**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**FACULTAD DE CIENCIAS**

# PROCESOS EPIDEMIOLÓGICOS EN POBLACIONES DE CONECTIVIDAD HETEROGÉNEA

**T E S I S**

**QUE PARA OBTENER EL TÍTULO DE:**

**LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN**

**P R E S E N T A :**

**ADAN EDOARDO HERRERA HIDALGO**


**DIRECTOR DE TESIS:**
**DRA NATALIA BÁRBARA MANTILLA BENIERS**
**CIUDAD DE MÉXICO, 2019**

*A mi madre, a mi padre,*

*y hermanos,*

*por su interminable presencia que alienta mi suspiro.*

# Índice general

**1. Introducción** **5**

**2. Epidemiología matemática** **14**

2.1. Epidemiología . . . . . 14

2.2. Epidemiologia matemática y el modelo $SIR$ . . . . . 17

2.3. Poblaciones heterogéneas . . . . . 20

**3. Teoría de gráficas y redes complejas** **22**

3.1. Definiciones y notación . . . . . 22

3.2. Medidas de centralidad . . . . . 24

3.2.1. Grado . . . . . 25

3.2.2. Cercanía (closeness) . . . . . 27

3.2.3. *Kshell* . . . . . 28

3.2.4. *PageRank* . . . . . 29

3.3. Algoritmo de centralidad k-shell ($\sigma$) . . . . . 30

3.3.1. Análisis del algoritmo . . . . . 33

3.4. Algoritmo de centralidad *PageRank* . . . . . 39

3.4.1. Algoritmo *PageRank* . . . . . 40

3.4.2. Consolidación del algoritmo . . . . . 47

**4. Simulación de procesos de contagio en redes** **53**

4.1. Dinámicas del proceso epidemiológico . . . . . 55
4.2. Medidas de los brotes epidémicos . . . . . 62
4.3. Análisis de los brotes epidémicos . . . . . 65

**5. Estrategias de vacunación** **80**
5.1. Procedimiento . . . . . 80
5.2. Inmunizando la red de colaboración de actores . . . . . 82

**6. Conclusiones** **90**

**A. Descripción de la red de colaboración de actores** **93**
A.1. Descripción de la red . . . . . 93
Referencias . . . . . 100

# Capítulo 1

# Introducción

El objetivo es estudiar procesos de contagio tomando en consideración que la cantidad de interacciones de cada individuo en una población es diversa. Esta heterogeneidad en el número de contactos se traduce en una diversidad de posibles roles en el transcurso de la epidemia. Las herramientas computacionales actuales nos permiten incorporar esta heterogeneidad al diseñar y validar estrategias de vacunación eficaces en dichas poblaciones. Para ello, simulamos un proceso de contagio en una red específica. Estas simulaciones nos dieron una forma de cuantificar la importancia que tiene cada nodo para propagar la infección en la población. Por otro lado caracterizamos la importancia de los nodos por medio de tres medidas de centralidad: grado, *kshell* y *PageRank*.

Estos valores los compararemos con los arrojados por otros algoritmos que buscan proveer a los nodos de la red de una métrica con la que puedan compararse. Dichos algoritmos son conocidos como *medidas de centralidad*, y en particular nosotros aplicaremos el grado, el *kshell* y el *PageRank*. Cabe señalar que cada una de estas medidas tiene su origen en contextos o sistemas reales específicos; así, el *PageRank* nace por la necesidad de consultar eficientemente el creciente número de páginas de Internet, el *kshell* como una caracterización topológica de las gráficas, y el grado es un concepto fundamental en la teoría de gráficas. A pesar de que tienen orígenes diferentes, y posiblemente tam-

bién modelen elementos importantes de distintos sistemas, todas buscan identificar a los nodos más importantes a través de asignarles valores numéricos. Aunque la importancia de un nodo no es absoluta, pues depende tanto del sistema que representa la red, como del proceso que se está estudiando. En nuestro caso, la importancia de un nodo estará dada por su capacidad de propagar una infección en la red, aunque también estudiaremos con mayor detalle los algoritmos de *kshell* y *PageRank* para tener mayor entendimiento sobre sus alcances.

Ya que conozcamos la importancia de cada nodo en la epidemia y calculemos su valor en cada una de las otras medidas, las utilizaremos para plantear estrategias de vacunación y ejecutaremos las simulaciones de contagio nuevamente para evaluar la eficacia de las mismas.

Como ya se mencionó, uno de los conceptos fundamentales que nos interesa abordar en este trabajo es la heterogeneidad en poblaciones. Por tal motivo, emplearemos el marco teórico de redes complejas, ya que con éstas, cada individuo de la población es representado por un nodo y el *contacto directo* de dos individuos de la población se modela con una arista que une a estos dos nodos. Con esta construcción tan sencilla (nodos y aristas que unen dos nodos), las redes capturan una gran diversidad de topologías propias de poblaciones heterogénas. Así, si construimos una red que modele las relaciones de colaboración entre investigadores del departamento de matemáticas, en donde un nodo representa a un matemático y una arista indica si los dos matemáticos que une colaboraron juntos en alguna publicación, es posible que el número de contactos de investigadores con trayectorias más largas sea mayor que la de un investigador nuevo. Por otro lado, si el investigador novato colabora con un investigador reconocido, su red de contactos indirectos se amplía, teniendo mayor alcance que la red de un investigador novato que colabora con otro investigador novato. Éstas son apenas dos situaciones distintas, pero en una comunidad científica real, la topología sería heterogénea. Las regularidades que pudieran existir en la comunidad de colaboraciones

entre matemáticos seguramente serán diferentes a una comunidad de químicos de la misma universidad y a su vez éstos podrían ser diferentes a las poblaciones de otras universidades, y otros países. A su vez, es posible que estas redes de interacciones sean muy diferentes a las de colaboración de actores de películas, especies de animales, usuarios de plataformas virtuales (como *Twitter*, *Facebook*, etc), aunque también es posible que haya rasgos universales. Ahora bien, emplear redes concretas para modelar procesos de contagio puede llevar a una sobreabundancia de modelos específicos que hagan difícil identificar y capturar las propiedades esenciales del proceso de interés. Esto se debe a que al comparar la evolución de un proceso en distintas poblaciones, como el de colaboración de investigadores en poblaciones de matemáticos, químicos de distintas universidades, las redes resultantes podrían tener topologías distintas entre sí que lleven a resultados contrastantes. Cada una de estas redes, además de retratar las dinámicas del proceso de interés en esta población, también estaría capturando, de manera indirecta, sus relaciones con otros procesos reales que intervienen en la dinámica del proceso; en este ejemplo, podrían ser los procesos administrativos, la normativa para publicar que incluso podría involucrar la edad en algunas regiones, los procesos que siguen las diferentes disciplinas de estudio. La modelación de poblaciones específicas por medio de redes captura detalles sobre las conexiones en las poblaciones, introduciendo mayor complejidad y haciendo más difícil obtener resultados universales. Actualmente, se han incrementado las investigaciones que estudian la generalización de propiedades en redes complejas, sin descartar la heterogeneidad de las poblaciones y se ha observado que muchas redes reales de conexiones de comunidades humanas tienen similitudes topológicas que podrían llevar a que el proceso de interés se comporte en forma parecida en todas ellas.

La forma en que se simula el proceso de contagio es otro aspecto central de este trabajo. Para definirla nos basaremos en los modelos de epidemiología clásicos.

Así pues, esta tesis está basada en tres ejes. El primero es la epidemiología matemáti-

ca clásica. El segundo son las redes complejas que nos permitirán modelar el contagio en poblaciones heterogéneas. La última son las computadoras, con las que se simulará y caracterizará la propagación epidémica y el impacto de estrategias alternativas de vacunación.

Estos temas están estructurados como sigue: en el Capítulo 1, además de dar un vistazo general del contenido del trabajo, damos una breve motivación sobre la complejidad y alcance del estudio de las enfermedades infecciosas. En el Capítulo 2 describimos conceptos que describen el fenómeno epidemiológico, en particular a las enfermedades infecciosas. Continuamos describiendo uno de los modelos esenciales de epidemiología matemática y concluimos con una breve discusión sobre los puntos débiles que presenta este modelo, y los cuales nos interesan abordar. El Capítulo 3 contiene definiciones y conceptos de teoría de gráficas, mismas que emplearemos en el resto del trabajo. En este caítulo también le dedicamos una sección para describir el concepto de *medida de centralidad* y explicamos a detalle dos medidas: el *PageRank* y el *kshell*. El Capítulo 4 contiene los resultados de los análisis y el proceso para obtenerlos, que incluye la definición de las simulaciones y su implementación en *Python*, el manejo de los datos, descripción de funciones que capturan el valor de la magnitud de las epidemias y concluimos con una discusión sobre la medida que mejor refleja la importancia que tienen los nodos en propagar una infección. En Capítulo 5 creamos diferentes estrategias de vacunación a partir de las diferentes medidas de centralidad, y las cuales, evaluaremos su eficiencia. Finalmente, en el Capítulo 6 presentamos nuestra conclusiones y en el Apéndice A damos otros detalles sobre la estructura de la red que empleamos a lo largo del trabajo.

## Contexto epidemiológico

Existen un sinfín de enfermedades infecciosas que afectan al humano y que le pueden provocar desde un simple resfriado hasta, en el peor de los casos, propiciar la muerte. A lo largo de su historia, la humanidad ha enfrentado enfermedades con consecuencias de distinta índole y en muchos casos letales. Por ejemplo, la peste es ocasionada por una bacteria que se transmite al humano por la mordida de pulgas o por contacto directo con fómites y si bien es una enfermedad que actualmente está controlada y tiene un tratamiento sencillo, ocasionó 3,248 casos, incluyendo 584 muertes a nivel mundial de 2010 a 2015 (OMS, 2018), en el siglo XIV la peste provocó una pandemia con consecuencias devastadoras. Conocida como la *Muerte Negra*, causó más de 50 millones de muertes en Europa (OMS, 2018) y motivó procesos sociales de gran magnitud, como la migración masiva de distintos grupos en Europa. Otra enfermedad que más recientemente causó temor es el VIH/SIDA, que ha tomado más de 35 millones de vidas humanas y al finalizar el 2016 contabilizaba aproximadamente 36.7 millones de personas contagiadas con el virus (OMS, 2018). A diferencia de la peste, el VIH se transmite directamente de humano a humano por medio del intercambio de ciertos fluidos corporales (como sangre o leche materna) de un individuo infectado a uno que no lo está. Actualmente, los servicios mundiales de salud continúan dedicando un gran esfuerzo para controlar la transmisión de VIH, ya que no existe una cura definitiva, y tan sólo en el 2016 hubo 1.8 millones de contagios y 1.0 millones de personas murieron a causa de esta enfermedad (OMS, 2018).

Recientemente se han propagado otras enfermedades con una tasa de mortalidad alta; por ejemplo, en 2014 inició un brote del virus del ébola que terminó hasta el 2016 (OMS, 2018). Éste se propagó en al menos tres países distintos del Oeste de África y se reportaron cerca de 29 mil casos y más de 11 mil muertes. El ébola se transmite por contacto directo a través de secreciones corporales tanto de animales como de

humanos infectados, o de manera indirecta por contacto con superficies y materiales contaminados con fluidos infectados (OMS, 2018).

Existen sectores de la población que son más vulnerables a ciertos tipos de enfermedades infecciosas. Tal es el caso de la neumonía, que en 2015 representó el 16 % de muertes en niños menores de 5 años de edad y dejó un saldo de 920,136 niños fallecidos (OMS, 2018). La neumonía es una enfermedad que se caracteriza por afectar el sistema respiratorio y puede ser causada por agentes tan diferentes como son virus, bacterias y hongos. Su transmisión ocurre por diversas rutas, siendo un medio común los aerosoles que se emiten mediante la tos o estornudos, aunque también se llega a propagar por vía sanguínea. Un factor importante en el sector salud es el económico; por ejemplo, el tratamiento con antibióticos tuvo un costo alto para tratar la neumonía de los niños de 66 de los países que consideró la iniciativa *Cuenta regresiva 2015*, estimado en 109 millones de dólares al año (OMS, 2018).

