arista-intermediación de todas las aristas después de cada eliminación. Esto hace que sea costoso computacionalmente.

Radicchi *et al.* (2004) [130] propusieron una modificación del algoritmo de Newman en que en lugar de calcular la arista-intermediación se calcula el *coeficiente de agrupamiento para aristas* (*edge-clustering coefficient*), introducido en el mismo trabajo. A su vez, observaron la necesidad de contar con una definición de comunidad no operacional –es decir, que no esté basada en el resultado de determinado algoritmo o método–. Esto permitiría evaluar los métodos de descubrimiento, compararlos e incluso decidir si la estructura comunitaria que obtienen es significativa o no. Radicchi *et al.* ofrecen dos definiciones de comunidad:

1. *Comunidad en sentido fuerte*. Un conjunto $C \subset V(G)$ es una comunidad en sentido fuerte si:

$$\forall v \in C : d_C^{in}(v) > d_C^{out}(v) .$$

2. *Comunidad en sentido débil*. Un conjunto $C \subset V(G)$ es una comunidad en sentido débil si:

$$\sum_{v \in C} d_C^{in}(v) > \sum_{v \in C} d_C^{out}(v) .$$

Símbolo	Significado
$\mathcal{C} = (C_1, C_2, \dots, C_{ \mathcal{C} })$	partición en comunidades de una red
$\sigma(v)$	subíndice de la comunidad de pertenencia de v
$C(v)$	comunidad de pertenencia de v
$d_C^{in}(v)$	grado interno de v respecto a C
$d_C^{out}(v)$	grado externo de v respecto a C
$\mu(v) = \frac{d_{C(v)}^{out}(v)}{d(v)}$	parámetro de mezclado de v

Cuadro 3.2: *Resumen de la notación relativa a estructuras comunitarias (Parte 1)*. Las cantidades $d_C^{in}(v)$ y $d_C^{out}(v)$ representan, respectivamente, la cantidad de vecinos de v dentro y fuera de C . La notación se aplicará para vértices v tanto dentro como fuera de la comunidad C .

La noción de comunidad en sentido fuerte se corresponde con la de *comunidad web* en Flake *et al.* [68] y con la de conjunto LS [98]. El método de Radicchi *et al.* construye un dendrograma a partir del coeficiente de agrupamiento para aristas de manera análoga a Newman, y utiliza los criterios de *comunidad fuerte* o *débil* como criterio de parada.

En el mismo año Newman propuso otra variante en que el *peso* de cada arista se calcula efectuando un camino aleatorio y contabilizando las veces que la misma es utilizada en cada sentido [120]. Luego se aplica el mismo algoritmo de agrupamiento jerárquico eliminando primero las aristas de menor peso. La discusión sobre qué nivel del dendrograma representa mejor la estructura comunitaria llevó a Newman a proponer un funcional global conocido como *modularidad* que se convirtió por varios años en el estándar para medir la bondad de las estructuras comunitarias y evaluar el desempeño de los algoritmos propuestos.

La modularidad. Dada una partición del conjunto de vértices de un grafo en un conjunto de comunidades $\mathcal{C} = (C_1, C_2, \dots, C_{|\mathcal{C}|})$, la modularidad $Q_G(\mathcal{C})$ se define como [120]¹.

$$Q_G(\mathcal{C}) = \text{Tr}(\mathbf{e}) - \|\mathbf{e}^2\| \ ,$$

¹Obsérvese la semejanza de esta expresión con la de concordancia por categorías (pág. 41). Considerando a las comunidades como categorías, la definición de modularidad coincide con la de concordancia, salvo un factor divisivo.

en donde $\mathbf{e}$ es una matriz cuyas componentes e_{ij} representan la probabilidad de que una arista (u, v) vaya de un vértice en la comunidad C_i a un vértice en la comunidad C_j . Dichas probabilidades se pueden calcular como

$$e_{ij} = \frac{|(C_i, C_j)|}{2e(G)} = \frac{\sum_{(u,v) \in C_i \times C_j} \mathbf{1}\{u \rightarrow v\}}{2e(G)} .$$

De aquí se deriva [45]

$$Q_G(\mathcal{C}) = \frac{1}{2e(G)} \sum_{(v_i, v_j) \in V(G) \times V(G)} \left[A_{ij} - \frac{d(v_i)d(v_j)}{2e(G)} \right] \mathbf{1}\{\sigma(v_i) = \sigma(v_j)\} , \quad (3.1)$$

en donde $\mathbf{1}\{\sigma(v_i) = \sigma(v_j)\}$ vale 1 cuando v_i y v_j pertenecen a la misma comunidad, y 0 en caso contrario.

El primer término de la modularidad, determinado por $\text{Tr}(\mathbf{e})$, coincide con la proporción de aristas internas a las comunidades –es decir, que conectan vértices de una misma comunidad– respecto del total de aristas. El segundo término evalúa la proporción esperada de aristas internas en un grafo aleatorio con los mismos vértices, grados esperados y comunidades². Podemos decir entonces que *la modularidad mide la bondad de una estructura comunitaria comparando su densidad de aristas internas con la que tendría si las conexiones se realizaran aleatoriamente*.

A partir del supuesto de que una estructura comunitaria es mejor que otra si su modularidad es mayor, Newman sugirió que la mejor partición en comunidades de un grafo sería aquella que maximice el valor de Q . El problema de maximización de la modularidad es un problema de optimización combinatoria³ computacionalmente costoso; de hecho Brandes *et al.* demostraron que es NP-completo [31]. Sin embargo, puede ser abordado por diversos métodos heurísticos.

La modularidad pasó entonces de ser meramente un evaluador de estructura comunitaria a ser un funcional global a optimizar. Entre los numerosos métodos de maximización de la modularidad, destacamos: el algoritmo goloso de Clauset-Newman-Moore (CNM, 2004) [45], el de Guimerà *et al.* basado en *simulated annealing* (2004) [85], el de *extremal optimization* de Duch y Arenas (2005) [63], el de Danon *et al.* (2006) [53], el de bisección por descomposición espectral de Newman [117], el de Wakita y Tsurumi (2007) [152], el de Blondel *et al.* (2008) [24] y el algoritmo multinivel de Noack y Rotta (2009) [121]. También se han desarrollado extensiones de la modularidad para grafos dirigidos [99] y pesados [10].

²Construido de acuerdo al modelo de grafo aleatorio con grados esperados específicos (ver pág. 65).

³Recae dentro de la categoría de problemas de *asignación cuadrática* (*quadratic assignment*).

Las limitaciones encontradas en la modularidad y que desarrollamos en la Sección 3.4 –en particular, el límite de escala– señalaron la necesidad de desarrollar *métodos locales* de búsqueda de comunidades. Uno de los primeros fue el denominado *Clique Percolation Method (CPM)* propuesto por Palla *et al.* (2005) [123]. Este método construye las comunidades a través de un proceso de percolación de cliques de orden k , y no encuentra particiones sino *cubrimientos*, en los que las comunidades pueden solaparse.

Raghavan *et al.* (2007) [131] propusieron un algoritmo local que encuentra una partición en comunidades utilizando un algoritmo de *propagación de etiquetas*. Se comienza asignando a cada vértice una etiqueta distinta, y luego, a través de un proceso iterativo, los vértices reemplazan su etiqueta por aquella que tiene mayoría de entre sus vecinos⁴. El criterio de terminación consiste en que todos los vértices tengan al menos tantas aristas hacia sus comunidades como hacia cada una de las demás⁵. Si bien el algoritmo podría ser inestable, y de hecho los autores no prueban la convergencia, en las redes complejas analizadas converge en pocas iteraciones. En este método se encuentra implícita la idea de que las comunidades juegan un rol importante en los procesos de difusión, idea que también está presente en los métodos de percolación y espectrales. Tibély y Kertész mostraron que el proceso de Raghavan *et al.* es equivalente a encontrar un mínimo local del hamiltoniano de un modelo de Potts [149].

En 2009 Lancichinetti *et al.* propusieron un método local basado en el concepto de *comunidad natural* [96]. La comunidad natural de un vértice se define constructivamente, partiendo del vértice e incorporando (y a veces eliminando) vértices con el objetivo de incrementar la función objetivo (*fitness function*) de la comunidad, definida como: [96]

$$f_L(C) = \frac{d^{in}(C)}{(d^{in}(C) + d^{out}(C))^\alpha} \quad , \quad (3.2)$$

en donde $d^{in}(C)$ y $d^{out}(C)$ representan la suma de los grados internos y externos de los vértices de C (véase esta notación en el Cuadro 3.3).

Una de las características del método de Lancichinetti *et al.* es que encuentra cubrimientos del grafo, dado que los vértices pueden pertenecer a más de una comunidad natural. Por otra parte, la función objetivo ofrece una medida cuantitativa de la significatividad de una comunidad.

Varios investigadores han analizado las distribuciones de tamaños de las comunidades y han encontrado distribuciones de cola larga. El fenómeno ya había sido observado en 2002 por Guimerà *et al* en la red de intercambio de correo electrónico [87], por Gleiser y

⁴En caso de empate, se elige un vecino al azar de entre los que realizan el máximo, y se asigna la etiqueta de ese vecino.

⁵Este criterio es similar a la definición de *comunidad en sentido fuerte* de Radicchi, aunque con un signo $\geq$ en lugar de $>$.

Danon en 2003 en la red de bandas de *jazz* [78] y por Newman en la red de colaboraciones científicas [113] en 2003. En todos estos casos los resultados se obtuvieron aplicando métodos basados en la maximización de la modularidad, y mostraron leyes de potencias en un rango de aproximadamente 3 décadas de la escala logarítmica, con exponentes de entre 1,5 y 2. El tamaño limitado de aquellas redes no permitía observar los efectos del límite de escala de la modularidad, que se hacen evidentes en redes de mayor tamaño. Los métodos locales de Lancichinetti *et al.* [96] y Palla *et al.* [123], en cambio, mostraron el mismo fenómeno en un rango mayor de valores. En conclusión, la existencia de un límite de resolución para la modularidad pone en duda su capacidad de encontrar estructuras comunitarias con distribuciones libres de escala en redes heterogéneas. En la Sección 3.5.7 mostraremos a partir del banco de prueba de Lancichinetti-Fortunato-Radicchi [97] los efectos del límite de escala de la modularidad sobre la distribución de grados de las comunidades.

Por último, dentro de los métodos globales mencionamos InfoMAP, basado en una novedosa idea propuesta por Rosvall y Bergstrom (2008) [139]. En este trabajo los autores sugirieron que la mejor estructura comunitaria es aquella que minimiza la *longitud de la descripción* (*description length*), que es la cantidad de información de una codificación conjunta de la estructura comunitaria y del grafo. Dicho de otra forma, en una estructura comunitaria óptima, a partir de la asignación de comunidades a los vértices debería poder recuperarse el grafo completo con poca información adicional, al mismo tiempo que la cantidad de información de la asignación de comunidades no debería ser excesiva. Los autores han minimizado este funcional global utilizando métodos variados como *simulated annealing* [139] y caminos aleatorios [140].

La longitud de la descripción. Para calcular la longitud de la descripción de una partición $\mathcal{C}$ se debe partir de: (i) una codificación del grafo en que a cada comunidad de la partición se le asigna un código, y (ii) un conjunto de codificaciones internas a las comunidades, que asignan un código a cada vértice de la comunidad. La longitud de la descripción representa entonces la longitud media de la descripción de un camino aleatorio infinito sobre el grafo a partir de este conjunto de codificaciones, que se calcula en el estado estacionario del proceso de Markov asociado al grafo. La mínima longitud de descripción, $L(\mathcal{C})$, es la longitud media mínima de entre todas las codificaciones, que corresponde al límite de Shannon. El desarrollo de su fórmula, conocida como *map equation*, se puede encontrar en [138]. Aquí nos limitamos a transcribirla en términos de

nuestras medidas m_V y c_E , para grafos no dirigidos:

$$L(\mathcal{C}) = \left(\sum_{C \in \mathcal{C}} c_E(C) \right) \log \left(\sum_{C \in \mathcal{C}} c_E(C) \right) - 2 \sum_{C \in \mathcal{C}} c_E(C) \log(c_E(C)) - \\ - \sum_{v \in V(G)} m_V(v) \log(m_V(v)) + \sum_{C \in \mathcal{C}} (c_E(C) + m_V(C)) \log(c_E(C) + m_V(C))$$

Cantidad	Notación	Definición	Equivalencias
Tamaño	$s(C_i)$	$ C_i $	
Grado	$d(C_i)$	$\sum_{v \in C_i} d(v)$	$ C_i, V(G) $
Medida del grado	$m_V(C_i)$	$\frac{d(C_i)}{2e(G)}$	
Grado interno	$d^{in}(C_i)$	$\sum_{v \in C_i} d_{C_i}^{in}(v)$	$ (C_i, C_i) $
Medida del grado interno	$m_E(C_i)$	$\frac{d^{in}(C_i)}{2e(G)}$	$\frac{ (C_i, C_i) }{2e(G)}$
Grado externo	$d^{out}(C_i)$	$\sum_{v \in C_i} d_{C_i}^{out}(v)$	$ (C_i, V(G) \setminus C_i) $
Medida del grado externo	$c_E(C_i)$	$\frac{d^{out}(C_i)}{2e(G)}$	$\frac{ (C_i, V(G) \setminus C_i) }{2e(G)}$
Medida del corte	$m_E(C_i \times C_j)$	$\frac{ (C_i, C_j) }{2e(G)}$	
Parámetro de mezclado	$\mu(C_i)$	$\frac{m_V(C_i) - m_E(C_i)}{m_V(C_i)}$	$\sum_{v \in C_i} \frac{\mu(v) \cdot d(v)}{d(C_i)}$

Cuadro 3.3: *Resumen de la notación relativa a estructuras comunitarias (Parte 2).*

3.3. Métricas de comparación

Dado que el concepto de comunidad no tiene una definición unívoca sino que más bien depende de cada contexto, es necesario establecer criterios para medir la bondad de los métodos de descubrimiento de comunidades. En principio, en la evaluación del desempeño de los diversos métodos se abren dos grandes caminos:

- Cuantificar la bondad de una estructura comunitaria a partir de un funcional global. Ejemplos de ello son la modularidad [45] y la mínima longitud de descripción [139]. En estos casos, podríamos decir que es el mismo funcional el que de

alguna manera *impone* una definición de estructura comunitaria, como “aquella que maximiza lo que el funcional evalúa”.

- En redes con comunidades conocidas *a priori*, se puede comparar a las dos estructuras comunitarias –la estructura *a priori* y aquella encontrada– a través de alguna métrica de comparación. Aquí es posible:
 - Utilizar redes reales. Son pocas las redes reales en que se conoce la estructura comunitaria; algunos ejemplos de ellas son: la red de karate, la red de delfines y la red de football.
 - Utilizar como bancos de prueba grafos aleatorios con estructura comunitaria. Destacamos el banco de prueba de Girvan-Newman, que es un caso particular del modelo *planted l-partition* (ver pág. 68), y el banco de prueba de Fortunato-Lancichinetti-Radicchi (ver pág. 68).

En esta sección discutiremos las siguientes métricas de comparación utilizadas en el problema de detección de comunidades: la *información mutua*, el *índice de Jaccard* y la *fracción de vértices correctamente clasificados*⁶.

Información mutua (*mutual information*) La información mutua se utiliza en la Teoría de la Información para cuantificar la cantidad de información que tienen en común un conjunto de dos o más variables aleatorias. Para utilizarla como medida de comparación entre dos estructuras comunitarias, definiremos dos variables aleatorias, X_1 y X_2 , asociadas a dos particiones $\mathcal{C}_1 = (C_{11}, C_{12}, \dots, C_{1n})$ y $\mathcal{C}_2 = (C_{21}, C_{22}, \dots, C_{2m})$ de un grafo G [54]. Consideremos un proceso aleatorio en el que se toma un vértice al azar de $V(G)$ con distribución uniforme, y se observa el subíndice de su comunidad en la partición $\mathcal{C}_1$, $\sigma_{\mathcal{C}_1}(v)$. Definimos entonces la variable X_1 como el subíndice de la comunidad a la que pertenece el vértice, que podrá tomar valores entre 1 y n . La distribución de probabilidad de X_1 será entonces:

$$\mathbb{P}[X_1 = i] = p_i = \frac{|C_{1i}|}{n(G)} ,$$

⁶El término métrica no se utiliza aquí en su estricta definición matemática. La definición formal de una métrica requiere del cumplimiento de condiciones como la positividad, la simetría y la desigualdad triangular. En este sentido: (i) la información mutua es una métrica si se la normaliza de cierta forma, pero no de la forma en que lo hacemos aquí; (ii) el índice de Jaccard, $JI(x, y)$, genera una métrica si se considera $1 - J(x, y)$. $J(x, y)$, en cambio, sería una medida de similaridad; (iii) la fracción de vértices correctamente clasificados no es una métrica, por no cumplir con la condición de simetría.

con $i = 1, 2, \dots, n$. La *entropía* de la partición $\mathcal{C}_1$ se define como:

$$H(\mathcal{C}_1) = - \sum_{i=1}^n p_i \cdot \log(p_i) .$$

Definimos la variable aleatoria X_2 a partir de la partición $\mathcal{C}_2$ en forma análoga, y definimos la siguiente distribución conjunta de X_1, X_2 :

$$\mathbb{P}[X_1 = i, X_2 = j] = p_{ij} = \frac{|C_{1i} \cap C_{2j}|}{n(G)} ,$$

con $i = 1, 2, \dots, n, j = 1, 2, \dots, m$. Definimos ahora la *entropía conjunta* de $\mathcal{C}_1$ y $\mathcal{C}_2$:

$$H(\mathcal{C}_1, \mathcal{C}_2) = - \sum_{i=1}^n \sum_{j=1}^m p_{ij} \cdot \log(p_{ij}) ,$$

y la *información mutua* como:

$$MI(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - H(\mathcal{C}_1, \mathcal{C}_2) .$$

La *información mutua normalizada* entre $\mathcal{C}_1$ y $\mathcal{C}_2$ se define de la siguiente manera: [54]

$$\begin{aligned} NMI(\mathcal{C}_1, \mathcal{C}_2) &= \frac{2MI(\mathcal{C}_1, \mathcal{C}_2)}{H(\mathcal{C}_1) + H(\mathcal{C}_2)} = \\ &= -2 \cdot \frac{\sum_{i=1}^n \sum_{j=1}^m p_{ij} \cdot \log\left(\frac{p_{ij}}{p_i \cdot p_j}\right)}{\sum_{i=1}^n p_i \cdot \log(p_i) + \sum_{j=1}^m p_j \cdot \log(p_j)} . \end{aligned} \quad (3.3)$$

La información mutua normalizada se encuentra entre 0 y 1, y da una idea de la similaridad entre dos particiones en términos de la información sobre una de ellas que subyace en la otra. Sólo toma el valor 1 cuando las dos particiones son coincidentes⁷.

Índice de Jaccard (Jaccard index) El *índice de Jaccard* calcula la relación entre los pares de vértices que están asignados a una misma comunidad en ambas particiones $\mathcal{C}_1$ y $\mathcal{C}_2$, respecto a la cantidad de pares de vértices (v, w) que, en alguna o ambas particiones, están asignados a una misma comunidad. Definimos:

- a_{11} : Cantidad de pares (v, w) asignados a la misma comunidad en $\mathcal{C}_1$ y en $\mathcal{C}_2$.
- a_{01} : Cantidad de pares (v, w) asignados a la misma comunidad sólo en $\mathcal{C}_2$.
- a_{10} : Cantidad de pares (v, w) asignados a la misma comunidad sólo en $\mathcal{C}_1$.

⁷Para un desarrollo más completo de las propiedades de la entropía y la información mutua en el contexto de la Teoría de la Información sugerimos consultar el libro de Cover y Thomas [52].

- a_{00} : Cantidad de pares (v, w) asignados a distintas comunidades en $\mathcal{C}_1$ y en $\mathcal{C}_2$.

El índice se define como:

$$JI(\mathcal{C}_1, \mathcal{C}_2) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} .$$

Observando que a_{11} , a_{01} y a_{10} se pueden calcular de la siguiente manera:

$$\begin{aligned} a_{11} &= \sum_{C_{1i} \in \mathcal{C}_1} \sum_{C_{2j} \in \mathcal{C}_2} \frac{|C_{1i} \cap C_{2j}|(|C_{1i} \cap C_{2j}| + 1)}{2} , \\ a_{10} &= \sum_{C_{1i} \in \mathcal{C}_1} \frac{|C_{1i}|(|C_{1i}| + 1)}{2} - a_{11} , \\ a_{01} &= \sum_{C_{2j} \in \mathcal{C}_2} \frac{|C_{2j}|(|C_{2j}| + 1)}{2} - a_{11} , \end{aligned}$$

obtenemos la siguiente fórmula para el cálculo del índice:

$$JI(\mathcal{C}_1, \mathcal{C}_2) = \frac{\sum_{C_{1i} \in \mathcal{C}_1} \sum_{C_{2j} \in \mathcal{C}_2} \frac{|C_{1i} \cap C_{2j}|(|C_{1i} \cap C_{2j}| + 1)}{2}}{\sum_{C_{1i} \in \mathcal{C}_1} \frac{|C_{1i}|(|C_{1i}| + 1)}{2} + \sum_{C_{2j} \in \mathcal{C}_2} \frac{|C_{2j}|(|C_{2j}| + 1)}{2} - \sum_{C_{1i} \in \mathcal{C}_1} \sum_{C_{2j} \in \mathcal{C}_2} \frac{|C_{1i} \cap C_{2j}|(|C_{1i} \cap C_{2j}| + 1)}{2}} .$$

Fracción de vértices correctamente clasificados Esta métrica fue propuesta por Newman [113] y requiere definir una función f cuyo dominio es una partición en comunidades *a-priori*, $\mathcal{C}_{ap} = (C_{a1}, C_{a2}, \dots, C_{an})$ y su codominio es la partición descubierta por determinado método M , $\mathcal{C}_M = (C_{M1}, C_{M2}, \dots, C_{Mm})$. Para cada comunidad *a priori* C_{ai} se asigna aquella C_{Mj} con la que comparte la mayor cantidad de vértices⁸:

$$f(C_{ai}) = \arg \max_{C_{Mj} \in \mathcal{C}_M} \{C_{ai} \cap C_{Mj}\} .$$

f no es necesariamente una biyección, puesto que varias comunidades *a-priori* pueden tener asignada la misma comunidad en la partición encontrada. Los vértices minoritarios de C_{ai} —es decir, aquellos que no pertenecen a $f(C_{ai})$ — se considerarán incorrectamente clasificados. Los vértices en $C_{ai} \cap f(C_{ai})$ se considerarán correctamente clasificados sólo cuando no exista otra comunidad *a priori* con la misma comunidad asignada. Introducimos entonces $g(C_{ai})$ como la cantidad de vértices de la intersección si $f(C_{ai})$ tiene una

⁸En [113] Newman no aclara que sucede si existen varias comunidades que comparten una cantidad máxima de vértices. Para resolver la dificultad consideraremos que se escoge una al azar, por lo que la métrica no será determinística. El *survey* de Fortunato sostiene en cambio que la imagen de C_{ai} debe contener a la mayoría de sus vértices, o de lo contrario no se cuenta dentro de la fracción de vértices correctamente clasificados ([70], pág.74).

única preimagen, y 0 en caso contrario:

$$g(C_{ai}) = |C_{ai} \cap f(C_{ai})| \cdot \mathbf{1}\{\forall C \neq C_{ai} \in \mathcal{C}_{ap} : f(C) \neq f(C_{ai})\}$$

Se define la *fracción de vértices de $\mathcal{C}_{ap}$ correctamente clasificados en $\mathcal{C}_M$* como:

$$FCCV(\mathcal{C}_{ap}|\mathcal{C}_M) = \sum_{C_{ai} \in \mathcal{C}_{ap}} \frac{g(C_{ai})}{n(G)} .$$

Este coeficiente no debería aplicarse entre particiones obtenidas con distintos métodos, porque supone que una de las dos particiones es la estructura comunitaria verdadera. Por este motivo es asimétrico: $FCCV(\mathcal{C}_{ap}|\mathcal{C}_M) \neq FCCV(\mathcal{C}_M|\mathcal{C}_{ap})$.

3.4. Análisis del funcional Q (modularidad)

A partir de su expresión original en la Ecuación (3.1) la modularidad ha tenido varias interpretaciones. A continuación presentamos dos de ellas, seguidas por nuestra interpretación como medida signada, a partir de la cual se deducirán fácilmente muchas de sus propiedades.

Interpretación como problema de asignación cuadrática Smith y White (2005) [147] reformularon el problema de maximizar la modularidad como un problema de asignación cuadrática. Dada una partición $\mathcal{C}$ y definiendo por cada comunidad $C \in \mathcal{C}$ un vector $\mathbf{x}_C$ de N elementos que contiene un 1 en la i -ésima posición cuando el vértice v_i está asignado a la comunidad C , y 0 en caso contrario, la modularidad se puede reescribir como:

$$Q_G(\mathcal{C}) = - \sum_{C \in \mathcal{C}} \mathbf{x}_C^T L_Q \mathbf{x}_C ,$$

en donde la matriz L_Q tiene las siguientes componentes:

$$l_{ij} = \frac{d^2(v_i)}{4e^2(G)} - \frac{A_{ij}}{2e(G)} .$$

Si unimos ahora todos los vectores $\mathbf{x}_C$ en una matriz de asignación X cuyas componentes x_{ic} representan la asignación de la comunidad C_c al vértice i , entonces llegamos a la siguiente expresión:

$$Q_G(\mathcal{C}) = -\text{Tr}(X^T L_Q X) .$$

La maximización de la modularidad se reduce entonces al problema de minimizar la traza de $X^T L_Q X$ sujeto a la restricción de que X sea una matriz de asignación, esto es,

que $X^T X$ sea una matriz diagonal a valores discretos $\{0, 1\}$, y con traza igual a $n(G)$.

Esta escritura como problema de asignación cuadrática abre las puertas a los métodos de descomposición espectral, que permiten optimizar la modularidad construyendo una partición a partir de las componentes de los vectores propios principales de la matriz L_Q . Como la descomposición espectral arroja vectores propios con componentes reales y no discretas, es muy habitual optimizar el problema continuo y luego aplicar algún algoritmo de agrupamiento de datos como el k -means. La Figura 3.1 ilustra este enfoque con la red de *football*.

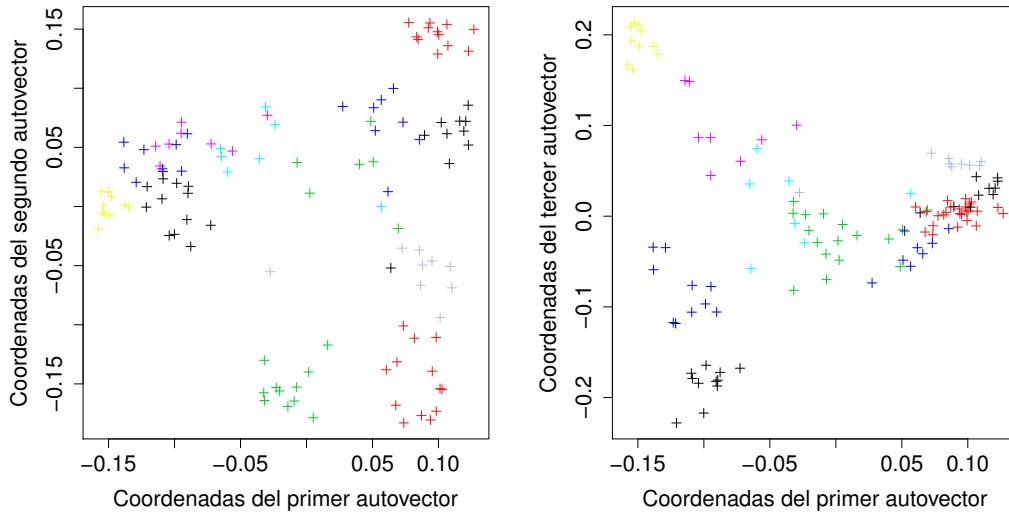


Figura 3.1: *Métodos espectrales de descubrimiento de comunidades. Red de football.* Aplicamos la descomposición espectral de la matriz L_Q para la red de *football*. Las coordenadas de los vértices de la red están determinadas por las componentes de los vectores propios principales de la matriz. En las figuras consideramos sólo aquellos asociados a los 3 valores propios de mayor módulo. Los colores de los vértices indican las comunidades establecidas *a priori* en la red.

En 2006 Newman propuso una aproximación similar para el caso de bisecciones (particiones con dos comunidades) utilizando el *laplaciano* del grafo [117].

Interpretación como modelo de Potts sobre un vidrio de espín [133]. Reichardt y Bornholdt mostraron que la modularidad es proporcional al hamiltoniano de un modelo de Potts en el que los valores de espín $\sigma(v_i)$ de los vértices representan los subíndices de sus comunidades de pertenencia en una partición $\mathcal{C} = (C_1, C_2, \dots, C_n)$:

$$\mathcal{H}_\gamma(\{\sigma(v_i)\}) = - \sum_{i,j} J_{ij} \mathbf{1}\{\sigma(v_i) = \sigma(v_j)\} ,$$

en donde: el término de la izquierda indica que el valor del hamiltoniano $\mathcal{H}$ es función del conjunto de todos los espines; la matriz J representa el acoplamiento entre los vértices y

se define como $J_{ij} = A_{ij} - \gamma \frac{d(v_i)d(v_j)}{2e(G)}$; $\mathbf{1}\{\sigma(v_i) = \sigma(v_j)\}$ toma el valor 1 cuando i y j tienen el mismo espín y 0 cuando su espín es distinto; y γ está vinculada con la temperatura del modelo de Potts. Puesta en estos términos, la modularidad se expresa como:

$$Q_G(\mathcal{C}) = -\frac{\mathcal{H}_1(\{\sigma(v_i)\})}{2e(G)} .$$

La partición que maximiza la modularidad se corresponde entonces con el estado de mínima energía (*ground state*) del vidrio de espín. En dicho estado, cada comunidad se ve reflejada en el conjunto de vértices que tienen un mismo espín. El control sobre la temperatura a partir del parámetro γ permitiría generar estructuras comunitarias a distintos niveles de resolución. Sin embargo, se ha mostrado que el ajuste de γ no subsana la existencia de un límite de escala en la modularidad [95].

Interpretación como medida signada. Nuestra interpretación de la modularidad como medida signada surge de la definición de dos *medidas*, m_E y m_V . La primera de ellas es una medida en el conjunto $V(G) \times V(G)$, mientras que la segunda lo es en $V(G)$. Definimos m_E a partir de su valor para cada par $(u, v) \in V(G) \times V(G)$ y de la aditividad, y m_V a partir de su valor para cada $v \in V(G)$ y de la aditividad, de la siguiente manera:

$$m_E(u, v) = \frac{\mathbf{1}\{u \rightarrow v\}}{2e(G)} \quad (3.4)$$

$$m_V(v) = \frac{d(v)}{2e(G)} . \quad (3.5)$$

Finalmente, a partir de m_V definimos la medida producto m_{VV} como

$$m_{VV}(u, v) = m_V(u)m_V(v) .$$

De estas definiciones se deduce en particular que $m_{VV}(C \times C) = \frac{d^2(C)}{4e^2(G)}$ y que $m_E(C \times C) = \frac{d^{in}(C)}{2e(G)}$ para $C \subset V(G)$. Para simplificar la notación, las llamaremos $m_V^2(C)$ y $m_E(C)$. Estas igualdades se resumen en el Cuadro 3.3.

A partir de estas dos medidas y de la observación de la Ecuación (3.1) la modularidad puede ser reescrita como

$$Q_G(\mathcal{C}) = \sum_{C_i \in \mathcal{C}} m_E(C_i) - m_V^2(C_i) .$$

Entonces, si definimos $D(\mathcal{C}) = \cup_i C_i \times C_i$, y aplicando las propiedades de toda medida,

la modularidad es simplemente

$$Q_G(\mathcal{C}) = \tilde{m}(D(\mathcal{C})) = m_E(D(\mathcal{C})) - m_{VV}(D(\mathcal{C})) , \quad (3.6)$$

de donde se observa que $Q_G(\mathcal{C})$ es una *medida signada* –por ser la diferencia de dos medidas–.

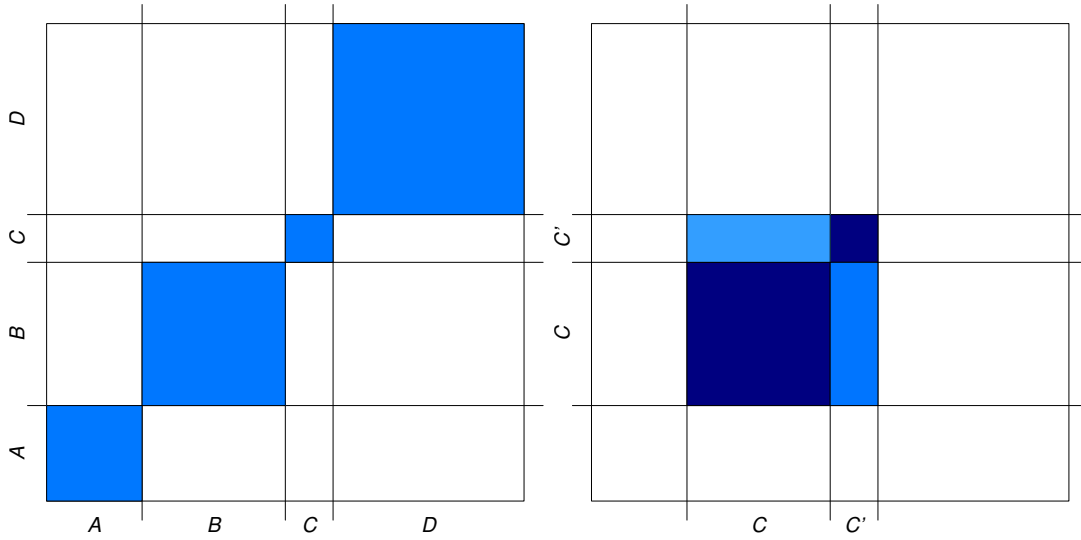


Figura 3.2: *Interpretación de la modularidad como medida signada.* Consideremos una partición $\mathcal{C} = (A, B, C, D)$. Graficamos el espacio $V(G) \times V(G)$ en una cuadrícula, ubicando en forma contigua a los vértices que pertenecen a una misma comunidad de $\mathcal{C}$, y asignando a cada vértice v_i una longitud de $m_V(v_i)$. Como m_V es una medida unitaria, obtenemos la cuadrícula $[0, 1] \times [0, 1]$. A la izquierda vemos la definición de la región $D(\mathcal{C}) = \sum_{C \in \mathcal{C}} C \times C$. A la derecha observamos que al unir dos comunidades C y C' se obtiene una nueva partición $\mathcal{C}'$ y una nueva región $D(\mathcal{C}')$ en que la modularidad registra una variación de $\Delta Q = \tilde{m}(D(\mathcal{C}')) - \tilde{m}(D(\mathcal{C})) = 2\tilde{m}(C \times C')$.

De nuestra interpretación de la Q llegamos fácilmente a los siguientes resultados:

- *Unión de dos comunidades.* A partir de una partición $\mathcal{C}$, la obtención de una partición $\mathcal{C}'$ mediante la unión de dos comunidades C_i y C_j produce una modularidad de:

$$Q(\mathcal{C}') = Q(\mathcal{C}) + 2\tilde{m}(C_i \times C_j) .$$

Observamos entonces que la modularidad aumenta si y sólo si

$$\tilde{m}(C_i \times C_j) = m_E(C_i \times C_j) - m_V(C_i)m_V(C_j) \geq 0 .$$

- *Límite de resolución.* Esta cuestión fue plantada por Fortunato y Barthélemy en 2007, quienes estudiaron la maximización de la modularidad en algunos grafos

sencillos y mostraron que la misma se ve afectada por un *límite de resolución*. Dicho límite implica que las comunidades reconocidas por la modularidad tienen un “nivel de detalle” que depende de características globales del grafo y no solamente de su estructura local. Este fenómeno está vinculado con el hecho de que la modularidad es un funcional global [71]. Los autores lo ejemplifican con algunos casos sencillos, como un anillo de *cliques* o un grafo con dos comunidades pequeñas y una grande, conectadas entre sí (ver Figura 3.3). En el caso del anillo de R *cliques* de orden k , llegan a la siguiente condición de separación de los *cliques*:

$$R < k(k-1) + 2 \quad [71].$$

Poco tiempo después Kumpula *et al.* [95] mostraron que el fenómeno también se

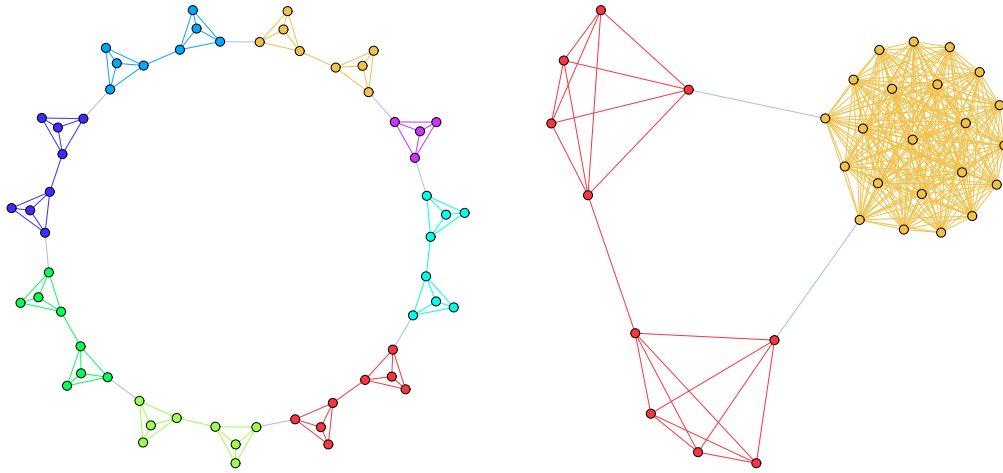


Figura 3.3: *Límite de resolución de la modularidad. Ejemplos.* A la izquierda, un total de R *cliques* de orden k conectados formando un anillo. El grafo tiene una cantidad de aristas $e(G) = Rk(k-1)/2 + R$. Cuando la cantidad de cliques, R , es mayor a $k(k-1) + 2$, la modularidad prefiere unir algunos *cliques* entre sí. La figura ilustra el óptimo de la modularidad para $k = 4$ y $R = 15$. Cada color representa una comunidad en el óptimo. A la derecha, una situación en que dos *cliques* de tamaño p tienen una conexión entre ellos y con un tercer *clique* de tamaño $k > p$. El grafo tiene $n(G) = k + 2p$ vértices y $e(G) = k(k-1)/2 + p(p-1)$ aristas. Si se cumple la condición $k(k-1) > (p(p-1) + 1)^2 + 7$, entonces la modularidad prefiere unir las dos comunidades pequeñas. Por ejemplo, para $p = 5$ la condición se cumple si $k \geq 22$. En la figura se muestra la situación para $k = 22$ y $p = 5$. Los dos resultados generales de las figuras se desprenden fácilmente de nuestra expresión del límite de resolución (Ecuación (3.7)).

produce cuando se utiliza el parámetro de resolución γ de Reichardt y Bornholdt,

y lograron generalizar la condición para el anillo de cliques a una resolución γ :

$$\frac{R}{\gamma} < k(k-1) + 2 \quad [95].$$

A mayor valor de γ la condición es más flexible y permite distinguir *cliques* de menor orden. Por lo tanto un mayor γ se asocia con una mayor resolución, equivalente a una menor temperatura. Lamentablemente este incremento también tendrá el efecto de romper las comunidades más grandes, y por eso no soluciona eficazmente el problema del límite de escala.

Tanto el trabajo de Fortunato y Barthélemy como el de Kumpula analizan casos particulares y no logran una formalización general del problema. El trabajo de Kumpula, particularmente, es impreciso al mencionar que “las comunidades con una cantidad de aristas internas menor a cierto valor no son resueltas” ([95], pág. 1). Veremos luego que esto no es cierto.