Desde luego que existen también enfermedades que reportan una mortalidad baja, aunque en algunos casos son difíciles de controlar o erradicar. Tal el es caso del resfriado común y la influenza estacional. Esta última es una infección viral aguda que se propaga fácilmente de humano a humano y anualmente afecta a la población mundial, independientemente de su edad. Comúnmente, los brotes de influenza inician en otoño, pero bajo ciertas condiciones, puede estallar un brote en otra época del año. Alrededor del mundo, se estima que esta enfermedad resulta en cerca de 3 a 5 millones de casos severos, y entre 290 mil y 650 mil muertes (OMS, 2018). La influenza estacional también suele tener consecuencias económicas, debido a la disminución de la productividad por la reducción temporal de la fuerza de trabajo y el estrés de los servicios de salud.

Las condiciones geográficas específicas de los países y regiones, así como las costumbres y tradiciones de sus habitantes, pueden aumentar o disminuir la tasa de morbilidad de las enfermedades. En México, el monitoreo epidemiológico del chikungunña inició formalmente en 2014, cuando el total de casos reportados fue de 222; en 2015 hubo

un incremento de casi 55 veces con un total de 11,394 casos (Nava-Frías, Searcy-Pavía, Juárez-Contreras, y Valencia-Bautista, 2016). Así como el chikungunña, la enfermedad del Chagas afecta regiones tropicales y subtropicales del país, principalmente zonas rurales. Ambas enfermedades se transmiten por medio de vectores. El ciclo de contagio del chikungunña requiere que los mosquitos hembra transmitan el virus al picar individuos sanos; mientras que la enfermedad del Chagas se transmite a humanos principalmente por las heces de insectos triatomíneos (conocidos como chinches) (OMS, 2018). A pesar de que el chikungunña presenta una tasa de mortalidad muy baja, no existe un tratamiento específico o vacuna para curarla, lo que incrementa su importancia como problema de salud. Lo mismo sucede con enfermedades como el Chagas y el dengue. Éstas, junto con el Zika y la malaria, pertenecen a un grupo importante de enfermedades transmitidas por vectores, que comprende el 17 % de las enfermedades infecciosas a nivel mundial (OMS, 2018).

Los datos descritos anteriormente evidencian la enorme diversidad de enfermedades infecciosas que pueden transmitirse de forma directa (eg influenza) o indirecta (por fómite: cólera, por vector: chikungunña) [1]. Actualmente, las epidemias fácilmente alcanzan difusión global, y son propias tanto de zonas rurales como urbanas. Cuando estas características se mezclan con los fenómenos sociales propios de cada región, se generan procesos de transmisión de enfermedades infecciosas complejos, que suelen estudiarse de forma multidisciplinaria por involucrar fenómenos sociales, biológicos y económicos.

La complejidad de estos procesos de transmisión de infecciones crea vulnerabilidades en las distintas poblaciones de la sociedad, a pesar del gran avance en las ciencias médicas y en la tecnología. Por tal motivo, el estudio multidisciplinario fundamentado en resultados analíticos de modelos matemáticos, la caracterización sistemática de epi-

[1] El modo de transmisión indirecta puede ser mediante vehículos de transmisión o fómites, que son objetos o materiales contaminados tales como juguetes, pañuelos, instrumentos quirúrgicos, agua, alimentos, leche o productos biológicos. O bien, por intermedio de un vector (OPS, 2018).

demias y el uso de tecnologías virtuales innovadoras resultan obligados hoy en día. Todo ello tiene el fin de generar conocimiento nuevo que podrá ser aprovechado, entre otras cosas, en la prevención y el control de enfermedades contagiosas nuevas o existentes, así como en la reducción de los daños a poblaciones expuestas a alguna infección.

La ***epidemiología*** es el área que estudia la ocurrencia de una enfermedad en una población (Esteva, 1991); la ***epidemiología matemática*** emplea el lenguaje de las matemáticas para construir modelos que nos permitan predecir la dinámica epidémica a partir de las particularidades de una infección y su población hospedera. Esta área estudia, también, el comportamiento a largo plazo de la población patógena, así como el posible impacto de una campaña de vacunación en la propagación de una infección.

En 1760 Daniel Bernoulli presentó el primer trabajo de epidemiología teórica (Bernoulli, 1760). Sin embargo fue hasta principios del siglo XX que Hamer realizó la primera contribución importante a la epidemiología teórica (Esteva, 1991) y es a partir del trabajo de Kermark y Mckendrick (1927) que se establecen los pilares de la epidemiología matemática. Desde su formalización y hasta la fecha, se han realizado gran cantidad de investigaciones en esta área, y a pesar de las dificultades, se han podido construir modelos útiles, que nos han aportado comprensión y hasta cierta capacidad de predicción. Estos modelos han derivado en una mejor comprensión de los mecanismos que dan lugar a distintas dinámicas infecciosas y han impulsado el desarrollo de la medicina preventiva.

A los modelos que constituyen cimientos se les conoce comúnmente como modelos clásicos de la epidemiología matemática. Muchos de ellos se basan en el supuesto de que las poblaciones son uniformes y se mezclan de manera homogénea. Pero esto no refleja en muchas ocasiones la realidad, pues lo que ocurre normalmente alrededor del mundo es que las personas nunca interactúan con otras personas que viven lejos de su lugar de residencia o trabajo. Además, no todas las personas conocen la misma cantidad de personas, ni todas frecuentan a personas de diferentes regiones. Tampoco todas tienen

la misma susceptibilidad, infecciosidad, etc.

En los últimos años se ha investigado el efecto de heterogeneidades en red de contactos sobre las dinámicas epidémicas por medio de modelos matemáticos (Kitsak et al., 2010; M. E. Newman, 2002). Al mismo tiempo, la capacidad del humano para obtener, almacenar y procesar datos ha aumentado exponencialmente y ha fomentado el desarrollo de nuevas disciplinas que los aprovechan en sus estudios.

La obtención de datos, el surgimiento de nuevas tecnologías y la investigación de métodos formales y herramientas para manipular y procesar la información han permitido describir y estudiar distintos fenómenos, ya sean procesos biológicos o virtuales, con aplicaciones a diversas ciencias modernas.

# Capítulo 2

# Epidemiología matemática

## 2.1. Epidemiología

En este trabajo nos interesan las enfermedades infecciosas causadas por microorganismos patógenos como bacterias, virus, microparásitos u hongos. Estas enfermedades pueden transmitirse, directa o indirectamente, de una persona a otra. Las ***zoonosis*** son enfermedades infecciosas de animales, que ocasionalmente pueden transmitirse a humanos (OMS, 2018).

En (Jaramillo Arango, Martínez Maya, et al., 2010) se describe la existencia de tres elementos fundamentales en un proceso infeccioso y se conceptualizan como sigue:

**Agente etiológico** (o infeccioso): Organismo, elemento o sustancia cuya presencia o ausencia en un hospedero bajo condiciones ambientales apropiadas sirve como estímulo para iniciar o perpetuar una enfermedad.

**Hospedero** : Organismo que en circunstancias naturales puede alojar un agente etiológico.

**Ambiente** : Entorno físico, biológico y socioeconómico en el cual el hospedero y el agente habitan e interactúan.

Jaramillo Arango et al. describen el ciclo de infección como las diferentes *fases de transición* que permiten la evolución de la enfermedad en sus diferentes periodos (*prepatogénico* y *patogénico*). Las *fases de transición* y los periodos de la enfermedad los describimos a continuación.

Durante el periodo prepatogénico, el agente etiológico, el hospedero y el ambiente interactúan. Este periodo precede a la infección pues aún no ocurre un contacto efectivo entre el agente y el hospedero. La infección, así como los cambios relacionados con la enfermedad, están determinados por las propiedades de los tres elementos mencionados arriba. Al agente biológico lo caracterizan su morfología (determina su ruta de penetración y tipo de infección), su infecciosidad (capacidad de invadir y multiplicarse en el organismo hospedero), su patogenicidad (capacidad para provocar daños específicos al hospedero), su virulencia (severidad de los daños), su inmunogenicidad (capacidad para inducir una respuesta protectora en el hospedero), su variabilidad (que le proporciona adaptabilidad a los cambios) y su viabilidad (capacidad de sobrevivir fuera del hospedero). Por su parte, el hospedero tiene características como especie, edad, sexo y edad fisiológica. El ambiente tiene componentes físicos, biológicos y socioeconómicos.

La modificación de alguna (o varias) de estas propiedades podría reducir las posibilidades de propagación de una infección (incluso nulificarlas) o por el contrario, podría aumentarlas.

El periodo patogénico inicia con el contacto efectivo entre el agente y el hospedero y puede derivar en una enfermedad clínica. Este periodo a su vez se divide en dos fases: en la primera, el agente penetra, se establece y encuentra un medio para multiplicarse (es necesario que al interior del hospedero la abundancia del agente rebase un umbral para que provoque síntomas (Keeling y Rohani, 2011)). Durante el periodo patogénico el hospedero presenta síntomas que derivan en enfermedad, y éste termina cuando se recupera o muere.

De manera general, el ciclo que explica la transición del periodo prepatogénico al

patogénico inicia cuando el agente se propaga por alguna vía, también llamada puerta de salida, sea ésta respiratoria, genitourinaria, digestiva, cutánea u otra. Una vez que el agente es expulsado, éste necesita una mecanismo de transporte que le permita encontrar la vía de entrada a un hospedero. Existen dos formas de transmisión: directa e indirecta. En la indirecta, juega un papel central la transmisión por vectores[1]. Finalmente el agente llega a alguna vía de entrada (como las de la puerta de salida) e ingresa a otro hospedero. En este ciclo, el lapso durante el cual un individuo funge como fuente de infección se denomina periodo infeccioso, y al transmisor de le conoce como enfermo o vector. Aquí vale la pena notar que un vector toma el rol de fuente de infección y modo de transmisión, ya que por un lado es un medio en el que un agente se mantiene y a partir del cual se puede transmitir de manera directa (por ejemplo, mediante mordedura) a un hospededero y al mismo tiempo, es un vehículo que le permite al agente transportarse hacia una vía de entrada de otro hospedero. Por esto, el estudio de los vectores toma gran relevancia en epidemiología, y las particularidades de su ciclo de vida impactan en la lógica subyacente de los modelos matemáticos.

Cuando ya está en marcha un proceso epidémico, las medidas usuales con las que se caracteriza su impacto son:

**Morbilidad** : Número de casos de una enfermedad en la población. Se mide a través de los siguientes indicadores:

**Prevalencia** Cociente de casos y la población total en un lugar y momento dados.

**Incidencia** Cociente de casos nuevos y la población inicial en un lapso de tiempo dado.

[1]En epidemiología, un ***vector*** es un organismo vivo que puede transmitir enfermedades infecciosas entre personas, o de animales a personas. Muchos de esos vectores son insectos hematófagos que ingieren los microorganismos patógenos junto con la sangre de un portador infectado (persona o animal), y posteriormente los inoculan a un nuevo portador al ingerir su sangre (OMS, 2018).

**Tasa de mortalidad** Número de muertes respecto de la población total.

**Tasa de letalidad** Número de muertes respecto del total de individuos enfermos.

## 2.2. Epidemiologia matemática y el modelo $SIR$

En esta sección describimos el modelo $SIR$, que forma parte de los modelos epidemiológicos clásicos, entre los que se encuentran los modelos $SI$, $SIRS$ y $SIS$.

De acuerdo con (Keeling y Rohani, 2011), el modelo *SIR* sin demografía se utiliza para estudiar las enfermedades infecciosas agudas, es decir, aquellas en las que la enfermedad dura un periodo breve comparado con la esperanza de vida del hospedero. El modelo $SIR$, además, considera sólo a las enfermedades en las que el hospedero genera inmunidad de por vida.