Ahora mostraremos que el límite de resolución se puede formalizar y demostrar muy sencillamente dentro de nuestro marco. Supongamos que $\mathcal{C}^*$ es una partición óptima de un grafo G . Entonces, las uniones de sus comunidades nunca mejorarán el valor de la modularidad, es decir que $\forall C_i, C_j, i \neq j$:

$$\tilde{m}(C_i \times C_j) = m_E(C_i \times C_j) - m_V(C_i)m_V(C_j) \leq 0 \quad .$$

Aplicando la aditividad de la medida y una simple desigualdad algebraica:

$$m_V^2(C_i \cup C_j) = (m_V(C_i) + m_V(C_j))^2 \geq 4m_V(C_i)m_V(C_j) \quad ,$$

obtenemos la siguiente condición para que una partición sea óptima:

$$4m_E(C_i \times C_j) \leq m_V^2(C_i \cup C_j) \quad . \quad (3.7)$$

El término de la derecha de la desigualdad es el grado de $C_i \cup C_j$ al cuadrado, normalizado por dos veces el tamaño del grafo, también al cuadrado. Este término decrece mucho más rápido que $m_E(C_i \times C_j)$, y como consecuencia *a medida que el tamaño del grafo aumenta, la modularidad no resiste a mantener las comunidades separadas salvo que las mismas estén desconectadas*. En otras palabras, *para todo par de comunidades conectadas C_i y C_j , haciendo crecer el grafo sin modificar las cercanías de C_i y C_j , en determinado momento la maximización de la modularidad preferirá unir las comunidades pequeñas entre ellas cuando estén conectadas*.

¿Existe entonces un tamaño mínimo de comunidad en el óptimo de la modularidad? Supongamos el caso de dos comunidades C_i y C_j conectadas por al menos una arista. En ellas se cumplirá:

$$m_V^2(C_i \cup C_j) \leq (m_V(C_i) + m_V(C_j))^2 \leq \frac{4 \max(d^2(C_i), d^2(C_j))}{4e^2(G)} \quad (3.8)$$

$$4m_E(C_i \times C_j) \geq \frac{4}{2e(G)} . \quad (3.9)$$

Entonces, las comunidades no podrán resolverse –es decir, serán unidas en la optimización de la modularidad– si se cumple que

$$\frac{4}{2e(G)} > \frac{4 \max(d^2(C_i), d^2(C_j))}{4e^2(G)} ,$$

lo que equivale a

$$\sqrt{2e(G)} > \max(d(C_i), d(C_j)) .$$

Esto implica que si ambas comunidades son suficientemente pequeñas, serán unidas. Sin embargo, una comunidad muy pequeña puede “sobrevivir” y ser resuelta cuando sólo está conectada a comunidades más grandes. Esta cuestión es tratada en forma ambigua en el trabajo de Kumpula, en que se lee que “las comunidades con una cantidad de aristas internas menor a $\frac{e(G)}{2}$ no son visibles” [95](pág 1)⁹.

- *Control de la temperatura.* Nuestra interpretación también abarca la generalización de Reichardt y Bornholdt [133]. Para una resolución γ definimos la modularidad generalizada como:

$$Q_\gamma(\mathcal{C}) = \tilde{m}_\gamma(D(\mathcal{C})) = m_E(D(\mathcal{C})) - \gamma m_{VV}(D(\mathcal{C})) .$$

Se verifica que esta definición es equivalente a la de $-\frac{\mathcal{H}_\gamma(\{\sigma(v_i)\})}{2e(G)}$. En ambas, para $\gamma = 1$ se recupera $Q(\mathcal{C})$. Los resultados anteriores pueden generalizarse de inmediato. En particular, el límite de resolución para un determinado γ se expresa como

$$4m_E(C_i \times C_j) \leq \gamma m_V^2(C_i \cup C_j) . \quad (3.10)$$

Otros resultados de esta interpretación de la modularidad se pueden encontrar en nuestro trabajo en [33]. En particular, proponemos allí un algoritmo goloso que encuentra particiones *débilmente óptimas*.

⁹La cantidad de aristas internas es $\frac{d^{in}(C)}{2}$. Como $d^{in}(C) < d(C)$, nuestra desigualdad implica que $\sqrt{\frac{e(G)}{2}} > \max\left(\frac{d^{in}(C_i)}{2}, \frac{d^{in}(C_j)}{2}\right)$.

3.4.1. Limitaciones

Para concluir mencionamos los siguientes dos resultados que plantearon, junto con el límite de resolución, la necesidad de encontrar métodos superadores del de maximización de la Q :

- En 2008 Brandes *et al.* demostraron que el problema de optimización de la modularidad es NP-completo [31]. Los métodos habituales optimizan el funcional Q empleando heurísticas.
- Más recientemente, en 2010, Good *et al.* [81] resaltaron la denominada *degeneración* de los extremos de la modularidad. Dicha degeneración implica que cerca del óptimo existe una cantidad de picos exponencial en la cantidad de comunidades, correspondientes a particiones cuyos valores de modularidad son muy cercanos al del óptimo. Este resultado cuestiona la significatividad de las particiones que maximizan la Q .

3.5. El método FGP

En esta sección presentaremos nuestro método local de detección de comunidades, denominado *FGP* (*Fitness Growth Process o Proceso de Crecimiento de Función Objetivo*).

Nuestro método es una extensión del propuesto por Lancichinetti *et al.* (2009) [96], en el que se define un proceso a partir de una función objetivo (*fitness function*) f_L con un parámetro α :

$$f_L(C) = \frac{d^{in}(C)}{(d^{in}(C) + d^{out}(C))^\alpha} . \quad (3.11)$$

El proceso comienza con una comunidad C constituida por un único vértice v y realiza las siguientes etapas:

1. Escoge un vértice w cuya inserción produce el mayor incremento en la función objetivo de la comunidad, y lo inserta en ella.
2. Elimina todos aquellos vértices cuya eliminación incrementa la función objetivo de la comunidad.
3. Vuelve al paso 1.

El proceso termina cuando no logra insertar ningún vértice. A la comunidad obtenida al final del proceso se la denomina *comunidad natural del vértice v* . El coeficiente α juega el papel de parámetro de resolución. Para mayores valores de α se obtienen comunidades

naturales más grandes. Para $\alpha = 1$ la función objetivo se vincula con la noción de comunidad en sentido débil de Radicchi [130], introducida en nuestra Sección 3.2.

Una vez terminada la primera comunidad natural, se comienza una nueva con alguno de los vértices que quedaron fuera. Esta nueva comunidad puede abarcar luego vértices de la primera, generando un *solapamiento*. El proceso se repite hasta que no quede ningún vértice sin comunidad asignada en el grafo. El resultado final es un *cubrimiento por comunidades del grafo*.

Nuestra contribución consistirá en definir un *proceso de crecimiento uniforme* que recorre el grafo completo a través de sus distintas comunidades. Haremos una modificación en la función objetivo, manteniendo un parámetro de resolución, y propondremos un algoritmo que aumenta en forma monótona la función objetivo a medida que recorre el grafo, actualizando el parámetro de resolución de manera dinámica. Finalmente, empleando un método de corte, partiremos el recorrido hecho a través del grafo para obtener una partición en comunidades.

3.5.1. Formalización del algoritmo de Lancichinetti *et al.*

Realizaremos a continuación una formalización del procedimiento descrito en Lancichinetti *et al.* [96] para obtener la comunidad natural de un vértice v , generalizado para cualquier función objetivo f . Lo denominaremos *proceso de crecimiento para f* .

El proceso de crecimiento posee una sucesión de etapas de inserción y eliminación de vértices. En la etapa de inserción siempre debe insertarse un vértice –de lo contrario el procedimiento termina–, mientras que en la de eliminación puede que no sea necesario eliminar ninguno. Se producirán entonces secuencias que contienen una inserción y un conjunto de eliminaciones (que puede estar vacío). Denotaremos la evolución de la comunidad a lo largo de estas secuencias con dos subíndices que llamaremos m y k . El primero, m , se incrementará luego de cada par de secuencias inserción–eliminaciones, desde 0 hasta M , y el segundo, k , se incrementará por cada vértice insertado o eliminado dentro de ese par de secuencias, desde 0 hasta k_m . De esta forma, la sucesión de comunidades a lo largo del algoritmo será:

$$(C_{mk}) = (C_{00}, C_{10}, \dots, C_{1k_1}, C_{20}, \dots, C_{2k_2}, \dots, C_{M0}, \dots, C_{Mk_M}) .$$

Obsérvese que:

- Para $m = 0$, sólo existe la comunidad C_{00} que contiene al vértice inicial, que no será eliminado.
- Para cualquier otro par de secuencias inserción–eliminaciones m , la comunidad

C_{m0} será igual a la última comunidad de la secuencia anterior, $C_{(m-1)k_{m-1}}$, porque todavía no se ha realizado la inserción. Luego, C_{m1} será igual a la unión de C_{m0} con el vértice insertado. Los C_{mk} restantes para $2 \leq k \leq k_m$ se obtendrán cada uno de ellos a partir de la eliminación de un vértice de la comunidad anterior, $C_{m(k-1)}$.

- En la última comunidad, C_{Mk_M} ninguna inserción ni eliminación aumenta la función objetivo, y por lo tanto el proceso termina.

La formalización del procedimiento se describe en el Algoritmo 1. En particular, para $f = f_L$ se obtiene el procedimiento descrito en Lancichinetti *et al.* [96] y la última comunidad, C_{Mk_M} , es la *comunidad natural del vértice* v^{10} . En el cuadro 3.4 se muestra un ejemplo.

En el caso de la función objetivo de Lancichinetti, f_L , hacemos la siguiente observación: Si bien la línea 1.4 considera a todos los vértices w que se encuentran fuera de la comunidad C_{m0} , sólo los vértices que pertenecen a la frontera de ella –es decir, que no están en C_{m0} pero poseen alguna conexión con algún vértice en ella– pueden incrementar la función objetivo. No es necesario entonces considerar a los vértices que no pertenecen a la frontera.

La complejidad computacional de este proceso –asumiendo que las eliminaciones son poco frecuentes– es del producto del orden del grafo por el tamaño de la comunidad final: $O(n(G) \cdot |C_{Mk_M}|)$, porque en cada inserción se debe considerar a toda la frontera, cuyo tamaño acotamos con $n(G)$, y la cantidad de inserciones es del orden de C_{Mk_M} , bajo el supuesto de que las eliminaciones son poco frecuentes. En el peor caso, la complejidad computacional de hallar la comunidad natural de un vértice es de $O(n(G)^2)$. Como el procedimiento de Lancichinetti *et al.* debe encontrar un cubrimiento del grafo, la complejidad puede acotarse como $n(G)^2$ multiplicado por la cantidad de comunidades del cubrimiento. Esto hace que la complejidad final sea de $O(n(G)^3)$ en el peor caso (si las comunidades tienen mucho solapamiento), y de $O(n(G)^2)$ en casos de poco solapamiento.

¹⁰Existen diferencias mínimas entre los dos procedimientos, que mencionamos a continuación:

- 1. Lancichinetti *et al.* omiten indicar qué sucede si en algún momento del proceso el mismo vértice semilla v cumple la condición de eliminación, lo cual es factible. En ese caso no resulta razonable eliminar al vértice v y llamar al resultado la *comunidad natural de v* . Consideramos que esto es una omisión, y optamos entonces por restringir las eliminaciones a vértices distintos de v .
- 2. Lancichinetti *et al.* escogen al momento de la inserción a aquel vértice que produce el mayor incremento en la función objetivo; nosotros en cambio tomamos cualquier vértice que la incremente. Sin embargo esta elección de tipo goloso no tiene una fundamentación particular y el mismo trabajo de Lancichinetti *et al.* ([96], pág. 4) señala la posibilidad de explorar otros mecanismos de elección.

Algoritmo 1: Comunidades naturales

Entrada: Un grafo G , una función objetivo f , un vértice inicial $v \in V(G)$
Salida: Un proceso de crecimiento $C_{00}, C_{10}, \dots, C_{a0}, \dots, C_{ak_a}, \dots, \dots, C_{Mk_M}$

1.1 **inicio**
1.2 $D_{00} = \{v\}$
1.3 $m = 0$
1.4 **mientras** *existe algún w fuera de C_{m0} tal que $f(C_{m0} + w) > f(C_{m0})$* **hacer**
1.5 $C_{m1} = C_{m0} + w$
1.6 $k = 1$
1.7 **mientras** *existe algún $w \in C_{mk}, w \neq v : f(C_{mk} - w) > f(C_{mk})$* **hacer**
1.8 $C_{m(k+1)} = C_{mk} - w$
1.9 $k = k + 1$
1.10 **fin**
1.11 $C_{(m+1)0} = C_{mk}$
1.12 $m = m + 1$
1.13 **fin**
1.14 **fin**

3.5.2. Funciones objetivo

El trabajo de Lancichinetti *et al.* sugiere la posibilidad de explorar otras funciones objetivo para la definición de comunidad natural. Trabajaremos con dos familias paramétricas de funciones objetivo, basadas en nuestras definiciones de las medidas m_V y c_E (Cuadro 3.3):

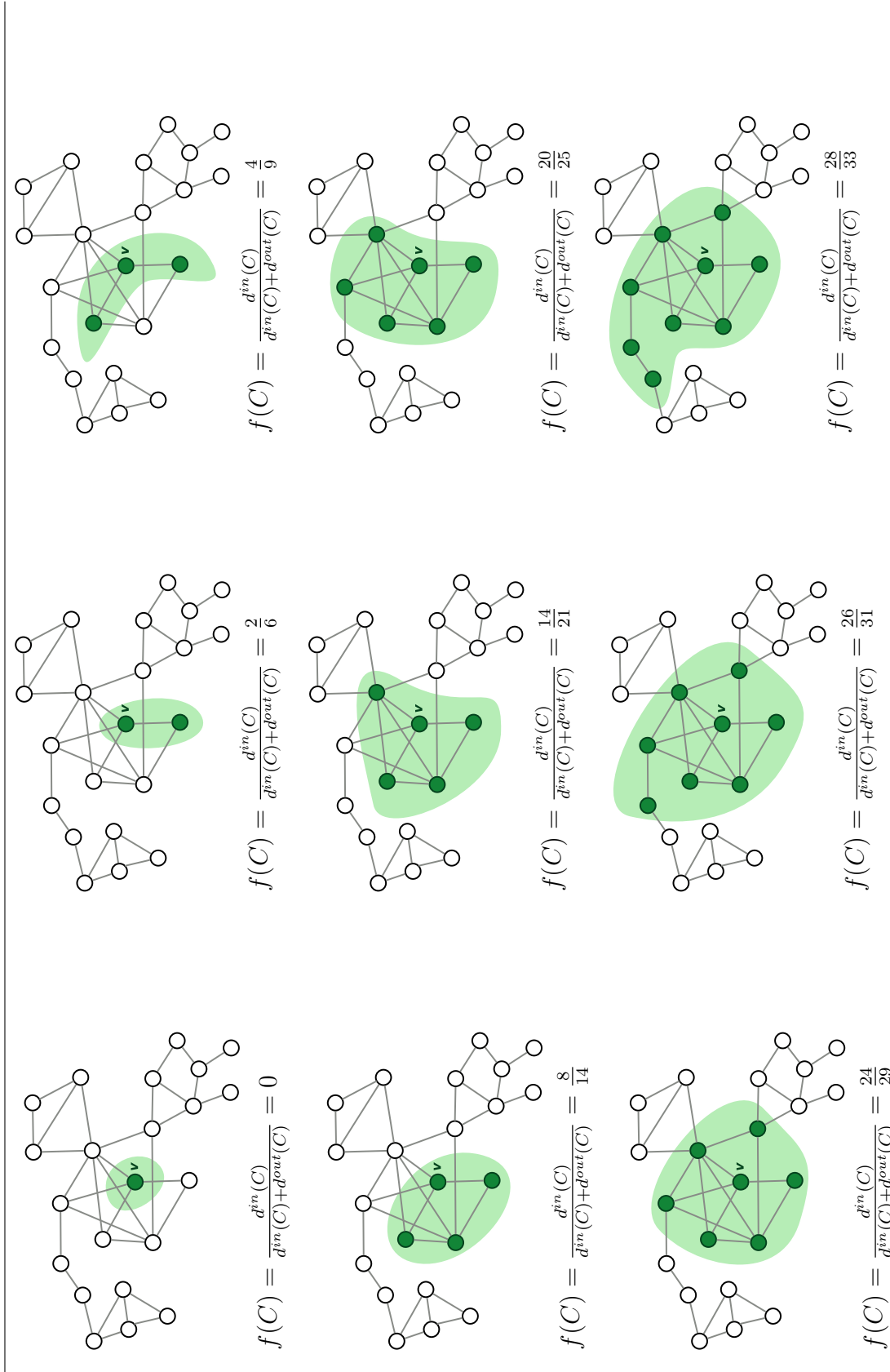
$$L_t = \frac{m_V - c_E}{m_V^{1/t}} \quad (3.12)$$

$$H_t = m_V(1 - m_V/2t) - c_E, \quad (3.13)$$

con $t > 0$. La primera de ellas es proporcional a la función objetivo de Lancichinetti *et al.* para $\alpha = 1/t$. El parámetro t , como veremos luego, tendrá el rol de parámetro de resolución.

Un análisis diferencial. A continuación probaremos que:

- Tanto en L_t como en H_t , la modificación del parámetro de resolución t no afecta esencialmente la evolución del proceso, sino que más bien define el criterio de terminación. Por lo tanto, aquellos vértices que son candidatos a ser insertados o eliminados para algún valor de t seguirán siendo candidatos para una resolución menor –es decir, cuando se llega a comunidades naturales más grandes–.
- Ambas funciones objetivo son esencialmente equivalentes, en el sentido de que



Cuadro 3.4: *Comunidad natural de un vértice para $\alpha = 1$. Construcción de la comunidad natural de un vértice, según el procedimiento descrito en Lancichinetti et al.. En este ejemplo particular no hay necesidad de realizar eliminaciones. La comunidad natural posee 9 vértices y el valor de la función objetivo es $\frac{28}{33}$. La incorporación de cualquier vértice externo disminuiría el valor de la función objetivo.*

quienes son candidatos a ser insertados o eliminados bajo L_t también lo son bajo H_t .

A los efectos de probarlo, consideremos una comunidad C_{mk} y un vértice w . Si $w \notin C_{mk}$ entonces estaremos considerando su inserción, y en caso contrario su eliminación. Obtendremos en ambos casos una nueva comunidad $C_{mk}^+ = C_{mk} \pm w$ ¹¹.

Llamemos ahora $\Delta m_V = m_V(C_{mk}^+) - m_V(C_{mk})$ y $\Delta c_E = c_E(C_{mk}^+) - c_E(C_{mk})$, y supongamos que $s, t > 0$ son dos valores fijos del parámetro de resolución. El cociente diferencial de L_t se puede expresar en forma aproximada como:

$$\frac{\Delta L_t}{\Delta m_V} \approx L'_t = \frac{1}{m_V^{1/t}} \left(1 - \frac{\Delta c_E}{\Delta m_V} - \frac{L_1}{t} \right) .$$

Mientras que para el cociente diferencial de H_t tenemos:

$$\frac{\Delta H_t}{\Delta m_V} \approx H'_t = \left(1 - \frac{\Delta c_E}{\Delta m_V} - \frac{m_V}{t} \right) .$$

Obsérvense entonces las siguientes relaciones:

$$H'_t = H'_s + \frac{t-s}{ts} m_V \quad (3.14)$$

$$m_V^{1/t} L'_t = m_V^{1/s} L'_s + \frac{t-s}{ts} L_1 \quad (3.15)$$

$$H'_t = m_V^{1/t} L'_t + (L_1 - m_V)/t . \quad (3.16)$$

La Ecuación (3.14) nos muestra que si $t > s$ y $H'_s > 0$, entonces $H'_t > 0$, lo cual significa que *si el vértice w es un candidato a ser insertado en C_{mk} bajo la función H_s , entonces también lo es para el proceso H_t .*

La Ecuación (3.15) nos muestra análogamente que si $t > s$ y $L'_s > 0$, entonces $L'_t > 0$, lo que implica que *si el vértice w es un candidato a ser insertado en C_{mk} bajo la función L_s , entonces también lo es para la función L_t .*

Esto demuestra que el parámetro t no juega un rol esencial en la construcción de la comunidad natural bajo ninguna de las dos funciones, sino que simplemente establece el *criterio de terminación*.

La Ecuación (3.16) muestra un hecho delicado: si un vértice w es candidato a ser insertado (eliminado) bajo la función L_t y $m_V < L_1$ entonces también es un candidato a ser insertado (eliminado) bajo la función H_t . La condición $m_V < L_1$ es habitualmente cierta teniendo en cuenta que cuando $m_V > L_1$ se cumple que $c_E > m_V(1 - m_V)$, lo que

¹¹Llamaremos C_{mk}^+ al elemento siguiente a C_{mk} en la secuencia. En un abuso de notación, $C_{mk} + w$ será equivalente a $C_{mk} \cup \{w\}$, mientras que $C_{mk} - w$ será equivalente a $C_{mk} - \{w\}$.

contradice la noción de comunidad (porque el segundo término sería el valor esperado del primero si los vértices que forman la comunidad se hubieran escogido al azar). Por lo tanto, ambos procesos son esencialmente equivalentes, diferenciándose únicamente en el criterio de terminación. Si bien hemos hecho aproximaciones para obtener resultados cuantitativos, nuestra experiencia al utilizar ambas funciones objetivo ha confirmado nuestras conclusiones.

3.5.3. El *fitness growth process* (FGP)

El algoritmo de construcción que describimos anteriormente obtiene comunidades naturales para distintos valores del parámetro t . Hemos visto que a mayor valor de t se obtienen comunidades mayores que en términos generales incluyen a las anteriores, y que por lo tanto t funciona como parámetro de resolución. Nos preguntamos entonces si es posible obtener las comunidades naturales para todos los valores de t con un único proceso. Veremos que esto es realmente posible utilizando nuestra familia de funciones H_t .

A la frontera de C_{mk} , formada por aquellos vértices fuera de C_{mk} pero con alguna conexión hacia ella, la denotaremos $\partial(C_{mk})$.

Consideraremos ahora una comunidad C_{mk} y su frontera $\partial(C_{mk})$, formada por aquellos vértices fuera de C_{mk} pero con alguna conexión hacia ella, y veremos qué sucede si incorporamos a C_{mk} un vértice w que pertenece a su frontera, o bien lo eliminamos si pertenece a C_{mk} ¹². El nuevo valor de la función objetivo en cada caso ($\pm$) será

$$\begin{aligned} H_t(C \pm w) &= (m_V + \Delta m_V)(1 - (m_V + \Delta m_V)/2t) - (c_E + \Delta c_E) \\ &= m_V(1 - m_V/2t) - c_E \\ &\quad - \frac{\Delta m_V}{t}(m_V + \Delta m_V/2) + \Delta m_V - \Delta c_E \\ &= H_t(C) - \frac{\Delta m_V}{t}(m_V + \Delta m_V/2) + \Delta m_V - \Delta c_E . \end{aligned}$$

La variación de la función objetivo es

$$\Delta H_t = -\frac{\Delta m_V}{t}(m_V + \Delta m_V/2) + \Delta m_V - \Delta c_E ,$$

de donde se observa que para algún valor de t suficientemente grande o chico, de acuerdo

¹²No es necesario considerar vértices que no pertenezcan a C_{mk} ni a su frontera, ya que para ellos el ΔH_t es negativo para todo t .

al signo de Δm_V , ΔH_t será positivo. Este valor crítico de t es:

$$t_c(C_{mk}, w) = \frac{\Delta m_V(m_V + \Delta m_V/2)}{\Delta m_V - \Delta c_E} .$$

Se satisface entonces que cuando w se está incorporando, $t > t_c \rightarrow \Delta H_t > 0$, mientras que cuando w se está eliminando $t < t_c \rightarrow \Delta H_t > 0$.

Supongamos entonces que al llegar al criterio de terminación de la comunidad natural a una cierta resolución t , incrementamos el parámetro t lo mínimo posible como para llegar a algún $t' = t_c(C_{mk}, w)$ que permita incorporar a algún w sin disminuir la función objetivo H_{t_c} . El resultado será un proceso de crecimiento uniforme para H_{t_c} . Si continuamos de esta forma hasta abarcar todo el grafo, entonces tendremos una secuencia de comunidades naturales (C_{mk}) a distinta resolución.

Cada comunidad natural C_{mk} tendrá una resolución t_{mk} asociada, que se actualizará cada vez que se produce una inserción, de la siguiente manera:

$$t_{mk}^+ = \text{máx}\{t_{mk}, t_c(C_{mk}, w)\} ,$$

en donde t_{mk}^+ es la resolución asociada a $C_{mk}^+ = C_{mk} \cup \{w\}$. La secuencia (t_{mk}) será entonces una secuencia no decreciente, y cada comunidad de la secuencia $C_{00}, \dots, C_{mk}$ será un proceso de crecimiento para H_t , $\forall t > t_{mk}$. La secuencia de comunidades naturales (C_{mk}) construída de esta forma es un *proceso de crecimiento uniforme* para H .

En el Algoritmo 2 describimos este proceso formalmente.

3.5.4. Extracción de las comunidades

Nuestra hipótesis es que el proceso de crecimiento uniforme recorre una tras otra las comunidades hasta cubrir todo el grafo. En cada paso, el proceso de crecimiento uniforme tiende a escoger el vértice a ser insertado en función de su cohesión con la comunidad construída hasta ese momento. Así, dos vértices insertados en forma consecutiva en el proceso debieran pertenecer a la misma comunidad, o bien ser vértices *de borde* de sus comunidades. Nuestro método de detección de comunidades incluye una técnica para “cortar” las comunidades a lo largo del proceso.

Partimos de una secuencia (C_{mk}) en la cual existen inserciones y eliminaciones. Como al final del proceso todo el grafo está incluído en la comunidad natural, cada vértice aparece una cantidad de veces impar en la secuencia, contando k inserciones y $k - 1$ eliminaciones. El primer paso del procesamiento consiste en dejar sólo la última inserción de cada vértice, considerando que la misma refleja la posición del vértice en la comunidad natural final. De esta forma obtenemos una nueva secuencia, que denominaremos $\mathcal{S}$, en la

Algoritmo 2: Proceso de crecimiento uniforme para H

Entrada: Un grafo G , un vértice inicial $v \in V(G)$ **Salida:** Un proceso de crecimiento uniforme para H : $C_{00}, C_{10}, \dots, C_{a0}, \dots, C_{ak_a}, \dots, C_{M0}, \dots, D_{Mk_M}$ **2.1 inicio****2.2** $C_{00} = \{v\}$ **2.3** $t_a = 0$ **2.4** $m = 0$ **2.5 mientras** *exista algún* $w \in \partial(C_{m0})$ **hacer****2.6** sea w_0 tal que $t_c(C_{m0}, w_0) = \min_{w \in \partial(C_{m0})} (t_c(C_{m0}, w))$ **2.7** $t_a = \max\{t_a, t_c(C_{m0}, w_0)\}$ **2.8** $C_{m1} = C_{m0} + w_0$ **2.9** $k = 1$ **2.10 mientras** *exista algún* $w \in C_{mk}, w \neq v : t_c(C_{mk}, w) > t_a$ **hacer****2.11** $C_{m(k+1)} = C_{mk} - w$ **2.12** $k = k + 1$ **2.13 fin****2.14** $C_{(m+1)0} = C_{mk}$ **2.15** $m = m + 1$ **2.16 fin****2.17 fin**

cual cada vértice aparece una única vez, y que representa por lo tanto un ordenamiento del conjunto $V(G)$.

La transformación de esta secuencia $\mathcal{S}$ en un conjunto de comunidades finales $\mathcal{C} = (C_1, C_2, \dots, C_N)$ se realiza observando el comportamiento de la siguiente función:

$$S(w) = \frac{c_E(C(w))}{m_V(C(w))} , \quad (3.17)$$

en donde los $C(w)$ son las subsecuencias de $\mathcal{S}$, desde el comienzo de la última comunidad que hemos extraído hasta w . El criterio para determinar la clausura de una comunidad C y el comienzo de una nueva es *que se produzca un incremento en la función $S(w)$* .

En otras palabras, la función $S(w)$ considera el conjunto de vértices insertados desde el comienzo de la comunidad que se está construyendo, y va calculando la evolución del cociente entre el grado externo normalizado de la comunidad –representado por c_E y el grado normalizado, m_V . En la siguiente sección ofrecemos una fundamentación estadística de esta técnica de corte.

3.5.5. Fundamentación del método en el límite termodinámico

Para entender el comportamiento estadístico de la función $S(w)$, consideremos una comunidad $C = (v_1, v_2, \dots, v_n)$ en que los vértices tienen un *parámetro de mezclado* $\mu(v_i)$ homogéneo, que dentro de la comunidad llamaremos directamente μ . Esto es, cada vértice comparte una fracción μ de sus aristas con otras comunidades, y una fracción $1 - \mu$ con su propia comunidad C . Llamaremos C_i a las comunidades *parciales* de C obtenidas desde la inserción de v_1 hasta la inserción de v_i . La evolución de $S(v_i)$ estará dada por

$$S_i = S(v_i) = \frac{m_E(C_i \times (V \setminus C_i))}{m_V(C_i)} = 1 - L_1(C_i) .$$

Nuestro análisis estadístico estará basado en las siguientes relaciones:

$$\begin{aligned} m_E(C_i \times (V \setminus C)) &= \mu m_V(C_i) \\ m_E(C_i \times C_i) &= \lambda_i m_E(C_i \times C) . \end{aligned}$$

La primera surge de la hipótesis de que todos los vértices de C tienen un mismo μ . La segunda es simplemente la definición de un parámetro λ_i en el intervalo $[0, 1]$.

A partir de estas ecuaciones se deduce a partir de simples cálculos, y usando la aditividad de la medida m_E , que

$$\begin{aligned} S_i &= \mu + (1 - \mu)(1 - \lambda_i) \\ (1 - \mu)\lambda_i &= L_1(C_i) . \end{aligned}$$

Asumimos que L_1 tiene un comportamiento monótono creciente a lo largo de la construcción de la comunidad¹³, lo que implica un comportamiento monótono decreciente en S_i , incluso sin asumir un μ constante en los vértices. Obsérvese también que para el último vértice de la comunidad, v_n , se cumple que $S = \mu$ (porque $\lambda = 1$).

Ahora, veamos qué sucede cuando la comunidad se concluye y se intenta incorporar al proceso un vértice externo, de una comunidad C' , con un parámetro de mezclado μ' . Llamaremos $C^+ = C \cup \{v\}$, y definiremos ϵ a través de la relación

$$m_E(\{v\} \times C) = \epsilon m_E(\{v\} \times (V \setminus C')) = \epsilon \mu' m_V(\{v\}) ,$$

que representa la proporción de aristas externas desde $v \in C'$ hacia vértices en C .

¹³Recordemos que la función objetivo L_1 está vinculada con el concepto de comunidad débil de Radicchi.

El nuevo valor de la función S es:

$$S^+ = \frac{m_E(C^+ \times (V \setminus C^+))}{m_V(C^+)}$$

y se puede mostrar que

$$S^+ = \mu + \frac{(1 - 2\epsilon\mu' - \mu)m_V(\{v\})}{m_V(C^+)} .$$

Si los parámetros de mezclado no son demasiado elevados –lo que implicaría comunidades muy poco cohesivas– o bien ϵ es pequeño, lo cual es razonable, entonces este nuevo valor S^+ romperá el comportamiento decreciente de S produciendo la clausura de la comunidad C y el comienzo de una nueva comunidad C' con v' como su primer nodo, v'_1 .

Sintetizamos entonces el comportamiento de la función $S(w)$ de la siguiente manera:

- La función inicia en $S(w) = 1$ al incorporar el primer vértice de la comunidad ($w = v_1$).
- La función $S(w)$ decrece desde 1 hasta μ a lo largo de la construcción de la comunidad.
- La función $S(w)$ aumenta cuando la comunidad concluye y el proceso incorpora un vértice externo w' .
- Bajo esta condición, se inicia una nueva comunidad C' y $S(w')$ toma el valor 1.
- Aún si el μ de los vértices de la comunidad no es homogéneo, el mínimo de $S(w)$ alcanzado al terminar la comunidad equivale matemáticamente al promedio de los $\mu(v)$ de todos los vértices de la comunidad, ponderado por su grado, $d(v)$. Lo denominaremos $\mu(C)$.

Ejemplo: La red de *football*. Ilustraremos el procedimiento de corte en la Figura 3.5 mostrando la evolución de la función $S(v)$ a lo largo del proceso de crecimiento para la red de *football*. Se observa claramente el comportamiento decreciente de la función dentro de cada comunidad. La Figura 3.4 visualiza la partición en comunidades encontrada.

3.5.6. Complejidad computacional

En esta sección probaremos que nuestro método de detección de estructura comunitaria tiene una complejidad computacional temporal de $O(n(G) \cdot d_{\max} + e(G) \cdot \log(n(G)))$, y espacial de $O(n(G) + e(G))$.

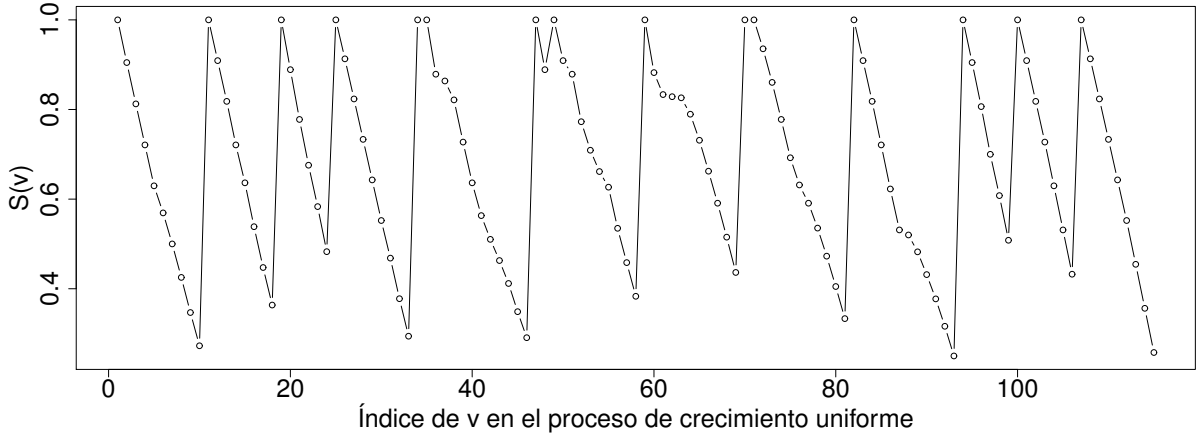


Figura 3.4: El proceso de crecimiento uniforme en la red de football.

Comenzamos analizando la complejidad temporal. Consideremos una comunidad C_{mk} en el proceso, con un parámetro t_{mk} asociado, en la que se está apunto de incorporar un nuevo vértice. La línea 2,6 del algoritmo 2 indica que debemos incorporar al vértice w en la frontera de C_{mk} con el menor $t_c(C_{mk}, w)$. Observamos, a partir de la expresión de t_c , que

$$t_c(C_{mk}, w) = \frac{\Delta m_V}{\Delta m_V - \Delta c_E} \cdot (m_V + \Delta m_V/2) .$$

De entre aquellos vértices de la frontera que tienen el mismo grado que w , el que minimiza el t_c es aquel que tiene el menor $\frac{\Delta m_V}{\Delta m_V - \Delta c_E}$. Si todos los vértices tienen el mismo grado, esto es equivalente a minimizar Δc_E , que a su vez es proporcional al $d_C^{out} - d_C^{in}$ del vértice. Por lo tanto, si agrupamos los vértices de la frontera en listas según su grado, y ordenamos éstas listas por valor creciente de $d_C^{out} - d_C^{in}$, entonces podemos afirmar que el vértice de la frontera que minimiza t_c debe estar a la cabeza de alguna de estas listas. Proponemos entonces mantener a lo largo del proceso una estructura actualizada con la frontera $\partial(C_{mk})$ (véase la Figura 3.6). Necesitaremos también una estructura análoga para los vértices de la comunidad C_{mk} con el fin de acelerar las eliminaciones; esta estructura se muestra en la misma figura. Reducimos así la complejidad de cada paso de tener que analizar toda la frontera o comunidad C_{mk} a analizar $d_{\max}$ vértices como máximo.

Denominaremos $l_{\max}$ a la longitud de la lista más extensa. Las listas se implementarán con una estructura de acceso directo y ordenada, como un mapa o un árbol. Una operación de inserción ordenada dentro de este tipo de estructuras tiene una complejidad de $O(\log(l_{\max}))$, mientras que el acceso tiene complejidad $O(1)$. Estamos listos entonces para analizar la complejidad del paso r :

1. Buscar al vértice w que minimiza $t_c(C_{mk}, w)$ implica encontrar el mínimo entre los

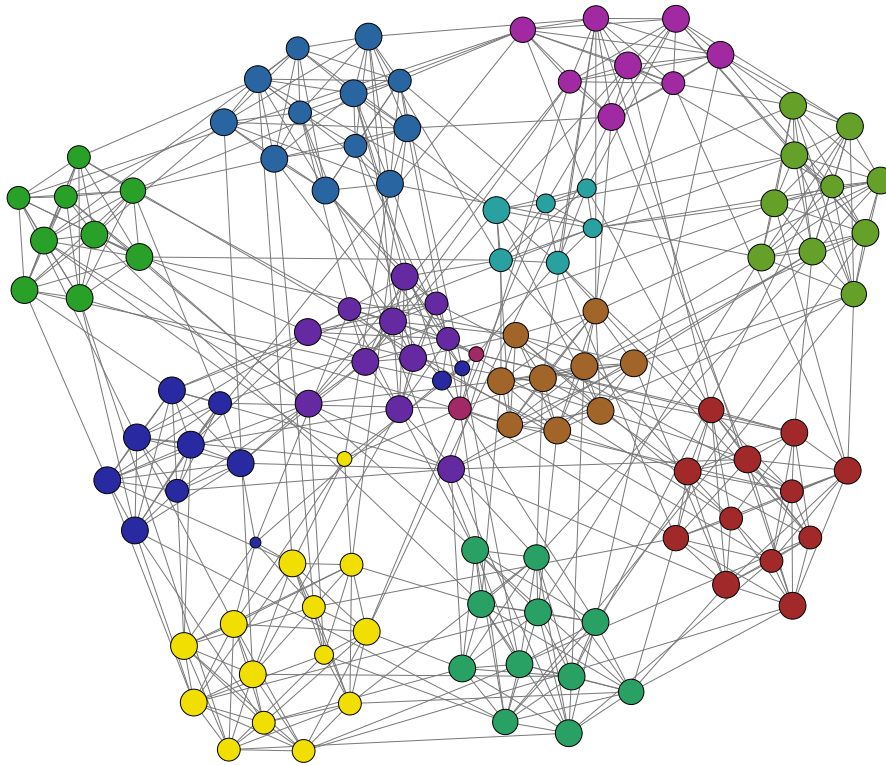


Figura 3.5: Método FGP. Comunidades descubiertas en la red de football. Imagen generada con el software Gephi.

vértices que se encuentran a la cabeza de las listas. Esto tiene una complejidad $O(d_{\text{máx}})$.

2. Actualizar las estructuras implica:

- a) Eliminar w de su lista en la estructura $\partial(C_{mk})$. Complejidad $O(1)$.
- b) Actualizar el Δ_{c_E} de w a $(-\Delta_{c_E})$. Complejidad $O(1)$.
- c) Insertar w en la lista $k(w)$ -ésima de la estructura C_{mk} . Complejidad $O(\log(l_{\text{máx}}))$.
- d) Actualizar Δ_{c_E} para los vecinos de w , es decir, para todo $v \in \mathcal{N}(w)$:
 - 1) Si $v \notin C_{mk}$, actualizar Δ_{c_E} a $\Delta_{c_E} - 2/(2e(G))$. Complejidad $O(1)$.
 - 2) Si $v \in C_{mk}$, actualizar Δ_{c_E} a $\Delta_{c_E} + 2/(2e(G))$. Complejidad $O(1)$.
- e) Reinsertar (o insertar) a los vecinos de w en las listas:
 - 1) Si $v \in C_{mk}$, reinsertarlo en la lista $k(v)$ -ésima de la estructura para C_{mk} , ordenado según su nuevo valor de Δ_{c_E} . Complejidad $O(\log(l_{\text{máx}}))$.

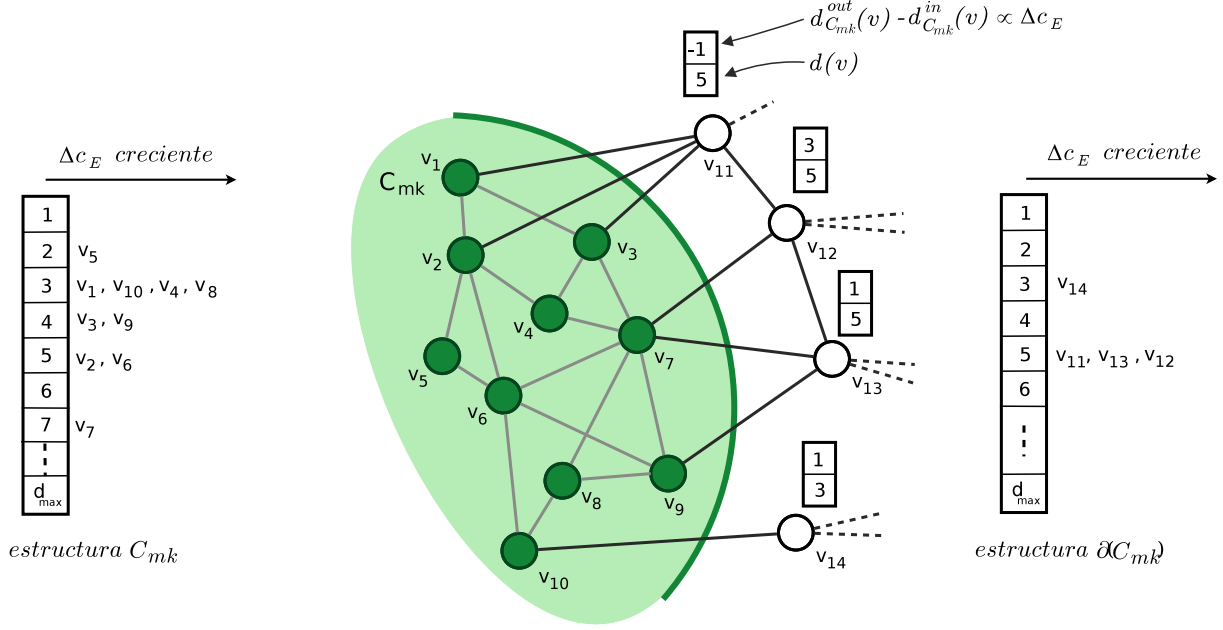


Figura 3.6: *Método FGP. Estructuras mantenidas para optimizar el proceso.* Estructuras mantenidas a lo largo del proceso para la comunidad natural C_{mk} y su frontera $\partial(C_{mk})$. En cada una de ellas los vértices se agrupan por grados (representados por las columnas con valores 1, 2, ..., d_{max}). Los vértices de igual grado se mantienen en una estructura lógica ordenada por $\Delta c_E(v)$ creciente (lo que es equivalente a $d_C^{out}(v) - d_C^{in}(v)$ creciente), como por ejemplo un árbol o un mapa. En la imagen mostramos el $d_C^{out}(v)$ y el grado $d(v)$ recuadrado –sólo para los vértices de la frontera–. En cada paso sólo es necesario considerar a los vértices que se encuentran a la cabeza de las estructuras para cada grado. En este ejemplo, se considera incorporar v_{11} y v_{14} , y se elige v_{11} porque minimiza el t_c . Utilizando estas estructuras, la complejidad del proceso de crecimiento se redujo a $O(n(G) \cdot d_{máx} + e(G) \cdot \log(n(G)))$.

- 2) Si $v \notin C_{mk}, v \notin \partial C_{mk}$, insertarlo en la lista $k(v)$ -ésima de la estructura para ∂C_{mk} , ordenado por su nuevo valor de Δc_E . Complejidad $O(\log(l_{máx}))$.
- 3) Si $v \notin C_{mk}, v \in \partial C_{mk}$, reinsertarlo en la lista $k(v)$ -ésima de la estructura para ∂C_{mk} , ordenado por su nuevo valor de Δc_E . Complejidad $O(\log(l_{máx}))$.

En resumen, la complejidad del paso r es $O(d_{máx} + |\mathcal{N}(w)| \cdot \log(l_{máx}))$.

Ahora bien, los pasos durante el proceso de crecimiento consisten no sólo de inserciones sino también de eliminaciones. La condición de eliminación se sintetiza en la línea 2.10 del Algoritmo 2.

La lógica de las eliminaciones es exactamente la misma: se analizan los vértices a la cabeza de las listas en la estructura para C_{mk} , en busca de un valor de t_c mayor que el actual t_a . Si se lo encuentra, el vértice se elimina de C_{mk} y sus vecinos se actualizan de

manera análoga y con igual complejidad computacional que en el caso de las inserciones.

A lo largo de nuestros experimentos hemos verificado que las eliminaciones son poco frecuentes, y asumiremos que son como máximo del mismo orden que las inserciones. Entonces podemos considerar que el proceso consiste sólo de inserciones a los efectos de calcular la complejidad. Bajo esta hipótesis, cada vértice se inserta una única vez en el proceso, y la complejidad del proceso de crecimiento puede expresarse como:

$$O \left(\sum_{w \in V} (d_{\text{máx}} + \mathcal{N}(w) \cdot \log(l_{\text{máx}})) \right) .$$

Sumar sobre todos los vecinos de $\mathcal{N}(w)$ puede interpretarse como que *cada arista del grafo se considera una única vez*. Con respecto a $l_{\text{máx}}$ no podemos hacer ninguna suposición. Teniendo en cuenta que en distribuciones con leyes de potencias la cantidad de vértices con un grado pequeño puede ser de $O(n(G))$, sólo nos queda acotar $l_{\text{máx}}$ con $n(G)$. Así, podemos expresar la complejidad como

$$O(n(G) \cdot d_{\text{máx}} + e(G) \cdot \log(n(G))) .$$

Mencionamos también que la inicialización del proceso requiere calcular los Δ_{c_E} y Δm_V de todos los vértices con sus grados. Estos cálculos no modifican la complejidad final.

Por otra parte, la técnica de corte que genera la partición en comunidades recorre el conjunto $\mathcal{S}$ una única vez, calculando el $S(w)$ de cada vértice a partir de los valores de c_E y m_V , que ya habían sido calculados durante el proceso de crecimiento. La complejidad de este paso es lineal.

En conclusión, la complejidad está dominada por el proceso de crecimiento, y es de $O(n(G) \cdot d_{\text{máx}} + e(G) \cdot \log(n(G)))$. A través del uso de estructuras de datos adecuadas logramos reducir la complejidad original del proceso, que es de $O(n(G)^2)$. Las mismas estructuras permitirían mejorar también la complejidad del algoritmo de cubrimiento de Lancichinetti *et al.*, que se encuentra originalmente entre $O(n(G)^2)$ y $O(n(G)^3)$, como discutimos en la Sección 3.5.1.

En cuanto a la complejidad espacial, no es más que $O(n(G) + e(G))$, que es la complejidad espacial del almacenamiento del grafo en memoria. Las estructuras de datos sobre la comunidad y su frontera sólo contienen una lista de grados de orden $O(d_{\text{máx}})$, y una cantidad de $d_{\text{máx}}$ listas que guardan información sobre los vértices con cada valor de grado. Para cada vértice se almacena una cantidad de información de $O(1)$, y por lo tanto el conjunto de todas estas listas tiene una extensión de $O(n(G))$. De esta manera, la complejidad espacial de las estructuras no supera a la complejidad espacial del grafo.

3.5.7. Resultados obtenidos

Hemos probado nuestro método de descubrimiento de comunidades en algunas redes reales y en instancias de grafos aleatorios generados con el banco de prueba LFR de Lancichinetti *et al.*. La velocidad de ejecución compite con la de métodos reconocidos, y nos permite aplicarlo a redes de varios millones de aristas. También hemos puesto el código fuente a disposición de la comunidad científica desde <https://code.google.com/p/commugp/>.

A continuación mostraremos los resultados obtenidos y realizaremos comparaciones con los siguientes métodos:

- InfoMAP, de Rosvall y Bergstrom, basado en la minimización de la longitud de la descripción [139].
- Louvain, de Blondel *et al.*, un algoritmo goloso muy eficiente de optimización de la modularidad [24].
- LPM, el método de propagación de etiquetas de Raghavan *et al.* [131].

Las instancias generadas con el modelo LFR tienen entre 1000 y 100000 vértices y parámetros de mezclado que oscilan entre 0,05 y 0,80. Los bancos de prueba BENCH1, BENCH2, BENCH3 y BENCH4 contienen *sets* de 1600 instancias cada uno. Cada *set* de 1600 instancias se subdivide en grupos de 100 en los que el parámetro de mezclado μ avanza de 0,05 a 0,80 en pasos de 0,05. De esta forma permiten seguir el desempeño de los métodos ante estructuras comunitarias de distinta cohesión. Una descripción más completa de los bancos de prueba generados se encuentra en el Cuadro 3.5. Entre las redes reales, hemos analizado la red de actores, la red de bandas de *jazz* y la red Web de *stanford.edu* (véase el Cuadro 3.6).

En el Cuadro 3.7 observamos el desempeño para el banco de prueba BENCH5: un grafo de 100000 vértices con parámetro de mezclado $\mu = 0,25$. Observamos que el tamaño de la partición obtenida (2331 comunidades) es muy cercano al tamaño de la partición *a priori* del banco de prueba. La información mutua entre nuestra partición y la partición *a priori* también refleja esta similaridad. Por otra parte es interesante analizar los valores de la modularidad para las particiones obtenidas a través de diversos métodos. Se manifiesta claramente el fenómeno de la degeneración de los picos de la modularidad observado por Good *et al.* [81]: particiones cualitativamente distintas —obsérvese por ejemplo el tamaño de la partición obtenida por Louvain— presentan valores de modularidad muy cercanos. También en la mínima longitud de descripción hay una diferencia significativa entre Louvain y el resto de los métodos, si la comparamos con una partición trivial. Por último, la distribución de tamaños de las comunidades muestra en forma clara las consecuencias

	BENCH1	BENCH2	BENCH3	BENCH4	BENCH5	BENCH6
Instancias	1600	1600	1600	1600	1	1
Tipo	<i>heterog.</i>	<i>homog.</i>	<i>heterog.</i>	<i>homog.</i>	<i>heterog.</i>	<i>heterog.</i>
α_d (vértices)	2.0	-	2.0	-	2.0	2.0
α_s (comunid.)	3.0	-	3.0	-	2.0	2.0
$n(G)$	1000	1000	5000	5000	100000	100000
$\bar{d}$	10	10	10	10	50	50
$d_{\text{máx}}$	50	50	50	50	1000	1000
$s_{\text{mín}}$	-	-	-	-	10	10
$s_{\text{máx}}$	-	-	-	-	1000	1000
$cc(G)$	-	-	-	-	0.40	-
μ	<i>variable</i> 0,05–0,80	<i>variable</i> 0,05–0,80	<i>variable</i> 0,05–0,80	<i>variable</i> 0,05–0,80	0.25	0.60

Cuadro 3.5: *Listado de bancos de prueba y sus parámetros.*

del límite de resolución. Mientras que los métodos FGP, InfoMAP y LPM logran extraer una estructura comunitaria en que la distribución de tamaños de las comunidades es heterogénea y responde a la partición *a priori*, Louvain se ve “forzado” a obtener una estructura comunitaria con tamaños mucho más homogéneos. Por esta razón, la partición obtenida tiene una cantidad pequeña de comunidades respecto a los demás métodos. El Cuadro 3.8 también confirma estos resultados en BENCH6, cuyo parámetro de mezclado es de $\mu = 0,60$.

La Figura 3.7 muestra a través de una serie de *boxplots* valores estadísticos de los resultados para 4 *sets* de bancos de prueba de 1000 y 5000 vértices. Cada *set* posee 1600 instancias de grafos en las que el parámetro de mezclado varía entre 0,05 y 0,80. Los gráficos reflejan la información mutua entre las particiones obtenidas y las particiones *a priori* en función del parámetro μ . En la Figura 3.8 se comparan los resultados con InfoMAP y Louvain. Se observa que InfoMAP obtiene los mejores resultados. En la misma figura evidenciamos la tendencia de la modularidad a generar particiones pequeñas, tendencia que se acentúa a valores mayores del parámetro μ .

En el Cuadro 3.9 se muestran los resultados para una red real: la red de bandas de jazz, constituida por 198 bandas cuyas conexiones indican que han tenido algún músico en común. Como no disponemos de una partición de referencia para calcular las métricas, hemos comparado los valores de modularidad, longitud mínima de descripción

	<i>football</i>	<i>jazz</i>	<code>stanford.edu</code>	<i>LiveJournal</i>
$n(G)$	115	198	255265	4843953
$e(G)$	613	2742	1941926	42845684
$\bar{d}$	10.66	27.70	15.21	17.69
d_{max}	12	100	38625	20333
$\overline{cc}(G)$	0.403	0.633	0.653	0.351
Referencia	[76]	[78]	[103]	[103]

Cuadro 3.6: *Listado de redes reales y sus parámetros.* Todas las redes se han considerado como grafos no dirigidos.

e información mutua normalizada. Si bien los dos primeros son muy similares entre los distintos métodos –a excepción de Louvain–, la información mutua revela que las particiones son estructuralmente distintas.

Hemos analizado también una porción del grafo de la Web correspondiente al dominio `stanford.edu`. Esta red contiene 281903 páginas web enlazadas por 2312497 hipervínculos¹⁴. El Cuadro 3.10 muestra los resultados.

El caso de la red LiveJournal, de 5 millones de vértices, es particularmente interesante. Por su tamaño y las limitaciones de *hardware* sólo hemos logrado procesarla con los métodos FGP y Louvain. El Cuadro 3.11 nos muestra que en ambos casos las distribuciones de los grados de las comunidades ($d(C)$), siguen leyes de potencias. El fenómeno del límite de escala no se hace manifiesto en Louvain en este caso, porque las comunidades pequeñas no están conectadas entre sí, sino que se encuentran conectadas a las pocas comunidades grandes de la red. Sin embargo las diferencias entre los métodos son notables. FGP detecta 127058 comunidades mientras que Louvain detecta 8491. En FGP, la mayor comunidad tiene 839473 vértices mientras que en Louvain posee 23993. Destacamos que en FGP el ajuste de la ley de potencias, a juzgar por la figura del Cuadro 3.11, es muy bueno.

Para comprobar esta singularidad respecto a la forma en que están conectadas las comunidades de Louvain, hemos tomado las 8 comunidades más grandes de la partición (en términos del grado, $d(C)$), y las más pequeñas –aquellas cuyo grado es menor o igual a 5–, y las hemos visualizado con nuestro software SnailVis [19]. La Figura 3.10 muestra que las comunidades pequeñas no tienen aristas entre ellas.

En conclusión, hemos mostrado que nuestro método FGP, basado en un proceso de crecimiento uniforme, obtiene estructuras comunitarias a partir de un concepto local de

¹⁴Sólo hemos considerado la componente conexa más grande, que contiene al 90 % de las páginas.

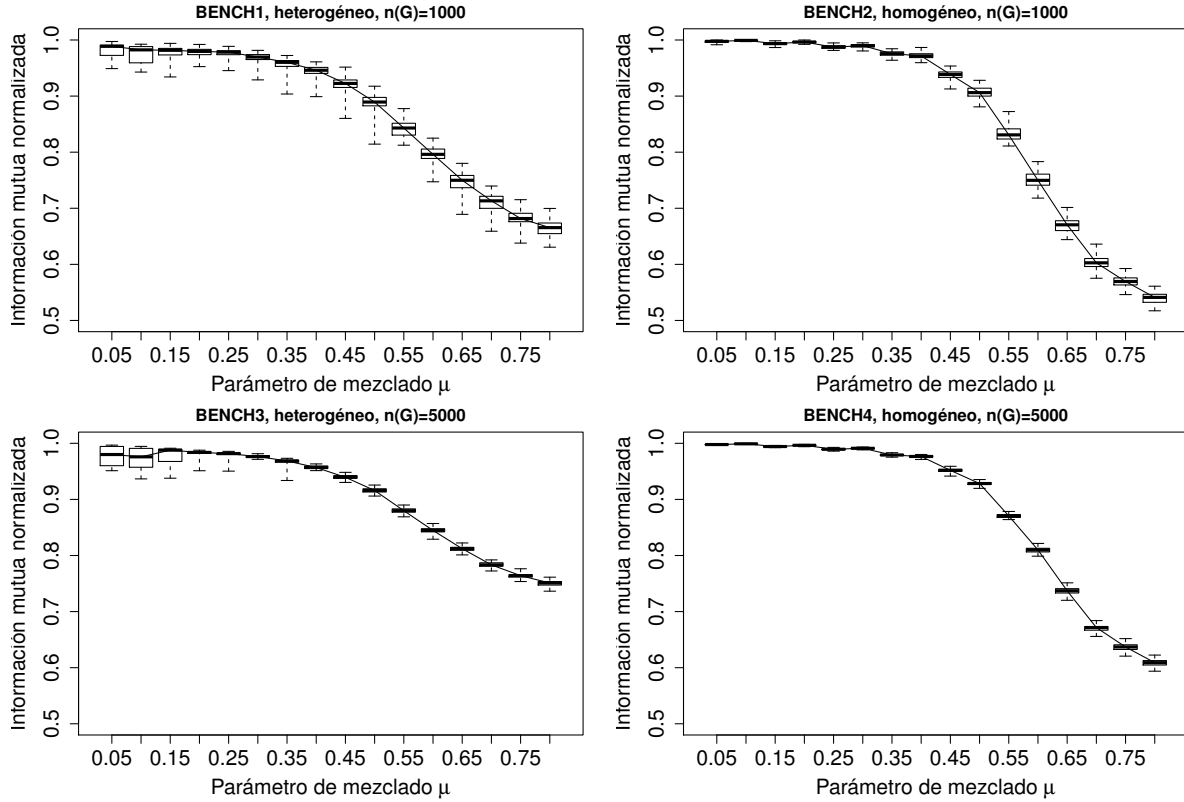
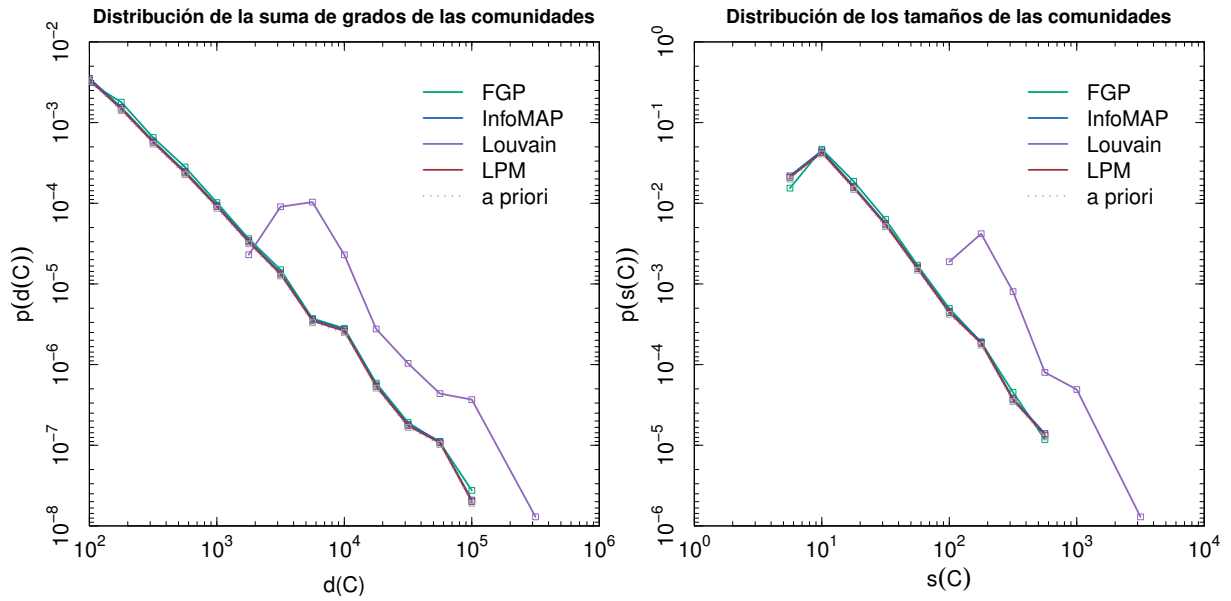


Figura 3.7: *Resultados de los bancos de prueba BENCH1–4 (Parte I).* Comparación de los valores de información mutua normalizada de las particiones obtenidas con el método FGP en los bancos de prueba BENCH1, BENCH2, BENCH3 y BENCH4, en función del parámetro de mezclado μ . Cada caja contiene información estadística de la información mutua para las 100 instancias del *set* correspondientes a cada valor de μ . La raya horizontal dentro de cada caja representa la mediana de las 100 muestras, mientras que los extremos de la caja corresponden al primer y tercer cuartil. El intervalo completo (*whiskers*) abarca desde el mínimo hasta el máximo de la muestra.

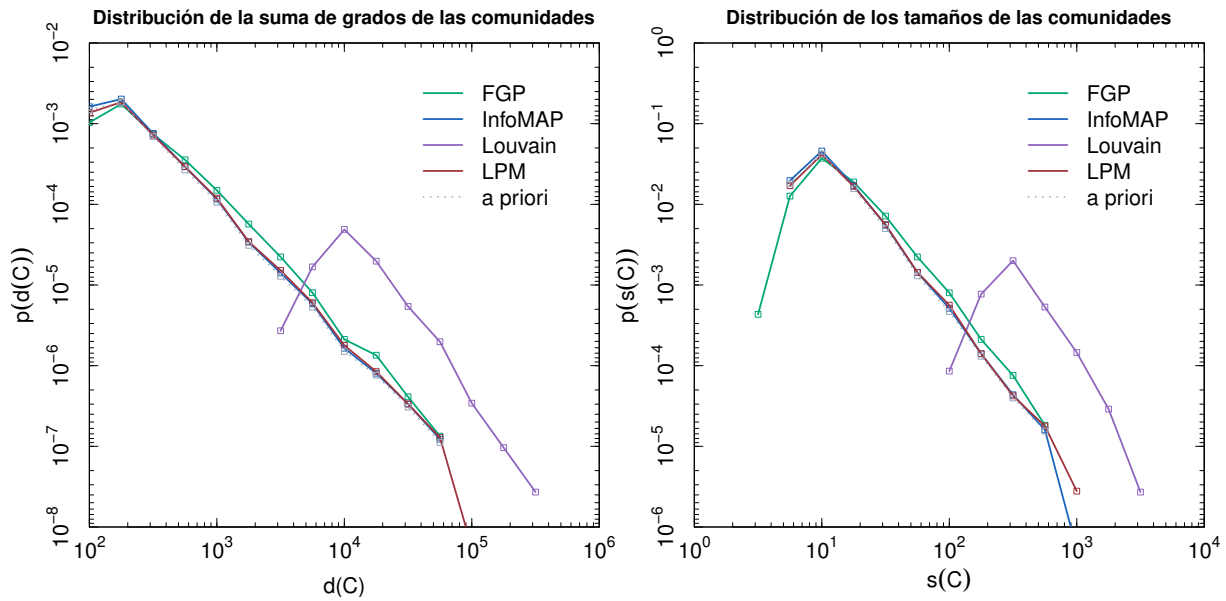
comunidad. Cuando los grados de las comunidades de la red siguen distribuciones de cola larga nuestro método permite detectarlas y no presenta el problema del límite de escala. En los bancos de prueba LFR nuestro método es superado por LPM e InfoMAP, mientras que en redes reales hemos encontrado resultados bastante similares. Creemos que una ventaja interesante de nuestro método es su complejidad acotada. Tanto en LPM como en InfoMAP es difícil hacer un análisis de complejidad. En el primero no está garantizada formalmente la convergencia, mientras que en InfoMAP se requiere de métodos heurísticos para minimizar la longitud de la descripción, y la complejidad computacional depende entonces del criterio de detención.

	FGP	InfoMAP	Louvain	LPM	<i>a priori</i>
$ \mathcal{C} $	2331	2346	314	2336	2346
$MI(\mathcal{C}, \mathcal{C}_{apriori})$	0.977	1.000	0.882	0.999	1.000
$L(\mathcal{C})$ (long.min.desc.)	10.44	10.21	11.15	10.21	10.21
$Q(\mathcal{C})$ (modularidad)	0.708	0.731	0.727	0.731	0.731
$JI(\mathcal{C}, \mathcal{C}_{apriori})$	0.897	1.000	0.354	0.992	1.000
$FCCV(\mathcal{C}, \mathcal{C}_{apriori})$	0.920	1.000	0.000	0.945	1.000
$\overline{\mu(\mathcal{C})}$	0.298	0.252	0.249	0.252	0.252



Cuadro 3.7: Resultados del banco de prueba BENCH5. (Arriba) Comparación de las particiones obtenidas con FGP, InfoMAP, Louvain y LPM para una instancia del banco de prueba LFR con 100000 vértices. La descripción de la red se encuentra en el Cuadro 3.5. A los efectos de interpretar los valores de la mínima longitud de descripción, mencionamos que para una partición trivial con una única comunidad, su valor sería de 12,82. La última fila, $\overline{\mu(\mathcal{C})}$, representa el parámetro de mezclado promedio de las comunidades de la partición. (Abajo) Distribución de los tamaños de las comunidades para las particiones obtenidas con FGP, InfoMAP, Louvain y LPM, y para la partición *a priori*. La distribución se aproximó con un *binning* logarítmico. La semejanza entre las distribuciones para métodos tan diversos como FGP, InfoMAP y LPM es sorprendente.

	FGP	InfoMAP	Louvain	LPM	<i>a priori</i>
$ \mathcal{C} $	1878	2314	150	2104	2315
$MI(\mathcal{C}, \mathcal{C}_{apriori})$	0.914	0.999	0.814	0.989	1.000
$L(\mathcal{C})$ (long.min.desc.)	14.09	13.56	14.37	13.61	13.56
$Q(\mathcal{C})$ (modularidad)	0.343	0.390	0.389	0.391	0.391
$JI(\mathcal{C}, \mathcal{C}_{apriori})$	0.635	0.978	0.189	0.814	1.000
$FCCV(\mathcal{C}, \mathcal{C}_{apriori})$	0.589	0.989	0.000	0.706	1.000
$\overline{\mu(C)}$	0.664	0.601	0.595	0.601	0.601



Cuadro 3.8: *Resultados del banco de prueba BENCH6.* (Arriba) Comparación de las particiones obtenidas con FGP, InfoMAP, Louvain y LPM para una instancia del banco de prueba LFR con 100000 vértices. La descripción de la red se encuentra en el Cuadro 3.5. (Abajo) Distribución de los tamaños de las comunidades para las particiones obtenidas con FGP, InfoMAP, Louvain y LPM, y para la partición *a priori*. La distribución se aproximó con un *binning* logarítmico.

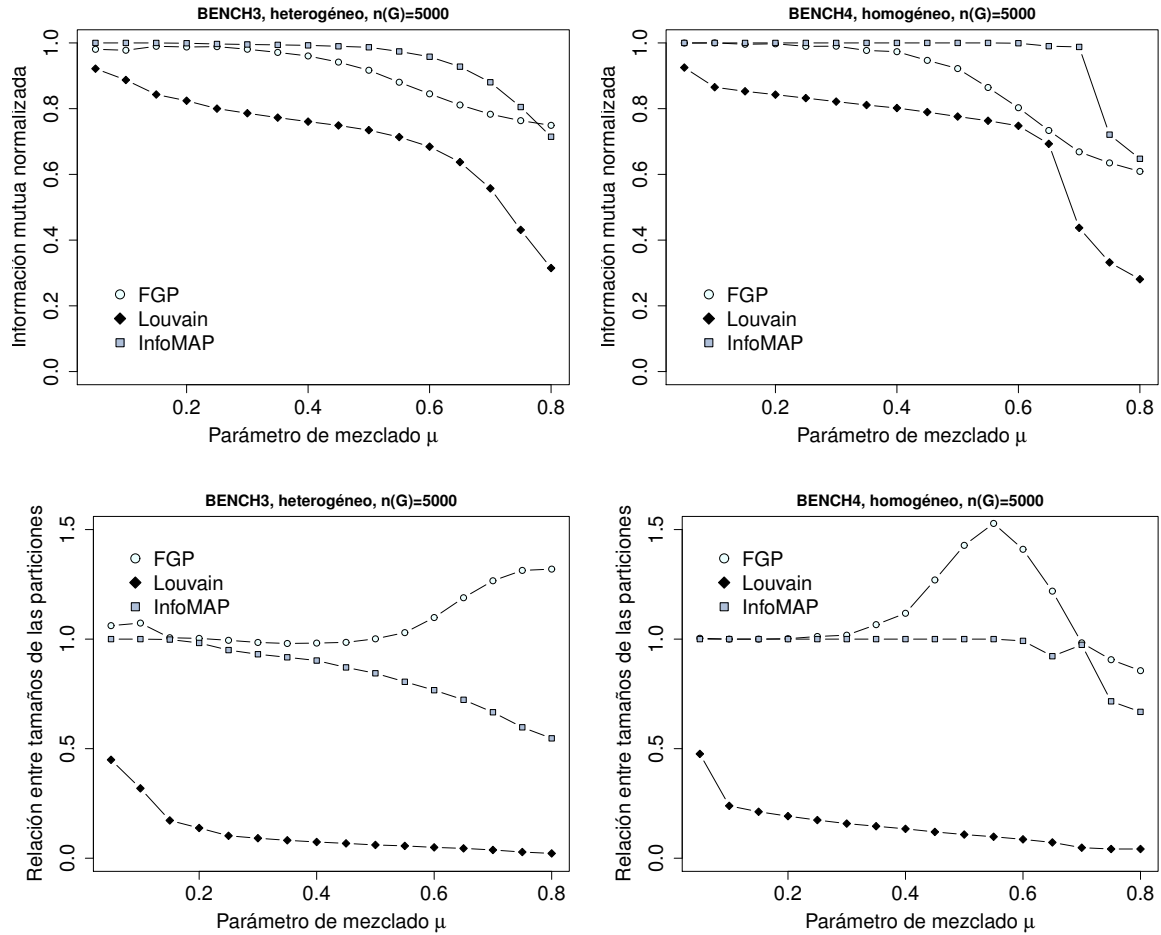
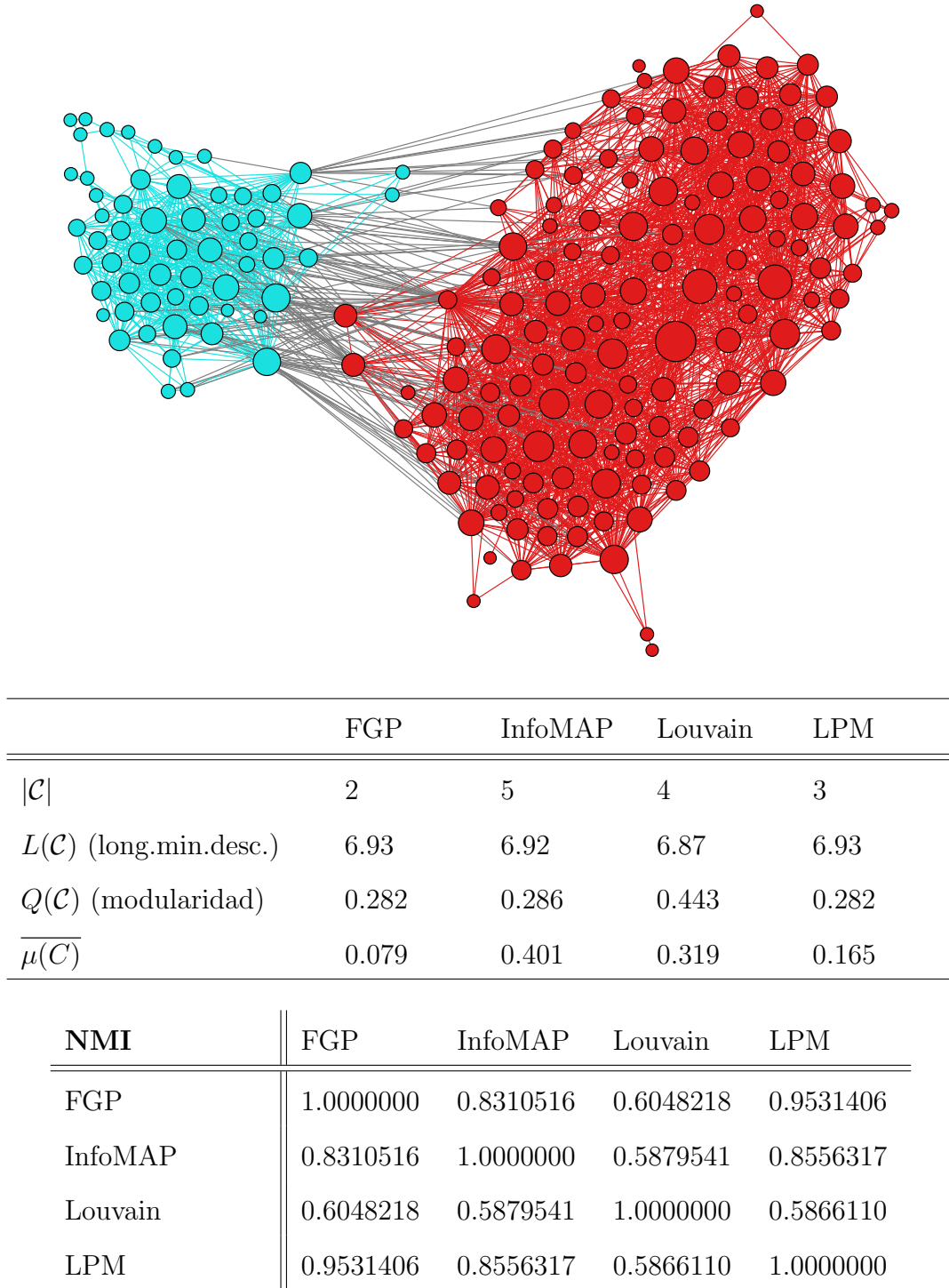


Figura 3.8: Resultados de los bancos de prueba BENCH1-4 (Parte II). (Arriba) Comparación de los valores de información mutua normalizada de las particiones obtenidas con los métodos FGP, Louvain e InfoMAP en los bancos de prueba BENCH1, BENCH2, BENCH3 y BENCH4, en función del parámetro de mezclado μ . Cada punto representa la mediana de la información mutua para las 100 instancias del *set* correspondientes a cada valor de μ . La información mutua normalizada siempre es calculada tomando la partición *a priori* generada por el banco de prueba como referencia. (Abajo) Una estadística similar para la relación entre los tamaños de las particiones, tomando a la partición *a priori* como referencia.



Cuadro 3.9: *Resultados obtenidos para la red de bandas de jazz. (Arriba)* Visualización de la partición obtenida con el método FGP. La visualización fue realizada en Gephi y el posicionamiento de los vértices se hizo a través de un método dirigido por fuerzas. Los colores de los vértices representan la comunidad asignada y los tamaños son proporcionales a sus grados. *(Centro)* Caracterización de las particiones obtenidas por los distintos métodos. *(Abajo)* Matriz de comparación de la información mutua normalizada entre las particiones.