Estos modelos segmentan a la población hospedera en compartimientos, que describen su condición infecciosa. En particular, el modelo $SIR$ tiene tres compartimientos: **S**usceptible, donde están quienes no han sido expuestos a la enfermedad y la pueden contraer; **I**nfectados, aquellos que contrajeron y desarrollaron la enfermedad de tal forma que la pueden transmitir; y **R**ecuperados, quienes ya se curaron de la enfermedad y no pueden volverse a contagiar ni transmiten la enfermedad. Al clasificar a los individuos de la población en estos tres compartimientos, se acotan los posibles estados epidemiológicos que un individuo puede asumir.

Para concluir la construcción del modelo $SIR$ es necesario determinar las fases de transición de un compartimiento a otro: el sistema inicia con una cantidad inicial $S_0$, de individuos susceptibles y cuando no se consideran procesos demográficos, no pueden generarse susceptibles nuevos; de forma que no ingresan individuos al compartimiento de susceptibles y únicamente salen de éste al infectarse. Así, entran al compartimiento de los infectados los susceptibles que contraen la enfermedad, y salen de él quienes se recuperan. Este esquema retrata a aquellas enfermedades que generan una inmunidad

que dura el resto de la vida del hospedero (como la viruela), suprimiendo la posibilidad de que los individuos en el compartimiento de los recuperados se desplacen al de los susceptibles.

Como se mencionó antes, omitimos los cambios demográficos en la población, por lo que nacimientos y muertes no alteran el flujo antes descrito, quedándonos con una población constante y cerrada (pues no hay flujo hacia o desde ella). Notamos que, en caso de incluir fenómenos demográficos, que son comúnmente nacimientos y muertes, tendríamos que realizar algunos ajustes al flujo de los estados epidemiológicos de los individuos. La modelación de nacimientos determinaría la incorporación de nuevos individuos a la población que, en ausencia de una vacuna, ingresarían directamente al compartimiento de susceptibles, ya que en el caso epidemiológico más sencillo las personas no nacen con enfermedades infecciosas o no las pueden transmitir. Una estrategia típicamente adoptada para incluir demografía en los modelos es suponer una población constante pero no cerrada, es decir, restar individuos (muertes) de uno o más compartimientos con la misma tasa con la que se introducen personas nuevas (nacimientos) a los susceptibles.

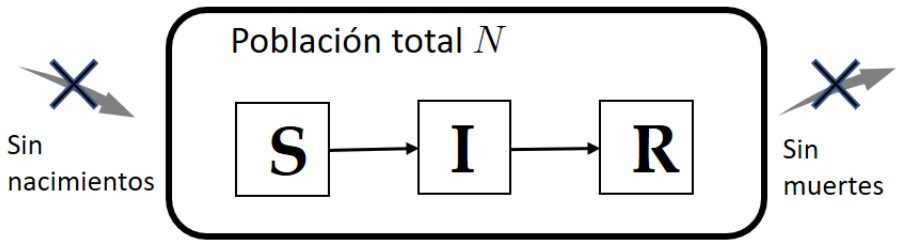



Figura 2.1: Diagrama que muestra los caminos posibles de individuos de la población para transitar de un compartimiento a otro. El tamaño de la población, $N$, se mantiene constante y no se permiten nacimientos ni muertes.

A fin de modelar el proceso de contagio, denotamos con $N$ a la población total y con $X$, $Y$, y $Z$ al número de susceptibles, infectados y recuperados respectivamente,

así, definimos los siguientes términos: $S = X/N$, $I = Y/N$ y $R = Z/N$, que expresan las fracciones de individuos en cada compartimiento. Además, definimos el término $\beta$ como el producto de la tasa de contactos y la probabilidad de transmisión (Keeling y Rohani, 2011). Por lo tanto, el número de infectados nuevos está dado por los individuos susceptibles que entran en contacto con los individuos infectados y que se infectan, lo que queda determinado por la ecuación diferencial, $\frac{dS}{dt} = -\beta SI$

Análogamente, la variación en la fracción infectada tendrá un incremento causado por las nuevas transmisiones entre susceptibles e infectados y un decremento de los individuos infectados que se recuperan a una tasa por la probabilidad de recuperación $\gamma$. Su razón instantánea de cambio se describe mediante la ecuación, $\frac{dI}{dt} = \beta SI - \gamma I$.

Por último, la fracción de recuperados tendrá un incremento como consecuencia de las recuperaciones, $\frac{dR}{dt} = \gamma I$

El modelo $SIR$, junto con modelos que retratan infecciones crónicas ($SI$) o que no generan buena inmunidad ($SIS$), conforman los modelos clásicos dentro de epidemiología matemática, aunque existe una diversidad de modelos con variaciones significativas que consideran diferentes características del proceso infeccioso y otros relacionados (Keeling y Rohani, 2011). En términos generales, estos modelos añaden o eliminan compartimientos que describen fases particulares de la enfermedad, retratan heterogeneidades o capturan el flujo de los individuos entre estos compartimientos mediante términos alternativos. Así, los modelos clásicos de epidemiología matemática abstraen el ciclo de infección epidemiológico que se explicó al inicio de este capítulo con base en dos pilares: primero, incorporar las clases adecuadas a la enfermedad y acordes al objetivo de la investigación; segundo, determinar las posibles transiciones de los individuos de un compartimiento a otro. De esta manera, por ejemplo, para modelar las enfermedades que se transmiten por un vector, podemos emplear el modelo $SIR$, tomando como una sola fase la transmisión hospedero-vector-hospedero. Alternativamente, podemos crear compartimientos que modelen los estados del vector y sus transiciones respectivas.

## 2.3. Poblaciones heterogéneas

Desde luego que el modelo $SIR$ falla en capturar la enorme diversidad de procesos y situaciones existentes. Esta exposición no retrata la diversidad de técnicas empleadas en epidemiología matemática, pero es un punto de partida e ilustra los siguientes supuestos de estos modelos:

1. Mezclado homogéneo. Todos los individuos interactúan con la misma probabilidad con el resto de la población.

2. Susceptibilidad e infectividad homogéneas: no se reconocen diferencias individuales en estos rasgos.

3. Las tasas de transmisión y recuperación son constantes, lo que se traduce en una distribución exponencial de los tiempos de espera en la case infectada (Wearing, Rohani, y Keeling, 2005)

4. Tamaño de población constante y cerrada. Los individuos iniciales de la población son los mismos durante todo el proceso, permitiendo únicamente sus transiciones de un compartimiento a otro. El modelo descrito en la sección anterior tampoco incluye elementos demográficos como nacimientos y muertes, aunque esta característica se puede modelar fácilmente [ver por ejemplo (Keeling y Rohani, 2011, Capítulo 2.1.2)].

El primer supuesto llama la atención ya que en una población real los individuos solamente están en contacto con una fracción de ésta, y durante un brote epidémico real un individuo infectado puede contagiar o transmitir la enfermedad únicamente a un número relativamente pequeño de personas de la población, que además varía con cada persona. Por lo tanto, la hipótesis de mezclado homogéneo elimina diferencias potencialmente importantes para el proceso epidemiológico.

Actualmente, los supuestos anteriores han sido modificados y estudiados de diferentes maneras, dependiendo del aspecto que se desea estudiar del proceso epidémico. Sin embargo, es con la aparición de la ciencia de redes y el incremento de nuestras capacidades de cómputo que se consideró la modelación de la transmisión de enfermedades en una comunidad desde otra perspectiva, preservando la estructura contactos en poblaciones reales.

En este trabajo, seguiremos una línea de investigación que se basa en simular el ciclo de infección de la enfermedad y su propagación mediante procesos estocásticos de contagio dentro de redes y posteriormente estudiaremos diferentes aspectos del proceso simulado.

Para ello, construimos una red específica, que describimos a detalle en el Sección A.1, y que utilizamos para simular el proceso de contagio conforme se detalla al inicio del Capítulo 4 y con base en los conceptos explicados en esta sección realizamos las simulaciones. Finalmente, analizamos la relación entre el tamaño final de una epidemia y la caracterización de sus nodos que dan distintas medidas de centralidad, y evaluamos el éxito de campañas de vacunación que se definen a partir de dicha caracterización.

Una parte importante del análisis de los datos, así como la selección de los algoritmos que utilizamos para estudiar la propagación de la enfermedad en las redes, se basa en el artículo de (Kitsak et al., 2010).

# Capítulo 3

# Teoría de gráficas y redes complejas

En este capítulo se establece la notación y se dan las definiciones de teoría de gráficas que utilizaremos. En particular describiremos brevemente algunas de las medidas de centralidad más comunes, y explicaremos a mayor detalle las dos medidas que elegimos para el estudio de los procesos epidemiológicos que son el *PageRrank* y el *kshell*.

## 3.1. Definiciones y notación

Una **red** (o gráfica) es una colección de ***nodos*** (o ***vértices***), $V = (v_1, \ldots, v_n)$, unidos mediante ***ligas*** (o ***aristas***), $E = \{\overline{uv} | u, v \in V\}$. Denotamos a la red como $G(V|E)$ o simplemente $G$. Comúnmente se denotan con $n$ y $m$ al número de nodos y al número de ligas de la red, respectivamente.

Cuando un par de nodos está unido por más de una arista, decimos que tiene ***aristas múltiples***. Un **bucle** es una arista que sale y regresa a un mismo nodo. Cuando la red no presenta aristas múltiples ni bucles se le conoce como **red simple**. Trabajaremos únicamente con redes simples en esta investigación.

Una red también puede ser **dirigida** (*digráfica*) o **no dirigida**. Cuando la red es dirigida, una arista que va de $u$ a $v$ no sirve para ir de $v$ a $u$, mientras que en la red no

dirigida, la dirección puede ir en ambos sentidos.

Definimos el **grado** de un nodo, $v$, como el número de ligas de la red en las que aparece $v$ y lo denotamos como $\delta(v)$. Cuando se necesita especificar la gráfica en la que se cuentan las ligas, se añade la etiqueta de la red (por ejemplo $G$) como subíndice: $\delta_G$. En el caso de redes dirigidas se especifica el grado de entrada, $\delta_{\text{in}}(v)$ y salida, $\delta_{\text{out}}(v)$, que cuentan las aristas que llegan a $v$ y las que salen de $v$ respectivamente. El **grado mínimo** de la red, $\delta(G)$, indica el grado más pequeño de los nodos de la red $G$.

La **matriz de adyacencia** permite representar a una red en su totalidad. Para redes no dirigidas, las entradas de esta matriz, $A$, son 0 o 1 y se definen como $A_{ij} = 1$ si hay una arista que va del nodo $j$ a $i$ y $A_{ij} = 0$ si los nodos $i$ y $j$ no están conectados mediante una liga o arista.

Una **supergráfica** es la gráfica formada de agregar nodos, aristas o ambos a una gráfica dada. Decimos que $H$ es una ***subgráfica*** de $G$ si $G$ es una supergráfica de $H$. Una subgráfica, $H$, de $G$ es **maximal** para cierta propiedad si tiene la propiedad y ninguna otra supergráfica de $H$ que sea subgráfica de $G$ también cumple la propiedad.

A continuación damos otras definiciones y notación de conceptos esenciales en teoría de gráficas. Estas definiciones las tomamos del texto de (Cormen, Leiserson, Rivest, y Stein, 2009, pág. 1168):

1. Un **camino** de longitud $k$ de un vértice $u$ a un vértice $v$ es una secuencia $\langle v_0, v_1, \ldots, v_k \rangle$ de vértices tal que $u = v_0$ y $v = v_k$ y $\overline{v_{i-1}v_i} \in E$. Un camino es **simple** si todos los vértices en el camino son diferentes. Si hay un camino $p$ de $u$ a $v$, decimos que $v$ es **alcanzable** desde $u$ por $p$.

2. En gráficas no dirigidas un camino $\langle v_0, v_1, \ldots, v_k \rangle$ forma un **ciclo** si $k > 0$ y $v_0 = v_k$ y todas las aristas en el camino son diferentes.

3. Una gráfica no dirigida es **conexa** si todos los vértices son alcanzables desde los otros vértices. Una gráfica es **acíclica** si no contiene ningún ciclo.