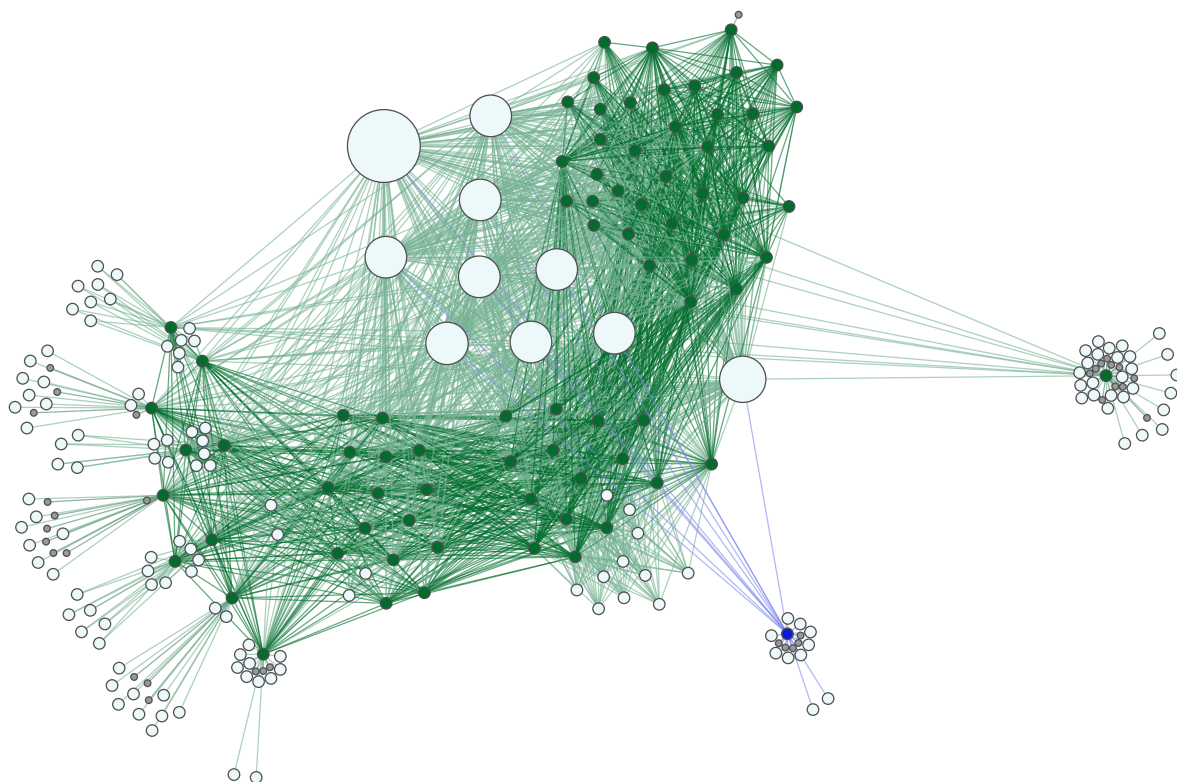
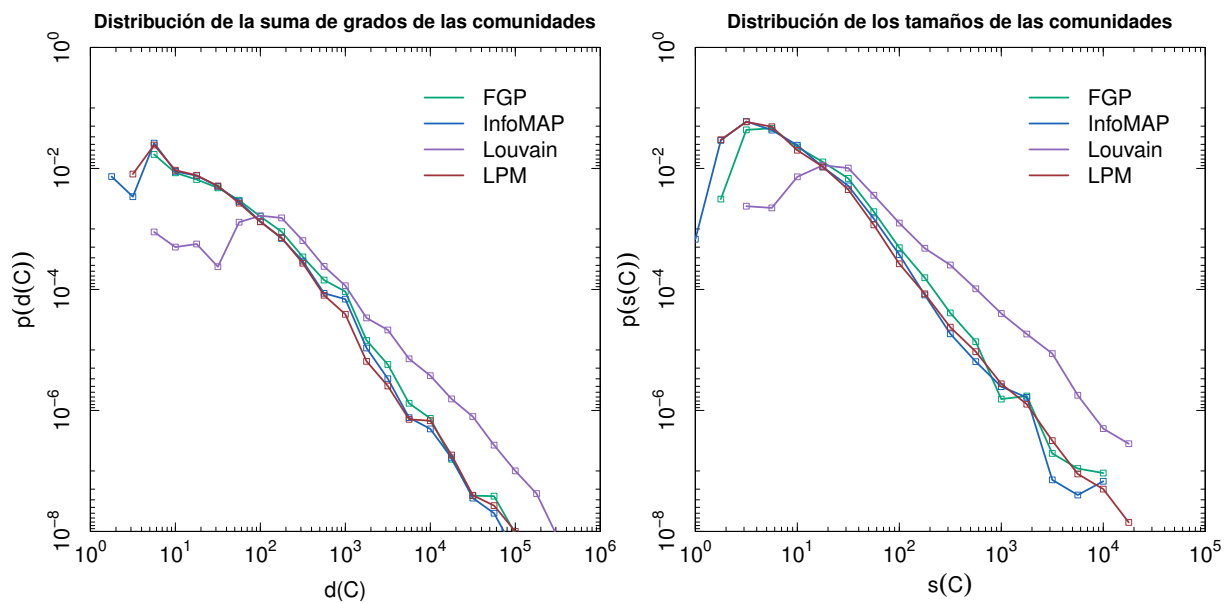


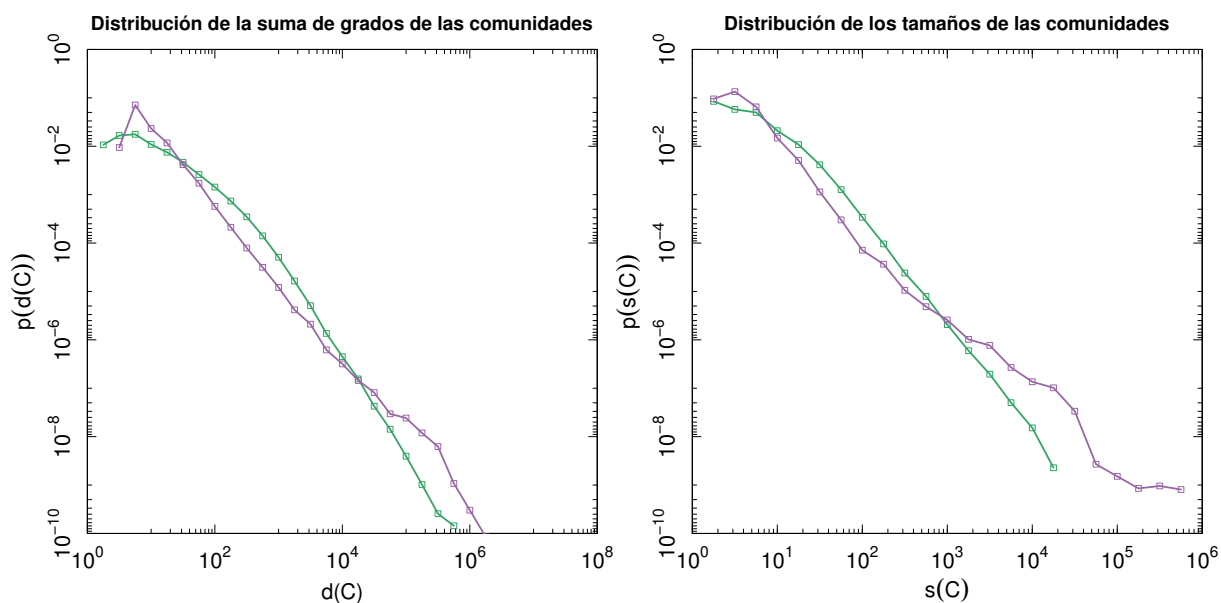
Figura 3.9: *Método FGP. Una comunidad del grafo Web de **stanford.edu**.* En la figura se muestran en verde los vértices que pertenecen a la comunidad (a excepción del vértice azul, que también pertenece a la comunidad), y en blanco o gris sus vecinos (es decir, vértices a distancia 1 de la comunidad) de otras comunidades. Sólo se muestran las aristas internas a la comunidad (verde oscuro) y aquellas que conectan a la comunidad con los vértices vecinos (verde claro), pero no se muestran las posible aristas de los vecinos entre sí. El vértice en azul es el primer vértice de la comunidad encontrado por el proceso. Obsérvese que se trata de un vértice de borde de la misma. El tamaño de los vértices en la imagen es proporcional a su grado. Mientras que los vértices de la comunidad tienen un grado medio de 40 con un desvío de 10, los vértices vecinos de gran tamaño que se observan están entre los 15 de mayor grado de todo el grafo, con grados que oscilan entre 20000 y 40000. La imagen fue generada con el software Gephi y los vértices han sido posicionados con un algoritmo dirigido por fuerzas.

	FGP	InfoMAP	Louvain	LPM
$ \mathcal{C} $	4173	5454	513	4678
$L(\mathcal{C})$ (long.min.desc.)	10.13	9.15	10.47	9.66
$Q(\mathcal{C})$ (modularidad)	0.769	0.846	0.920	0.861
$\overline{\mu(\mathcal{C})}$	0.201	0.198	0.010	0.151



Cuadro 3.10: Resultados obtenidos para la porción del grafo Web de `stanford.edu`. (Arriba) Comparación de las particiones obtenidas con FGP, InfoMAP, Louvain y LPM. (Abajo) Distribución de los tamaños de las comunidades para las particiones obtenidas con FGP, InfoMAP, Louvain y LPM. La distribución se aproximó con un *binning* logarítmico.

	FGP	Louvain
$ \mathcal{C} $	127058	8491
$L(\mathcal{C})$ (long.min.desc.)	18.05	17.66
$Q(\mathcal{C})$ (modularidad)	0.304	0.727
$\overline{\mu(C)}$	0.551	0.126



Cuadro 3.11: Resultados obtenidos para el grafo de la red social LiveJournal. (Arriba) Comparación de las particiones obtenidas con FGP y Louvain. (Abajo) Distribución de los tamaños de las comunidades para las particiones obtenidas con FGP (verde) y Louvain (violeta). La distribución se aproximó con un *binning* logarítmico.

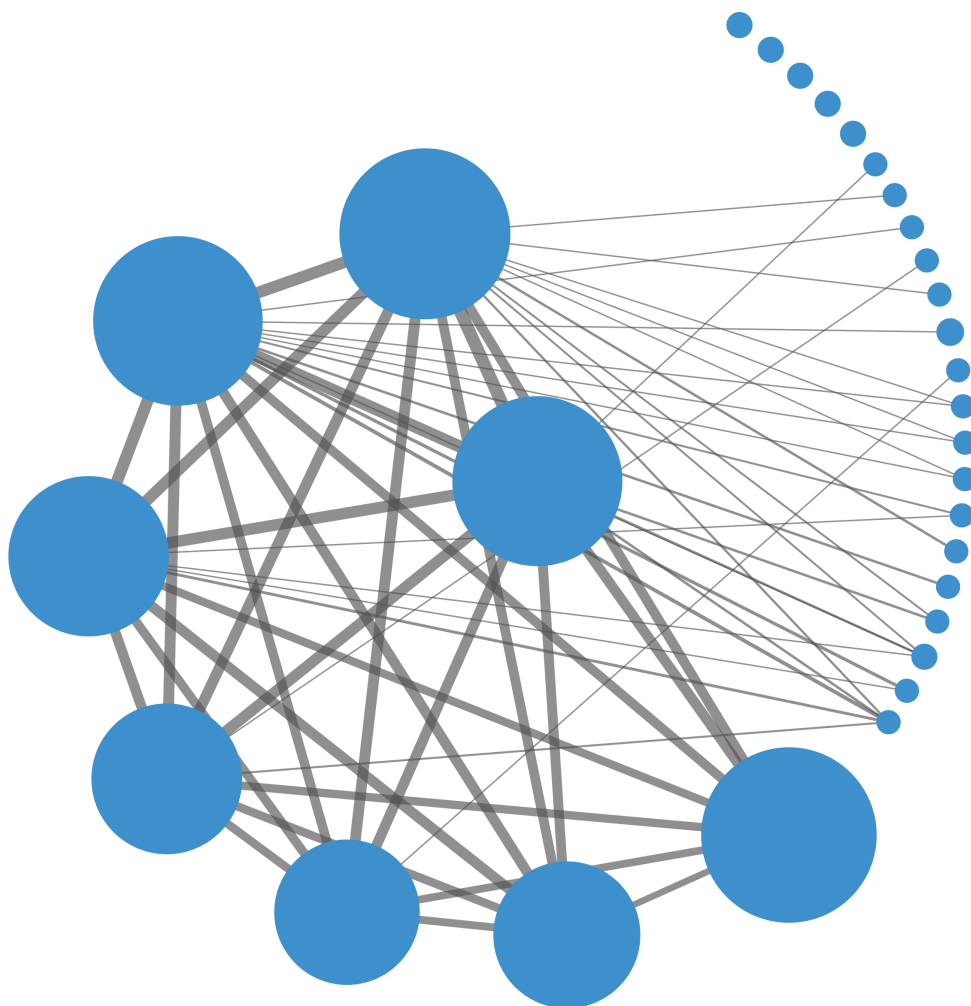


Figura 3.10: *Comunidades obtenidas por Louvain en LiveJournal*. Visualización de las 8 comunidades más grandes, de todas aquellas con grado menor o igual a 5 y del corte por aristas entre ellas en la red *LiveJournal*, de acuerdo a la partición obtenida por Louvain. Cada círculo representa una comunidad C , cuyo radio es proporcional al logaritmo de su grado, $d(C)$. Las aristas entre comunidades tienen un grosor proporcional al logaritmo del corte por aristas. Observamos que, mientras que las comunidades grandes forman un *clique* entre ellas, las comunidades pequeñas no tienen aristas entre sí. Las comunidades dibujadas pueden además tener aristas hacia otras comunidades de tamaño intermedio que no fueron representadas. La imagen fue generada con el software SnailVis [19].

Capítulo 4

Estudio de la Conectividad en Internet

En este capítulo nos proponemos abordar el estudio de Internet como sistema complejo. Comenzaremos explicando la importancia tecnológica de su estudio y mencionando los resultados más importantes hasta la fecha. En la Sección 4.2 presentaremos nuestro aporte, que vincula la arista-conectividad de la red con la descomposición en k -núcleos, y mostraremos los resultados que obtuvimos en recientes exploraciones.

4.1. Introducción

En sus comienzos, Internet estaba formada por una serie de enlaces troncales de larga distancia que constituían el llamado *backbone*. Hacia 1995 este *backbone* era la red NSF-Net, cuyos enlaces de 45 Mbps pertenecían al gobierno de los Estados Unidos. En 1995 la red NSFNet fue retirada e Internet pasó a ser una red completamente descentralizada. Sin embargo hoy en día las grandes compañías de telecomunicaciones a nivel mundial son las que proveen de conectividad a la red a través de enlaces de alta velocidad, que a su vez son contratados por compañías más pequeñas, revendedoras del servicio. Esta organización hace que Internet tenga cierta estructura jerárquica en la que algunos nodos están más cerca del centro de la red o *backbone* y mejor ubicados que otros¹.

Por otra parte, las compañías de comunicaciones en cualquiera de los Tiers poseen una organización interna *autónoma* del resto de la red. Esto da lugar al concepto de *Sistema Autónomo* (*Autonomous System, AS*). Un *Sistema Autónomo* de Internet es una

¹El concepto de *Tier*, si bien es algo difuso en su definición, está vinculado con dicha estructura jerárquica. Se suele decir que un *Tier 1* de Internet es una subred que forma parte del *backbone* de Internet. Los *Tier 2* en cambio se conectan a los *Tier 1* y los utilizan para alcanzar otras partes de la red, y a su vez ofrecen el servicio a otros *Tiers*. Por último, los *Tier 3* compran el servicio a los *Tiers 2*, también se conectan entre sí, y suelen ser los proveedores de acceso a Internet para los usuarios finales.

porción de la red que se encuentra bajo el control de una o más compañías de comunicaciones, que definen una política de ruteo dentro del sistema. Es decir, los *Sistemas Autónomos* controlan la forma en que se realiza el *encaminamiento* de la información dentro de ellos, y tienen una visión completa de su propia estructura. La estructura de un Sistema Autónomo está determinada por un grafo de red formado por *routers* (vértices) conectados a través de enlaces (aristas).

Podemos distinguir entonces dos niveles del estudio de Internet como red compleja:

- El nivel de *Sistemas Autónomos (ASes)*, en que los nodos de la red son los Sistemas Autónomos y los enlaces son los enlaces entre ASes, que surgen a partir de acuerdos comerciales entre los mismos.
- El nivel de *routers* (IR, por *inter-router level*) , de mayor nivel de detalle, formado por los *routers* y sus enlaces.

En ambos niveles, resulta de alto valor tecnológico entender la vinculación entre la estructura de la red y su funcionamiento. Algunos de los aspectos más importantes del estudio de Internet son:

- La *latencia*: Es la demora en la comunicación entre dos nodos de la red. Está relacionada con la demora en los enlaces de comunicaciones y fundamentalmente con la demora de procesamiento en los nodos, afectada por la *congestión*.
- El *ancho de banda*: Es la cantidad de información transmitida entre dos nodos por unidad de tiempo. Si bien depende de la capacidad física de los enlaces –que se vuelve cada vez mayor con el desarrollo de las tecnologías de comunicaciones–, también es enormemente afectada por la *congestión*.
- La *robustez o resiliencia (resilience)*: Es la capacidad de la red de soportar una falla local sin que se produzcan efectos graves en su funcionamiento. Un factor fundamental que afecta a la robustez es la *redundancia*, estrechamente vinculada a la *conectividad*, es decir, la existencia de múltiples caminos entre los nodos.
- La *topología*: Internet es un sistema complejo, y presenta distribuciones libres de escala, comportamiento emergente y carencia de un control centralizado. En particular, Internet parece estar diseñada para maximizar la tolerancia a fallos (como sugiere el mecanismo HOT) y el flujo de información [127].

Como vemos, la topología de Internet y su constitución como sistema complejo afectan a la *congestión* y la *robustez*. De allí la importancia de conocer la estructura del grafo de red subyacente.

Dado que Internet es una red dinámica, es imposible obtener de ella una instantánea completa. Por otra parte, como no es un sistema centralizado, tampoco existe una institución o ente que disponga de un registro global de lo que sucede en ella. Por este motivo uno de los problemas iniciales en el estudio de Internet fue la *exploración de la red*.

Exploraciones de Internet Existen hoy en día diversas instituciones que llevan a cabo esta tarea. Aquí trabajaremos con las siguientes:

- *CAIDA Association*²: Las exploraciones realizadas por esta asociación consisten en enviar paquetes IP –denominados *sondas (probes)*– desde sitios controlados de la red –los *monitores*– hacia distintos destinos. Como el ruteo IP provee información sobre el camino recorrido por los paquetes, es posible a partir de esta información reconstruir parcialmente el grafo. A la fecha (julio de 2013) el sistema cuenta con alrededor de 80 monitores dispersos por el mundo.
- *DIMES Project*³: Es un sistema distribuido compuesto por nodos que participan voluntariamente. Desde cada nodo se envían paquetes IP con una baja frecuencia. Actualmente (julio de 2013) cuenta con cerca de 400 agentes activos, la gran mayoría de ellos en los Estados Unidos.
- *Route Views Project*⁴: A diferencia de los anteriores, no realiza mediciones *activas* sino pasivas. El proyecto observa las tablas de ruteo BGP de algunos *routers* de borde de los sistemas autónomos. Como el protocolo BGP almacena el camino completo para llegar desde un AS a otro, es posible a partir de las tablas reconstruir la topología de la red a nivel de ASes. Sin embargo el método es sesgado porque algunas rutas entre ASes son ocultas (debido a políticas o acuerdos) y no son mostradas por BGP.

CAIDA y DIMES proveen información a nivel de *routers* de la red. Como los *routers* se identifican por direcciones IPs que están asociadas públicamente a los Sistemas Autónomos, es posible deducir el grafo a nivel de ASes a partir del grafo a nivel de *routers*. En el caso de Route Views, en cambio, como las tablas BGP rutean entre ASes, sólo se tiene una visión a nivel de Sistemas Autónomos.

Antes de que surgieran estos proyectos, los primeros trabajos sobre topología de Internet se realizaron a partir de la observación de algunas tablas BGP. Este es el caso de Govindan y Reddy (1997) [82], que observaron que frente al crecimiento en tamaño de la

²<http://www.caida.org/home/> [34].

³<http://www.netdimes.org/new/> [56].

⁴<http://www.routeviews.org/> [151].

red en sus primeros años, el diámetro prácticamente no mostraba variaciones. En 1998, Pansiot y Grad reconstruyeron el grafo de Internet a nivel de *routers* enviando paquetes IP entre 11 nodos en distintos puntos de la red [124]. Un año después, Govindan y Tangmunarunkit logran construir un mapa mucho más completo explotando la opción de *source-routing* del protocolo IP [83].

En 1999 Faloutsos *et al.* presentaron su reconocido trabajo en el que mostraron la existencia de leyes de potencias en las distribuciones de algunos parámetros del grafo de Internet, como los grados de los vértices y las distancias entre pares de vértices [66]. Para obtener estos resultados, se basaron en observaciones de tablas BGP provistas por el NLANR⁵ y en la exploración a nivel de *routers* de Pansiot y Grad [124].

El trabajo de Faloutsos *et al.* tuvo muchas repercusiones. Entre ellas, Pastor-Satorras *et al.* confirmaron las distribuciones libres de escala y a la vez observaron un comportamiento discordante (*disassortative*) de los grados de los vértices en el nivel de Sistemas Autónomos [126], con datos también provistos por el NLANR. Este resultado está estrechamente vinculado con la estructura de Internet: como mencionan Catanzaro *et al.* [39], la combinación de distribuciones libres de escala en los grados con un comportamiento discordante evita la formación de una estructura auto-semejante y da lugar a una estructura jerárquica formada por *hubs* (nodos densamente conectados) que se conectan con otros *hubs*, y nodos periféricos que necesitan de los *hubs* para conectarse entre ellos.

La estructura jerárquica característica de Internet en el nivel de Sistemas Autónomos se refleja en modelos conceptuales como el *jellyfish* de Siganos *et al.* (2006) [146] y el *MEDUSA* de Carmi *et al.* (2007) [38]. En ambos se modela la red con una estructura en capas. El modelo *jellyfish* es exigente respecto a la densidad de conexiones dentro de las capas: se requiere que las mismas constituyan *cliques* o *k-plexos* (véase su definición en la Figura 3.1). En cambio el modelo *MEDUSA* se inspira en la menos estricta descomposición en *k*-núcleos, que hemos introducido en la Sección 2.1.3.4.

La descomposición en *k*-núcleos es una herramienta útil en el estudio de la estructura de Internet. Alvarez-Hamelin *et al.* [7] mostraron que los *k*-núcleos de Internet conservan el comportamiento libre de escala de la red completa: al observar la distribución de grados dentro de un núcleo, se encuentra una ley de potencias con el mismo exponente que el de toda la red. Lo mismo sucede con la distribución de grados de los vecinos y el coeficiente de agrupamiento de los vértices en función del grado. Por último, los autores confirman la presencia de un comportamiento discordante.

Por otra parte, los *k*-núcleos están estrechamente vinculados con la conectividad. Los trabajos de Carmi *et al.* (2006) [37] y de Alvarez-Hamelin *et al.* (2008) [7] mostraron

⁵*National Laboratory for Advanced Network Research*. El proyecto que lo sostenía finalizó en 2006 y sus recursos quedaron bajo la administración del proyecto CAIDA.

empíricamente que los k -núcleos de Internet son k -conexos.

Nuestro aporte en este capítulo consistirá en establecer condiciones suficientes para garantizar la núcleo-conectividad de una red, que definiremos como la k -arista-conectividad de sus k -núcleos. Mostraremos que estas condiciones se satisfacen en los grafos de Internet a nivel de Sistemas Autónomos. Los resultados de este trabajo se encuentran publicados en [6].

4.2. Estimación de la conectividad empleando k -núcleos

Recordemos que la arista-conectividad de un grafo conexo G , $\kappa'(G)$, es la mínima cantidad de aristas que deben ser eliminadas para transformarlo en un grafo no conexo, y es equivalente a la capacidad del corte mínimo por aristas (véase el Cap. 2, Sección 2.1.2.2). Decimos que G es k -arista-conexo cuando $\kappa'(G) \geq k$. Por otra parte, si G es k -arista-conexo, entonces existen al menos k caminos arista-disjuntos entre todo par de vértices en G .

4.2.1. Formalización del método

Comenzamos introduciendo un teorema de expansión sobre la noción de distancia:

4.2.1.1. Un teorema de expansión

Dado un grafo simple G , definiremos la distancia entre un vértice $x \in V(G)$ y un subconjunto $A \subset V(G)$, $d_G(x, A)$, como el mínimo de las distancias entre x y los vértices de A . Es decir, $d_G(x, A)$ es la distancia desde x hasta el vértice de A más cercano.

(Fig.4.1.a) Consideramos en este teorema dos subconjuntos disjuntos Q y C , no vacíos, del conjunto de vértices $V(G)$. Llamamos G' al grafo inducido por $C' = Q \cup C$; es decir,

(Fig.4.1.b) $G' = G[C']$ ⁶. Definimos la *distancia contraída* entre vértices $x, y \in Q$ como:

(Fig.4.1.c)

$$d_{C'/C}(x, y) = \min\{d_{G'[Q]}(x, y), d_{G'}(x, C) + d_{G'}(y, C)\} ,$$

(Fig.4.1.d) y entre vértices $x \in C', y \in C$ como:

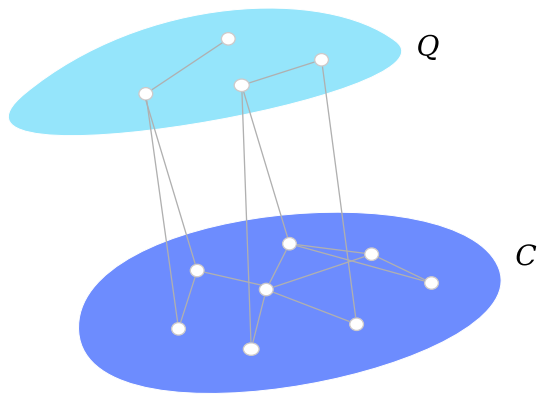
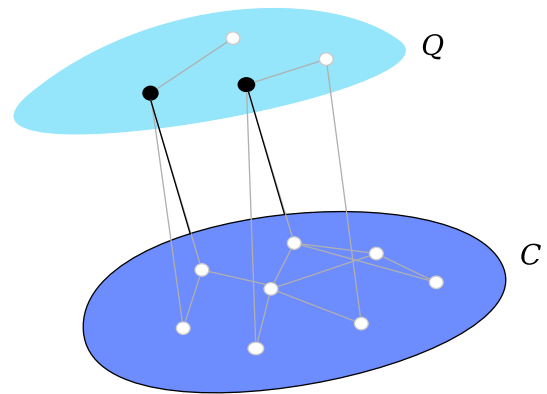
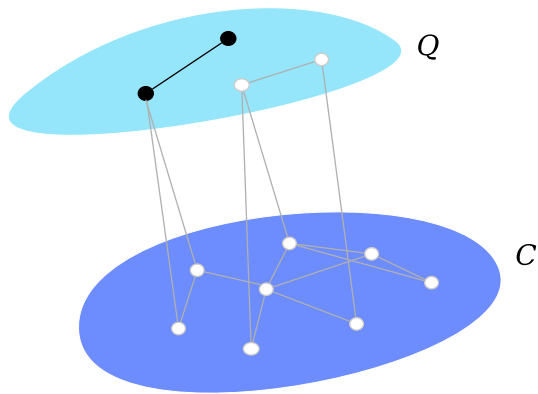
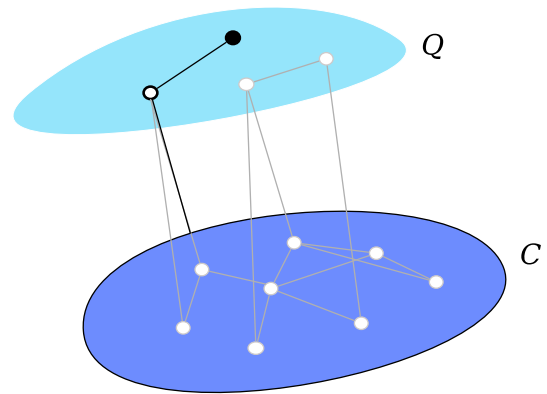
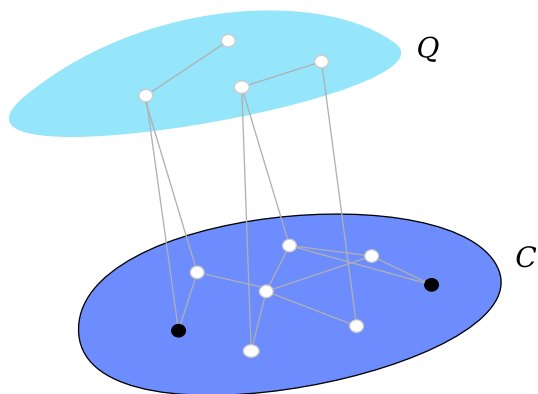
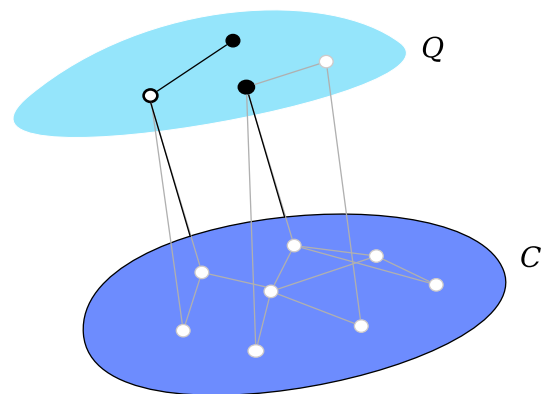
(Fig.4.1.e)

$$d_{C'/C}(x, y) = d_{C'/C}(y, x) = d_{G'}(x, C) .$$

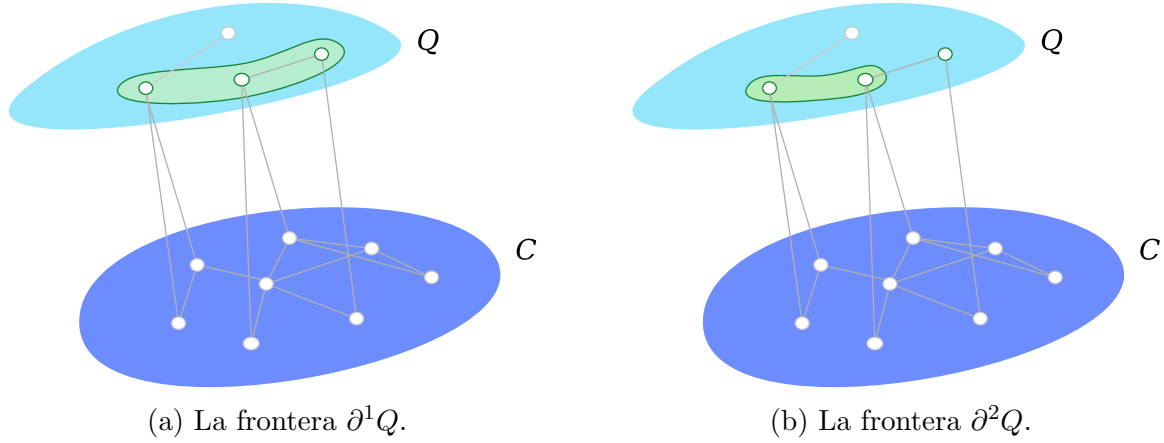
De esta forma nuestra noción de distancia contraída queda definida en todo C' ⁷.

⁶Adelantamos aquí que C representará a un k -núcleo más central que proveerá de conectividad a Q .

⁷El nombre *distancia contraída* puede interpretarse como la distancia en G' cuando el conjunto C colapsa a un único vértice, que se conecta a aquellos vértices en Q que eran vecinos de algún vértice en C .

(a) Grafo G' , inducido por $C' = C \cup Q$.(b) Dos vértices en Q con distancia contraída 2.(c) Dos vértices en Q con distancia contraída 1.(d) La distancia contraída entre el vértice relleno en negro y cualquier vértice de C es 2.(e) La distancia contraída entre vértices de C es 0.(f) El diámetro contraído de G' es 3.Figura 4.1: *La noción de distancia contraída.*

También definimos la distancia contraída entre un vértice $x \in C'$ y un subconjunto

Figura 4.2: Frontera de Q .

$A \subset C'$, como:

$$d_{C'/C}(x, A) = \min_{a \in A} d_{C'/C}(x, a) .$$

(Fig.4.1.f) Por último, introducimos la noción de *diámetro contraído* de $G' = G[C']$ respecto a C como

$$\text{diam}_{C'/C} = \max_{x, y \in C'} d_{C'/C}(x, y) .$$

Con estas definiciones se cumple que si $d_{C'/C}(x, y) = 2$ para algún par $x, y \in C'$, entonces existe un $z \in C'$ tal que $d_{C'/C}(x, z) = d_{C'/C}(z, y) = 1$.

También utilizaremos la siguiente notación:⁸

$$\begin{aligned} \partial^j Q &= \{x \in Q : |[x, C]| \geq j\} \\ \bar{\partial}^j Q &= \{x \in Q : |[x, C]| < j\} = Q \setminus \partial^j Q . \end{aligned}$$

(Fig.4.2.a) Estos conjuntos anidados $\partial^j Q$ organizan a los vértices *frontera* de Q en relación con
(Fig.4.2.b) la cantidad de conexiones que tienen con C .

Por último, consideraremos:

$$\Phi_{C'/C} = \sum_{x \in Q} \min\{\max\{1, |[x, \bar{\partial}^2 Q]|\}, |[x, C]|\}$$

Enunciamos entonces el siguiente teorema.

Teorema 1. *Dado un grafo simple G' tal que $V(G') = C'$ y $C \subset C'$, si $\text{diam}_{C'/C} \leq 2$, entonces para los cortes por aristas $[S, \bar{S}]$ en G' tales que $C \subset S$ se verifica que:*

1. Si $\max_{\bar{s} \in \bar{S}} d_{C'/C}(\bar{s}, S) = 1$, entonces $|[S, \bar{S}]| \geq \max_{\bar{s} \in \bar{S}} d(\bar{s})$.

⁸Haremos un pequeño abuso de notación al escribir $|[x, C]|$ en lugar de $|[\{x\}, C]|$.

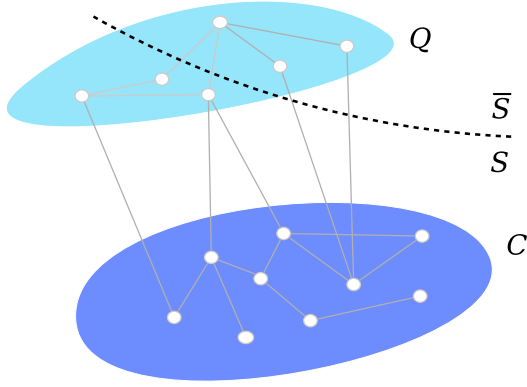
2. Si $\max_{\bar{s} \in \bar{S}} d_{C'/C}(\bar{s}, S) = 1$, entonces $|[S, \bar{S}]| \geq |\bar{S}|$.
3. Si $\max_{\bar{s} \in \bar{S}} d_{C'/C}(\bar{s}, S) = 2$, entonces $|\bar{S}| > \min_{\bar{s} \in \bar{S}} d(\bar{s})$.
4. Si $\max_{\bar{s} \in \bar{S}} d_{C'/C}(\bar{s}, S) = 2$, entonces $\max_{s \in S} d_{C'/C}(s, \bar{S}) = 1$.
5. Si $\max_{s \in S \cap Q} d_{C'/C}(s, \bar{S}) = 1$, entonces $|[S \cap Q, \bar{S}]| \geq \max_{s \in S \cap Q} (d(s) - d_C(s))$.⁹
6. Si $\max_{s \in S \cap Q} d_{C'/C}(s, \bar{S}) = 1$, entonces $|[S \cap Q, \bar{S}]| \geq |S \cap Q|$.

Demostración.

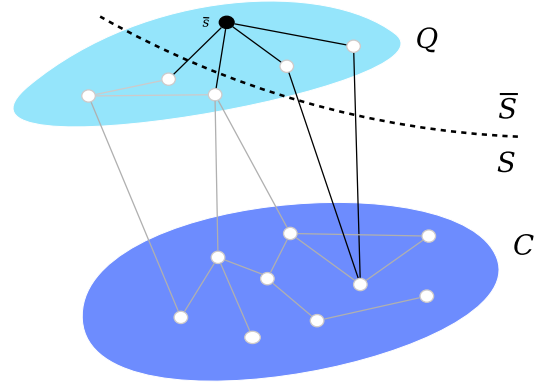
1. Sea $\bar{s} \in \bar{S}$. Dividimos el grado de $\bar{s}$ en dos componentes: $d_S(\bar{s}) = |[\bar{s}, S]|$ y $d_{\bar{S}}(\bar{s}) = |[\bar{s}, \bar{S}]|$. Por cada vecino de $\bar{s}$ en S , $\bar{s}$ aporta una unidad al corte por aristas $[\bar{s}, \bar{S}]$. Por otra parte, para cada uno de los vecinos de $\bar{s}$ en $\bar{S}$, que también satisfacen que su distancia a S es 1, se cumple que ese vecino tiene alguna arista hacia S . Luego: $|[S, \bar{S}]| \geq d_S(\bar{s}) + d_{\bar{S}}(\bar{s}) = d(\bar{s})$. Como esta afirmación vale para todo $\bar{s} \in \bar{S}$, obtenemos que $|[S, \bar{S}]| \geq \max_{\bar{s} \in \bar{S}} d(\bar{s})$. (Fig.4.3.a) (Fig.4.3.b)
2. Es inmediato si se observa que para cada $\bar{s} \in \bar{S}$ existe al menos una arista hacia S , la cual forma parte del corte por aristas $[S, \bar{S}]$. (Fig.4.3.c)
3. En este caso existe algún $\bar{s} \in \bar{S}$ que no posee aristas hacia S . Para este $\bar{s}$, $d_{\bar{S}}(\bar{s}) = d(\bar{s})$, y luego $|\bar{S}| \geq d(\bar{s}) + 1 > \min_{\bar{s} \in \bar{S}} d(\bar{s})$. (Fig.4.3.d) (Fig.4.3.e)
4. Siguiendo el razonamiento del punto anterior, si $\bar{s}$ no tiene aristas hacia S entonces el camino mínimo para llegar a él desde cualquier vértice $s \in S$ debe tener longitud 2 (porque el diámetro contraído es menor o igual a 2) y el vértice intermedio de ese camino debe estar en $\bar{S}$. Luego, $d(s, \bar{S}) = 1$. (Fig.4.3.f)
5. Si los vértices en S que pertenecen a Q tienen al menos una arista hacia $\bar{S}$, entonces por un argumento similar al del ítem 1, para cada $s \in S \cap Q$ las aristas que no van hacia C van o bien hacia $\bar{S}$ o bien hacia otros vecinos en $S \cap Q$ que también tienen al menos una arista hacia $\bar{S}$. Luego, $d(s) - d_C(s)$ es una cota inferior de $|[S \cap Q, \bar{S}]|$.
6. De forma similar al punto 2, esto es inmediato si se observa que para cada $s \in S \cap Q$ existe al menos una arista hacia $\bar{S}$.

□

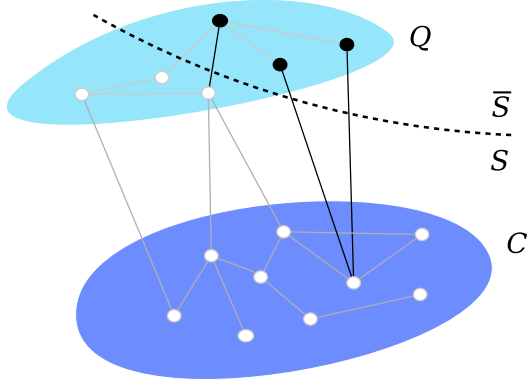
⁹La notación $d_C(s)$ designa el grado de s interno a C , conforme a la notación que usamos en el Capítulo 3. Es la cantidad de aristas que salen de s e inciden en vértices de C .



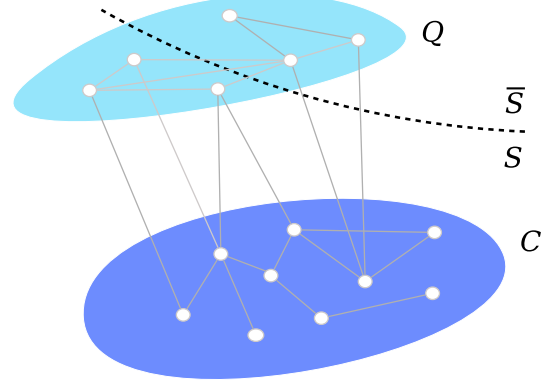
(a) Un grafo G' , inducido por $C' = C \cup Q$, cuyo diámetro contraído es 2, y un corte por aristas $[S, \bar{S}]$ tal que $C \subset S$. Para todo $\bar{s} \in \bar{S}$ se cumple que $d_{C'/C}(\bar{s}, S) = 1$.



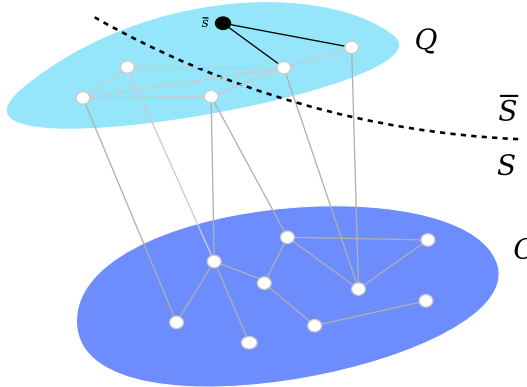
(b) Punto 1. El grado de $\bar{s}$ es una cota inferior de $||S, \bar{S}||$.



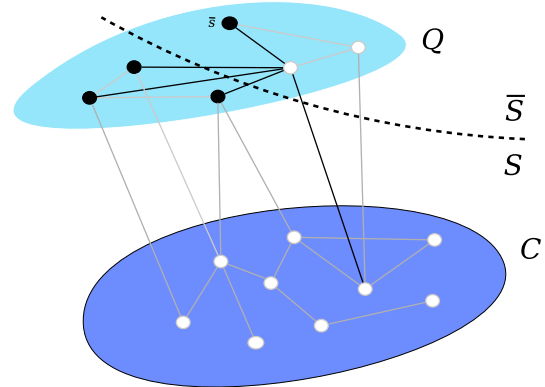
(c) Punto 2. El cardinal de $\bar{S}$ es también una cota inferior de $||S, \bar{S}||$.



(d) Modificamos las aristas de los vértices en Q . El diámetro contraído sigue siendo 2, pero ahora existen vértices en $\bar{S}$ que no tienen conexiones hacia S . Para todo $\bar{s} \in \bar{S}$ se cumple que $d_{C'/C}(\bar{s}, S) \leq 2$.



(e) Punto 3. $\bar{s}$ no tiene aristas hacia S . Entonces el grado de $\bar{s}$ más 1 es una cota inferior del cardinal de $\bar{S}$.



(f) Punto 4. Todo vértice en S está a distancia contraída 2 de $\bar{s}$. Entonces todo vértice en S está a distancia contraída 1 de $\bar{S}$.

Figura 4.3: Ilustración del Teorema 1.

Corolario 1. *Supongamos que junto a las hipótesis del Teorema 1 se cumple que*

$$|[S, \bar{S}]| < \min_{v \in Q} d(v) \ .$$

Entonces:

1. $\max_{\bar{s} \in \bar{S}} d_{C'/C}(\bar{s}, S) = 2$.
2. $\max_{s \in S} d_{C'/C}(s, \bar{S}) = 1$.
3. $|[C, \bar{S}]| \geq 1$.
4. $|S \cap Q| < |[S, \bar{S}]| < \min_{v \in Q} d(v) < |\bar{S}|$.
5. $S \cap Q \subset \partial^2 Q$, o lo que es igual, $\partial^2 Q \subset \bar{S}$.
6. $\Phi_{C'/C} \leq |[S, \bar{S}]|$.

Demostración.

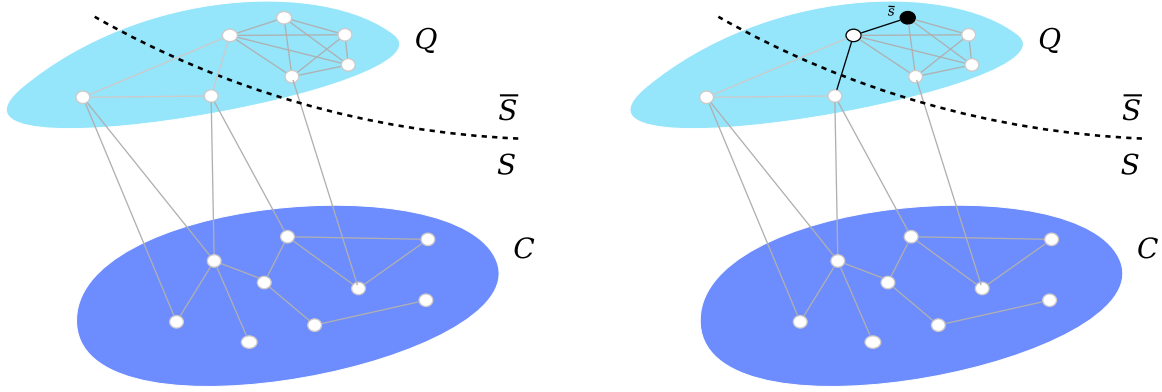
1. Es consecuencia del Punto 1 del Teorema 1. De lo contrario todos los vértices de $\bar{S}$ deberían tener una arista en el corte por aristas, y entonces la capacidad de éste sería mayor o igual al grado de cada $\bar{s}$. (Fig.4.4.a) (Fig.4.4.b)
2. Es consecuencia inmediata del Punto 4 del Teorema 1 y de la nueva hipótesis. (Fig.4.4.c)
3. De lo contrario, todos los vértices en $\bar{s} \in \bar{S}$ deberían tener una conexión hacia $S \cap Q$, y entonces resultaría $|[S, \bar{S}]| \geq d(\bar{s})$. (Fig.4.4.d)
4. De los Puntos 3 y 4 se deduce la primera desigualdad. La segunda es la hipótesis de este Corolario, y la última surge del Punto 3 del Teorema 1.
5. Del Punto 5 del Teorema 1 y el Punto 3 de este Corolario se sigue que:

$$|[S, \bar{S}]| = |[S \cap Q, \bar{S}]| + |[C, \bar{S}]| > \max_{s \in S \cap Q} (d(s) - d_C(s))$$

Entonces para todo $s \in S \cap Q$, utilizando la hipótesis:

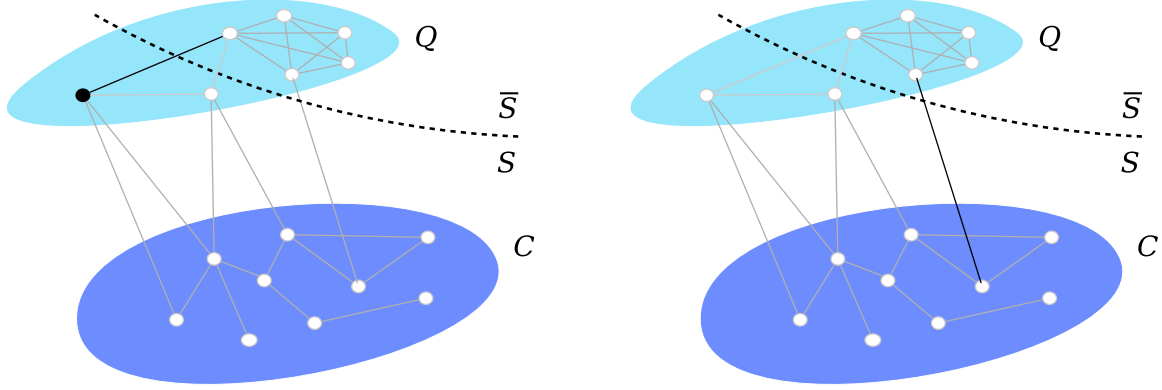
$$d(s) > |[S, \bar{S}]| > (d(s) - d_C(s))$$

de donde $d_C(s) \geq 2$, y concluimos que todos los vértices de $S \cap Q$ pertenecen a la frontera $\partial^2 Q$.

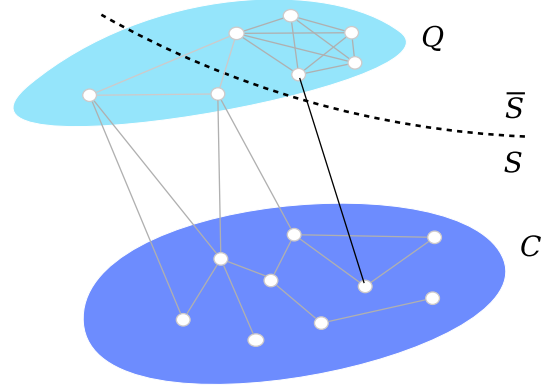


(a) Un grafo G' , inducido por $C' = C \cup Q$, de diámetro contraído 2, con un corte por aristas $[S, \bar{S}]$ tal que $C \subset S$. Se cumple la hipótesis adicional $|[S, \bar{S}]| < \min_{v \in Q} d(v)$.

(b) Punto 1. $d(\bar{s}, S) = 2$.



(c) Punto 2. Los vértices en $S \cap Q$ necesariamente tienen alguna arista hacia $\bar{S}$.



(d) Punto 3. La capacidad del corte por aristas $[C, \bar{S}]$ es al menos 1.

Figura 4.4: Ilustración del Corolario 1.

6. Como $\bar{\partial}^2 Q \subset \bar{S}$, se cumple que para $s \in S \cap Q$:

$$|[s, \bar{S}]| \geq \max\{1, |[s, \bar{\partial}^2 Q]|\}$$

mientras que para $\bar{s} \in \bar{S}$ se cumple que $|\bar{s}, S| \geq |\bar{s}, C|$. Entonces:

$$\begin{aligned} |[S, \bar{S}]| &= |[S \cap Q, \bar{S}]| + |[C, \bar{S}]| \\ &\geq \sum_{s \in S \cap Q} \max\{1, |[s, \bar{\partial}^2 Q]|\} + \sum_{\bar{s} \in \bar{S}} |\bar{s}, C| \\ &\geq \Phi_{C'/C} \end{aligned}$$

□

A continuación utilizaremos el Teorema y Corolario anteriores para enunciar un resultado sobre la k -arista-conectividad del grafo G' .

Corolario 2. Sea $k \leq d_{\min}(G')$. Si se cumple que:

1. $G'[C]$ es $d_{\min}(G')$ -arista-conexo
2. $\text{diam}_{C'/C} \leq 2$

Entonces cualquiera de las siguientes condiciones implica que G' es k -arista-conexo:

1. $\Phi_{C'/C} \geq k$
2. $|\partial^1 Q| \geq k$
3. $Q = \partial^1 Q$

Demostración. Sea $[S, \bar{S}]$ un corte por aristas en G' . Mostraremos que bajo las 2 hipótesis y cualquiera de las 3 alternativas, se cumple que $|[S, \bar{S}]| \geq k$.

Supongamos en un primer caso que C queda dividido por el corte por aristas, es decir, $S \cap C \neq \emptyset$ y $\bar{S} \cap C \neq \emptyset$. Entonces el corte por aristas $[S \cap C, \bar{S} \cap C]$ está incluido en $\subset [S, \bar{S}]$. Pero como asumimos que $G'[C]$ es k -arista-conexo, se sigue que:

$$|[S, \bar{S}]| \geq |[S \cap C, \bar{S} \cap C]| \geq k$$

Supongamos entonces que $C \subset S$ (sin pérdida de generalidad; sólo para seguir la notación de los resultados anteriores). Si sucediera que $|[S, \bar{S}]| < k$, entonces como $k \leq d_{\min}(G') \leq \min_{v \in Q} d(v)$, se verificaría la hipótesis del Corolario 1.

Sin embargo, la primera de las condiciones contradice el Punto 6 del Corolario.

Por otra parte, si $v \in \partial^1 Q$ entonces v tiene alguna arista hacia C . Luego, v contribuye en al menos una unidad a $\Phi_{C'/C}$. Entonces la segunda de nuestras condiciones implica la primera, y nuevamente contradice el Corolario.

Por último, si $Q = \partial^1 Q$ entonces todos los vértices de Q tendrán alguna arista hacia C , lo que entra en contradicción con el Punto 1 del Corolario. $\square$

Notación. Para resumir las tres condiciones del Corolario 2, utilizaremos la siguiente notación:

$$\Psi_{C'/C}(k) = \max\{\Phi_{C'/C} - k, |\partial^1 Q| - k, |\partial^1 Q| - |Q|\}, \quad \text{para } k \leq d_{\min}(G') .$$

De esta manera nuestras 3 condiciones se resumen en: $\Psi_{C'/C}(k) \geq 0$.

Observación: Nuestro Corolario 2 está intimamente vinculado con el Teorema de Plesník [128], que establece que en grafos simples de diámetro 2 la arista-conectividad es igual al grado mínimo. De hecho, la condición de diámetro contraído 2 asegura que el grafo que se obtendría a partir de G' contrayendo C a un vértice es k -arista-conexo para $k \leq d_{\min}(G')$. Sin embargo esto no asegura la k -arista-conectividad de G' , por ello se necesita alguna de las 3 condiciones adicionales.

4.2.1.2. Arista-conectividad en sentido estricto y en sentido amplio

Expandiremos a continuación la noción de arista-conectividad para subgrafos inducidos por subconjuntos de vértices $A \subset V(G)$.

Diremos que un subgrafo inducido $G[A]$ es *k -arista-conexo en sentido estricto* cuando $G[A]$ es k -arista-conexo, es decir, cuando todo corte por aristas en $G[A]$ tiene al menos k aristas o, lo que es lo mismo, existen k caminos arista-disjuntos de a a b entre todo par de vértices u, v en el grafo $G[A]$.

Diremos que un subgrafo inducido $G[A]$ es *k -arista-conexo en sentido amplio* cuando todo corte por aristas $[X, \bar{X}]$ en G que divide al conjunto A —es decir, tal que $X \cap A \neq \emptyset$ y $\bar{X} \cap A \neq \emptyset$ — tiene al menos k aristas. Esto equivale a la existencia en el grafo completo G de al menos k caminos arista-disjuntos de a a b entre pares de vértices u, v en A .

Es inmediato observar que si $G[A]$ es k -arista-conexo en sentido estricto, entonces también es k -arista-conexo en sentido amplio.

4.2.1.3. Construcción de conjuntos núcleo-conexos

Vincularemos ahora nuestras nociones de *arista-conectividad en sentido estricto y amplio* con la descomposición en k -núcleos. Recordemos que un k -núcleo es un subgrafo inducido de grado mínimo k , maximal con respecto a esta propiedad (véase la Sección 2.1.3.4). Nuestra hipótesis es que los k -núcleos suelen ser k -arista-conexos. Desarrollaremos entonces un algoritmo que recorre los k -núcleos, desde aquel más central (de mayor índice k) hacia afuera, y construye un subconjunto $C \subset V(G)$ tal que los k -núcleos del subgrafo inducido por C son k -arista-conexos en sentido estricto (amplio). A esta propiedad la denominaremos *núcleo-conectividad en sentido estricto (amplio)*:

Definición. Diremos que un grafo es *núcleo-conexo en sentido estricto (amplio)* cuando todos sus k -núcleos son k -arista-conexos en sentido estricto (amplio).

Sería deseable que todo el grafo G verifique la *núcleo-conectividad*. Cuando ello no sea posible, el algoritmo intentará extraer un subgrafo inducido núcleo-conexo lo más grande posible.

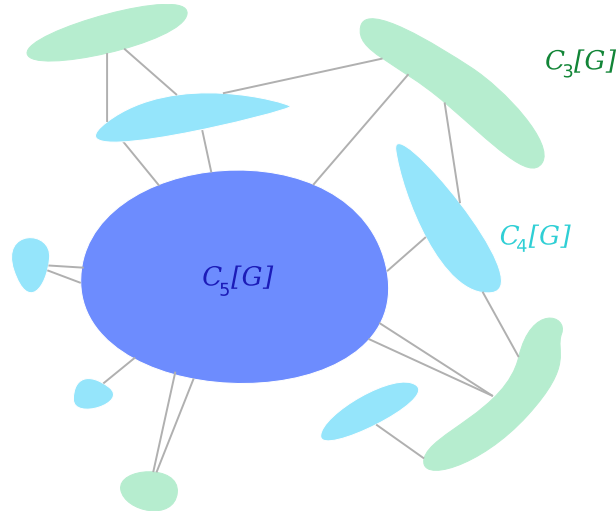


Figura 4.5: k -capas y clusters de un grafo. En este ejemplo el grafo tiene núcleo-profundidad 5. El k -núcleo central está incluido dentro del 4-núcleo (azul+celeste). Los vértices que pertenecen al 4-núcleo pero no al 5-núcleo forman la 4-capa (celeste). La 4-capa tiene 5 componentes conexas (clusters). A su vez, el 4-núcleo está inmerso en el 3-núcleo (azul+celeste+verde). La 3-capa (verde) está integrada por 4 clusters.

k -conectividad en sentido estricto El algoritmo requiere un subconjunto inicial con la máxima arista-conectividad posible, por ello comienza por el k -núcleo de mayor índice k . Como debe verificarse la condición de diámetro 2 del Teorema de Plesník, este $k_{\text{máx}}$ -núcleo debe tener una única componente conexa. En caso que el $k_{\text{máx}}$ -núcleo tenga varias componentes conexas, se considerará cada una individualmente.

Si no encuentra ninguna componente conexa de diámetro 2 en el $k_{\text{máx}}$ -núcleo, el algoritmo pasará al núcleo inmediatamente inferior, considerando a los vértices del $(k_{\text{máx}} - 1)$ -núcleo que no pertenecen al $k_{\text{máx}}$ -núcleo. A esta “corteza” de un k -núcleo la llamaremos k -capa (k -shell) (véase la Figura 4.5). La k -capa es el subgrafo inducido por los vértices que tienen índice de capa igual a k . Cada k -capa puede estar formada por varias componentes conexas, a las que denominaremos clusters. El algoritmo avanzará por las k -capas hasta encontrar un primer cluster de diámetro 2 y grado mínimo k . Los vértices de este cluster formarán el conjunto C inicial, y como el cluster es k -arista-conexo para k igual a su k -capa de pertenencia, el grafo $G[C]$ será núcleo-conexo.

Una vez terminada esta primera parte, se intentará anexar a C nuevos clusters¹⁰. Comenzando por la k -capa inmediatamente inferior, se intentará aplicar el Corolario 2 a cada uno de los clusters del mismo. El cluster de la k -capa ocupará el lugar de Q en el Teorema, mientras que el conjunto C satisface la hipótesis requerida de k -arista-

¹⁰Es fundamental comprender que, a medida que se agreguen nuevos vértices, la arista-conectividad de $G[C]$ irá en disminución, pero $G[C]$ seguirá siendo siempre núcleo-conexo.

conectividad¹¹. Para poder aplicar el teorema en $G[C']$, con $C' = C \cap Q$, el algoritmo verifica si se cumple alguna de las 3 condiciones del Corolario 2. En caso afirmativo, el *cluster* Q se incorporará a C ¹².

El procedimiento recorre todas las k -capas considerando sus distintos *clusters* hasta terminar con la 2-capa. En el caso de la 1-capa, las condiciones del Corolario 2 son demasiado estrictas y simplemente debe verificarse para cada uno de sus *clusters* que exista al menos alguna arista hacia C .

El resultado final será un subgrafo $G[C]$ que satisface la núcleo-conectividad, es decir que sus k -núcleos son k -arista-conexos en sentido estricto. La complejidad computacional del algoritmo es de $O(e(G))$ (véase [6]).

El procedimiento completo se muestra en el Algoritmo 3.

k-conectividad en sentido amplio El procedimiento se muestra en el Algoritmo 4. En este caso nuestro algoritmo cuenta con un conjunto temporal (*buffer*) $\mathfrak{B}$ en el que se almacenan los *clusters* que no pudieron ser añadidos. Si en algún momento uno de ellos satisface las condiciones de la línea 4.15, el mismo es agregado a un conjunto D . Estos *clusters* agregados tardíamente tienen una conectividad menor en $G[C \cup D]$ que el índice de la capa a la que pertenecían. Sin embargo el valor de k del paso en que se agregaron asegura la k -arista-conectividad de $G[C \cup D]$, que es la hipótesis requerida sobre $G[C \cup D]$ para poder seguir aplicando el teorema. De esta forma, si bien los vértices en D no forman parte del conjunto núcleo-conexo, pueden ser utilizados por otros *clusters* para establecer sus caminos. La conectividad así obtenida es una conectividad en sentido amplio, porque los caminos que conectan a los vértices en el conjunto núcleo-conexo C obtenido finalmente pueden atravesar el conjunto D final.

4.2.2. Resultados obtenidos

Hemos aplicado nuestros algoritmos al análisis de la *núcleo-conectividad* de grafos de Internet a nivel de Sistemas Autónomos (ASes). Los grafos de red se obtuvieron de exploraciones de CAIDA y DIMES, y se resumen en el Cuadro 4.1.

En el Cuadro 4.2 observamos los tamaños de los subgrafos núcleo-conexos extraídos a partir de los dos algoritmos. Observamos que en gran medida los vértices del grafo de Internet a nivel de ASes pertenecen al subgrafo núcleo-conexo. Dentro de este subgrafo podemos garantizar entonces que para todo par de vértices u y v , la arista-conectividad

¹¹Dado que C es núcleo-conexo y su grado mínimo es mayor o igual al k actual, C es k -arista conexo.

¹²Al incorporar a Q , C tendrá grado mínimo k y será, como consecuencia del teorema, k -núcleo-conexo. Pero dado que el $(k+1)$ -núcleo de C no incluye a ninguno de los vértices en Q , sigue teniendo el grado de arista-conectividad previo. Luego, C seguirá siendo *núcleo-conexo*.

Algoritmo 3: Núcleo-conectividad en sentido estricto

Entrada: $S_k[G] = \{Q_{k1}, Q_{k2}, \dots, Q_{kM_k}\}$, las k -capas de G (desde 1 hasta $k_{\text{máx}}$),
particionadas en sus componentes conexas (*clusters*)

Salida: $C \subset V$, núcleo-conexo en sentido estricto

```

3.1  $C \leftarrow \emptyset$ 
3.2  $k \leftarrow k_{\text{max}}$ 
3.3 inicio
3.4   mientras  $C = \emptyset$  y  $k \geq 1$  hacer
3.5     si existe algún  $Q \in S_k[G]$  tal que  $\text{diam}(G[Q]) \leq 2$  y  $d_{\min}(G[Q]) \geq k$ 
3.6       entonces
3.7          $C \leftarrow C \cup Q$ 
3.8       fin
3.8      $k \leftarrow k - 1$ 
3.9   fin
3.10  mientras  $k \geq 2$  hacer
3.11    mientras existe algún  $Q \in S_k[G]$  tal que:  $\left\{ \begin{array}{l} \text{diam}_{C \cup Q/C} \leq 2 \\ \Psi_{C \cup Q/C}(k) \geq 0 \end{array} \right\}$  hacer
3.12       $C \leftarrow C \cup Q$ 
3.13       $S_k[G] \leftarrow S_k[G] \setminus Q$ 
3.14    fin
3.15     $k \leftarrow k - 1$ 
3.16  fin
3.17  para cada  $Q \in S_1[G]$  hacer
3.18    si  $|\partial^1 Q| \geq 1$  entonces
3.19       $C \leftarrow C \cup Q$ 
3.20    fin
3.21  fin
3.22 fin

```

es de al menos el mínimo entre el índice de capa de los dos vértices.

Para comparar la cota inferior de arista-conectividad que asegura nuestro algoritmo con la arista-conectividad real del grafo, hemos generado las Figuras 4.7 y 4.8. En ambos gráficos se considera a todos los pares de vértices en el grafo y se los organiza de acuerdo al mínimo entre sus índices de capa, que se encuentra sobre el eje x . Para cada valor de este mínimo se muestra en el eje y un segmento con la media y el desvío estándar de la arista-conectividad entre los pares de vértices. La arista-conectividad se calcula en dos variantes: como arista-conectividad en el k -núcleo más pequeño que contiene a ambos vértices – lo que denominamos arista-conectividad hacia adentro– y como arista-conectividad en el grafo completo. En ambos casos graficamos simultáneamente la curva $f(x) = x$ que corresponde a la cota inferior de arista-conectividad garantizada por nuestro algoritmo para aquellos vértices que pertenecen al conjunto núcleo-conexo. Concluimos que esta cota es una muy buena aproximación de la arista-conectividad hacia adentro.

El cálculo de la arista-conectividad se efectuó construyendo un árbol de Gomory-Hu del grafo completo, o de cada k -núcleo en el caso de la arista-conectividad hacia adentro. El procedimiento se describe brevemente a continuación.

4.2.2.1. Árboles de Gomory-Hu

La arista-conectividad de un grafo está vinculada con el corte mínimo por aristas a través del teorema de Menger para aristas (véase la pág. 35). Esto implica que la arista-conectividad se puede calcular aplicando el algoritmo de Ford-Fulkerson de flujo máximo, colocando pesos unitarios en las aristas. A través de la aplicación reiterada del algoritmo de Ford-Fulkerson, Gomory y Hu mostraron que es posible construir un árbol pesado en las aristas que contiene toda la información de la conectividad en el grafo [80].

La Figura 4.6 muestra un árbol de Gomory-Hu de un grafo sencillo. La lectura de este árbol se realiza de la siguiente forma: la arista-conectividad entre dos vértices v y w equivale al mínimo de entre las capacidades de las aristas del único camino que conecta a v con w en el árbol.

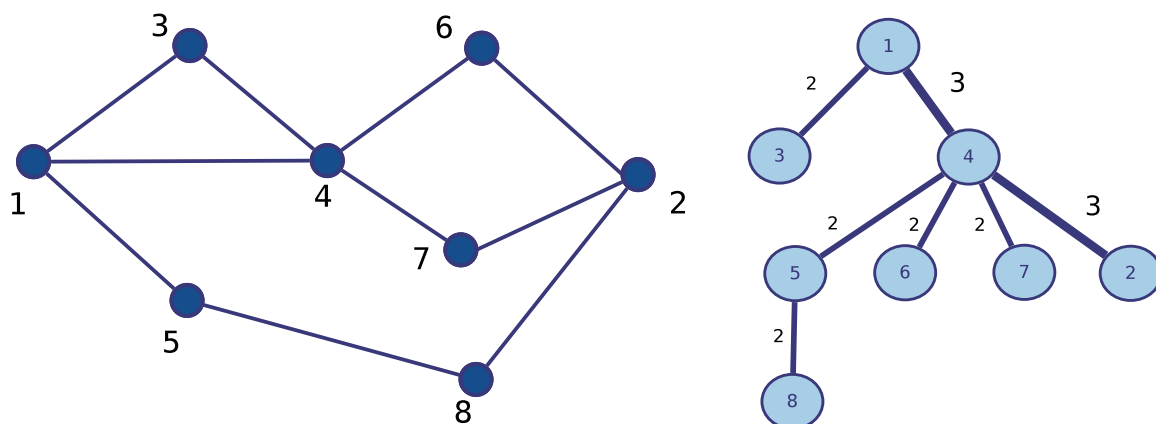


Figura 4.6: *Cálculo de la arista-conectividad con árboles de Gomory-Hu.* A la izquierda mostramos un grafo sencillo, y a la derecha un árbol de Gomory-Hu del mismo. El árbol contiene la información sobre la arista-conectividad entre todo par de vértices v y w como el mínimo de entre las capacidades de las aristas del camino entre v y w . En particular, el mínimo de entre las capacidades de todas las aristas del árbol equivale a la arista-conectividad del grafo.

4.3. Visualización de la conectividad en Internet

Hemos utilizado la herramienta de visualización LaNet-vi [5] para graficar la descomposición en k -núcleos de los grafos de Internet. Desde la versión 2.2.0 de LaNet-vi

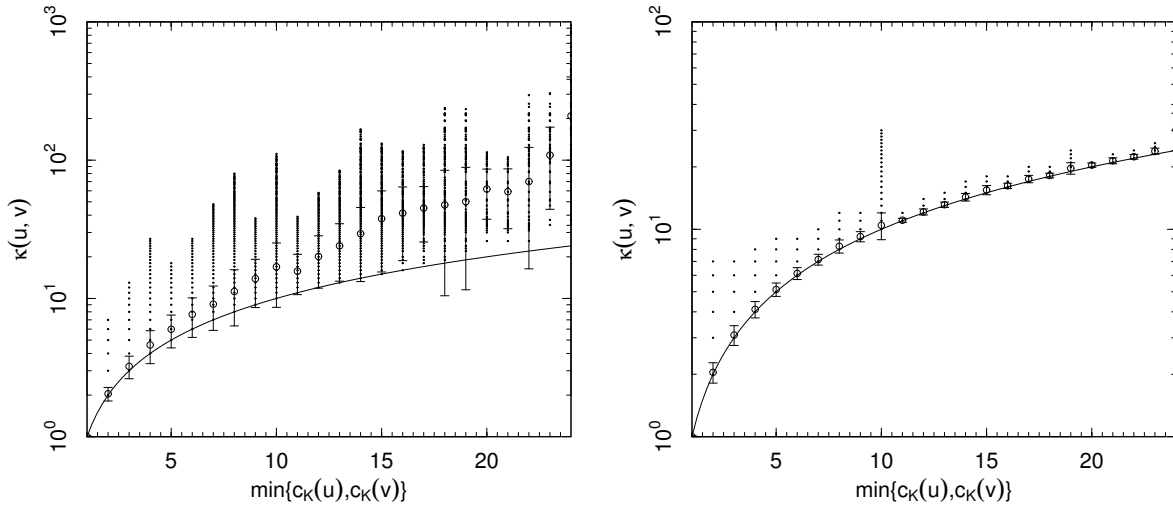


Figura 4.7: *Arista-conectividad en la red AS-CAIDA 2013*. El gráfico de la izquierda muestra la arista-conectividad entre todo par de vértices $\{u, v\}$ en la red, en función del mínimo entre sus índices de capa, $\min\{c_K(u), c_K(v)\}$. A la derecha se grafica lo que denominamos *arista-conectividad hacia adentro*, es decir la arista conectividad dentro del k -núcleo más pequeño que incluye a los dos vértices u y v . La línea continua representa la función $f(x) = x$. Los segmentos representan la media y la desviación estándar para cada valor de la abscisa. Observamos que el mínimo entre los índices de capa de dos vértices es muy un buen estimador de la arista-conectividad hacia adentro entre ambos. La arista-conectividad se calculó construyendo un árbol de Gomory-Hu del grafo [80].

incorporamos la posibilidad de encontrar subgrafos *núcleo-conexos* en sentido estricto y en sentido amplio utilizando los algoritmos aquí presentados.

En las Figuras 4.9 y 4.10 mostramos las descomposiciones en k -núcleos de las redes AS-CAIDA 2011 y AS-DIMES 2011. Los vértices que no pertenecen al subgrafo *núcleo-conexo* en sentido estricto se indican en color negro. Nuevamente observamos que son muy escasos, y vemos que se encuentran en las capas inferiores de la red.

Los gráficos nos muestran también que el nivel de Sistemas Autónomos de Internet presenta una núcleo-profundidad elevada, en aumento con el transcurso de los años. Entre las exploraciones de CAIDA de 2009 y de 2013, la núcleo-profundidad de la red pasó de 16 a 24. En la Figura 4.11 podemos observar cómo ha evolucionado el núcleo más profundo de Internet entre 2009 y 2013. Observamos a través de las etiquetas junto a los vértices que los ASes que forman parte del centro de la red suelen mantenerse en él, aunque se han agregado muchos nuevos. Los Sistemas Autónomos que se encuentran dentro de este núcleo son los mayores proveedores de conectividad en Internet.

Por último, vemos que las exploraciones de DIMES son bastante más detalladas que las de CAIDA. En 2011 se encuentra una núcleo-profundidad de 35, a comparación de la núcleo-profundidad de 20 en CAIDA. La k -arista-conectividad se sigue verificando, a

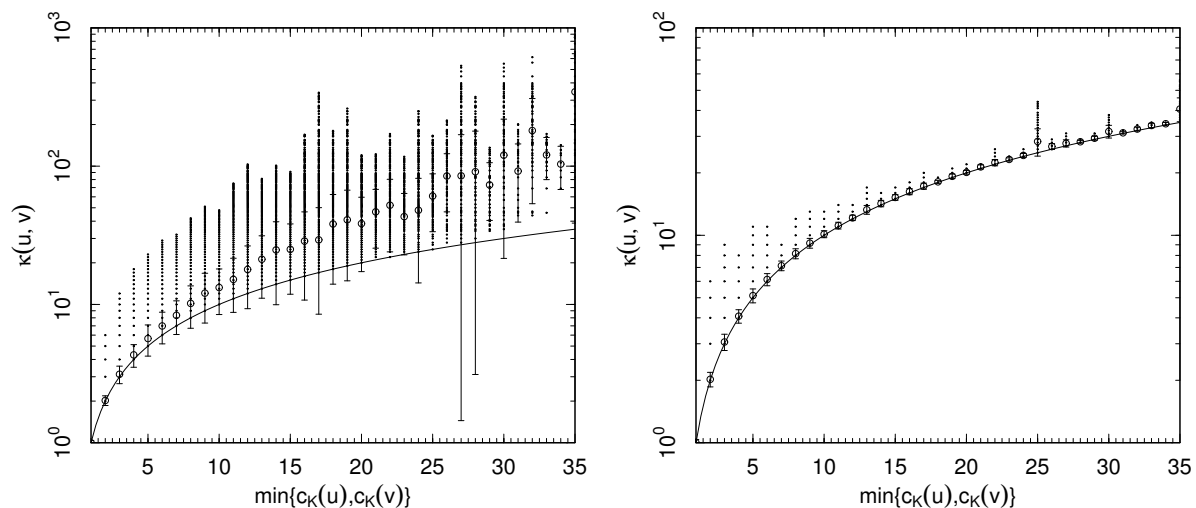


Figura 4.8: *Arista-conectividad en la red AS-DIMES 2011*. Arista-conectividad (*Izq*) y arista-conectividad hacia adentro (*Der*) entre todo par de vértices $\{u, v\}$ en la red, en función del mínimo entre sus índices de capa, $\min\{c_K(u), c_K(v)\}$. Para más detalles véase la descripción de la Figura 4.7.

excepción de en unos pocos vértices.

En este capítulo hemos mostrado de qué manera es posible obtener cotas inferiores para la arista-conectividad en tiempo lineal con el tamaño de grafo. Mostramos también que en los grafos de Internet a nivel de Sistemas Autónomos estas cotas se ajustan muy bien a la *arista-conectividad hacia adentro*.

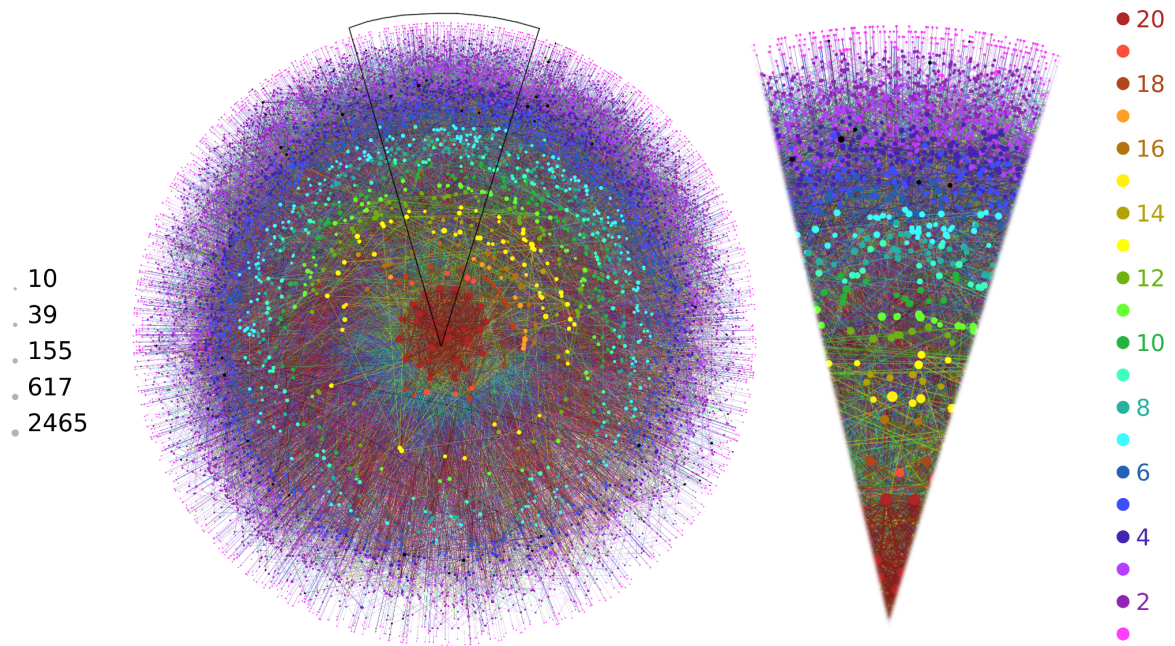


Figura 4.9: *Descomposición en k -cores y conjunto núcleo-conexo en sentido estricto de la red AS-CAIDA 2011*. La escala de la izquierda representa el grado de los vértices; la de la derecha corresponde al índice de capa.

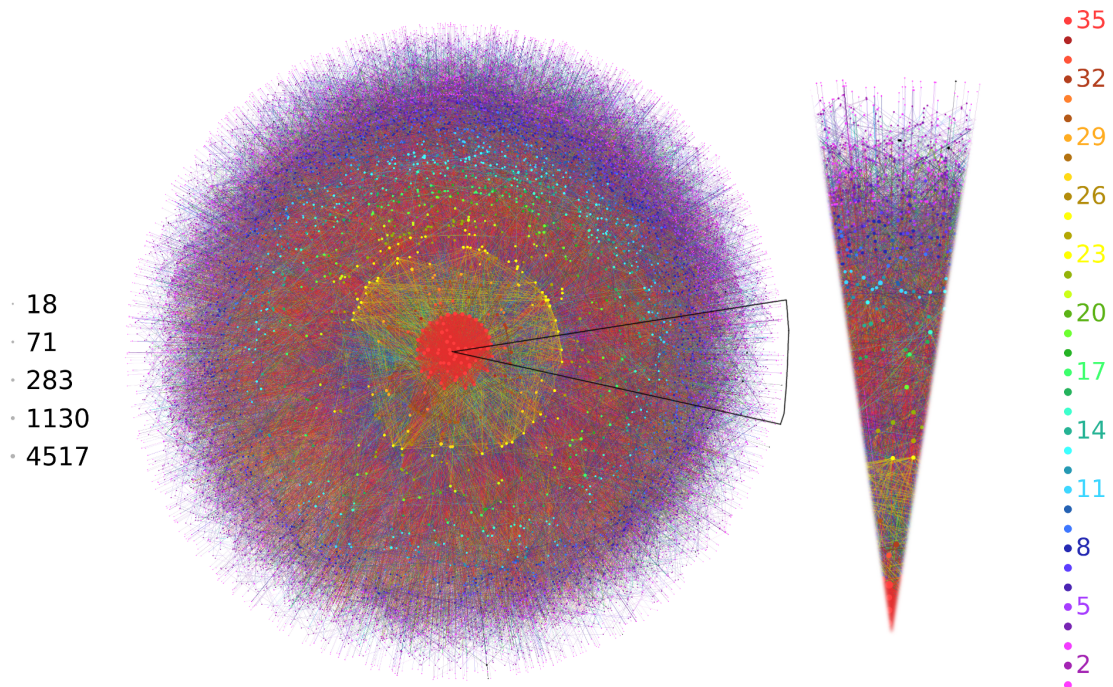


Figura 4.10: *Descomposición en k -cores y conjunto núcleo-conexo en sentido estricto de la red AS-DIMES 2011*. La escala de la izquierda representa el grado de los vértices; la de la derecha corresponde al índice de capa.

Algoritmo 4: Núcleo-conectividad en sentido amplio

Entrada: $S_k[G] = \{Q_{k1}, Q_{k2}, \dots, Q_{kM_k}\}$, las k -capas de G (desde 1 hasta $k_{\text{máx}}$),
particionadas en sus componentes conexas (*clusters*)**Salida:** $C \subset V$, núcleo-conexo en sentido amplio

```

4.1  $C \leftarrow \emptyset$ 
4.2  $D \leftarrow \emptyset$ 
4.3  $\mathfrak{B} \leftarrow \emptyset$ 
4.4  $k \leftarrow k_{\text{max}}$ 
4.5 inicio
4.6   mientras  $C = \emptyset$  y  $k \geq 2$  hacer
4.7     si existe algún  $Q \in S_k[G]$  tal que  $\text{diam}(G[Q]) \leq 2$  y  $d_{\min}(G[Q]) \geq k$ 
4.8       entonces
4.9          $C \leftarrow C \cup Q$ 
4.9          $S_k[G] \leftarrow S_k[G] \setminus Q$ 
4.10      fin
4.11       $\mathfrak{B} \leftarrow \mathfrak{B} \cup S_k[G]$ 
4.12       $k \leftarrow k - 1$ 
4.13   fin
4.14   mientras  $k \geq 2$  hacer
4.15     mientras existe algún  $Q' \in \mathfrak{B}$  tal que:  $\left\{ \begin{array}{l} \text{diam}_{(C \cup D \cup Q')/(C \cup D)} \leq 2 \\ \Psi_{(C \cup D \cup Q')/(C \cup D)}(k) \geq 0 \end{array} \right\}$  hacer
4.16        $D \leftarrow D \cup Q'$ 
4.17        $\mathfrak{B} \leftarrow \mathfrak{B} \setminus \{Q'\}$ 
4.18     fin
4.19     mientras existe algún  $Q \in S_k[G]$  tal que:  $\left\{ \begin{array}{l} \text{diam}_{(C \cup D \cup Q)/(C \cup D)} \leq 2 \\ \Psi_{(C \cup D \cup Q)/(C \cup D)}(k) \geq 0 \end{array} \right\}$ 
4.20       hacer
4.21          $C \leftarrow C \cup Q$ 
4.21          $S_k[G] \leftarrow S_k[G] \setminus \{Q\}$ 
4.22       fin
4.23        $\mathfrak{B} \leftarrow \mathfrak{B} \cup S_k[G]$ 
4.24        $k \leftarrow k - 1$ 
4.25     fin
4.26     para cada  $Q \in S_1[G]$  hacer
4.27       si  $|\partial^1 Q| \geq 1$  entonces
4.28          $C \leftarrow C \cup Q$ 
4.29       fin
4.30     fin
4.31 fin

```

	AS-CAIDA 2009	AS-CAIDA 2011	AS-CAIDA 2013	AS-DIMES 2011
$n(G)$	16117	19895	23779	26083
$e(G)$	32847	44560	54752	83305
$\bar{d}$	4,08	4,48	4,61	6,39
d_{max}	2012	2465	2818	4517
k_{max}	16	20	24	35
$cc(G)$	0,013	0,014	0,016	0,015

Cuadro 4.1: *Listado de grafos de exploraciones de Internet.* Para más detalles sobre las estadísticas de cada grafo consúltese el Apéndice B.