4. Un **árbol** es una gráfica no dirigida, conexa y acíclica.

5. Una gráfica simple es **completa** si cualquier par de vértices está conectado por una arista.

6. Una gráfica es ***k-regular*** si todos los nodos tienen grado $k$.

## 3.2. Medidas de centralidad

Las ***medidas de centralidad*** corresponden a distintas formas de calcular y asignar valores a los nodos para medir su importancia; entre mayor sea el valor de una medida para un nodo, mayor será su importancia o "centralidad dentro de la red. Hay varias medidas, ya que no existe un concepto universal de importancia, y éste dependerá de los objetivos de nuestro estudio y del sistema que se modela. Por tanto, un nodo puede ser muy importante de acuerdo con una medida y poco importante para otra. Dicho esto, las medidas de centralidad más conocidas y utilizadas están relacionadas entre sí, y frecuentemente observamos que los nodos con una medida centralidad alta tienen otras medidas altas también (aunque el orden exacto difiera o los valores no guarden una relación proporcional).

En esta sección definimos distintas medidas de centralidad que forman parte de una variedad de algoritmos y técnicas que se han desarrollado con el fin de caracterizar las redes a partir de asignarles valores a los nodos con base en diferentes propiedades y resultados. Estos algoritmos comúnmente devuelven valores numéricos positivos, ya sean reales o enteros.

El estudio de medidas de centralidad es de gran importancia para este trabajo, ya que éstas nos proveerán de un marco cuantitativo con el cual determinamos la posible influencia de cada nodo en el proceso de contagio.

### 3.2.1. Grado

Definimos previamente el grado de un nodo como el número de aristas de éste, en el caso de redes no dirigidas. En el caso de redes dirigidas, definimos el grado de salida y el grado de entrada como el número de aristas que salen o entran al nodo, respectivamente. En esta sección retomamos este concepto para discutir su valor como medida de centralidad.

Desde esta perspectiva, es la medida mejor conocida y más utilizada, y presenta distintas ventajas: es fácil de calcular, pues únicamente demanda contar; es un concepto muy intuitivo, pues si un nodo tiene muchas aristas, entonces tiene muchos vecinos, y por tanto es "importante"; la usabilidad: debido a que su cálculo es sencillo, se puede aplicar el grado en trabajos teóricos para derivar conceptos más sofisticados (como la distribución del grado del nodo) o en trabajos prácticos (como categorizar redes específicas). Finalmente, su cálculo con cualquier lenguaje de programación también resulta sencillo. Esta cualidad toma relevancia debido a que, como ya se mencionó al inicio de este capítulo, la ciencia de redes usa el cómputo como un pilar, y debe ser posible calcular estas medidas en redes de todo tipo que sea posible construir. Este enfoque conlleva una variedad de problemas que se deben solucionar, siendo el más evidente la necesidad de trabajar con redes masivas y cuyo análisis y manipulación requieren algoritmos o sistemas de cómputo más especializados.

En años recientes se extendió el concepto de nodo de concentración (mejor conocido como *hub* en inglés), que se refiere a aquellos nodos con un número de aristas mucho mayor al promedio que, por lo mismo, son nodos poco frecuentes en redes reales. Se ha visto que en distintas redes reales existen muchos nodos con pocas aristas y unos cuantos nodos con muchas aristas (los *hubs*) (M. Newman, 2010; Barabási et al., 2016)

Tomando en cuenta la naturaleza de estos *hubs*, se puede decir que éstos son los más importantes, pero se ha mostrado que existe una gran cantidad de situaciones en

que otras medidas de centralidad reflejan con mayor precisión la importancia de un nodo en un proceso dado. Por ejemplo, en este trabajo veremos que al menos para la red utilizada el *kshell* supera al grado del nodo en el contexto de los procesos epidemiológicos, pues un valor de *kshell* alto corresponde a nodos en que invariablemente inicia una epidemia. Otro ejemplo típico, para mostrar que los llamados *hubs* de la red no están necesariamente asociados con la centralidad de los nodos, son los procesos subyacentes a los motores de búsqueda en la red de Internet. Cuando realizamos una consulta en algún motor de búsqueda, esperamos que las páginas de Internet sugeridas por el motor sean relevantes para esta consulta. En estos procesos, los *hubs* no suelen ser páginas relevantes. En cambio, el motor de búsqueda más utilizado, *Google* se basó originalmente en el algoritmo de jerarquización *PageRank*, un algoritmo que ordena por relevancia a las páginas y mismo que explicamos más adelante.

El grado es un indicador pobre porque se asocia comúnmente a que éste es una medida meramente local, y como tal, únicamente considera información de la topología inmediata de la red, lo que constituye un nivel de profundidad mínimo, al tomar en cuenta sólo los nodos contiguos.

Finalmente, destacamos que la selección del grado o cualquier otra medida de centralidad para el estudio de una red dependerá completamente del problema que se estudia, y como es habitual, no existen recetas únicas para establecer un criterio. Por ejemplo, si lo que deseamos es identificar a los individuos que conocen a más personas, digamos en todos los estudiantes de nivel superior, entonces basta con crear una red de contactos de esta población, y el grado de los nodos responderá exactamente esta pregunta. Así que el grado como primer acercamiento para determinar la importancia de cada nodo dentro de una red puede ser una estrategia útil.

### 3.2.2. Cercanía (closeness)

La medida de centralidad conocida como ***cercanía*** $C$ de un nodo $v$ de la red $G(V, E)$ se define como el inverso multiplicativo del promedio de la distancia más corta, $d$, del nodo $v$ al resto de los nodos de la red. Se denota $C(v)$, y se define como

$$C(v) = \frac{n}{\sum_{u \in V(G)} d(v, u)}$$

Intuitivamente, asociamos mayor importancia a nodos con valores altos de centralidad. Puesto que nodos muy comunicados tienen promedios de distancia bajos, se toma el inverso de dichos promedios para definir esta medida.

Se ha visto que en redes reales, las distancias geodésicas de una red aumentan logarítmicamente con el tamaño de la red (M. Newman, 2010). Esto tiene como consecuencia que los valores de $C$ se ubiquen en un intervalo pequeño, que va desde 1 hasta $\log n$. Tomamos como ejemplo el descrito en (M. Newman, 2010, pág. 183), de una red de colaboración de actores construida a partir de datos obtenidos del sitio *Internet Movie Database*. Se encontró que en el componente más grande, con el 98 % de los actores, el mayor valor de *closeness* era 0.4143 del actor Christopher Lee, y el más bajo de la actriz Leia Zanganeh con 0.1154. La razón de estos valores es tan sólo de 3.6, pese a que hay casi medio millón de actores.

### Intermediación (betweenness)

La conceptualización de esta medida se da dentro del contexto de la teoría de la información, en donde los nodos intercambian información y ésta se empaqueta y viaja de nodo en nodo a través de sus aristas. Se busca que la información se envía recorriendo la mínima cantidad de nodos y aristas, es decir, seguirán el camino más corto disponible. (M. Newman, 2010) De esta manera, todos los nodos mandan y reciben información, y pueden ser parte de una o varias rutas de comunicación. La intermediación busca

identificar a los nodos que aparecen en el mayor número de rutas. Esta medida considera el número total de caminos geodésicos $L(u, v)$ entre dos nodos $u$ y $v$, así como el número de caminos geodésicos entre esos nodos que pasan por el nodo $x$, $L_x(u, v)$. De acuerdo con (Pastor-Satorras, Castellano, Van Mieghem, y Vespignani, 2015), la intermediación se define como

$$B(x) = \sum_{u \neq v} \frac{L_x(u, v)}{L(u, v)}$$

.

La importancia de cada nodo dependerá entonces del número de caminos más cortos en los que participa, mismos que se identifican tomando todas las posibles parejas de nodos en la red y trazando los caminos más cortos entre ellos.

### 3.2.3. *Kshell*

El algoritmo que se utiliza para calcular el *kshell* se describe detalladamente en Sección 3.3, ya que utilizamos esta medida para estudiar la importancia de cada nodo en procesos de contagio.

Para fines de esta descripción, proporcionamos la definición condensada que da Rodrigues (2019):

> El k-core es una subgráfica tal que todos sus vértices tienen grado al menos $k$. Esta medida de centralidad se obtiene por el algoritmo de descomposición de *kshell*, en el que se particiona la red mediante un proceso iterativo que elimina los nodos con grado menor a $k$. Después de eliminar estos nodos, el proceso reanaliza la red para verificar si quedaron nodos con grado menor que $k$. Si aún existen, también se eliminan. Este proceso se repite hasta que el grado mínimo en la red es $k$. La subgráfica resultante es llamada *k-core* de la red. Un nodo $i$ tiene *coreness* $\sigma(i) = k$ si pertenece al *k-core* pero

no pertenece al $(k+1)$*core*. De acuerdo con esta medida, los nodos más centrales tienen mayores valores de *coreness* (número *k-core*).

### 3.2.4. *PageRank*

El algoritmo de *PageRank* lo veremos en detalle a la Sección 3.4. Este algoritmo fue propuesto por Larry Page y Sergey Brin en su tesis de doctorado a finales de los años 90, como fundamento teórico para el motor de búsqueda *Google*, que también desarrollaron y que pronto tuvo una gran difusión.

El algoritmo está pensado para aplicarse a la red de Internet y mide la importancia de un nodo con base en dos conceptos: la cantidad de páginas que hacen referencia al nodo y la importancia de cada una de estas páginas. Por tanto, la importancia de un nodo dependerá del número de nodos importantes que hacen referencia él.

Actualmente el motor de búsqueda *Google* utiliza una variedad de algoritmos que resuelven otros problemas además de ordenar por relevancia las páginas web [1].

En esta sección introdujimos apenas algunas de las medidas de centralidad que se han propuesto en los últimos años, pero existen muchas otras de las que ya no hablaremos en mayor detalle, aunque vale la pena mencionar algunas. Así, por ejemplo, *TwitterRank*, que de acuerdo con sus creadores en (Weng, Lim, Jiang, y He, 2010), es una extensión de *PageRank* y mide la influencia de los usuarios en *Twitter* basándose en los temas preferidos de los usuarios y la estructura topológica de la red. Otras medidas utilizadas son: *Katz*, *Eigenvector*, *VoteRank*, *LocalCentrality*, *ClusterRank*, *LeaderRank* y *HITS* (M. Newman, 2010; Chen, Lü, Shang, Zhang, y Zhou, 2012; Zhang, Chen, Dong, y Zhao, 2016; Pastor-Satorras et al., 2015)

[1] Para conocer más detalles sobre el estado general actual del motor de búsqueda *Google* invitamos al lector a explorar la liga de `https://www.google.com/search/howsearchworks/`.

## 3.3. Algoritmo de centralidad k-shell ($\sigma$)

En esta sección estudiamos la noción de *k-core* que nos permite definir el algoritmo de *k-shell*. Ambos conceptos se usan para clasificar y comparar nodos tomando en cuenta la topología de la red de la que son parte, de lo que se deriva una medida de centralidad.

Para motivar estos conceptos nos planteamos las siguientes preguntas: De las gráficas en la Figura 3.1, ¿cuál será más difícil de *deshacer*? o dicho de otra forma ¿cuál gráfica tiene una estructura más "robusta"?

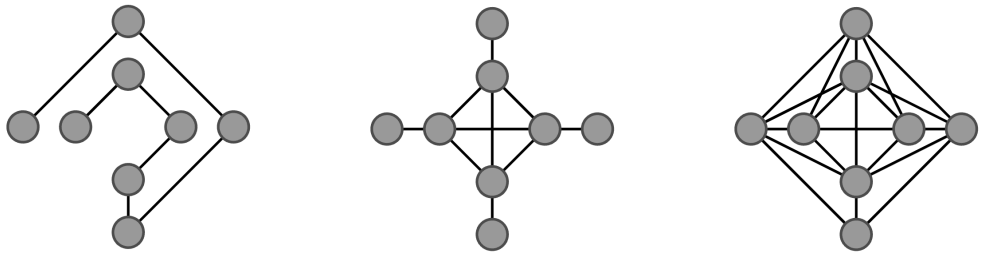


Figura 3.1: Se muestran tres gráficas con 8 nodos pero distintos números de aristas. El número de aristas aumenta de izquierda a derecha (7, 10 y 14 respectivamente), siendo la del extremo derecho la más parecida a una gráfica completa, y la que sugiere mayor robustez en su estructura topológica.

De la Figura 3.1, y de manera intuitiva, podemos decir que la gráfica más robusta es la que más se asemeja a una gráfica completa (derecha) y la más *frágil* la que tenga pocas conexiones entre sus nodos (izquierda). Por su parte, la gráfica de en medio contiene una subgráfica completa, compuesta por cuatro nodos, que le aporta una estructura más sólida; pareciera incluso que es mucho más robusta que la gráfica de la izquierda, a pesar de que ésta tiene tan solo 3 aristas menos.

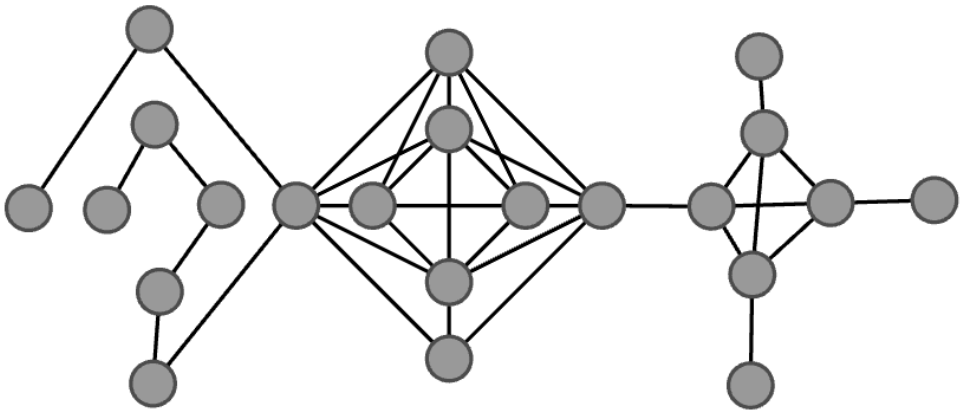


Figura 3.2: Gráfica que resulta de unir las mostradas en la Figura 3.1, obteniendo una topología heterogénea ya que contiene, al menos, una subgráfica "robusta" y una "débil".

Si consideramos ahora una gráfica en que se unen las tres gráficas de la Figura 3.1 (como lo muestra la Figura 3.2), resulta más complicado hacer una clasificación general de esta nueva gráfica en términos de su *robustez*, pero podríamos iniciarla dividiendo la gráfica de la Figura 3.2 en subgráficas como las de la Figura 3.1 que son más fáciles de clasificar. El algoritmo *kshell* utiliza una estrategia similar a ésta, que es la de buscar una clasificación general de la gráfica aprovechando características más fáciles de capturar de sus subgráficas.

A continuación definimos formalmente el concepto de *k-core* y el algoritmo del *kshell*. Estos conceptos permiten definir una medida de centralidad, ya que establecen una medida de importancia de los nodos cuantificando su cohesión en la gráfica: los nodos de cohesión alta son los que resultan más difíciles de eliminar de la gráfica y los de cohesión baja los que se desprenden más fácilmente. Estos dos conceptos fueron propuestos originalmente en (Seidman, 1983).

**Definición 3.3.1.** *K-core. Sea $G$ una gráfica y $H$ subgráfica de $G$. Decimos que $H$ es*

*un* k-core *de $G$ si $H$ es una subgráfica conexa maximal de $G$ con $\delta(H) \geq k$.*

De esta definición aparentemente sencilla obtenemos resultados interesantes que, entre otras cosas, permiten medir la cohesión de los nodos con base en los enlaces que forman distintos grupos de nodos.

**Proposición 3.3.1.** *Sea $G(V, E)$ una gráfica. Si $H(V_H, E_H)$, $I(V_I, E_I)$ son* k-cores *de $G$ para una misma $k \geq 1$, entonces no existe $\overline{uv} \in E$ tal que $u \in V_H$ y $v \in V_I$.*

*Demostración.* Supongamos que sí existe la arista $\overline{uv}$ con $u \in V_H$ y $v \in V_I$ y sean $n$, $m$ tales que $n = \delta_H(u)$ y $m = \delta_I(v)$. Como $H$ e $I$ son *k-core* de $G$ entonces $n$, $m \geq k$. Construimos $L(V_L, E_L)$ gráfica tal que $V_L = V_H \cup V_I$ y $E_L = E_H \cup E_I \cup \{\overline{uv}\}$. Como $V_L \subseteq V$ y $E_L \subseteq E$, $L$ es subgráfica de $G$, y como $H$ e $I$ son *k-core*, entonces $L$ es conexa. Además se tiene que $\delta_L(u) = n + 1 > k$ y $\delta_L(v) = m + 1 > k$, por lo que $\delta(L) \geq k$, y como $L$ es una supergráfica de $H$ e $I$, entonces $H$ e $I$ no son gráficas maximales, lo que contradice la hipótesis. Esta contradicción proviene de suponer que $\overline{uv} \in G$ tal que $u \in V_H$ y $v \in V_I$ por tanto, $H$ e $I$ no son adyacentes. □

**Proposición 3.3.2.** *Sea $G(V, E)$ conexa. Entonces $G$ tiene a lo más una gráfica* 2-core.

*Demostración.* Supongamos existen $H(V_H, E_H)$, $I(V_I, E_I)$ gráficas *2-core* de $G$. Por la conexidad de $G$, para cualesquiera nodos $u \in V_H$ y $v \in V_I$ existe una trayectoria en $G$ $t = uu_1 \dots u_i x_1 \dots x_n v_1 \dots v_j v$ tal que $u, u_i \in V_H$ y $v, v_j \in V_I$ y $x_i \in V - V_H - V_I$. Por el resultado **1** se cumple que $n \geq 1$ y además, $\delta(y_i) \geq 2$ $\forall y_i \in t$.

Construimos la gráfica $L(V_L, E_L)$, $V_L = V_H \cup V_I \cup \{y_i \in t\}$ y $E_L = E_H \cup E_I \cup \{\overline{yz} \in t\}$. De esta manera, $L$ es subgráfica de $G$, es conexa y $\delta(L) \geq 2$. Dado que $L$ es una supergráfica de $H$ e $I$, $H$, $I$ no son gráficas maximales, por lo que no son *k-cores* de $G$, que contradice nuestra suposición. Por lo tanto, no existe más de una dos *2-core* de $G$ gráfica conexa. □

**Proposición 3.3.3.** *Si $G$ es un árbol de grado $\delta$ entonces $G$ a lo más tiene un 1-core*

*Demostración.* Sea $H$ subgráfica conexa de $G$, entonces $H$ es conexa y acíclica, es decir, $H$ es también un árbol. Como el grado de las hojas de un árbol siempre es 1, entonces no existe $H$ tal que $\delta(H) \geq k$ con $k \geq 2$. Por lo tanto, no existe un *k-core* de $G$ con $k \geq 2$. □

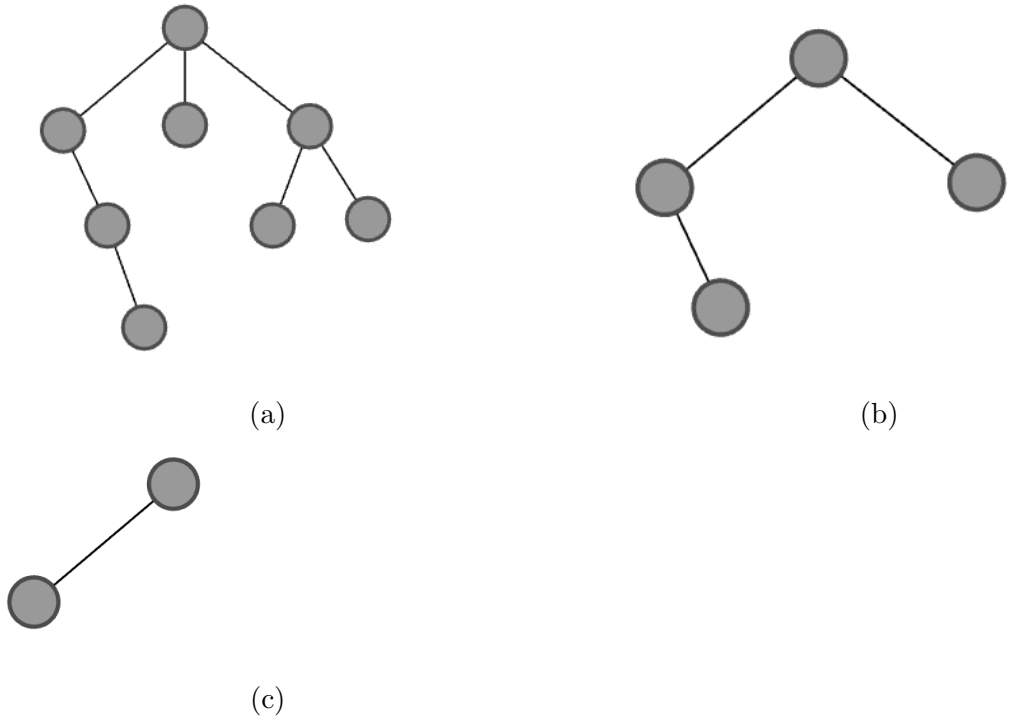
**Proposición 3.3.4.** *Sea $H(V_H, E_H)$ un k-core de G. Entonces $|V_H| \geq k + 1$.*

*Demostración.* Sea $v \in V_H$. Como $\delta(H) \geq k$ entonces $N(v) \geq k$, por lo tanto $|V_H| \geq k + 1$. □

Ahora estamos listos para explicar el algoritmo *k-shell* que le asigna a cada nodo un valor conocido en la literatura como índice *kshell* o en otros casos como *core number* y que denotamos con $\sigma$.

### 3.3.1. Análisis del algoritmo

La Figura 3.3 muestra un árbol de altura 4 y grado máximo 3. Empecemos por eliminar o *podar* los nodos de grado 1, que de manera intuitiva son aquellos con menor adherencia a la gráfica y en consecuencia los más fáciles de desconectar. Continuamos bajo el mismo criterio de podar los nodos con menor adherencia en la gráfica que resultó de la poda previa, lo que deja una subgráfica con otros nodos de grado 1. La gráfica resultante únicamente tiene dos nodos de grado 1, que al momento de podarlos nos deja la gráfica vacía, lo que constituye la condición de paro de este procedimiento. En este ejemplo, la poda de los nodos en cada iteración tuvo la misma "dificultad", es decir, involucró eliminar sólo una arista para cada nodo de grado 1. Así que podemos identificar el valor de *cohesión* de cada nodo en el árbol original con la cantidad de aristas que necesitamos podar (1) en una iteración dada.



(a) (b)

(c)

Figura 3.3: Secuencia de gráficas obtenidas de iterar el algoritmo de *kshell*. La Figura 3.3a muestra la gráfica original, un árbol de grado máximo 2. La Figura 3.3b resulta de eliminar a los nodos de menor cohesión, que este caso son los de grado 1. La Figura 3.3c nuevamente tiene nodos de grado 1, que se deberán eliminar. Detenemos el proceso al obtener la gráfica vacía y le asignamos a cada nodo el valor del número de aristas que le quitamos, en el momento del algoritmo en que lo eliminamos.

El pseudocódigo que formaliza este procedimiento se presenta en el recuadro siguiente:

**Algoritmo 1** Asigna el índice kshell, $\sigma$, a cada nodo de la gráfica $H(V, E)$

1: $k \leftarrow \delta(H)$ ▷ grado mínimo de $H$
2: **while** exista nodo en $H$ **do**
3: seleccionar a los nodos con grado menor o igual que $k$
4: asignar valor $\sigma \leftarrow k$ a los nodos seleccionados
5: eliminar nodos seleccionados
6: **if** $\delta(H) \leq k$ **then**
7: repetir desde 3
8: **else**
9: $k \leftarrow \delta(H)$
10: **end if**
11: **end while**

Este algoritmo le asigna un valor de *k-shell*, que denotamos como $\sigma$, a cada uno de los nodos de la gráfica, $H(V, E)$.