	$ V(G) $	$ V(G) \setminus C_{strict} $	$ V(G) \setminus C_{wide} $
AS-CAIDA 2009	16117	145	94
AS-CAIDA 2011	19895	111	72
AS-CAIDA 2013	23779	28	24
AS-DIMES 2011	26083	45	34

Cuadro 4.2: *Núcleo-conectividad de los grafos de Internet.* Nuestro algoritmo obtiene para cada exploración un subgrafo núcleo-conexo $G[C]$. La núcleo-conectividad implica que los diversos k -núcleos de este subgrafo son k -arista-conexos. En esta tabla se muestra en la segunda columna la cantidad de vértices de la exploración, y en las siguientes dos columnas la cantidad de vértices que no pertenecen al subgrafo núcleo-conexo en sentido estricto y en sentido amplio, respectivamente.

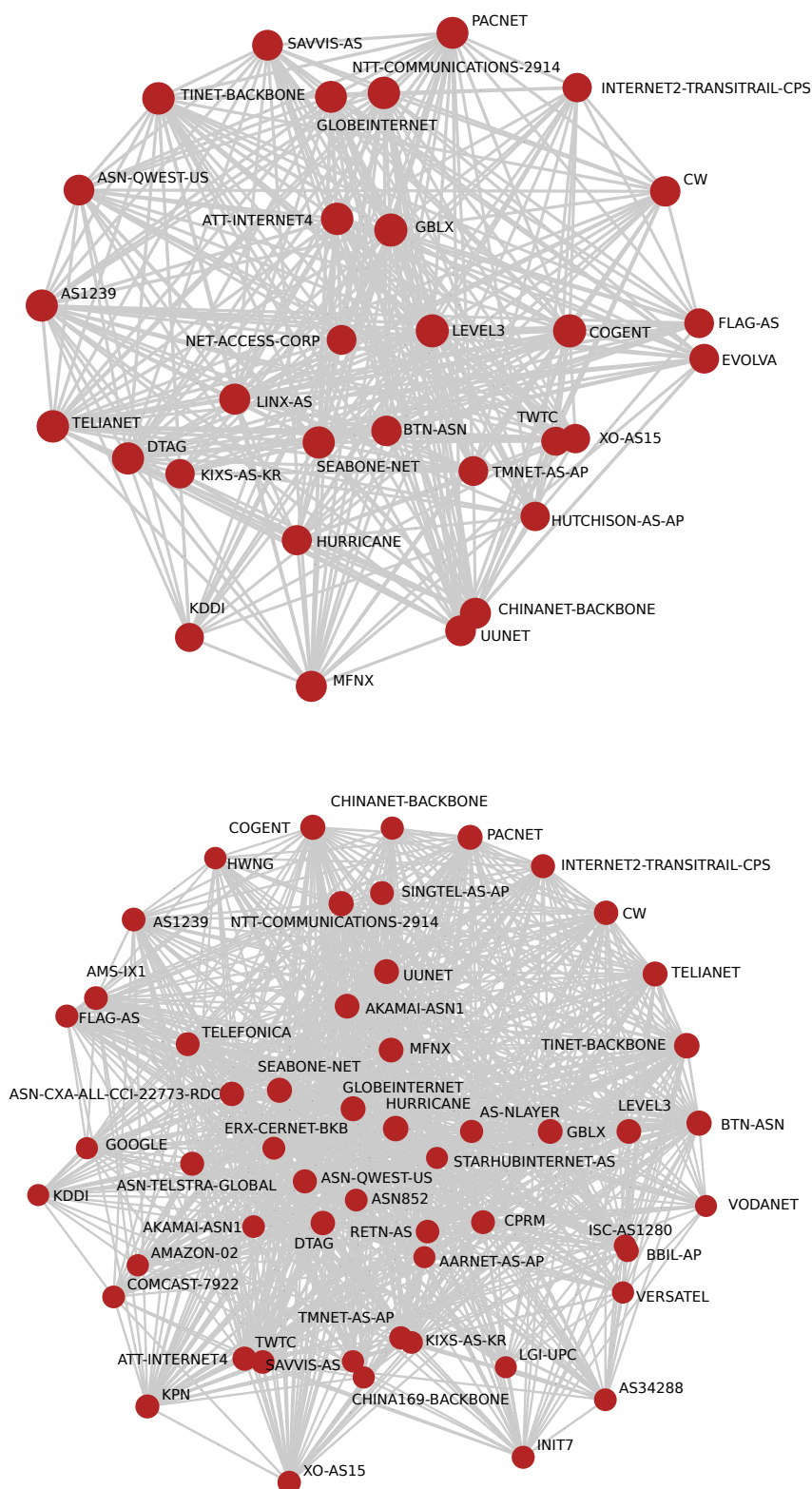


Figura 4.11: *Evolución del núcleo central de Internet según CAIDA entre 2009 (arriba) y 2013 (abajo).* La asignación de nombres a los Sistemas Autónomos a partir de su número se realizó utilizando datos de 2013.

Capítulo 5

Estudio del Agrupamiento en Redes Complejas

Los sistemas complejos se encuentran en un punto intermedio entre el orden y el desorden. Esto hace que exhiban características típicas como el fenómeno de mundo pequeño y distribuciones de grados libres de escala. El orden, en particular, se manifiesta fundamentalmente por la presencia de correlaciones en el grafo.

El concepto de orden se vincula a la existencia de una estructura métrica en la red. Como el coeficiente de agrupamiento (véase pág. 37) es uno de los invariantes más sencillos que captura la desigualdad triangular, es muy utilizado para estudiar el orden en las redes complejas.

En este capítulo discutiremos algunos de los modelos de agrupamiento existentes y utilizaremos la descomposición en k -densos para mostrar que algunos de ellos se ajustan mejor a las redes reales que otros.

Uno de los objetivos de este capítulo es resaltar la importancia de la visualización como herramienta en el estudio de los sistemas complejos. Implementamos una visualización de la descomposición en k -densos como variante de la descomposición en k -núcleos, en el software LaNet-vi 3.0 [5]. Mostraremos que las diferencias entre los modelos se hacen patentes a simple vista en las visualizaciones.

Los resultados descriptos aquí se encuentran publicados en [50].

5.1. Introducción

Los modelos de grafos aleatorios clásicos como el Erdos-Renyi y sus generalizaciones¹ no presentan correlaciones y por lo tanto generan grafos con un pobre agrupamiento;

¹Véase la Sección 2.3.3.

las vecindades de los vértices tienen más bien un aspecto de árbol, con pocas conexiones entre los vecinos. Sin embargo estos modelos presentan la ventaja de ser tratables matemáticamente por la sencillez de su descripción.

Los primeros modelos del agrupamiento intentaron entonces incorporar correlaciones en forma sencilla, de forma de poder calcular las propiedades en el límite termodinámico. El modelo de Newman [118] (2009) y el de Gleeson [77] (2009) utilizan métodos que denominamos *basados en cliques* (*CB*, por *clique-based clustering*).

El método de Gleeson toma como parámetro de entrada una distribución conjunta $\gamma(c, k)$, que representa la probabilidad de que un vértice tomado al azar tenga grado k y pertenezca a un *clique* de tamaño c . Utilizando esta distribución se construye un grafo formado por *cliques* que se encuentran embebidos dentro de un grafo mayor en que, si consideramos a los *cliques* como vértices, los mismos resultan conectados bajo el modelo de configuración clásico. Escogiendo la distribución $\gamma(c, k)$ adecuada, se obtiene un grafo con una distribución de grados esperada $p(k)$ y un coeficiente de agrupamiento promedio en función del grado.

Los métodos basados en *cliques*, con una estructura *modular* formada por *cliques* conectados entre sí, representan un alto nivel de ordenamiento en el grafo. Es posible sin embargo construir grafos con alto agrupamiento pero con la mínima correlación necesaria entre las aristas. A este tipo de métodos los denominamos *de agrupamiento aleatorio máximo* (*MR*, por *maximally random clustering*). El modelo general que proponemos aquí se basa en un conjunto de *grafos aleatorios exponenciales* [125]. Un grafo aleatorio exponencial bajo ciertos invariantes esperados es un grafo aleatorio en el que la distribución de probabilidades del grafo es aquella que maximiza la entropía para dicho valor esperado de los invariantes. En nuestro caso el invariante es la distribución del coeficiente de agrupamiento de los vértices, que tomamos de la red real, y la distribución de probabilidades queda entonces representada por el siguiente hamiltoniano:

$$H(G^*) = \sum_{k=1, p(k) \neq 0}^{k=d_{\max}(G)} |\overline{cc}^*(k) - \overline{cc}(k)| \ ,$$

en donde $\overline{cc}^*(k)$ es el coeficiente de agrupamiento promedio de los vértices de grado k en el grafo G^* mientras que $\overline{cc}(k)$ es el coeficiente de agrupamiento promedio en la red original. La minimización del hamiltoniano se realiza a través de un proceso de recocido simulado (*simulated annealing*). Mayores detalles sobre el recableado de los vértices durante el proceso pueden encontrarse en [50].

Ambos tipos de métodos –basados en *cliques* (CB) y de agrupamiento aleatorio máximo (MR)– son en cierta forma opuestos dentro del espacio de los grafos con una distribu-

ción de grados fija $p(k)$ y coeficiente de agrupamiento promedio $\overline{cc}(k)$. Nos preguntamos entonces a cuál de entre los dos responden las redes complejas reales. Para responder esta pregunta utilizaremos como herramienta la descomposición en k -densos, que fue presentada en la Sección 2.1.3.5.

5.2. Cálculo de la descomposición en k -densos

Recordemos que un k -denso es un subgrafo maximal cuyas aristas tienen multiplicidad al menos $k - 2$. Para calcular la descomposición en k -densos hemos desarrollado un enfoque novedoso. En el trabajo original de Saito *et al.* [141] se obtiene cada k -denso por eliminación sucesiva de las aristas de multiplicidad menor a $k - 2$, pero al eliminar cada arista se debe recalcular la multiplicidad de todas las aristas adyacentes a ella. Utilizando en cambio una estructura que almacena los triángulos asociados a cada arista, podemos realizar esta actualización en forma mucho más ágil.

Nuestro algoritmo de descomposición trabaja con un *hipergrafo* H que se construye a partir del grafo original. Un hipergrafo es una generalización del concepto de grafo en que cada arista está asociada a un subconjunto no vacío del conjunto de vértices, mientras que en un grafo estándar cada arista está asociada a exactamente dos vértices. En nuestro caso, H tendrá un vértice por cada arista del grafo original. Las aristas de H conectarán una terna de vértices cada una. Tres vértices del hipergrafo estarán conectados por una arista cuando las aristas asociadas a esos vértices en el grafo original constituyan un triángulo. En resumen, en nuestro hipergrafo cada arista del grafo original es un vértice, y cada triángulo del grafo original es una arista.

Bajo este esquema, hemos demostrado que la descomposición en k -densos del grafo original es equivalente a la descomposición en k -núcleos del hipergrafo (véase [50], *Supplementary Information*). El conjunto de vértices del k -núcleo del hipergrafo nos determina el conjunto de aristas del $(k + 2)$ -denso del grafo. La Figura 5.1 ilustra el procedimiento.

Dado que la complejidad computacional de la descomposición en k -núcleos es de $O(e(H))$ y que la cantidad de aristas de H es la cantidad de triángulos de G , concluimos que nuestro algoritmo tiene una complejidad del orden de la cantidad de triángulos en G .

5.3. Visualización de los modelos de agrupamiento

Hemos analizado 3 redes reales de diverso origen: una exploración de Internet en el nivel de Sistemas Autónomos obtenida por CAIDA en 2009, la red de confianza de PGP [25] y la red metabólica de la bacteria *E. Coli* [145]. Tomando la distribución real

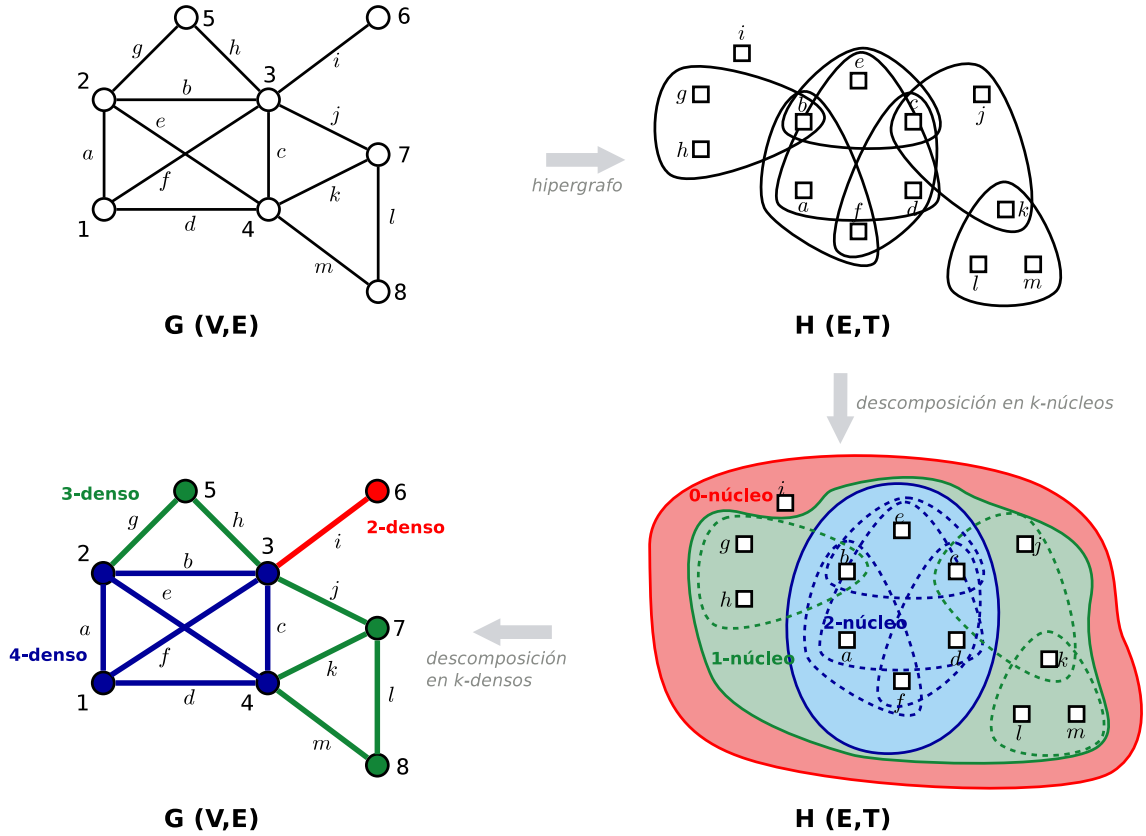


Figura 5.1: *Procedimiento para la descomposición en k -densos.* En una primera instancia se construye el hipergrafo H , cuyo conjunto de vértices es el conjunto de aristas $E(G)$, y cuyo conjunto de aristas es el conjunto T de los triángulos en G . Luego se calcula la descomposición en k -núcleos de H . Finalmente, se asigna a cada arista en G un índice de denso igual al índice de capa de su vértice correspondiente en H .

de los grados de los vértices y del coeficiente de agrupamiento promedio en función del grado, se construyeron instancias de redes de similar tamaño siguiendo: (a) el procedimiento basado en *cliques* de Gleeson [77]; y (b) nuestro modelo de agrupamiento aleatorio máximo.

En nuestras visualizaciones hemos utilizado como índice para los densos la multiplicidad m de sus aristas en lugar del valor de k . Una multiplicidad de m se corresponde con un $(m + 2)$ -denso o, dicho de otra forma, un k -denso tiene multiplicidad $k - 2$.

La lectura de los gráficos se realiza de la siguiente manera: cada k -denso tiene un espacio circular en el que se dibuja, aunque el borde de dicho círculo no se muestra. Los círculos que se observan en las imágenes corresponden a las componentes conexas dentro del k -denso correspondiente a su color.

Cuando dentro de un mismo k -denso existen numerosas componentes conexas pero una de ellas es la principal, las pequeñas componentes conexas se dibujan alrededor de la principal, de manera que se observan pequeños círculos alrededor de uno más grande.

Es el caso, por ejemplo, de la red original PGP y su modelo MR.

Cuando no existe una componente conexa principal, en cambio, se observa que los pequeños círculos de un k -denso rodean no ya a una componente conexa, sino al espacio circular del k -denso inferior, y se observan entonces pequeños círculos de distinto radio. Esto se ve claramente en el modelo CB de la red metabólica.

Comenzamos ahora la descripción de cada figura. En el caso de la red de Sistemas Autónomos vemos que la red original presenta una estructura jerárquica en la que la componente conexa principal de cada k -denso está contenida dentro de la componente conexa principal del k -denso inferior. Este hecho es bien reproducido por el modelo MR, mientras que el modelo CB genera una enorme cantidad de pequeñas componentes conexas disconexas en cada k -denso.

Lo que ocurre en la red de confianza de PGP resulta interesante. Al tratarse de una red social, la red original combina una estructura modular (determinada por la existencia de numerosas componentes conexas pequeñas en los k -densos) junto con una estructura jerárquica. La estructura jerárquica implica una importante densidad de conexiones “radiales” entre los k -densos. Como consecuencia de ella, cada k -denso tiene una componente conexa principal, inmersa en la componente conexa principal del k -denso inferior (el $(k - 1)$ -denso). Sin embargo el modelo CB produce una simple estructura modular sin jerarquía. Todas las componentes conexas son muy pequeñas.

Por último, en la red metabólica –mucho más pequeña que las anteriores– se observa claramente la estructura jerárquica de la red original, que no logra ser capturada por el modelo CB. Nuevamente este modelo obtiene una estructura modular que no es propia de este tipo de redes biológicas.

En conclusión, hemos mostrado a través de estas visualizaciones que el modelo CB, a pesar de reproducir correctamente el coeficiente de agrupamiento, no logra modelar correctamente aquellas redes que poseen una estructura jerárquica.

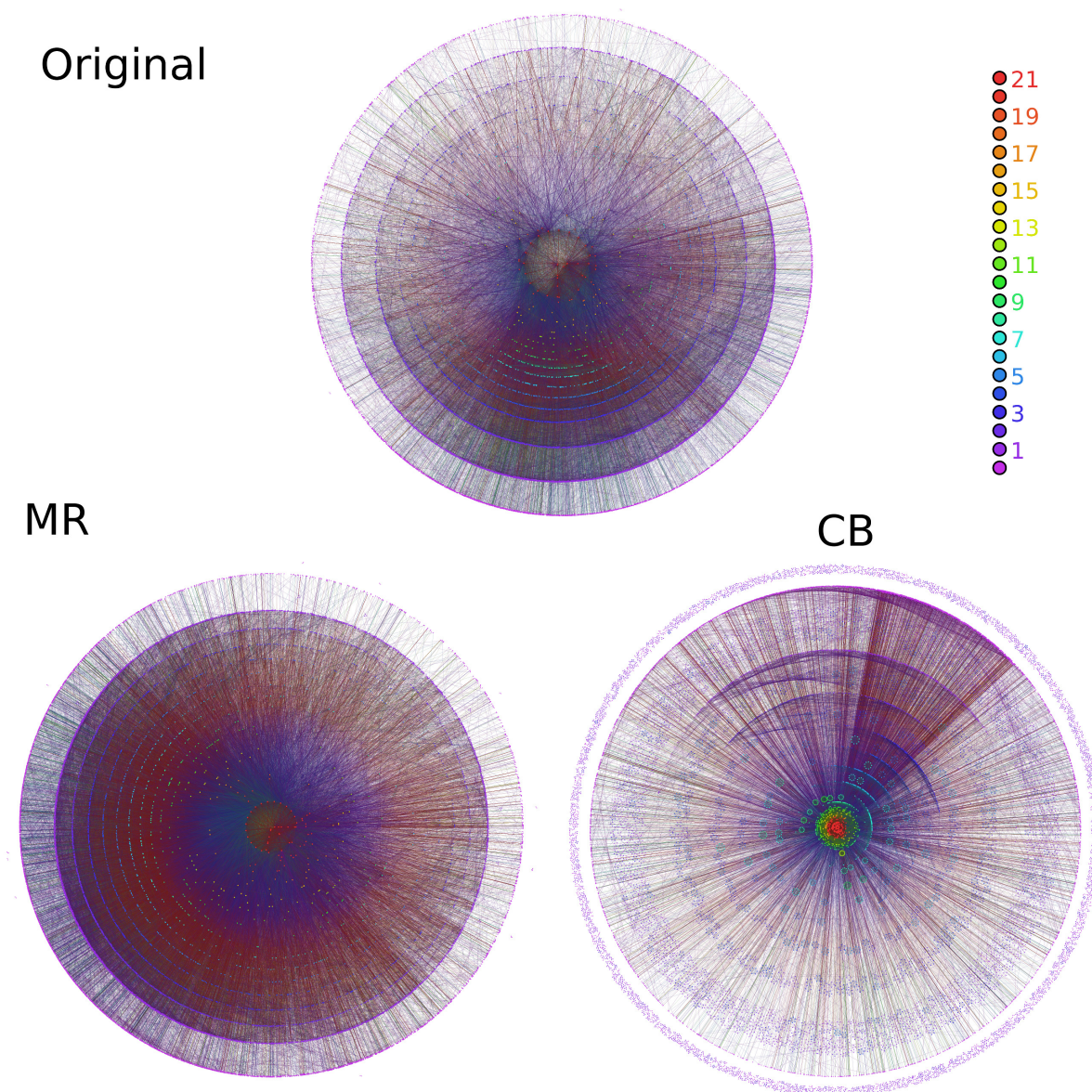


Figura 5.2: *Descomposición en k -densos del grafo de Sistemas Autónomos de Internet*. Se muestran la red original (*Arriba*), la obtenida bajo el modelo de agrupamiento aleatorio máximo (MR) (*Izquierda*) y la obtenida con el modelo basado en *cliques* (CB) (*Derecha*). La escala de colores se determina en base a la denso-profundidad de la red original, que en este caso es 21. En las redes obtenidas a través de los modelos, todos los vértices con índice de denso mayor ó igual a 21 se colorean también en rojo. Las denso-profundidades en los modelos son 27 (MR) y 58 (CB).

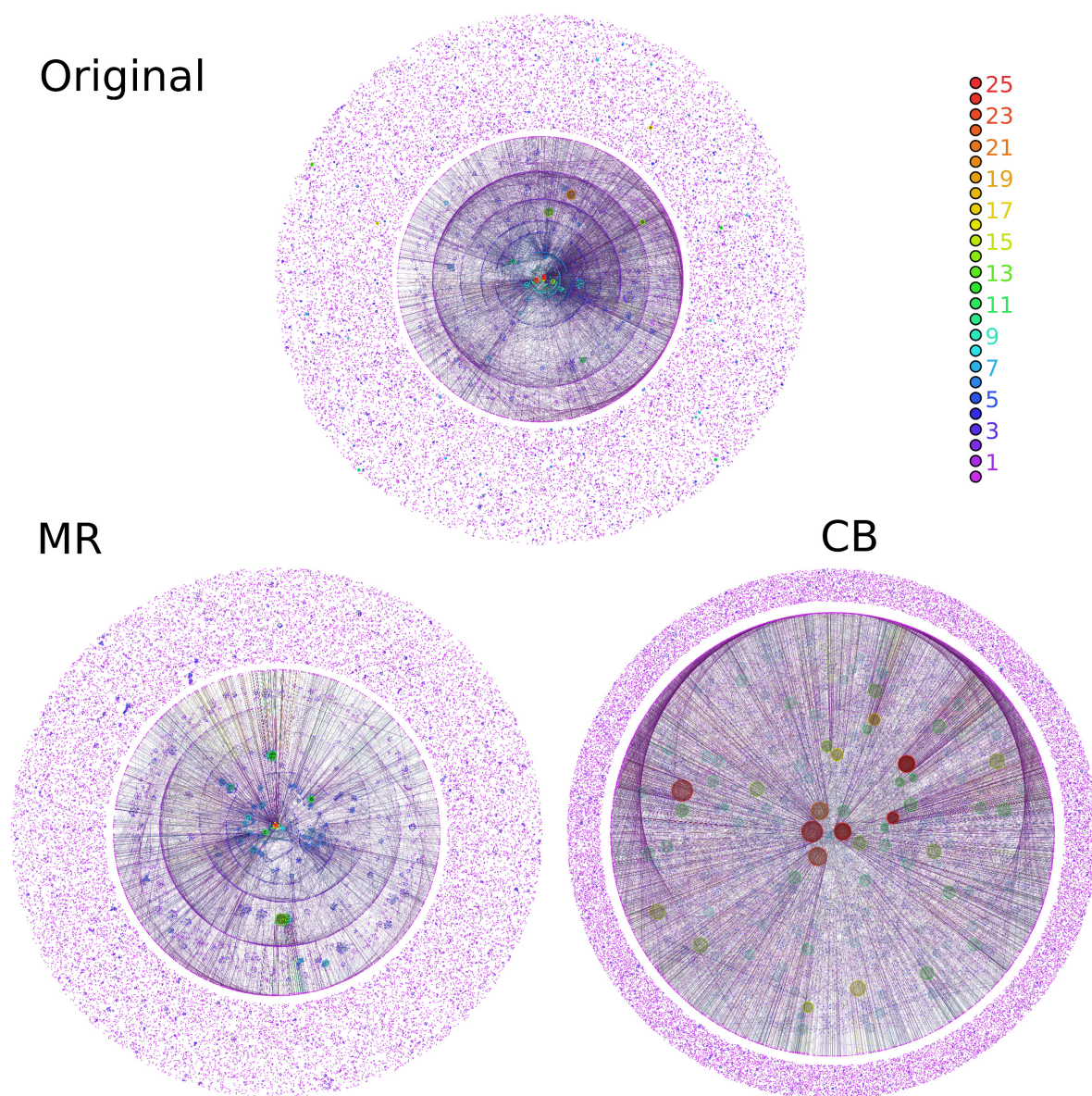
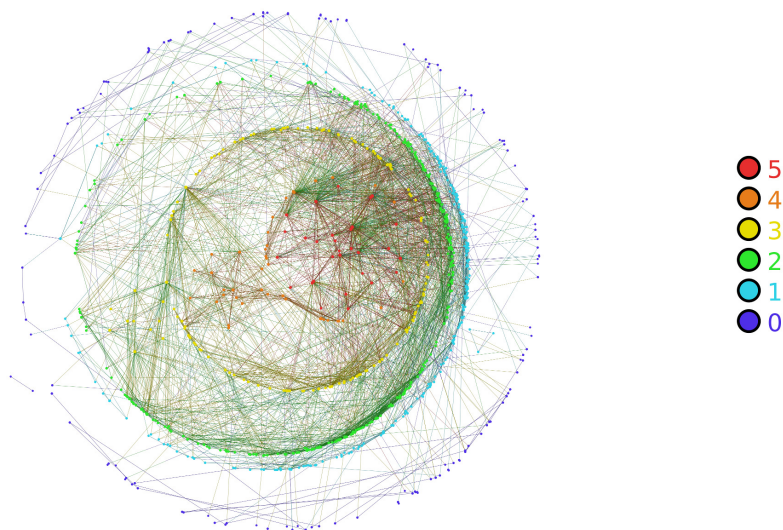
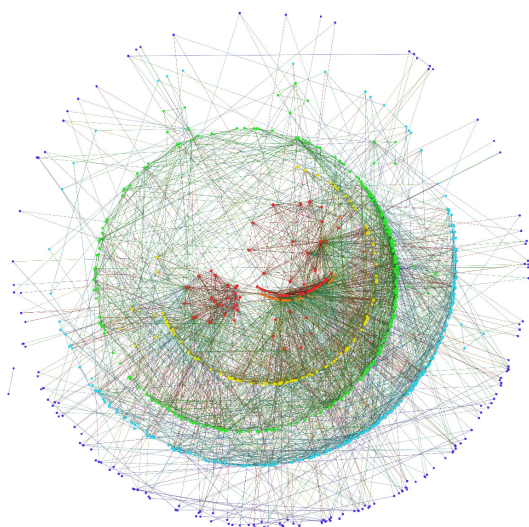


Figura 5.3: *Descomposición en k -densos del grafo de la red PGP*. Los gráficos corresponden a la red original (*Arriba*), el modelo de agrupamiento aleatorio máximo (MR) (*Izquierda*) y el modelo basado en *cliques* (CB) (*Derecha*). La escala de colores se determina en base a la denso-profundidad de la red original, que en este caso es 25. Las denso-profundidades en los modelos son 23 (MR) y 36 (CB).

Original



MR



CB

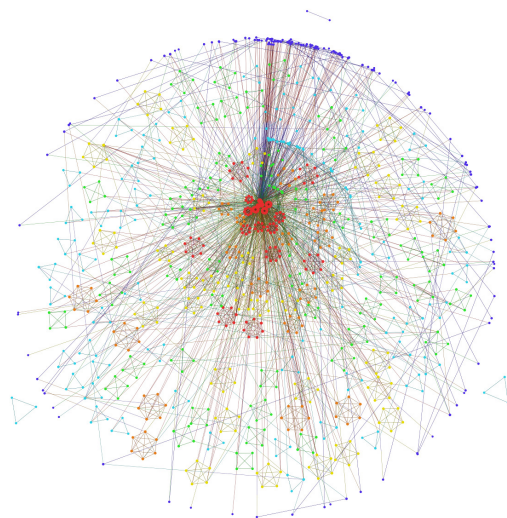


Figura 5.4: *Descomposición en k -densos del grafo de la red metabólica de *E. Coli*. Los gráficos corresponden a la red original (Arriba), el modelo de agrupamiento aleatorio máximo (MR) (Izquierda) y el modelo basado en cliques (CB) (Derecha). La escala de colores se determina en base a la denso-profundidad de la red original, que en este caso es 5. Las denso-profundidades en los modelos son 9 (MR) y 14 (CB).*

Capítulo 6

Conclusiones

En esta tesis hemos estudiado la problemática de la modelización combinatoria de los sistemas complejos y hemos presentado algunos modelos adecuados para caracterizar redes complejas.

Hemos puesto especial atención al problema de la complejidad computacional de los modelos, y en cada uno de nuestros aportes procuramos que las soluciones propuestas fueran escalables y pudieran aplicarse a redes de gran tamaño.

Los métodos que desarrollamos pueden ser clasificados en tres grupos:

- El descubrimiento de estructura comunitaria.
- La caracterización de invariantes de las redes complejas, como la arista-conectividad y el coeficiente de agrupamiento (*clustering*).
- La visualización de redes.

En el Capítulo 3 hemos caracterizado la denominada *estructura comunitaria* en las redes complejas. El desarrollo de modelos de estructura comunitaria es de importancia práctica porque permite explicar el comportamiento colectivo y predecir grupos de afinidad en las redes sociales. En las redes biológicas también es utilizado para inferir funcionalidad a partir de la estructura. Entre nuestros aportes dentro de esta área destacamos:

- *El desarrollo de una formalización de la modularidad*, que nos permitió expresar en forma precisa y simple el problema del límite de escala [33]. Una formalización similar fue luego utilizada para describir nuestro proceso de crecimiento [20].
- *La propuesta de un método local de descubrimiento de comunidades*. Nuestro método basado en un proceso de crecimiento de una función objetivo tiene como ventaja principal que puede aplicarse a redes de gran escala. Lo hemos comparado con los

métodos más conocidos de descubrimiento de comunidades y vimos que, respecto a aquellos basados en la optimización de la modularidad, el proceso de crecimiento soluciona el problema del límite de escala. Respecto a métodos como InfoMAP y LPM, cuyos resultados son cualitativamente muy buenos, nuestro método local tiene una complejidad baja y acotada.

- *La justificación del método.* Mostramos que el proceso de crecimiento tiene un comportamiento correcto en el límite termodinámico cuando los vértices de la comunidad poseen un parámetro de mezclado característico. Optimizamos el algoritmo y las estructuras de datos para obtener una complejidad de $O(n(G) \cdot d_{\text{máx}} + e(G) \cdot \log(n(G)))$, y lo aplicamos en redes de hasta 5 millones de nodos. Hemos observado que en muchas redes reales nuestro método obtiene comunidades de muy variado tamaño y con distribuciones que pueden ajustarse a leyes de potencias, de acuerdo con lo esperado [20].

En el Capítulo 4 investigamos la topología de Internet a través de la descomposición en k -núcleos e hicimos un estudio minucioso de la relación entre ellos y la arista-conectividad. Nuestro aporte fundamental ha sido desarrollar un algoritmo de baja complejidad que permite garantizar un mínimo de arista-conectividad entre los vértices de la red a partir de la verificación de condiciones simples. Hemos mostrado que estas condiciones se cumplen para la mayoría de los vértices del grafo de Internet, tanto a nivel de Sistemas Autónomos como a nivel de *routers*. La obtención de cotas inferiores para la conectividad en las redes de flujo de información como Internet es de gran importancia práctica, porque permite a los proveedores garantizar una cierta robustez ó calidad de servicio a los usuarios. Con nuestro algoritmo de núcleo-conectividad en sentido estricto podemos obtener estas cotas en un tiempo de $O(e(G))$ [6].

En el Capítulo 5, por último, estudiamos algunos modelos de agrupamiento y los comparamos con redes complejas reales a través de la descomposición en k -densos. Hemos propuesto un algoritmo eficiente para calcular la descomposición, cuya complejidad es del orden de la cantidad de triángulos en el grafo. Utilizamos la visualización de los k -densos para mostrar que el fenómeno del agrupamiento es mejor modelado por los métodos de agrupamiento aleatorio máximo que por aquellos basados en *cliques* [50].

A lo largo de toda la tesis hemos hecho énfasis en los modelos de visualización. Hemos mejorado y agregado funcionalidades a la herramienta de visualización LaNet-vi, incorporando la descomposición en k -densos y la visualización de la núcleo-conectividad, junto con otras mejoras menores. LaNet-vi fue utilizado ampliamente en el Capítulo 4 para visualizar los conjuntos núcleo-conexos que encontramos en el grafo de Internet, y en el Capítulo 5 para comparar modelos de agrupamiento a partir de la descomposición

en k -densos.

Todos los métodos desarrollados están publicamente disponibles a la comunidad científica desde los siguientes accesos:

- CommUGP (descubrimiento de comunidades locales utilizando un proceso de crecimiento uniforme): <https://code.google.com/p/commugp/>
- LaNet-vi (visualización de k -núcleos y k -densos y cálculo de conjuntos núcleo-conexos): <http://lanet-vi.fi.uba.ar/>
- SnailVis (visualización de estructura comunitaria): <http://cnet.fi.uba.ar/mariano.beiro/snailvis.tar.gz>
- DeltaCom (algoritmo de optimización golosa de la modularidad): <http://sourceforge.net/projects/deltacom/>

Los resultados de este trabajo fueron publicados en los siguientes artículos en revistas internacionales:

M.G. Beiró, J.R. Busch, S.P. Grynberg, and J.I. Alvarez-Hamelin. Obtaining communities with a fitness growth process. *Physica A: Statistical Mechanics and its Applications*, 392(9):2278 – 2293, 2013.

J.I. Alvarez-Hamelin, M.G. Beiró, and J.R. Busch. Understanding edge connectivity in the internet through core decomposition. *Internet Mathematics*, 7(1):45–66, 2011.

P. Colomer de Simón, M.A. Serrano, M.G. Beiró, J.I. Alvarez-Hamelin, and M. Boguñá. Deciphering the global organization of clustering in real complex networks. *Scientific Reports*, 3(2517), 2013.

También se puede consultar:

J.R. Busch, M.G. Beiró, and J.E. Alvarez-Hamelin. On weakly optimal partitions in modular networks. *CoRR*, abs/1008.3443, 2010.

M.G. Beiró, J.R. Busch, J.I. Alvarez-Hamelin. SnailVis: a paradigm to visualize complex networks. Simposio Argentino de Tecnología, 39° JAIIO (Jornadas Argentinas de Informática e Investigación Operativa), Buenos Aires, 2010.

Apéndice A

Leyes de Potencias

En los sistemas complejos se observan a menudo parámetros cuya función de densidad de probabilidad sigue una ley de la forma $f(x) \propto x^{-\alpha}$, que por su estructura suele denominarse *ley de potencias*. A diferencia de distribuciones más clásicas como la binomial o la normal, las leyes de potencias tienen una lenta atenuación con respecto a valores crecientes de la variable aleatoria. Esto da lugar a interesantes fenómenos, como ser que la concentración de probabilidad para grandes valores de la variable apartados de la media sea no despreciable, o que la media no tenga demasiada relevancia como estimador de una muestra, por ser la varianza demasiado grande.

Una de las primeras observaciones de este comportamiento la realizó V. Pareto en 1906 al estudiar la distribución de la riqueza en la población. Su observación de que “el 80 % de la riqueza en Italia estaba concentrada en el 20 % de la población” es una consecuencia de una ley de potencias en la distribución de la riqueza. Las leyes de potencias pueden observarse al estudiar la concentración de población en las ciudades [116], las magnitudes de los terremotos [88], las citas en las publicaciones científicas [55], o los hipervínculos entre páginas web [3]. En los sistemas complejos, los valores del exponente de las leyes de potencias observadas suelen estar en el rango $2 \leq \alpha \leq 3$ [116].

Muchas de las variables estudiadas en los sistemas complejos toman valores discretos. Es el caso de la cantidad de hipervínculos de una página web, la cantidad de coautores que han escrito trabajos científicos con un determinado autor, o la cantidad de aristas que confluyen en un nodo de una red (ya sea una red de tránsito, de comunicaciones, de personas, etc), lo que se conoce como *grado* del nodo. En estos casos, podemos modelar las variables como variables aleatorias discretas, o bien podemos realizar una aproximación continua –que funcionará muy bien cuando el número de muestras es muy elevado y las funciones de distribución continua y discreta se asemejen–. Comenzaremos analizando este último caso, es decir, que la variable tiene una distribución continua, y dejaremos para una sección posterior la discusión sobre leyes de potencias discretas.

A.1. Propiedades matemáticas de las leyes de potencias continuas

Diremos que una variable aleatoria continua X sigue una ley de potencias cuando su función de densidad es de la forma

$$f(x) = Cx^{-\alpha}, x \geq x_{\min} > 0,$$

con $\alpha > 1$. El soporte debe comenzar en algún $x_{\min} > 0$ porque $x^{-\alpha}$ tiene una singularidad no integrable en el origen¹. El valor de la constante C se deduce al exigir que la función de densidad tenga área 1:

$$\int_{x_{\min}}^{\infty} Cx^{-\alpha} = 1 \Rightarrow C = (\alpha - 1) \cdot x_{\min}^{\alpha-1}.$$

Las leyes de potencias tienen *momentos de orden m* finitos sólo para $m < \alpha - 1$. Por ejemplo, para el rango habitual de valores $2 < \alpha \leq 3$ la media es finita pero la varianza no. En los casos en que son finitas, la media μ y la varianza σ^2 valen:

$$\mu = \frac{(\alpha - 1)}{(\alpha - 2)} \cdot x_{\min} \quad \sigma^2 = \frac{(\alpha - 1)}{(\alpha - 3)} \cdot x_{\min}^2.$$

La cola de la función de distribución de X también sigue una ley de potencias, pero con un exponente β corrido en una unidad respecto a α :

$$G(x) = P[X > x] = \int_x^{\infty} Cx'^{-\alpha} dx' = \left(\frac{x}{x_{\min}} \right)^{-(\alpha-1)} = \left(\frac{x}{x_{\min}} \right)^{-\beta}, \quad x \geq x_{\min}, \beta = \alpha - 1.$$

La función de distribución inversa G^{-1} de X es:

$$G^{-1}(y) = x_{\min} \cdot y^{-1/\beta}.$$

Esta última fórmula es particularmente útil para generar muestras de la variable X a partir de muestras de una variable uniforme $U(0, 1)$.

Es habitual graficar las leyes de potencias en el plano cartesiano con ambos ejes en escala logarítmica. Así, si llamamos $y' = \log(y)$ y $x' = \log(x)$, tenemos que:

¹También se puede hablar de leyes de potencias con exponentes menores a 1, pero las mismas no son de importancia práctica en el estudio de sistemas complejos. En estos casos la función $x^{-\alpha}$ tiene una singularidad no integrable en infinito en vez de en 0.

$$\begin{aligned}
y' &= \log(y) \\
&= \log(f(x)) \\
&= \log(C \cdot x^{-\alpha}) \\
&= \log(C) - \alpha \log(x) \\
&= \log(C) - \alpha x' .
\end{aligned}$$

Es decir que en escala log-log observamos una recta que decrece con pendiente $-\alpha$. La figura A.1 ilustra la situación con una ley de potencias de exponente $\alpha = 3$ graficada en escala lineal y en escala log-log.

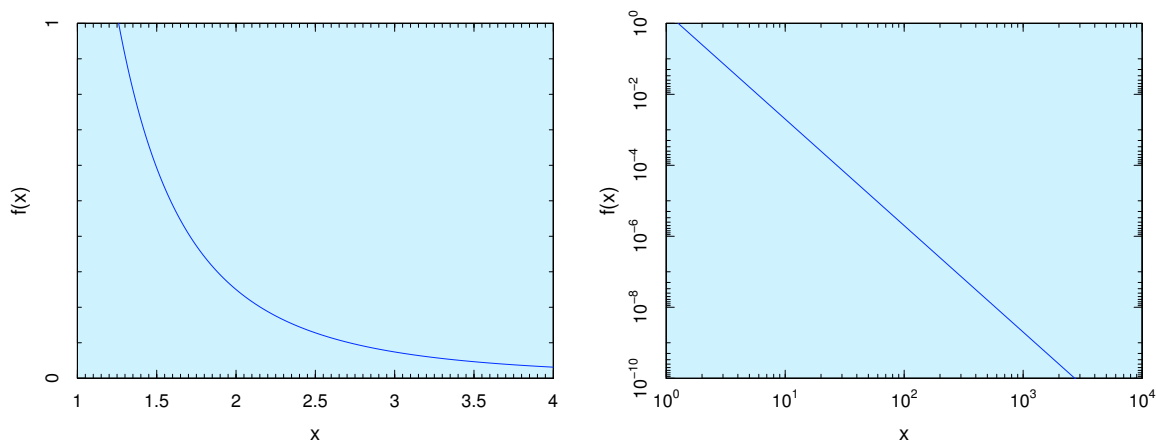


Figura A.1: *Leyes de potencias*. Ley de potencias con exponente $\alpha = 3$ y $x_{\min} = 1$ graficada en escala lineal (izq) y log-log (der).

A.2. Ajuste empírico de leyes de potencias continuas

Las leyes de potencias generalmente se observan a partir de tomar cierto número de *muestras* de la variable en el sistema en estudio. Por ello abordaremos el problema del ajuste de leyes de potencias a partir de datos empíricos.

Dada una muestra aleatoria $(X_1, X_2, \dots, X_N)$ de una variable aleatoria continua X que suponemos distribuida siguiendo una ley de potencias, podemos aproximar la función de densidad con un *histograma*. Un histograma es un conjunto de puntos (x_i, y_i) que se obtiene a partir del siguiente procedimiento denominado *binning*:

1. Se define una secuencia (m_i) de M intervalos o *bins* consecutivos de la forma $[a_i, b_i)$, con $i = 0, 1, \dots, M - 1$, de manera que:

$$a) \ a_0 = x_{\min}$$

$$b) \ a_i = b_{i-1} \text{ para } i = 1, 2, \dots, M - 1$$

$$c) \ b_{M-1} = x_{\max} \ .$$

2. Se cuenta la cantidad de muestras dentro de cada intervalo: $S_i = \sum_1^N \mathbf{1}\{X_j \in m_i\}$.

3. Por cada intervalo se define un punto del histograma $(x_i, y_i) = \left(a_i, \frac{S_i}{N \cdot (b_i - a_i)}\right)$.

El histograma tiene la propiedad de que los valores de y_i representan la probabilidad de que una de las muestras tomada al azar caiga en el intervalo m_i , normalizada por la longitud del intervalo, de manera de representar una aproximación por rectángulos de la función de densidad de probabilidad. Al construir un histograma se debe escoger una subdivisión en intervalos del soporte de la variable aleatoria. En otros contextos, es frecuente utilizar una división equiespaciada, y escoger la cantidad de intervalos según distintas reglas, como ser que los intervalos tengan determinada longitud, que haya un mínimo de muestras por intervalo, o que haya la misma cantidad de muestras promedio por intervalo que cantidad de intervalos, por ejemplo. Sin embargo, en las distribuciones que siguen leyes de potencia –o distribuciones de cola larga en general– construir el histograma con un binning equiespaciado en la escala lineal suele presentar dos problemas: (i) introduce mucho ruido para valores grandes de la variable aleatoria; y (ii) al transformarlo a escala log-log los *bins* del histograma se acumulan hacia la derecha del gráfico, dejando relativamente poco muestreados los valores pequeños de la variable, que son los más frecuentes. Lo más habitual es entonces realizar un *binning logarítmico*: esto logrará que los *bins* queden equiespaciados al visualizarlos en escala logarítmica.

Binning logarítmico. El binning logarítmico se construye de la siguiente forma:

$$\begin{aligned} a_0 &= x_{\min} \\ a_i &= a_{i-1} \cdot \frac{x_{\max}^{1/(M-1)}}{x_{\min}^{1/(M-1)}} = x_{\min} \cdot \frac{x_{\max}^{i/(M-1)}}{x_{\min}^{i/(M-1)}} \text{ para } i=1,2,\dots,M-1 \ . \end{aligned}$$

En la escala logarítmica los límites de los *bins* se encontrarán en:

$$\begin{aligned} a'_0 &= \log(x_{\min}) \\ a'_i &= \log(x_{\min}) + \frac{i}{M-1} \log\left(\frac{x_{\max}}{x_{\min}}\right) \text{ para } i=1,2,\dots,M-1 \ . \end{aligned}$$

Los puntos del histograma logarítmico serán $(x'_i, y'_i) = \left(a'_i, \log\left(\frac{S_i}{N \cdot (b_i - a_i)}\right)\right)$.

Estimación de parámetros. Cuando el histograma logarítmico parece revelar una ley de potencias el siguiente problema consiste en estimar los dos parámetros de la distribución, $\hat{x}_{\min}$ y $\hat{\alpha}$:

- El valor de $\hat{x}_{\min}$ suele desprenderse directamente del significado de la variable que estamos modelando. También es muy frecuente que se tome como valor de $\hat{x}_{\min}$ al mínimo entre todas las muestras obtenidas.
- El valor de $\hat{\alpha}$ en muchos casos es ajustado por regresión lineal, es decir, buscando la recta $y' = \log(C) - \hat{\alpha}x'$ que minimiza el error cuadrático medio de los pares de puntos (x'_i, y'_i) del histograma logarítmico. Sin embargo, se ha observado que la regresión lineal para ajustar leyes de potencias suele tener un elevado margen de error, y que es ampliamente superada por el método de máxima verosimilitud (*max-likelihood*) [116, 46].

Regresión lineal. La regresión lineal ajusta los puntos a una recta $y' = Ax' + B$. De acuerdo con lo visto anteriormente, $A = -\hat{\alpha}$ y $B = \log(\hat{C})$. Como la regresión lineal no está sujeta a la restricción $\hat{C} = (\hat{\alpha} - 1) \cdot \hat{x}_{\min}^{\hat{\alpha}-1}$, los valores de α y C que se deduzcan de aquí no corresponderán necesariamente a una distribución de probabilidad. Una solución es simplemente considerar el $\hat{\alpha}$ calculado por la regresión, y deducir $\hat{C}$ a partir de un $x_{\min}$ conocido a priori. Otra posibilidad es escoger el $\hat{x}_{\min}$, de manera que la función de distribución de probabilidad esté realmente comprendida en la recta de la regresión lineal.

En el planteo por regresión lineal, siendo $\mathbf{x}'$ el vector columna de las coordenadas x de los puntos del histograma logarítmico, e $\mathbf{y}'$ el vector columna de las coordenadas y :

$$\begin{pmatrix} A & B \end{pmatrix} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \cdot \mathbf{y}' ,$$

en donde $\mathbf{Z} = \begin{pmatrix} \mathbf{x}' & \mathbf{1} \end{pmatrix}$ y $\mathbf{1}$ es un vector columna de M unos. Operando, esto da como resultado:

$$A = \frac{M \sum x'_i y'_i - \sum x'_i \sum y'_i}{M \sum x'^2_i - (\sum x'_i)^2}$$

$$B = \frac{\sum y'_i (\sum x'^2_i) - \sum x'_i \sum x'_i y'_i}{\sum x'^2_i - (\sum x'_i)^2} .$$

Máxima verosimilitud. En el planteo por máxima verosimilitud se calcula la función de densidad conjunta de la muestra $(X_1, X_2, \dots, X_N)$, parametrizada con α y $x_{\min}$, y se la evalúa en el punto $(x_1, x_2, \dots, x_N)$ muestreado. Al resultado de este cálculo, en tanto función de los parámetros α y $x_{\min}$, lo llamamos *función de verosimilitud (likelihood)* $\mathcal{L}(\alpha x_{\min} | x_1 x_2 \dots x_N)$:

$$f_{\alpha, x_{\min}}(x_1 x_2 \dots x_N) = \prod_{i=1}^N f_{\alpha, x_{\min}}(x_i) \doteq \mathcal{L}(\alpha x_{\min} | x_1 x_2 \dots x_N) .$$

Siendo que las muestras pertenecen a variables independientes, idénticamente distribuidas y con una distribución ley de potencias, la función de verosimilitud se puede expresar de la siguiente manera:

$$\mathcal{L}(\alpha x_{\min} | x_1 x_2 \dots x_N) = (\alpha - 1)^N x_{\min}^{(\alpha-1)N} \prod_{i=1}^N x_i^{-\alpha} \quad \alpha > 1, x_{\min} \leq \min(x_1, x_2, \dots, x_N) .$$

Las estimaciones de α y $x_{\min}$ corresponderán al punto donde se produce el máximo de la función de verosimilitud:

$$(\hat{\alpha}, \hat{x}_{\min}) = \arg \max_{(\alpha, x_{\min})} \mathcal{L}(\alpha x_{\min} | x_1 x_2 \dots x_N) .$$

$\mathcal{L}(\alpha x_{\min} | x_1 x_2 \dots x_N)$ es estrictamente creciente en $x_{\min}$. Luego, su máximo tiene coordenada $\hat{x}_{\min} \leq \min(x_1, x_2, \dots, x_N)$, mientras que la coordenada $\hat{\alpha}$ es el resultado de la siguiente maximización:

$$\hat{\alpha} = \arg \max_{\alpha} \mathcal{L}(\alpha \hat{x}_{\min} | x_1 x_2 \dots x_N) .$$

Por conveniencia maximizaremos el logaritmo de la función $\mathcal{L}(\alpha \hat{x}_{\min} | x_1 x_2 \dots x_N)$:

$$\begin{aligned} \ln \mathcal{L}(\alpha \hat{x}_{\min} | x_1 x_2 \dots x_N) &= \ln \left((\alpha - 1)^N \hat{x}_{\min}^{N(\alpha-1)} \prod_{i=1}^N x_i^{-\alpha} \right) = \\ &= N \ln(\alpha - 1) + N(\alpha - 1) \ln(\hat{x}_{\min}) - \alpha \sum_{i=1}^N x_i . \end{aligned}$$

El valor de α que maximiza $\ln \mathcal{L}$ es

$$\hat{\alpha} = 1 + N \cdot \left(\sum_{i=1}^N \ln \left(\frac{x_i}{\hat{x}_{\min}} \right) \right)^{-1} .$$

Ejemplo. Para ilustrar los métodos desarrollados en este apartado, se generaron un millón de muestras de una ley de potencias continua con $x_{\min} = 1$ y $\alpha = 3$. En la figura A.2 se muestra el histograma en escala logarítmica, junto con el $\hat{\alpha}$ estimado por mínimos cuadrados y por máxima verosimilitud.

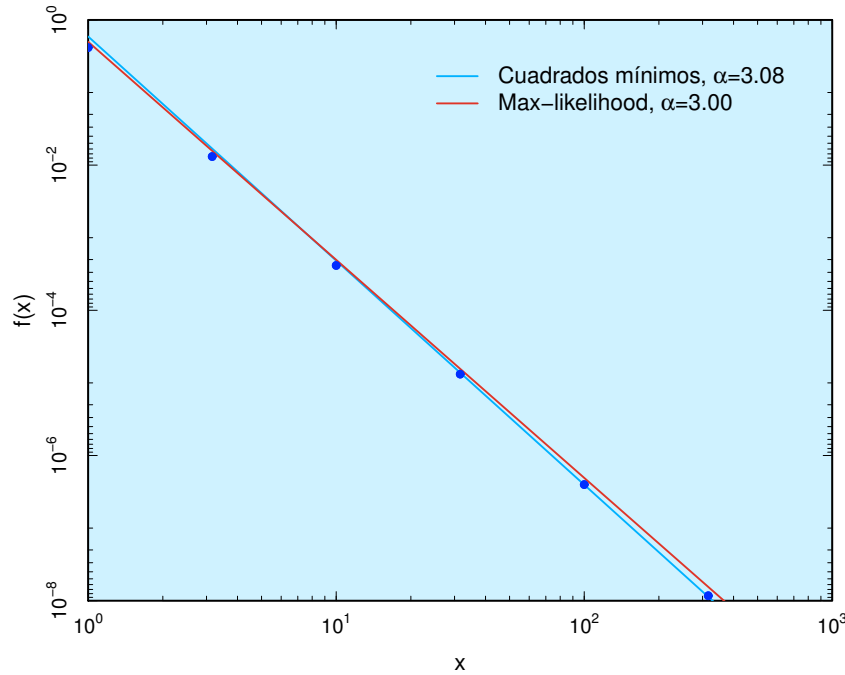


Figura A.2: *Estimación de leyes de potencias.* Histograma de un millón de muestras de una ley de potencias continua con $x_{\min} = 1$ y $\alpha = 3$. La estimación por cuadrados mínimos (*celeste*) dio un coeficiente $\alpha = 3,08$. Por máxima verosimilitud (*rojo*) se obtuvo $\alpha = 3,00$.

Estimación del máximo. Al muestrear una ley de potencias, suele ser muy útil estimar cuál será el valor máximo de entre todas las muestras. Es interesante la observación hecha en [115], según la cual el valor esperado del máximo de entre N muestras de una ley de potencias continua es cercano a aquel valor para el cual la probabilidad acumulada a derecha es igual a $1/N$, es decir:

$$E[X_{\max}] = E[\max(X_1, X_2, \dots, X_N)] \approx N^{\frac{1}{\alpha-1}} = N^{\frac{1}{\beta}}.$$

A.3. Propiedad libre de escala

Las funciones de densidad de las leyes de potencias presentan la interesante propiedad de *invariancia ante cambios de escala*. Esto quiere decir que un cambio de variable

$Z = cX$ conserva la función de densidad:

$$f_Z(z) = f_Z(cx) = \frac{1}{c} f_X(x) \propto f_X(cx) \quad , z \geq cx_{\min} .$$

En efecto, las distribuciones de las leyes de potencias son las únicas funciones continuas y derivables que presentan esta propiedad, como se mostrará a continuación. Supongamos que una función f satisface

$$f(x) = g(c)f(cx) \quad x > 0 .$$

Dado que este comportamiento se verifica para todo $c > 0$, derivamos respecto a c :

$$0 = g'(c)f(cx) + xg(c)f'(cx) .$$

Para $c = 1$:

$$xf'(x) = -\frac{g'(1)f(x)}{g(1)} .$$

La solución de esta ecuación diferencial es:

$$f(x) = Cx^{-\frac{g'(1)}{g(1)}} = Cx^{-\alpha} .$$

Finalmente la restricción de área 1 para ser una distribución de probabilidad restringe las posibilidades a $\alpha > 1$ y a considerar un $x_{\min} > 0$.

La consecuencia de la invariancia de escala es que –volviendo a uno de los ejemplos iniciales– si observamos la distribución de la riqueza, es indistinto que la midamos en dólares, millones de dólares, yenes, o su equivalente en oro: siempre encontraremos una ley de potencias *con el mismo exponente* α .

Comparemos este comportamiento con el que encontramos en las leyes exponenciales. Tomemos como ejemplo el tiempo de vida de un componente electrónico, que se modela habitualmente con una distribución exponencial $\lambda e^{-\lambda x}$. Esta distribución tendrá un exponente $\lambda_1 x$ si medimos el tiempo en meses, y un exponente distinto, $\lambda_2 = 12\lambda_1$, al medirlo en años. Es decir, la “forma de la función de densidad” se mantiene, pero no así sus parámetros. Las leyes de potencias, en cambio, conservan la distribución de la variable luego del escalado.

A.4. Leyes de potencias discretas

Como mencionamos al comienzo de este apéndice, también es posible trabajar con leyes de potencias discretas, de la forma²

$$p(k) = Ck^{-\alpha} \quad k \geq k_0 > 0, k \in \mathbb{N} ,$$

con $\alpha > 1$. El valor de la constante C es:

$$C = \frac{1}{\zeta(\alpha, k_0)} ,$$

en donde $\zeta(\alpha, k_0)$ es la función ζ de Hurwitz:

$$\zeta(\alpha, k_0) = \sum_{k=k_0}^{\infty} k^{-\alpha} .$$

La media es finita para $\alpha > 2$ y toma el mismo valor que en el caso continuo:

$$\mu = \frac{(\alpha - 1)}{(\alpha - 2)} \cdot k_0 .$$

La cola de la función de distribución es:

$$G(k) = \sum_{k'=k}^{\infty} Ck'^{-\alpha} = \frac{\zeta(\alpha, k)}{\zeta(\alpha, k_0)}, \quad k \geq k_0 .$$

Los métodos matemáticos para trabajar con leyes de potencias discretas suelen ser bastante más trabajosos. Por ejemplo, el ajuste por máxima verosimilitud arriba a una ecuación trascendente que involucra a la función $\zeta(\alpha)$, a ser maximizada por métodos numéricos.

A.4.1. Ajuste de leyes de potencias discretas con leyes continuas

Dadas estas dificultades, es muy habitual aproximar las leyes de potencias discretas con leyes continuas; este es el método empleado en el presente trabajo. Los ajustes con leyes de potencias continuas se construyen de la misma forma que se explicó en la sección A.2: el histograma, el *binning logarítmico* y la regresión lineal siguen el mismo procedimiento. Para la estimación de α por máxima verosimilitud, es conveniente el

²No es ésta la única generalización de la ley de potencias continua. Existen otras como la basada en la función Beta, o la distribución de Yule. Véanse las referencias [116, 46].

estimador

$$\hat{\alpha} = 1 + N \cdot \left(\sum_{i=1}^N \ln \left(\frac{x_i}{\hat{x}_{\min} - \frac{1}{2}} \right) \right)^{-1} ,$$

que difiere ligeramente de su versión para leyes de potencias continuas, y tiene mayor precisión [46].

A.5. Otras distribuciones de cola larga

Las leyes de potencias se enmarcan dentro de un grupo de distribuciones denominadas *de cola larga (heavy-tailed)*, que se caracterizan por tener una caída más lenta en $x \rightarrow \infty$ con respecto a la de una distribución exponencial. Es decir:

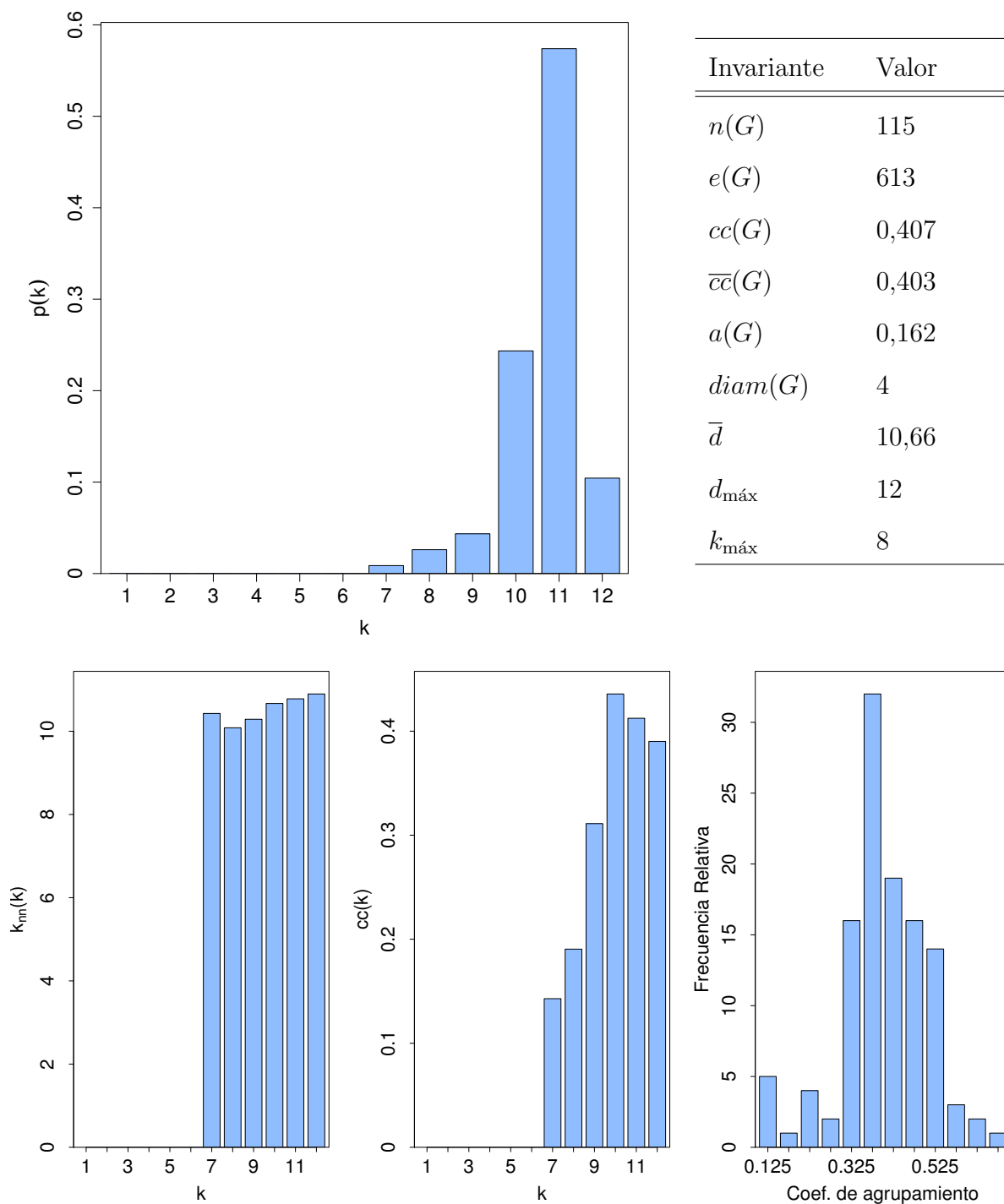
$$\lim_{x \rightarrow \infty} \frac{f(x)}{e^{-x}} = \infty .$$

Ejemplos de otras distribuciones de cola larga son la *distribución log-normal*, la *distribución de Lévy* y la *distribución t de Student*.

Apéndice B

Redes Utilizadas

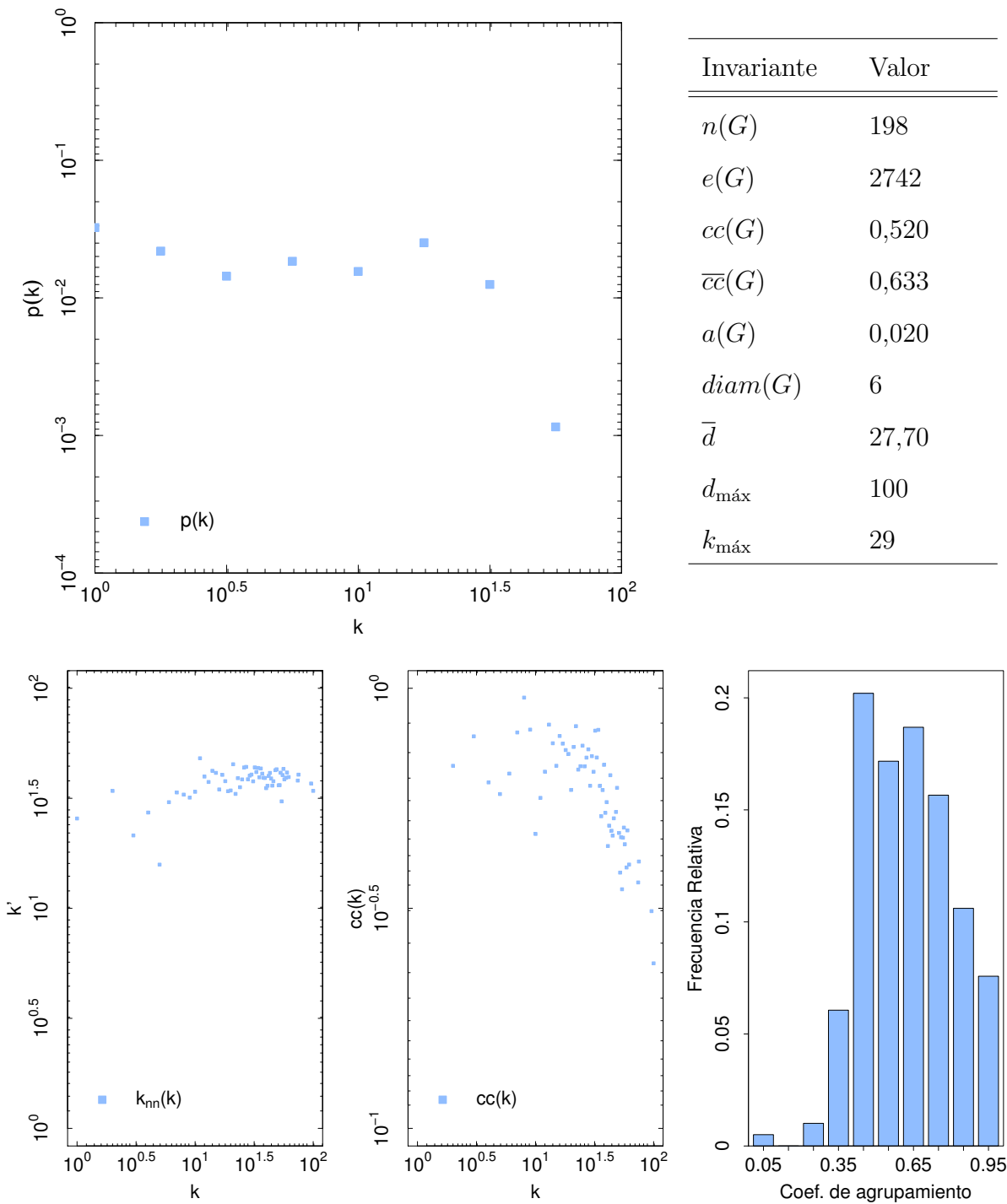
football



Cuadro B.1: *Red de football*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices. Abajo, de izquierda a derecha: el k_{nn} en función del grado, el coeficiente de agrupamiento promedio de los vértices en función del grado, y un histograma del coeficiente de agrupamiento de los vértices.

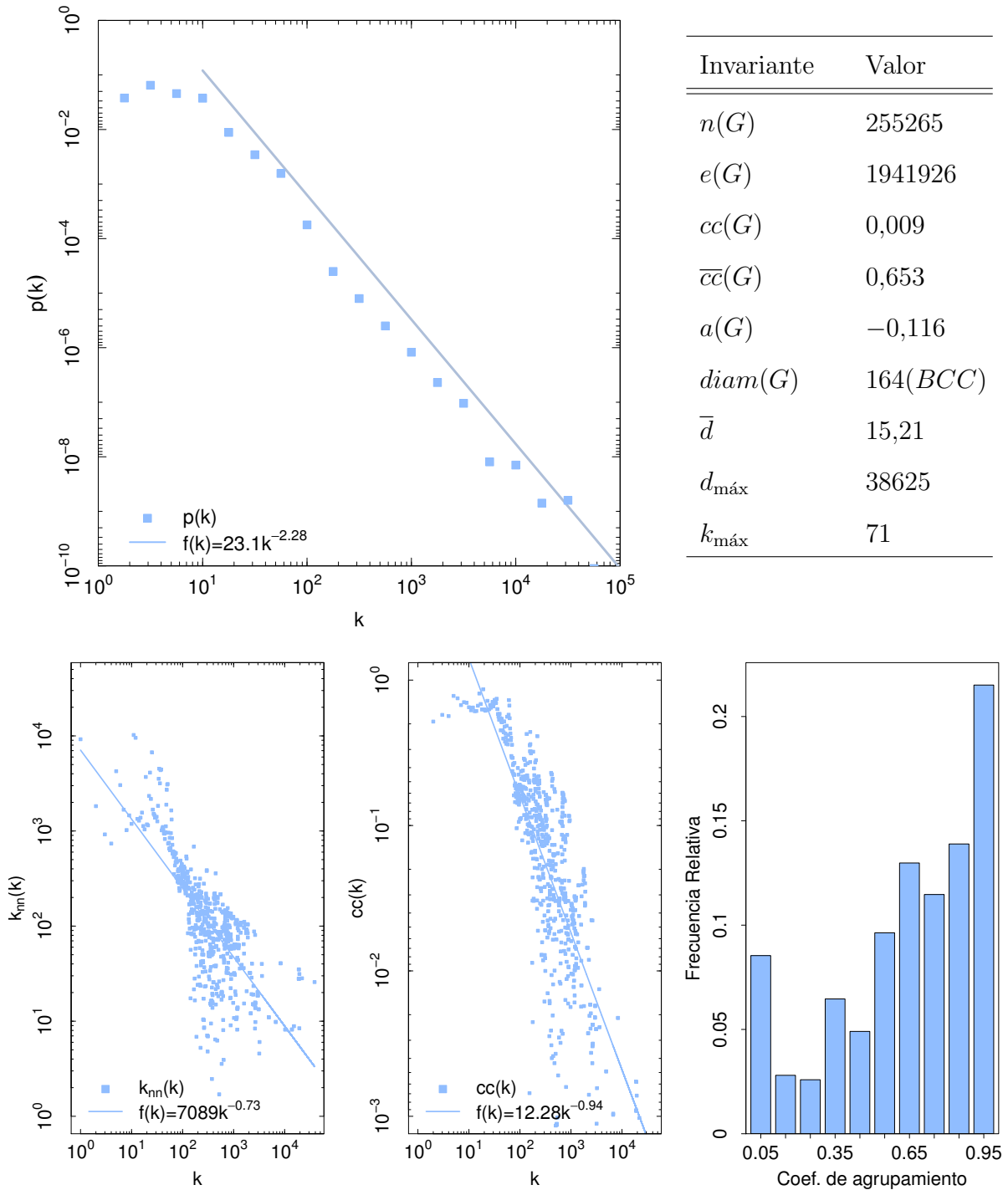
Fuente de los datos: [76].

Bandas de jazz



Cuadro B.2: *Red de bandas de jazz*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices. Abajo, de izquierda a derecha: el k_{nn} en función del grado, el coeficiente de agrupamiento promedio de los vértices en función del grado, y un histograma del coeficiente de agrupamiento de los vértices. Fuente de los datos: [78].

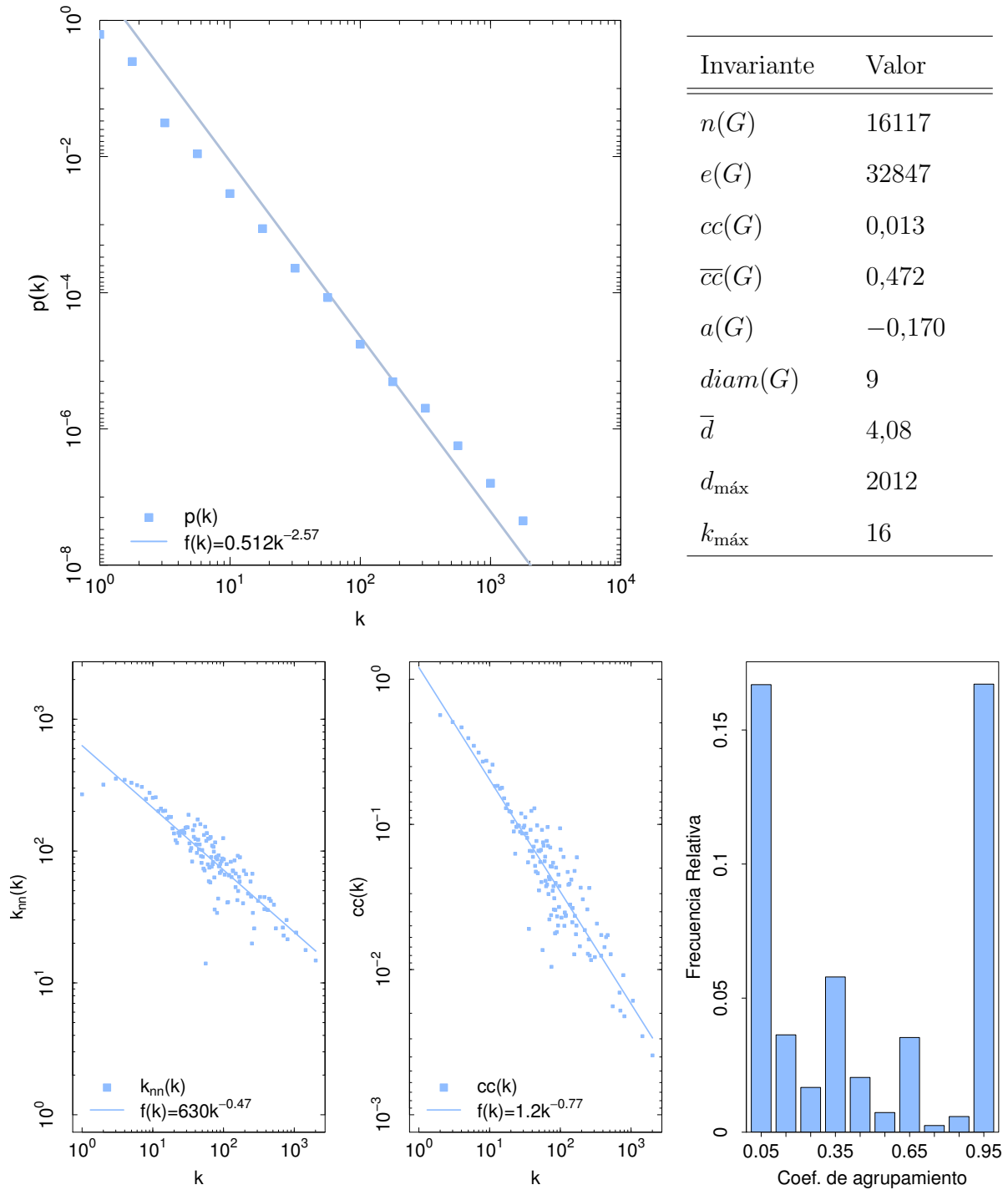
Web (stanford.edu)



Cuadro B.3: *Red de la Web de stanford.edu*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud a partir de $k = 10$. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices. Sólo se consideró la mayor componente conexa de la red (90,6 % del total de vértices).

Fuente de los datos: Stanford Large Network Dataset Collection <http://snap.stanford.edu/data/web-Stanford.html> [103].