A continuación aplicamos este algoritmo a la siguiente gráfica.

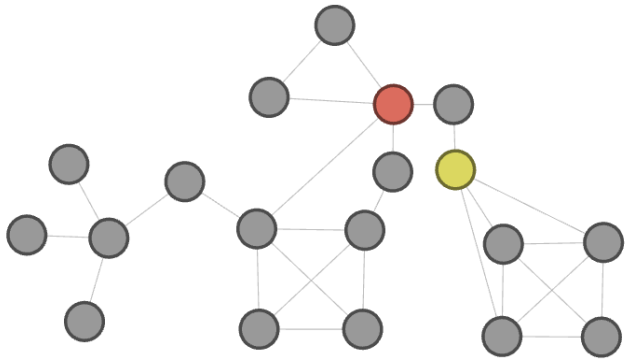


Figura 3.4: Obtenemos el valor de *kshell* de cada nodo al aplicar el algoritmo. Coloreamos dos nodos con grados diferentes, el rojo tiene grado 5 y el amarillo grado 4. El nodo amarillo es parte de una subgráfica 3-regular y el rojo no lo es.

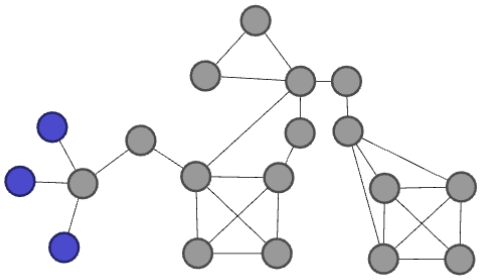


Figura 3.5

Primeramente seleccionamos los nodos con grado 1 (nodos azules en Figura 3.5). Les asignamos el valor $\sigma = 1$ y los eliminamos de la gráfica junto con sus aristas.

Continuamos eliminando los nodos de grado menor o igual a uno mientras sigan apareciendo (nodos azules en la Figura 3.6). En este caso, después de la poda de la Figura 3.5, asignamos $\sigma = 1$ al nodo azul en la imagen de arriba y lo podamos. Obtenemos la gráfica de la imagen de abajo, que nuevamente tiene un solo nodo de grado 1, al que le asignamos $\sigma = 1$ y lo podamos.

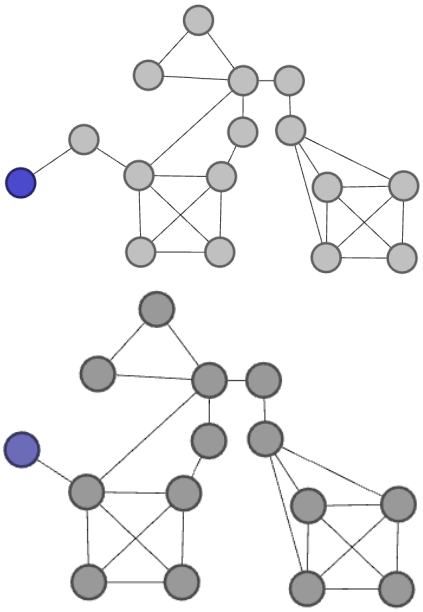


Figura 3.6

Debido a que ya no hay nodos de grado menor o igual a uno, seleccionamos los nodos de grado 2 (nodos azules en la Figura 3.7). Les asignamos el valor $\sigma = 2$ y los eliminamos de la red.

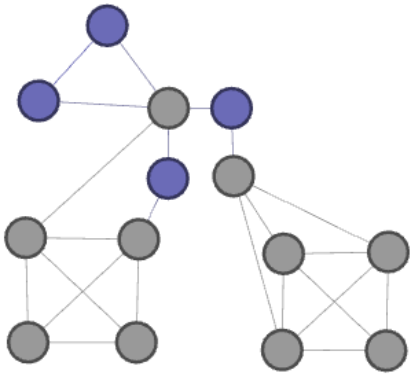


Figura 3.7

La Figura 3.8 muestra la gráfica resultante de la poda anterior. El nodo azul es el único con grado menor o igual a 2, por lo que también lo etiquetamos con $\sigma = 2$ y lo eliminamos. Después de esta poda, ya no quedan nodos con grado menor o igual que 2.

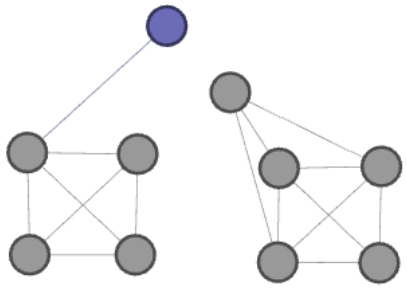


Figura 3.8

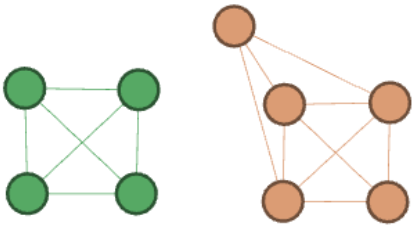


Figura 3.9

Continuamos seleccionando los nodos de grado 3 que etiquetamos con $\sigma = 3$ y eliminamos de la gráfica. En este caso, todos los nodos de la gráfica resultante del paso previo coloreados en verde y naranja en la Figura 3.9. La eliminación de estos nodos resulta en la gráfica vacía, con lo que detenemos el procedimiento, y cada nodo tiene asignado un valor de $\sigma$.

Una vez que concluida la ejecución del algoritmo, notamos que al igual que con el ejemplo anterior del árbol (ver Figura 3.3), los valores de $\sigma = 1$ no necesariamente corresponden a nodos de grado 1 en la gráfica original. Además, los nodos de grado tres en la última iteración del segundo ejemplo, mostrado en la Figura 3.9, representan todos los nodos de la gráfica organizados en dos subgráficas *3-regulares* y éstas son todas las subgráficas *3-core* que contiene la gráfica original (ver Figura 3.4). Por tal motivo, el valor de *kshell* que se asigna a los nodos corresponde al *k-core* más grande del que forma parte cada uno, logrando una jerarquización de los nodos por capas, tal que los nodos con valores bajos de *kshell* se encuentran en las capas superficiales y los de valores altos se encuentran al interior de la gráfica.

Otra observación interesante es que si nos fijamos en los nodos coloreados de rojo y amarillo, en la gráfica inicial de la Figura 3.4, tienen grados 5 y 4 respectivamente, pero sus valores $\sigma$ son 2 y 3. Es decir, a pesar de que el grado del nodo rojo es más grande que el del nodo amarillo, el nodo amarillo tiene un $\sigma$ mayor. Podemos realizar la siguiente reflexión: el nodo rojo es parte de una gráfica *3-regular*. Por ello es vecino de y vecino del vecino de una subgráfica *4-regular* y además es vecino de un nodo "solito". Por su parte, el nodo amarillo sólo es vecino del mismo nodo "solito" y de otra subgráfica *4-regular*, aunque tiene tres aristas que lo unen a ésta.

## 3.4. Algoritmo de centralidad *PageRank*

Pensemos en la red informática mundial (mejor conocida como *World Wide Web* o simplemente *la Web*) que usamos todos los días al *navegar en Internet*. La principal característica de *la Web* es que conjunta una cantidad inmensa de documentos interconectados mediante ligas de hipertexto, que además están montados en la infraestructura de *Internet*. Dicho de otra manera, tenemos muchísimas páginas a las que podemos acceder a través de las ligas que mandan de una a la otra.

*La Web* puede modelarse mediante una red dirigida, donde cada nodo de ésta representa una página web y hay una arista dirigida del nodo $u$ al nodo $v$ si existe una liga (o *link*) en la página $u$ que permita navegar hacia la página $v$.

Ahora supongamos que programamos un robot que recorre aleatoriamente esta gráfica, es decir, un robot que da un *click* aleatorio a alguna liga de la página en la que se encuentra en cada instante; de esta forma el robot navega por *la Web* de forma aleatoria.

Supongamos tambien que el robot "está sometido al tiempo", lo que significa que cada vez que el tiempo avanza una unidad, el robot se ve forzado a dar un *click* a un vínculo con lo que navega hacia otra página. El tiempo $i$ está indexado con los enteros no negativos, por lo que $i \in \{0, 1, 2, \ldots\}$. Este procedimiento establece la base del algoritmo *PageRank*.

Ahora estamos listos para plantear el problema de interés, y que también es la motivación para la construcción de este algoritmo: dado que el robot inició el recorrido en el nodo $u$, ¿en qué nodo se encontrará después de moverse una vez? ¿y después del segundo movimiento? Después de que el robot haya navegado durante un periodo prolongado ¿cuál será la frecuencia con que visitó cada uno de los nodos?

Para el desarrollo de esta explicación nos basamos en el texto (Rousseau, Saint-Aubin, Antaya, Ascah-Coallier, y Hamilton, 2008).

### 3.4.1. Algoritmo *PageRank*

A continuación se aplicará el proceso descrito anteriormente a una gráfica particular para fines ilustrativos y posteriormente se responderán las preguntas planteadas al inicio de este capítulo.

Dada una gráfica simple y dirigida $G(V, E)$:

- Denotaremos los nodos correspondientes a páginas *web* con mayúsculas o combinación de letras mayúsculas $\{A, B, \ldots, AA, \ldots\}$. Como etiquetas variables o indeterminadas para los nodos se utilizarán $u$, $v$, etc.
- El conjunto de todas las páginas *web* lo denotamos como $I = \{A, \ldots, \}$ y la gráfica inducida por $I$ y sus ligas es $G_I$
- La probabilidad de estar en una página $X$ al tiempo $i$ la denotamos como $P(X)_i$.
- La probabilidad de estar en una página $u$ al tiempo $t$ dado que se estuvo en otra página $v$ en el tiempo anterior $t-1$ la denotamos $P(u|v)$, la cual se calcula por medio de la fórmula $P(u|v) = \frac{1}{|\{\overrightarrow{vb} \ | v,w \in V\}|}$, ya que sólo puede haber una arista dados dos nodos diferentes y suponemos que el robot puede seguir cualquier liga que aparezca en $v$ con la misma probabilidad.

Para ilustrar el algoritmo de *PageRank* consideramos la gráfica de la Figura 3.10, que tiene cinco nodos, y las aristas dirigidas como se muestra. Fijamos el inicio del recorrido del robot en el nodo $A$ al tiempo 0.

De esta manera, para $t = 0$ la probabilidad de estar en cada uno de los nodos es: $P_0(A) = 1$ y $P_0(A) = P_0(B) = P_0(C) = P_0(D) = P_0(E) = 0$.

Al tiempo 1, el robot únicamente puede alcanzar las páginas $B$ o $C$, dado que $A$ tiene dos aristas que salen de éste hacia dichos nodos, y siempre tiene que dar un *click*. Por tanto, las probabilidades de que ocupe los distintos nodos de la gráfica al tiempo $t = 1$ son

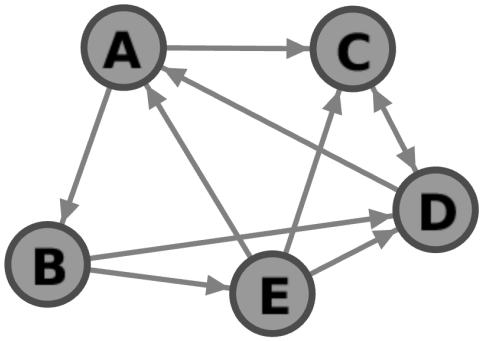
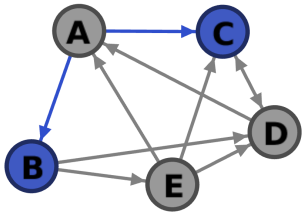


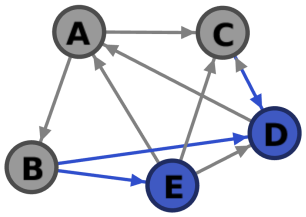

Figura 3.10: Gráfica simple dirigida que se utiliza para ejemplificar el algoritmo de *PageRank*.

$$P_1(A) = 0,$$
$$P_1(B) = P(B|A) = \frac{1}{|\{\overrightarrow{AC}, \overrightarrow{AB}\}|} = \frac{1}{2},$$
$$P_1(C) = P(C|A) = \frac{1}{|\{\overrightarrow{AC}, \overrightarrow{AB}\}|} = \frac{1}{2},$$
$$P_1(D) = P_1(E) = 0.$$



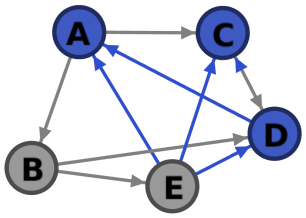

Al tiempo 2, el robot puede llegar a la página $D$ desde $B$ o $C$, o bien puede visitar $E$ pero únicamente si está en $B$. Para calcular estas probabilidades debemos considerar el nodo en el que estuvimos en el paso anterior. La primera posibilidad es haber partido desde $C$, del cual hay una sola arista que va a $D$, por lo que la probabilidad de estar en $D$ dado que estuvo en $C$ es de 1. Recordemos que llegamos a $C$ con probabilidad de $\frac{1}{2}$, por lo que la probabilidad de llegar a $C$ y luego a $D$ es el producto $\frac{1}{2} \cdot 1$. Pero también podríamos haber navegado a $D$ si en el tiempo anterior estuvimos en $B$, es decir, la probabilidad de visitar $B$ y luego $E$ es $\frac{1}{2} \times \frac{1}{2}$ por lo que la probabilidad de estar en $D$ en el tiempo $t = 2$ es la suma de $\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2}$, que es la suma de las probabilidades de cada uno de los dos caminos. Calculamos además la probabilidad de llegar a $E$ con el mismo razonamiento. Por lo tanto las probabilidades de estar en cada uno de los nodos

al tiempo $t = 2$ son




$$
\begin{aligned}
P_2(A) &= 0 \\
P_2(B) &= 0 \\
P_2(C) &= 0 \\
P_2(D) &= P_1(B) \times P(D|B) \\
&\quad + P_1(C) \times P(D|C) \\
&\quad = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times 1 = \frac{3}{4} \\
P_2(E) &= P_2(B) \times P(E|B) = \frac{1}{2} \times \frac{1}{2} \quad = \frac{1}{4}
\end{aligned}
$$

Desde $D$ o $E$ el robot podría ir a $A$, $C$ o $D$. Calculamos las probabilidades de que llegue a cada una de ellas usando el mismo argumento antes utilizado y obtenemos




$$
\begin{aligned}
P_3(A) &= P_2(E) \times P(A|E) \\
&\quad + P_2(D) \times P(A|D) \\
&\quad = \frac{1}{4} \times \frac{1}{3} + \frac{3}{4} \times \frac{1}{2} = \frac{11}{24} \\
P_3(B) &= 0 \\
P_3(C) &= P_2(E) \times P(C|E) \\
&\quad + P_2(D) \times P(C|D) \\
&\quad = \frac{1}{4} \times \frac{1}{3} + \frac{3}{4} \times \frac{1}{2} = \frac{11}{24} \\
P_3(D) &= P_2(E) \times P(D|E) \\
&\quad = \frac{1}{4} \times \frac{1}{3} = \frac{2}{24} \\
P_3(E) &= 0.
\end{aligned}
$$

De esta manera hemos ilustrado el procedimiento para calcular las probabilidades de visitar cada nodo de la red para los tiempos 1, 2 y 3 cuando iniciamos el recorrido en

$A$. De la misma forma podríamos calcular las probabilidades para cualquier tiempo $n$. Sin embargo, este método es ineficiente y es fácil cometer errores; por lo que conviene mejorarlo y generalizarlo.

Antes de generalizar este método, observamos que para calcular la probabilidad $P_i(u)$ de estar en un nodo cualquiera $u$ al tiempo $t = i$ dado que en el tiempo anterior estuvimos en un nodo diferente $v$, basta con conocer dos valores: la probabilidad de estar en el nodo en $v$ en el tiempo anterior $t = i - 1$, $P_{i-1}(v)$, para cualquier $v$, y la probabilidad de navegar en un tiempo de $v$ a $u$, $P(u|v)$, de tal manera que queda bien determinado $P_i(u|v) = P_{i-1}(v)P(u|v)$. Aquí las probabilidades $P(u|v)$ no cambian con el tiempo y se obtienen directamente de la gráfica; mientras que el valor de $P_i(v)$ sí cambia a lo largo del tiempo, y es éste el valor que irá acumulando la memoria de los estados anteriores.

Denotamos al vector de probabilidades al tiempo $t$, dado nuestro universo de páginas de Internet, $I$, como

$$P_t(X) = \begin{pmatrix} P_t(u_1) \\ \vdots \\ P_t(u_n) \end{pmatrix} \text{ , } t \in \{1, 2, \ldots\} \text{ y } u_i \in G_I(V).$$

Resaltamos el hecho de que las entradas del vector $P_t(X)$ indican la probabilidad de estar en cada una de las páginas (o nodos de la gráfica).

También conviene definir la matriz de probabilidades condicionales:

**Definición 3.4.1.** *Dada una gráfica simple y dirigida $G(V|E)$, llamamos* ***matriz de transición*** *$T$ de $G$, a la matriz $P_{n\times n}$ tal que $p_{ij} \in [0, 1] \quad \forall i, j \in \{1, ..., n\}$ y $\sum_i^n p_{ij} =$*

1 $\forall j \in \{1, ..., n\}$ :

$$T = \begin{array}{c} \\ u_1 \\ u_2 \\ \\ u_n \end{array} \begin{array}{c} \begin{array}{ccccc} u_1 & u_2 & u_3 & \cdots & u_n \end{array} \\ \begin{pmatrix} P(u_1|u_1) & P(u_1|u_2) & P(u_1|u_3) & \cdots & P(u_1|u_n) \\ P(u_2|u_1) & P(u_2|u_2) & P(u_2|u_3) & \cdots & P(u_2|u_n) \\ \vdots & & & & \\ P(u_n|u_1) & P(u_n|u_2) & P(u_n|u_3) & \cdots & P(u_n|u_n) \end{pmatrix} \end{array}$$

*y donde $u_i$ son nodos de $G$ con $i \in \{1, \ldots, n\}$.*

**Propiedad 3.4.1.** *Dado un vector $x \in R^n$ tal que $x_j \geq 0$, $j, i \in \{1, ..., n\}$ y $T$ una matriz de transición, se tiene que $\sum_{j=1}^{n}(T \cdot \overline{x}) = \sum_{j=1}^{n}(\overline{x})$ y $\sum_{j=1}^{n}(T \cdot \overline{x}) \geq 0$.*

*Demostración.*

$$\begin{aligned}
\sum_{j=1}^{n}(T \cdot x) &= \sum_{j=1}^{n}\left(\begin{pmatrix} \sum_{i=1} p(u_1|v_i) \cdot x_1 \\ \vdots \\ \sum_{1=1} p(u_n|v_i) \cdot x_n \end{pmatrix}\right) \\
&= \sum_{j=1}\sum_{i=1} p(u_j|v_i) \cdot x_i \\
&= \sum_{i=1}\sum_{j=1} p(u_j|v_i) \cdot x_i \\
&= \sum_{i=1}\Big(\sum_{j=1} p(u_j|v_i)\Big) \cdot x_i \\
&= \sum_{i=1}(1) \cdot x_i \text{ , por ser } T \text{ matriz de transición} \\
&= \sum_{i=1} x_i \\
&= \sum(\overline{x})
\end{aligned}$$

Como $p(u_i|v_j), x_i \geq 0$ $\forall i, j \in \{1, \ldots, n\}$ entonces $T \cdot x \geq 0$ □

Por el ejemplo trabajado, sabemos que para calcular la probabilidad de estar en cada una de las páginas al tiempo $n$, para $n \geq 1$ basta con conocer la matriz de transición y el vector de probabilidades al tiempo anterior, pues

$$P_n(X) = T * P_{n-1}(X).$$

De forma análoga podemos escribir el estado del sistema en $t = n-1$ como función del que tenía en el tiempo anterior e iterar eso hasta verlo como función del estado inicial:

$$\begin{aligned} P_n(X) &= T * (T * (P_{n-2}(X))) \\ &\vdots \\ &= T * (T * \cdots * (T * (P_0(X)))) \text{ por asociatividad} \\ &= T^n * P_0(X). \end{aligned}$$

Resumiendo, el vector de probabilidades del sistema al tiempo $t = n$ se calcula con la matriz de adyacencias y el vector inicial:

$$P_n(X) = T^n * P_0(X) \tag{3.1}$$

La Ecuación 3.1 indica que podemos calcular las probabilidades de estar en cada página únicamente operando la matriz de transición y el vector de probabilidades inicial.

Apliquemos estos resultados al ejemplo que desarrollamos anteriormente. Para ello, reescribimos las probabilidades de estar en cada nodo en el tiempo cero de la gráfica de Figura 3.10, lo que queda expresado como un vector $P(\overline{X})_0$,:

$$P_0(X) = \begin{pmatrix} P_0(A) \\ P_0(B) \\ P_0(C) \\ P_0(D) \\ P_0(E) \end{pmatrix}$$

Recordemos que en el ejemplo desarrollado iniciamos en el nodo $A$, por lo que el vector de probabilidades de estar en cada página en el tiempo 0 era $P_0(X) = (1, 0, 0, 0, 0)$

[2].

A fin de encontrar las probabilidades de que el robot esté en cada página en los tiempos 1, 2 y 3, obtenemos la matriz de transición $T$ de la gráfica Figura 3.10:

$$T = \begin{pmatrix} P(A|A) & P(A|B) & P(A|C) & P(A|D) & P(A|E) \\ P(B|A) & P(B|B) & P(B|C) & P(B|D) & P(B|E) \\ P(C|A) & P(C|B) & P(C|C) & P(C|D) & P(C|E) \\ P(D|A) & P(D|B) & P(D|C) & P(D|D) & P(D|E) \\ P(E|A) & P(E|B) & P(E|C) & P(E|D) & P(E|E) \end{pmatrix}$$

Calculamos las probabilidades de interés por medio de $T$ y $P_{i-1}(X)$ con $i \in \{1, 2, 3\}$ para compararlo con los resultados obtenidos antes:

$$P_1(X) = T \cdot P_0(X)$$

$$= \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}$$

[2]Es importante notar que este vector $P_0(X)$ es en realidad el transpuesto, pero no lo escribimos para no introducir más notación, por lo que el lector tendrá que tener presente este abuso de notación en el resto de esta sección

$$P_2(X) = T \cdot P_1(X)$$

$$= \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}$$

y por último,

$$P_3(X) = T \cdot P_2(X)$$

$$= \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{3}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{11}{24} \\ 0 \\ \frac{11}{24} \\ \frac{2}{24} \\ 0 \end{pmatrix}$$

### 3.4.2. Consolidación del algoritmo

Del proceso desarrollado en la sección anterior obtuvimos la Ecuación 3.1, que nos permite calcular la probabilidad de estar en cada una de las páginas al tiempo $t = n$. Sin embargo, este algoritmo presenta dos dificultades: primera, para los vértices que no tienen aristas de salida, la suma de las entradas de la columna que les corresponde en la matriz de transición es 0, lo que viola la restricción de que la suma de las columnas debe

ser 1. Desde la perspectiva del recorrido del robot, lo podemos interpretar como sigue: cuando el robot inicia en una página que no tiene ligas a otras, se queda atrapado en ese vértice a partir de que llega a él, y la probabilidad de visitar ese vértice es, entonces, 1 para todo tiempo futuro, dejando en 0 al resto de las probabilidades. Aunque en la teoría podemos simplemente ignorar estos casos, ya sea eliminando estos vértices o trabajando únicamente con gráficas que no los tengan, en la realidad no conviene desecharlos, ya que existen sistemas reales que sí los incluyen. Por ejemplo, en la red de *Internet*, la ausencia de hipervínculos en una página se puede presentar porque las páginas de destino aún no han sido indexadas por *Google* o simplemente no tienen hipervínculos de salida. Esto ocurre frecuentemente cuando la página tiene fines educativos o de investigación (lo que en ocasiones vuelve innecesarios los hipervínculos de salida). También sucede cuando hay algún error en la liberación de la página, cuando es un sitio fraudulento (en cuyo caso podría importar únicamente atraer a los usuarios, sin importar si tiene ligas de salida), o en general porque el contenido o propósito de la página no lo requiere.