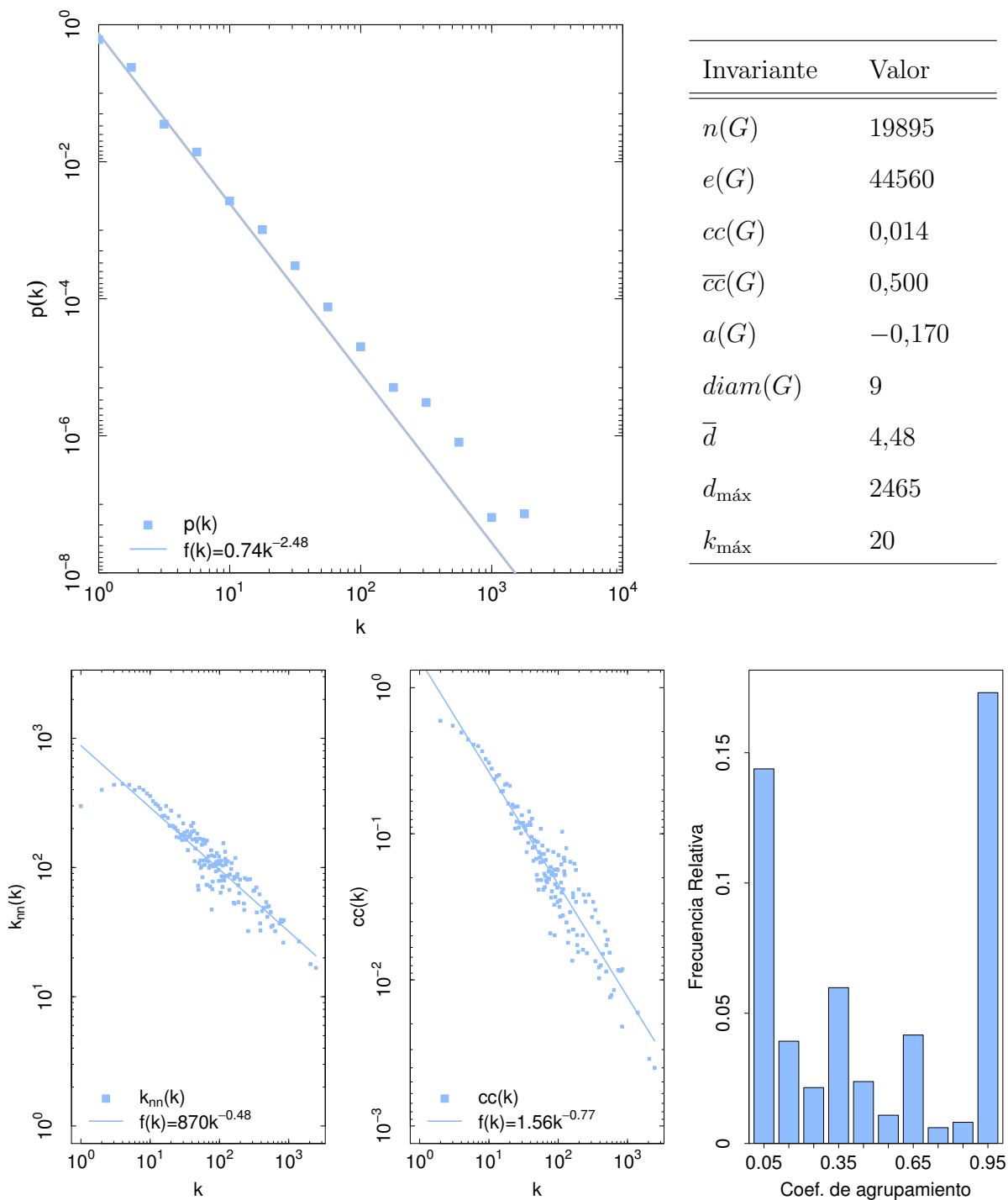
AS-CAIDA 2009



Cuadro B.4: *Red AS-CAIDA 2009*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

Fuente de los datos: The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 2009-07-02, http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

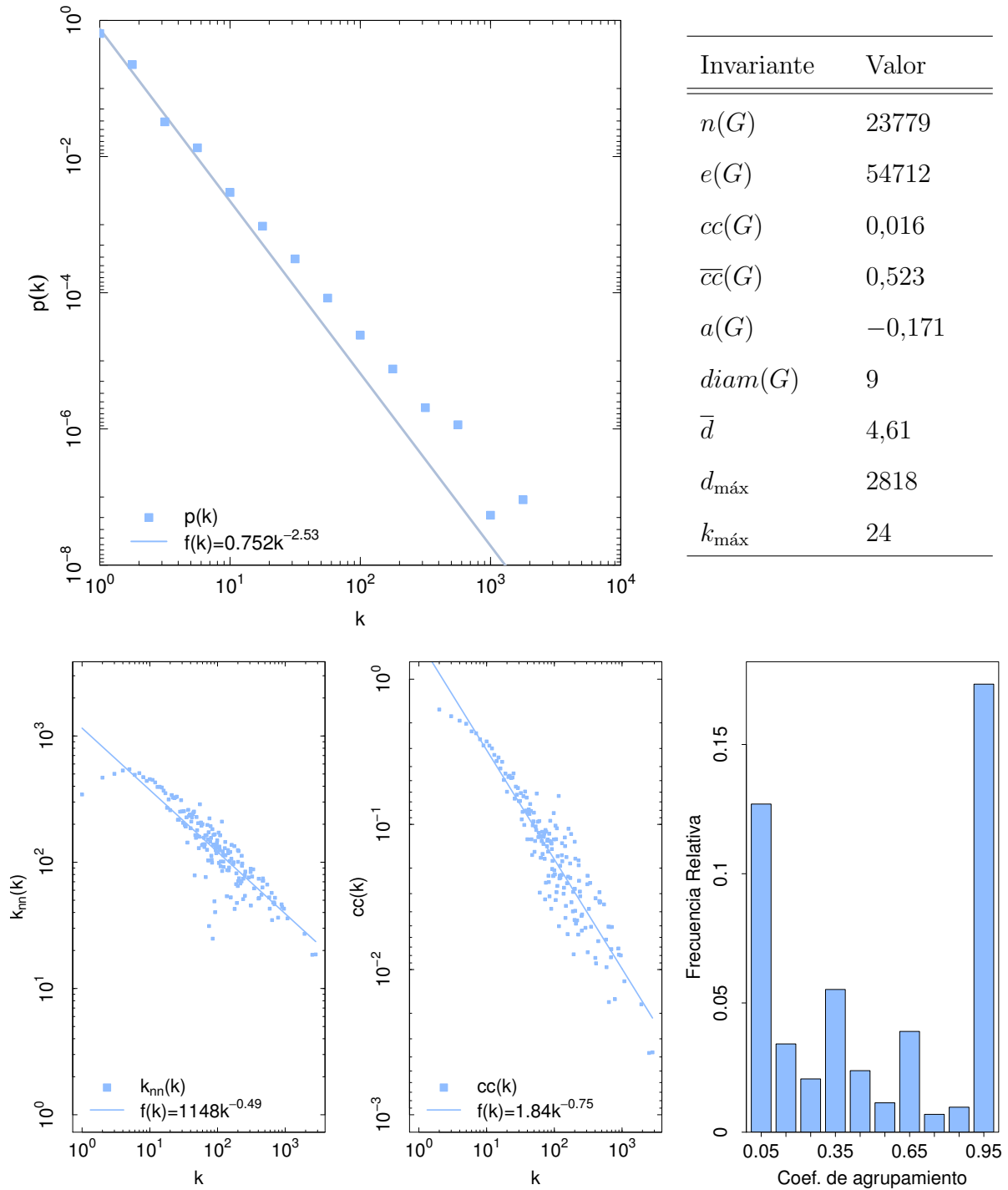
AS-CAIDA 2011



Cuadro B.5: *Red AS-CAIDA 2011*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

Fuente de los datos: The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 2011-06-30, http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

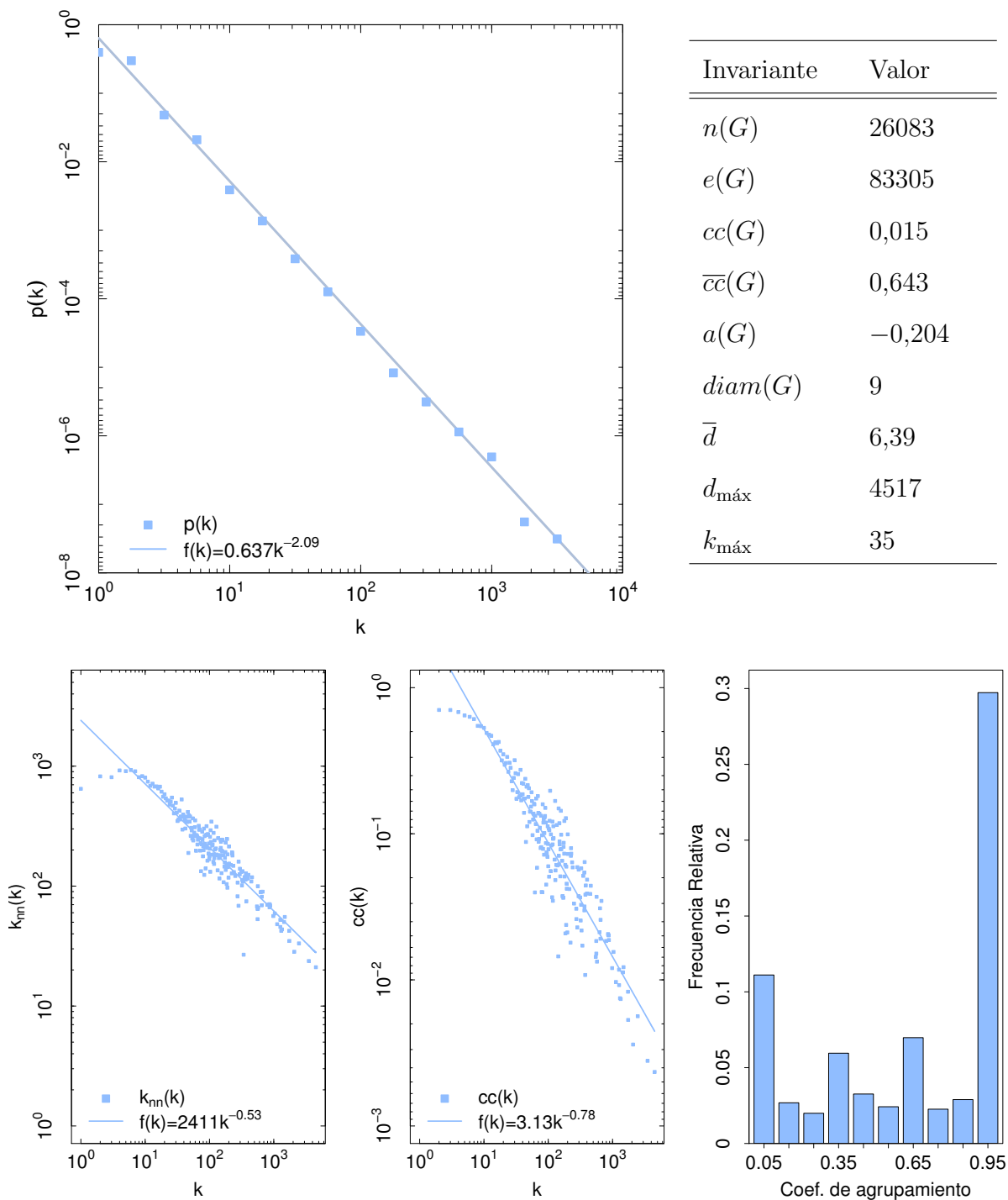
AS-CAIDA 2013



Cuadro B.6: *Red AS-CAIDA 2013*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

Fuente de los datos: The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 2013-07-03, http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

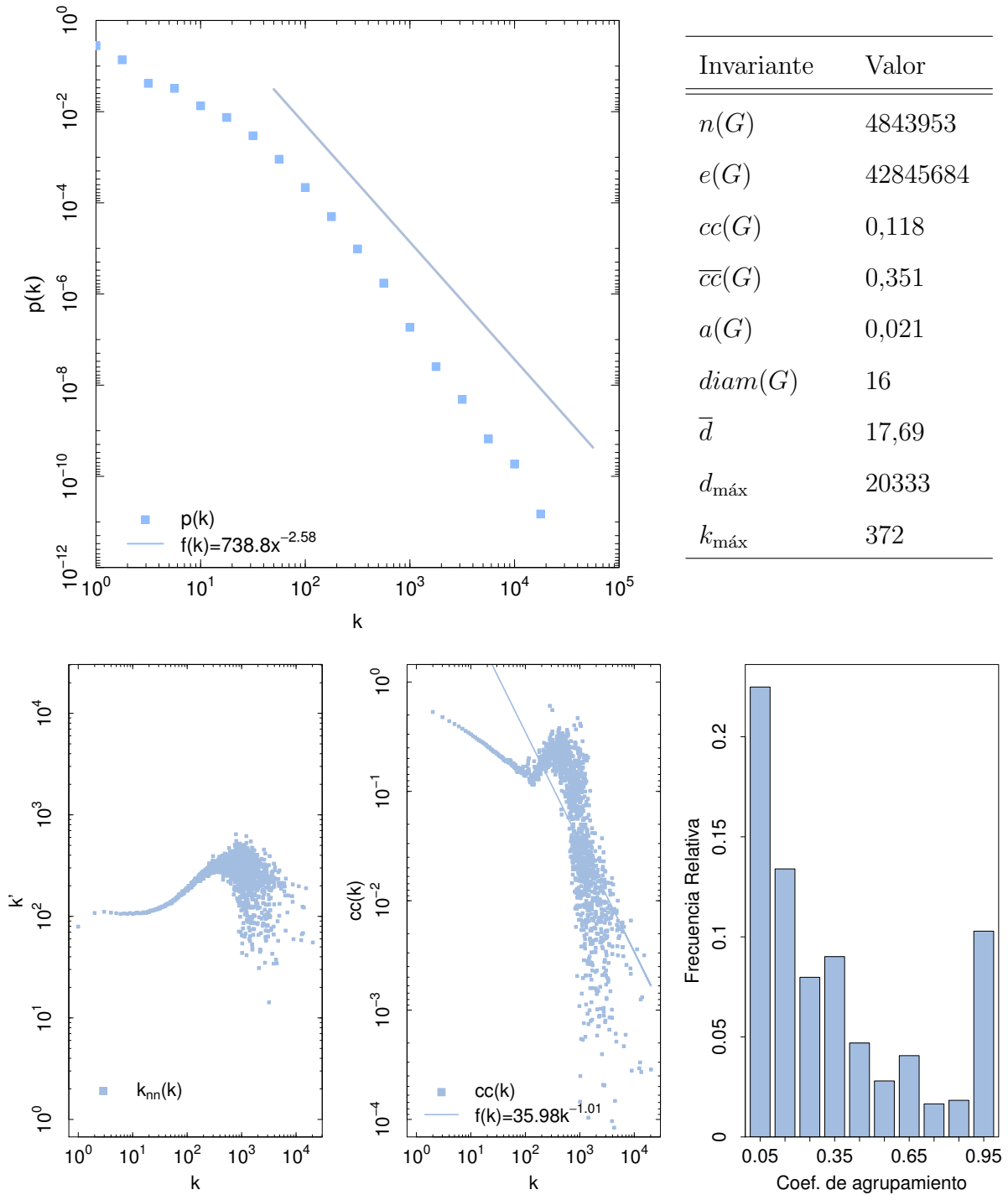
AS-DIMES 2011



Cuadro B.7: Red AS-DIMES 2011. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

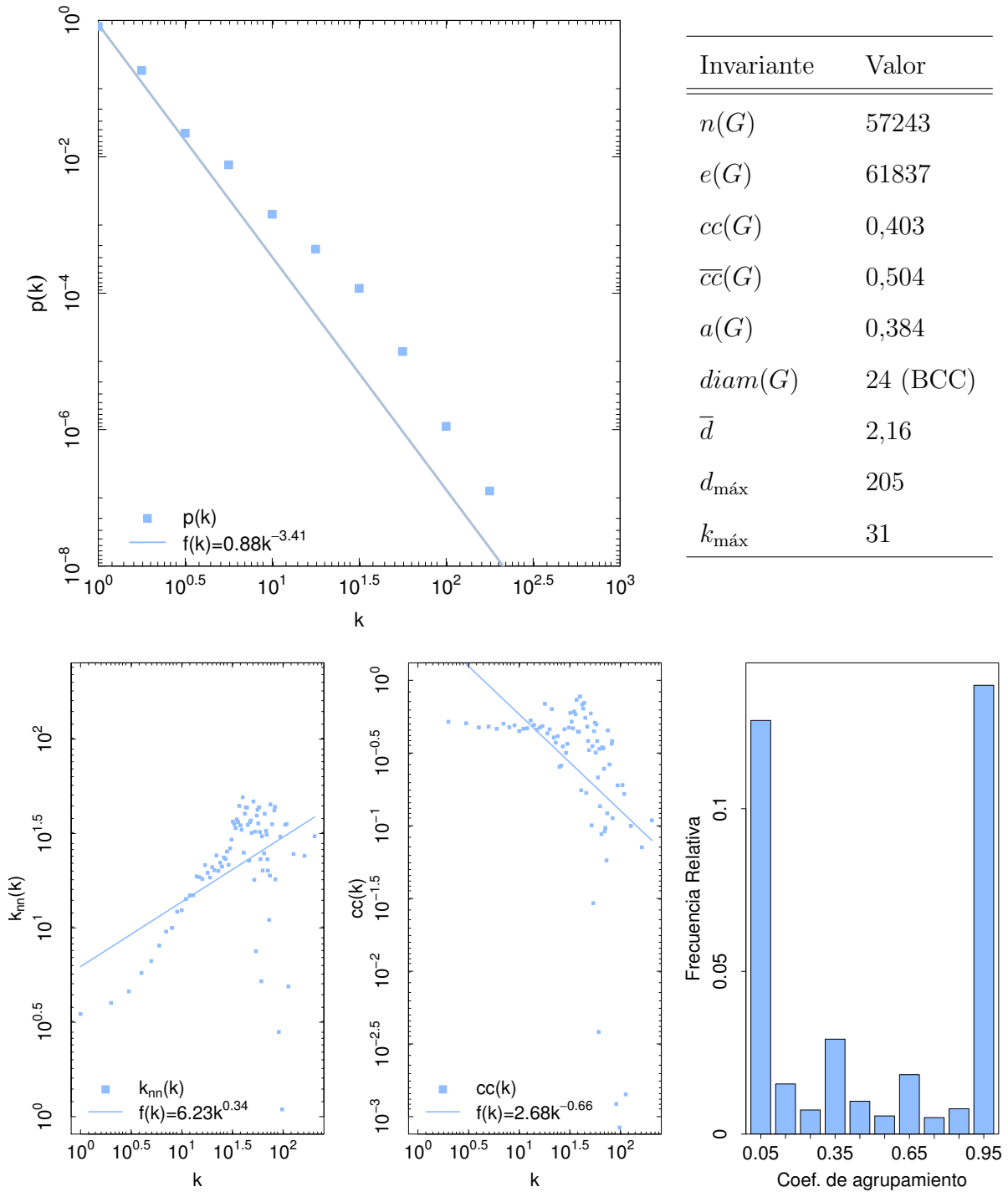
Fuente de los datos: DIMES, Distributed Internet MEasurements and Simulations, <http://www.netdimes.org/>.

LiveJournal



Cuadro B.8: *Red LiveJournal*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud a partir de $k = 50$. Abajo, de izquierda a derecha: el k_{nn} en función del grado, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices. Sólo se consideró la mayor componente conexas de la red (99,9 % del total de vértices). Fuente de los datos: Stanford Large Network Dataset Collection <http://snap.stanford.edu/data/soc-LiveJournal1.html> [103].

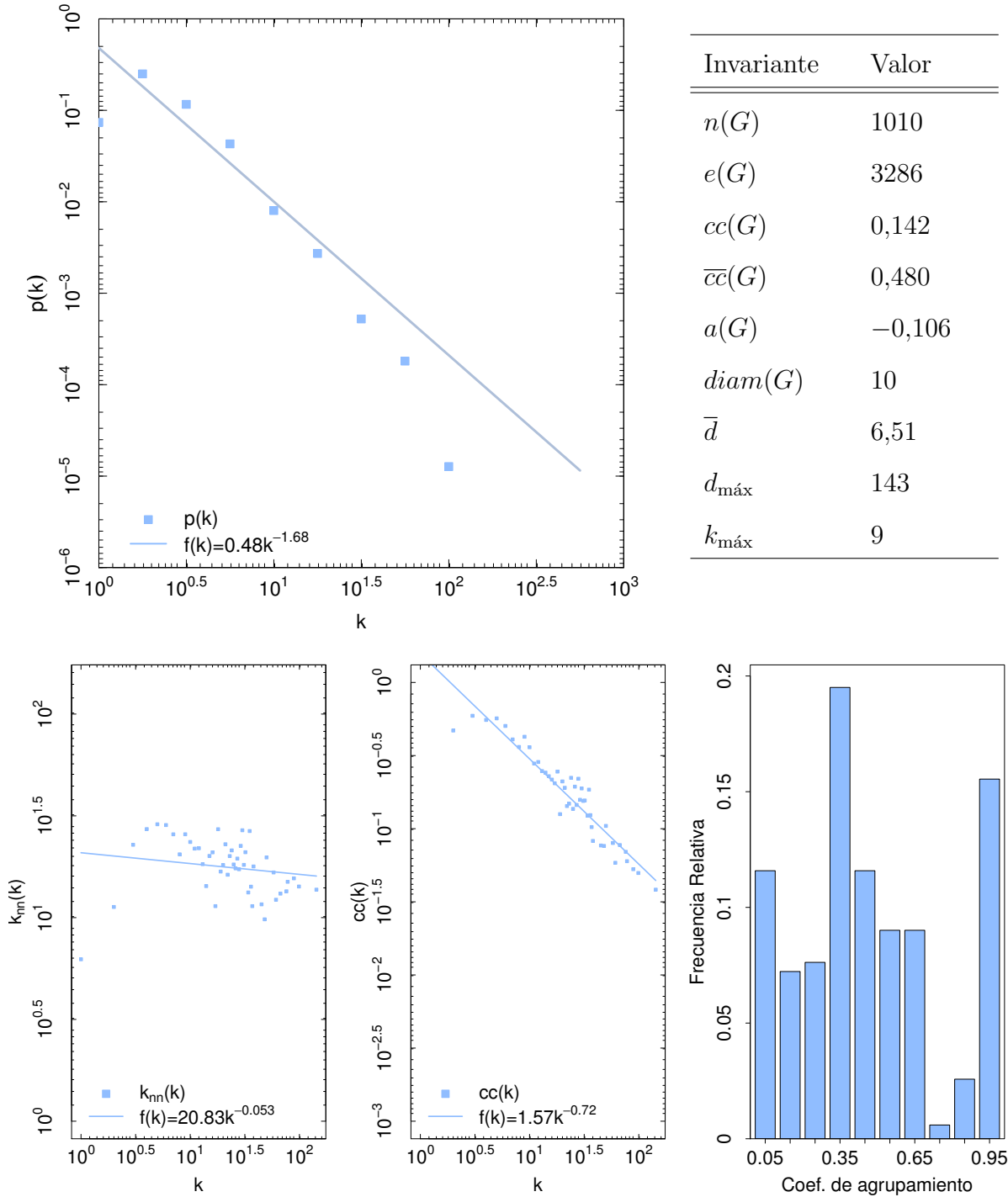
PGP



Cuadro B.9: *Red de confianza de PGP*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

Fuente de los datos: [25].

E. Coli



Cuadro B.10: *Red metabólica de E. Coli*. Arriba a la izquierda, una histograma de la distribución de grados de los vértices, y el ajuste por máxima verosimilitud. Abajo, de izquierda a derecha: el k_{nn} en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, el coeficiente de agrupamiento promedio de los vértices en función del grado y su ajuste por cuadrados mínimos a una ley de potencias, y un histograma del coeficiente de agrupamiento de los vértices.

Fuente de los datos: [145].

Bibliografía

- [1] R.D. Alba. A graph-theoretic definition of a sociometric clique. *The Journal of Mathematical Sociology*, 3(1):113–126, 1973. 75
- [2] R. Albert and A-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002. 50
- [3] R. Albert, H. Jeong, and A-L. Barabási. The diameter of the world wide web. *Nature*, 401:130–131, 1999. 13, 20, 48, 49, 157
- [4] R. Albert, H. Jeong, and A-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:200–0, 2000. 50
- [5] J.I. Alvarez-Hamelin, M.G. Beiró, A. Barrat, L. Dall’Asta, and A. Vespignani. Lanet-vi: Large network visualization tool. <http://lanet-vi.fi.uba.ar/>. 137, 145
- [6] J.I. Alvarez-Hamelin, M.G. Beiró, and J.R. Busch. Understanding edge connectivity in the internet through core decomposition. *Internet Mathematics*, 7(1):45–66, 2011. 125, 135, 154
- [7] J.I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and Vespignani A. k-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *Networks and Heterogeneous Media*, 3(2):371, 2008. 50, 51, 124
- [8] J.I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and A. Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In *NIPS*, 2005. 50
- [9] J.I. Alvarez-Hamelin and N. Schabanel. An internet graph model based on trade-off optimization. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):231–237, 2004. 63
- [10] A. Arenas, J. Duch, A. Fernandez, and S. Gomez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9(6):176, 2007. 79

- [11] D. Bailer-Jones. *Scientific models in philosophy of science*. University of Pittsburgh Press Pittsburgh, Pa, 2009. 22
- [12] P. Bak, K. Chen, and C. Tang. A forest-fire model and some thoughts on turbulence. *Physics Letters A*, 147(5-6):297–300, 1990. 18, 20
- [13] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. an explanation of $1/f$ noise. *Physical Review Letters*, 59:381–384, 1987. 16, 20
- [14] A-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999. 20, 48, 54, 58, 62
- [15] A-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187, 1999. 60
- [16] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747, 2004. 37, 40
- [17] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13(3):547–560, January 2000. 66
- [18] V. Batagelj and M. Zaversnik. An $o(m)$ algorithm for cores decomposition of networks. *arXiv*, 2001. 45
- [19] M.G. Beiró, J.R. Busch, and J.I. Alvarez-Hamelin. Snailvis: a paradigm to visualize complex networks. In *39 Jornadas Argentinas de Informática e Investigación Operativa (JAIIO)*, pages 1682–1693. SADIO, 2010. <http://cnet.fi.uba.ar/mariano.beiro/snailvis.tar.gz>. 110, 119
- [20] M.G. Beiró, J.R. Busch, S.P. Grynberg, and J.I. Alvarez-Hamelin. Obtaining communities with a fitness growth process. *Physica A: Statistical Mechanics and its Applications*, 392(9):2278 – 2293, 2013. 74, 153, 154
- [21] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, May 1978. 54, 64
- [22] G. Bianconi and A-L. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436, 2001. 54

- [23] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer, 2007. 41
- [24] V.D. Blondel, J-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 79, 108
- [25] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70(5):056122+, November 2004. 147, 176
- [26] B. Bollobás. *Graph Theory, An Introductory course*. Springer-Verlag, New York, Heidelberg, Berlin, 1979. 30
- [27] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001. 53
- [28] B. Bollobás. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks*, pages 1–37. Wiley, 2003. 53, 62
- [29] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, January 2004. 62
- [30] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Struct. Algorithms*, 18(3):279–290, May 2001. 62
- [31] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, February 2008. 79, 93
- [32] M. Bunge. *Epistemología*. Ariel, Barcelona, 1980. 9
- [33] J.R. Busch, M.G. Beiró, and J.I. Alvarez-Hamelin. On weakly optimal partitions in modular networks. *CoRR*, abs/1008.3443, 2010. 74, 92, 153
- [34] CAIDA. The cooperative association for internet data analysis. <http://www.caida.org/>. 50, 123
- [35] G. Caldarelli and A. Vespignani. *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2007. 40, 48
- [36] J. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. *Physical Review E*, 60(2):1412–1427, 1999. 18, 20, 62

- [37] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. Medusa - new model of internet topology using k-shell decomposition. *arXiv*, January 2006. 51, 124
- [38] S. Carmi, S. Havlin, S. Kirkpatrick, and E. Shir. A model of internet topology using k-shell decomposition. *PNAS*, 104:11150–11154, 2007. 124
- [39] M. Catanzaro, G. Caldarelli, and L. Pietronero. Assortative model for social networks. *Physical Review E*, 70(3), 2004. 54, 124
- [40] D.J. Chalmers. *Strong and Weak Emergence, on The Re-Emergence of Emergence*. Oxford University Press, 2006. 8
- [41] J. Chen and B. Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, September 2006. 74
- [42] F. Chung and L. Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, May 2001. 55
- [43] F.R.K. Chung and L. Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1(1):91–113, 2003. 54, 65
- [44] A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008. 55
- [45] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, December 2004. 79, 82
- [46] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, November 2009. 161, 165, 166
- [47] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37, 1960. 41
- [48] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Physical Review Letters*, 86(16):3682–3685, April 2001. 50
- [49] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5):058701+, February 2003. 62
- [50] P. Colomer de Simón, M.A. Serrano, M.G. Beiró, J.I. Alvarez-Hamelin, and M. Boguñá. Deciphering the global organization of clustering in real complex networks. *Scientific Reports*, 3(2517), 2013. 145, 146, 147, 154

- [51] A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001. 55, 68
- [52] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991. 84
- [53] L. Danon, A. Díaz-Guilera, and A. Arenas. Effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics*, 2006(11):P11010, 2006. 79
- [54] L. Danon, A.D. Guílera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(9):P09008–09008, September 2005. 83, 84
- [55] D.J. De Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, July 1965. 157
- [56] DIMES. Distributed internet measurements and simulations. <http://www.netdimes.org/>. 50, 123
- [57] M.B. Doar. A better model for generating test networks. In *Global Telecommunications Conference, 1996. GLOBECOM '96. 'Communications: The Key to Global Prosperity*, pages 86–93, 1996. 54
- [58] S. Dorogovtsev. *Lectures on Complex Networks*. Oxford University Press, Inc., New York, NY, USA, 2010. 48
- [59] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, 80:1275–1335, Oct 2008. 50
- [60] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85:4633–4636, 2000. 62
- [61] J.C. Doyle, D.L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. The “robust-yet-fragile” nature of the internet. *Proceedings of the National Academy of Sciences*, 102(41):14497–14502, October 2005. 50
- [62] B. Drossel and F. Schwabl. Self-organized critical forest-fire model. *Physical Review Letters*, 69:1629–1632, September 1992. 18
- [63] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005. 79

- [64] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959. 54, 55
- [65] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming, ICALP '02*, pages 110–122. Springer-Verlag, 2002. 20, 54, 62
- [66] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, SIGCOMM '99*, pages 251–262, New York, NY, USA, 1999. ACM. 20, 48, 58, 124
- [67] R.A. Fiesner. *Advances in Chemical Physics, Computational Methods for Protein Folding*. Wiley-Interscience, 2001. 11
- [68] G.W. Flake, S. Lawrence, and C.L. Giles. Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '00*, pages 150–160, New York, NY, USA, 2000. ACM. 74, 75, 78
- [69] L.R. Ford and D.R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956. 35
- [70] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010. 77, 85
- [71] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, January 2007. 90
- [72] L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March 1977. 41
- [73] L.C. Freeman. Centrality in social networks: conceptual clarification. *Social Networks*, 1(3):215–239, 1979. 42
- [74] R. Garcia. *Sistemas complejos. Conceptos, método y fundamentación epistemológica de la investigación interdisciplinaria*. Gedisa, Barcelona, 2006. 10
- [75] M. Gardner. Mathematical games: The fantastic combinations of john conway’s new solitaire game “life”. *Scientific American*, pages 120–123, 1970. 9, 20

- [76] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. 42, 68, 77, 110, 168
- [77] J.P. Gleeson. Bond percolation on a class of clustered random networks. *Physical Review E*, 80(3):036107+, September 2009. 146, 148
- [78] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573, July 2003. 81, 110, 169
- [79] T. Gneiting and M. Schlather. Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 46(2):pp. 269–282, 2004. 25
- [80] R.E. Gomory and T.C. Hu. Multi-terminal network flows. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):551–570, 1961. 137, 138
- [81] B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010. 93, 108
- [82] R. Govindan and A. Reddy. An analysis of internet inter-domain topology and route stability. In *Proceedings of the INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, INFOCOM '97, pages 850–, Washington, DC, USA, 1997. IEEE Computer Society. 123
- [83] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *Proceedings of the INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 1371–1380, Washington, DC, USA, 2000. IEEE Computer Society. 124
- [84] C.W.J. Granger and Roselyne Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980. 24
- [85] R. Guimerà and L.A.N. Amaral. Cartography of complex networks: modules and universal roles. *J. Stat. Mech.-Theory and Exp.*, 2:02001+, February 2005. 79
- [86] R. Guimerà and L.A.N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, February 2005. 74
- [87] R. Guimerà, L. Danon, Díaz A. Guilera, F. Giralt, and A. Arenas. Self-similar community structure in organisations. *Physical Review E*, 68, 2002. 80

- [88] B. Gutenberg and C. Richter. *Frequency of Earthquakes in California*. Bulletin of the Seismological Society of America. Seismological Society of America, 1944. 157
- [89] F. A. Hayek. Degrees of explanation. *The British Journal for the Philosophy of Science*, 6(23):pp. 209–225, 1955. 22
- [90] H. Jeong, B. Tombor, R. Albert, Z.N. Oltval, and A-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, October 2000. 49, 51
- [91] S. Johnson, J.J. Torres, J. Marro, and Miguel A. Muñoz. Entropic origin of disassortativity in complex networks. *Physical Review Letters*, 104(10):108702+, March 2010. 52
- [92] J. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000. 54
- [93] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85:4629–4632, November 2000. 60
- [94] A.E. Krause, K.A. Frank, D.M. Mason, R.E. Ulanowicz, and W.W. Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, November 2003. 74
- [95] J.M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with potts model approach. *The European Physical Journal B*, 56(1):41–45, 2007. 88, 90, 91, 92
- [96] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009. 80, 81, 93, 94, 95
- [97] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, October 2008. 55, 68, 81
- [98] E.L. Lawler. Cutsets and partitions of hypergraphs. *Networks*, 3(3):275–285, 1973. 75, 78
- [99] E.A. Leicht and M.E.J. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703+, March 2008. 79
- [100] W.E. Leland, M.S. Taqqu, Willinger W., and D.V. Wilson. On the self-similar nature of ethernet traffic. In *In Proceedings of the ACM SIGCOMM'93*, 1993. 20, 24

- [101] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 915–924. ACM, 2008. 12
- [102] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM. 55
- [103] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009. 110, 170, 175
- [104] C. Levinthal. How to Fold Graciously. In J. T. P. Debrunner and E. Munck, editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pages 22–24. University of Illinois Press, 1969. 11
- [105] N. Litvak and R. van der Hofstad. Degree-degree correlations in random graphs with heavy-tailed degrees, October 2012. 52
- [106] R.D. Luce and A.D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949. 75
- [107] T. Łuczak. Size and connectivity of the k-core of a random graph. *Discrete Mathematics*, 91(1):61 – 68, 1991. 51
- [108] D. C. Mikulecky. The emergence of complexity: science coming of age or science growing old? *Computers and Chemistry*, 25(4):341–348, 2001. 5
- [109] R.J. Mokken. Cliques, clubs and clans. *Quality & Quantity*, 13(2):161–173, April 1979. 75
- [110] E. Morin. *El Método I. La naturaleza de la naturaleza*. Cátedra, Madrid, 1977. 9
- [111] M.E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, January 2001. 73, 75
- [112] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, October 2002. 40
- [113] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, September 2003. 81, 85

- [114] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2), 2003. 41, 51, 52
- [115] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003. 48, 163
- [116] M.E.J. Newman. Power laws, pareto distributions and zipf’s law. *Contemporary Physics*, 46(5):323–351, May 2005. 157, 161, 165
- [117] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), 2006. 79, 87
- [118] M.E.J. Newman. Random Graphs with Clustering. *Physical Review Letters*, 103(5):058701+, July 2009. 146
- [119] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010. 42
- [120] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(026113), 2004. 78
- [121] A. Noack and R. Rotta. Multi-level algorithms for modularity clustering. In *Proceedings of the 8th International Symposium on Experimental Algorithms*, SEA ’09, pages 257–268, Berlin, Heidelberg, 2009. Springer-Verlag. 79
- [122] L. Page. Method for node ranking in a linked database. United States patent 6,285,999, 2001. 52
- [123] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005. 80, 81
- [124] J-J. Pansiot and D. Grad. On routes and multicast trees in the internet. *Computer Communication Review*, 28(1):41–50, January 1998. 124
- [125] J. Park and M.E.J. Newman. The statistical mechanics of networks. *Physical Review E*, 70(066117), 2004. 146
- [126] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87:258701+, 2001. 39, 50, 124
- [127] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, April 2001. 50, 122

- [128] J. Plesník. Critical graphs of a given diameter. *Acta Facultatis Rerum Naturalium Universitatis Comenianae: Mathematica*, 30:71–93, 1975. 133
- [129] D.D. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976. 48, 59
- [130] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658, 2004. 77, 94
- [131] U.N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106+, September 2007. 80, 108
- [132] E. Ravasz and A-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, February 2003. 52
- [133] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, July 2006. 87, 92
- [134] H. Reittu and I. Norros. On the power-law random graph model of massive data networks. *Perform. Eval.*, 55(1-2):3–23, January 2004. 65
- [135] L. G. Rodríguez Zoya and J.L. Aguirre. Teorías de la complejidad y ciencias sociales; nuevas estrategias epistemológicas y metodológicas. *Nómadas. Revista Crítica de Ciencias Sociales y Jurídicas*, 30(2), 2011. 10
- [136] R. Rosen. *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life (Complexity in Ecological Systems)*. Columbia University Press, July 2005. 8, 22
- [137] A. Rosenblueth and N. Wiener. The role of models in science. *Philosophy of Science*, 12(4):pp. 316–321, 1945. 22, 23
- [138] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009. 81
- [139] M. Rosvall and C.T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007. 81, 82, 108

- [140] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. 81
- [141] K. Saito, T. Yamada, and K. Kazama. Extracting communities from complex networks by the k-dense method. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, E91-A(11):3304–3311, November 2008. 45, 147
- [142] S.B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983. 44
- [143] S.B. Seidman and B.L. Foster. A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, 6:139–154, 1978. 75
- [144] E. Seneta. *Non-Negative Matrices and Markov Chains*. Springer, 2006. 42, 43
- [145] M.A. Serrano, M. Boguñá, and F. Sagues. Uncovering the hidden geometry behind metabolic networks. *Molecular BioSystems*, 8:843–850, 2012. 147, 177
- [146] G. Siganos, S.L. Tauro, and M. Faloutsos. Jellyfish: A conceptual model for the as internet topology. *Journal of Communications and Networks*, 8(3):339–350, 2006. 124
- [147] S. Smyth and S. White. A spectral clustering approach to finding communities in graphs. *Proceedings of the 5th SIAM International Conference on Data Mining*, pages 76–84, 2005. 86
- [148] C. Song, S. Havlin, and H.A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, January 2005. 52
- [149] G. Tibély and J. Kertész. On the equivalence of the label propagation method of community detection and a potts model approach. *Physica A: Statistical Mechanics and its Applications*, 387(19-20):4982–4984, 2008. 80
- [150] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969. 12, 20
- [151] Route Views. University of oregon route views project. <http://www.routeviews.org/>. 50, 123

- [152] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1275–1276, New York, NY, USA, 2007. ACM. 79
- [153] D.J. Watts. *Small worlds: The dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, 1999. 55
- [154] D.J Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998. 12, 14, 20, 37, 54, 66
- [155] B.M. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, 1988. 54, 57
- [156] W. Weaver. Science and complexity. *American Scientist*, 36(4):536–544, 1948. 5, 6, 7, 19
- [157] D.B. West. *Introduction to Graph Theory (2nd Edition)*. Prentice Hall, 2000. 30, 35, 37, 46
- [158] S. Wolfram. *A New Kind of Science*. Wolfram Media, 2002. 16
- [159] W. Y. Yang and M. Gruebele. Folding at the speed limit. *Nature*, 423:193–197, 2003. 11
- [160] S.H. Yook, F. Radicchi, and H. Meyer-Ortmanns. Self-similar scale-free networks and disassortativity. *Physical Review E*, 72(4):045105, 2005. 52
- [161] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977. 12
- [162] E.W. Zegura, K.L. Calvert, and M.J. Donahoo. A quantitative comparison of graph-based models for internet topology. *IEEE/ACM Transactions on Networking*, 5(6):770–783, December 1997. 54

Índice alfabético

- acoplamiento preferencial, 20, 54, 59
- arista-conectividad, 37, 125
 - en sentido amplio, 133
 - en sentido estricto, 133
- assortativity*, véase concordancia
- autómata celular, 9, 15
- auto-organización, 8
- auto-semejanza, 14, 17

- betweenness*, véase intermediación
- binning, 159
 - logarítmico, 160

- centralidad, 41
- centralidad de vector propio, 42, 52
- cercanía, 42
- cibernética, 19
- closeness*, véase cercanía
- coeficiente de agrupamiento, 37, 145
- comportamiento
 - colectivo, 8
 - concordante, 41, 51, 52
 - discordante, 41, 50, 52, 54
- comunidad
 - en sentido débil, 77, 94
 - en sentido fuerte, 77
 - natural, 80, 93, 95
 - web, 75
- concordancia, 40, 52, 54
 - por grados, 40
- conectividad, 36, 122
- cubrimiento, 80, 94

- dendrograma, 77
- dependencia de largo alcance, 16, 17, 24
- diámetro, 37
- distribución
 - de cola larga, 14, 166
 - libre de escala, 14, 49, 163

- emergencia, 7
- emergentismo, 8
 - débil, 9
 - fuerte, 8
- escuela austríaca, 19
- estructura comunitaria, 52, 55
- experimento de Asch, 11
- experimento de Milgram, 12, 49

- fracción de vértices correctamente clasificados, 86
- función objetivo, 80, 93, 94

- grafo
 - aleatorio, 53
 - definición, 30

- highly optimized tolerance (HOT), 18, 20, 62, 122
- hipergrafo, 147
- histograma, 159
- holismo, 8

- índice de Jaccard, 84
- información mutua, 84
 - normalizada, 84

- intermediación, 41, 77
- juego de la vida, 9, 15, 20
- k -capa, 134
- k -core, véase k -núcleo
- k -denso, 45, 147
- k -núcleo, 43
- k -shell, véase k -capa
- ley de potencias, 14, 17, 48
- máxima verosimilitud, 162
- mínima longitud de descripción, 81
- modelo
 - Barabási-Albert (BA), 20, 58
 - de configuración, 64
 - de la pila de arena, 16, 20
 - de Waxman, 57
 - definición, 22
 - Erdős-Rényi, 55
 - FKP, 20, 62
 - forest-fire, 18, 20
 - LFR, 68, 83
 - planted l -partition, 68, 83
 - Watts-Strogatz, 12, 14, 20, 65
- modelos de sistemas complejos, 21
 - basados en agentes, 9, 20, 22
 - combinatorios, 21
 - de autómatas celulares, 9, 15, 25
 - de campo medio, 23
 - en ecuaciones de recurrencia, 24
 - en ecuaciones diferenciales, 23
 - en series temporales, 17, 24
- modularidad, 78
- mundo pequeño, 12, 14, 49, 66, 145
- patrones de mezclado, 40, 51
- plegamiento de proteínas, 10
- proceso de crecimiento, 94
 - uniforme, 100
- realimentación, 19, 50
- red
 - compleja, 21, 26
 - de bandas de jazz, 109, 169
 - de CAIDA(2009), 171
 - de CAIDA(2011), 172
 - de CAIDA(2013), 173
 - de confianza de PGP, 147, 176
 - de DIMES(2011), 174
 - de football, 87, 103, 168
 - de interacciones entre proteínas, 49, 74
 - de Internet, 48
 - de karate (Zachary), 12, 22
 - de la Web (Barabási), 13, 48, 49
 - de la Web (Stanford), 110, 117, 170
 - de LiveJournal, 110, 175
 - metabólica, 74
 - metabólica de E. Coli, 147, 177
 - trófica, 74
- reduccionismo científico, 8
- regresión lineal, 161
- self-organized criticality (SOC), 17, 20
- sistema, 9
- Sistema Autónomo, 121
- sistema caótico, 6
- sistema complejo
 - adaptativo, 20
 - definición, 9
- teoría fractal, 14, 20