El segundo problema lo podemos ver como un caso general del primero, y es más difícil de detectar. Tiene que ver con que el robot quede "atrapado" en un subconjunto propio del total de páginas. Esto puede ocurrir incluso si la gráfica es conexa, cuando existen subconjuntos de nodos que tienen una estructura similar a una isla dentro de la red, en donde la entrada y salida a éstas es difícil. Este problema se presenta en una variedad de casos muy amplio. Ilustramos un caso específico. Supongamos que existe un subconjunto de nodos que forman un ciclo, $C$, y las conexiones dentro de $C$ son muchas. Pero existen pocas aristas que permitan la entrada a $C$, y pocas aristas de salida. El efecto que tendría la existencia de ciclos como éste, es que, cuando el recorrido conduzca al robot entrar a $C$, al cabo de un tiempo y debido a su alta interconectividad, la probabilidad de estar en los nodos de $C$ es alta, reduciendo casi a 0 la probabilidad de estar en los nodos fuera de $C$, con lo que el robot queda atrapado en $C$ por el bajo número de aristas de salida. Si continuamos avanzando en el tiempo, el robot

eventualmente logrará salir de $C$, y después de continuar el recorrido, las probabilidades se invertirían: la de estar en los nodos fuera de $C$ sería alta y la estar dentro de los nodos de $C$, baja.

Para resolver estos dos inconvenientes, que se presentan con frecuencia en la realidad y en una gran variedad de casos, el algoritmo *PageRank* brinda al robot la posibilidad de navegar al azar a cualquier otra página con probabilidad uniforme en cada momento. Es decir, cada vez que que el robot se encuentra en una página cualquiera, puede navegar a alguna de las páginas accesibles vía hipervínculos o a una página cualquiera de la red al azar con probabilidad $p \in (0, 1)$.

Esta propiedad está expresada analíticamente en la siguiente ecuación:

$$P' = \beta P + (1 - \beta)Q,\ \beta \in [0, 1], \tag{3.2}$$

donde $P \in M_{n\times n}$ es la matriz de transición de la red y $Q \in M_{n\times n}$ es otra matriz de transición cuyas entradas son $\frac{1}{n}$. Esto modela la probabilidad uniforme, sin sesgo, de navegar aleatoriamente hacia cualquier otra página. Ahora bien, la decisión de navegar sólo a los vecinos o hacia cualquier parte de la red está ponderada con el factor $1 - \beta$, de forma que si $\beta$ es cero el robot ignorará por completo la topología de la red, navegándola aleatoriamente, y en caso contrario (para $\beta = 1$) el robot no daría nunca *saltos aleatorios*. De esta manera, el algoritmo de *PageRank*, en vez de utilizar la matriz de transición obtenida directamente de la red para iterar el proceso, utiliza la matriz de transición calculada con la Ecuación 3.2.

Implementaciones de *PageRank* desarrolladas por la comunidad *open source* típicamente usan el método iterativo para calcular las probabilidades de visita de cada página y utilizan como estado inicial el vector cuyas entradas son todas $\frac{1}{n}$ (Gephi, 2019), (NetworkX, 2019). Finalmente, en el artículo original los autores proponen un valor de $\beta = 0.85$, aunque las distintas implementaciones toman este valor como parámetro.

A continuación calculamos la medida de centralidad *PageRank* de cada nodo de la

red mostrada en la Figura 3.11, a partir de iterar como lo indica la Ecuación 3.1 a la matriz de transición obtenida con la Ecuación 3.2.

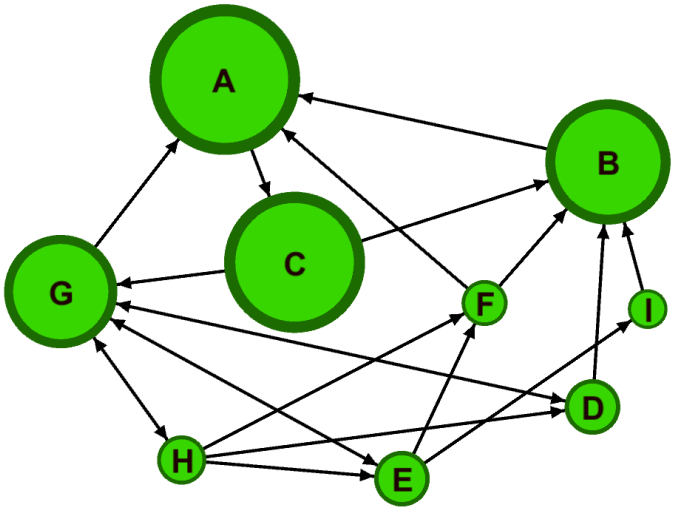



Figura 3.11: El tamaño de cada vértice es proporcional a su *PageRank*

Calculamos la matriz de transición $T$ de la red y fijamos el estado inicial $P_0(V) = (\frac{1}{n}, ..., \frac{1}{n})$

$$T = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{array} \begin{array}{c} \begin{array}{ccccccccc} A & B & C & D & E & F & G & H & I \end{array} \\ \begin{pmatrix} 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.25 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 & 0.5 & 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.25 & 0.25 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.25 & 0.25 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.33 & 0.0 & 0.0 & 0.25 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.33 & 0.0 & 0.0 & 0.25 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.25 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.33 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix} \end{array}$$

Calculamos la matriz de transición de *PageRank*:

$$T' = 0.85 \cdot T + (1 - 0.85) \cdot Q$$

donde $q_{ij} = \frac{1}{9} \in Q$, para toda $i, j \in \{1, \ldots, 9\}$

Finalmente, obtenemos el valor de *PageRank* de cada nodo sustituyendo en la Ecuación 3.1, $P_n(X) = T'^{\,n} * P_0(V)$, con el vector de probabilidades iniciales, $P_0(V)$, cuyas entradas son $v_i = \frac{1}{9}$, y la matriz de transición $T'$, calculada en el párrafo anterior. A continuación mostramos los resultados. La columna derecha muestra el vector que contiene los valores de *PageRank* de los nodos, y la columna izquierda la matriz de transición elevada a la potencia de $n$ igual a 5, 10 y 50 (de arriba a abajo).

$$\begin{pmatrix} 0.172 & 0.113 & 0.333 & 0.333 & 0.214 & 0.252 & 0.159 & 0.213 & 0.391 \\ 0.280 & 0.128 & 0.115 & 0.115 & 0.221 & 0.114 & 0.191 & 0.201 & 0.100 \\ 0.099 & 0.376 & 0.158 & 0.158 & 0.138 & 0.238 & 0.262 & 0.184 & 0.099 \\ 0.051 & 0.050 & 0.077 & 0.077 & 0.054 & 0.070 & 0.049 & 0.056 & 0.091 \\ 0.051 & 0.050 & 0.077 & 0.077 & 0.054 & 0.070 & 0.049 & 0.056 & 0.091 \\ 0.036 & 0.064 & 0.040 & 0.040 & 0.037 & 0.047 & 0.045 & 0.038 & 0.031 \\ 0.231 & 0.141 & 0.097 & 0.097 & 0.203 & 0.111 & 0.168 & 0.170 & 0.081 \\ 0.045 & 0.030 & 0.067 & 0.067 & 0.046 & 0.058 & 0.038 & 0.048 & 0.085 \\ 0.030 & 0.044 & 0.030 & 0.030 & 0.029 & 0.035 & 0.035 & 0.030 & 0.025 \end{pmatrix} \rightarrow P_5(V) = \begin{pmatrix} 0.2427 \\ 0.1632 \\ 0.1908 \\ 0.0643 \\ 0.0643 \\ 0.0425 \\ 0.1450 \\ 0.0543 \\ 0.0325 \end{pmatrix}$$

$$\begin{pmatrix} 0.190 & 0.250 & 0.218 & 0.218 & 0.201 & 0.231 & 0.224 & 0.211 & 0.212 \\ 0.174 & 0.153 & 0.193 & 0.193 & 0.178 & 0.179 & 0.164 & 0.175 & 0.204 \\ 0.235 & 0.196 & 0.175 & 0.175 & 0.216 & 0.182 & 0.210 & 0.209 & 0.167 \\ 0.056 & 0.065 & 0.059 & 0.059 & 0.058 & 0.061 & 0.061 & 0.059 & 0.058 \\ 0.056 & 0.065 & 0.059 & 0.059 & 0.058 & 0.061 & 0.061 & 0.059 & 0.058 \\ 0.047 & 0.043 & 0.042 & 0.042 & 0.045 & 0.042 & 0.045 & 0.045 & 0.041 \\ 0.158 & 0.136 & 0.167 & 0.167 & 0.159 & 0.156 & 0.147 & 0.156 & 0.176 \\ 0.044 & 0.054 & 0.050 & 0.050 & 0.046 & 0.052 & 0.050 & 0.048 & 0.049 \\ 0.035 & 0.033 & 0.032 & 0.032 & 0.034 & 0.032 & 0.034 & 0.034 & 0.032 \end{pmatrix} \rightarrow P_{10}(V) = \begin{pmatrix} 0.2176 \\ 0.1796 \\ 0.1965 \\ 0.0601 \\ 0.0601 \\ 0.0438 \\ 0.1585 \\ 0.0498 \\ 0.0335 \end{pmatrix}$$

$$\begin{pmatrix} 0.217 & 0.217 & 0.217 & 0.217 & 0.217 & 0.217 & 0.217 & 0.217 & 0.219 \\ 0.175 & 0.175 & 0.175 & 0.175 & 0.175 & 0.175 & 0.175 & 0.175 & 0.175 \\ 0.201 & 0.201 & 0.201 & 0.201 & 0.201 & 0.201 & 0.201 & 0.201 & 0.201 \\ 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 \\ 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 & 0.060 \\ 0.044 & 0.044 & 0.044 & 0.044 & 0.044 & 0.044 & 0.044 & 0.044 & 0.044 \\ 0.155 & 0.155 & 0.155 & 0.155 & 0.155 & 0.155 & 0.155 & 0.155 & 0.155 \\ 0.049 & 0.049 & 0.049 & 0.049 & 0.049 & 0.049 & 0.049 & 0.049 & 0.049 \\ 0.033 & 0.033 & 0.033 & 0.033 & 0.033 & 0.033 & 0.033 & 0.033 & 0.033 \end{pmatrix} \rightarrow P_{50}(V) = \begin{pmatrix} 0.2179 \\ 0.1756 \\ 0.2019 \\ 0.0603 \\ 0.0603 \\ 0.0443 \\ 0.1558 \\ 0.0497 \\ 0.0337 \end{pmatrix}$$

La observación más interesante es que los valores de *PageRank* parecen converger a un valor conforme aumentamos el valor de $n$. En realidad, se puede demostrar utilizando álgebra lineal que este proceso es una cadena de *Markov* y si se cumplen ciertas

condiciones, cuando hacemos tender $n$ al infinito, el vector resultante converge, dando como resultado que las probabilidades de estar en cada uno es única. Aunque en este trabajo ya no estudiaremos con ese nivel de profundidad este algoritmo, aunque sugerimos leer el trabajo de (Rousseau et al., 2008).

# Capítulo 4

# Simulación de procesos de contagio en redes

Imaginemos que tenemos acceso a un laboratorio y creamos un virus parecido al que produce la viruela (al que nombramos *Virus*). Lo diseñamos de forma que no sea en absoluto letal para el humano, de forma que su tasa de mortalidad sea de cero por ciento, aunque sí muy contagioso. De hecho, durante su creación en el laboratorio, podemos modificar a *Virus* para hacerlo más o menos contagioso y estamos interesados en estudiar la manera en que se propaga *Virus* en cierta población, por lo que buscamos algún voluntario que nos permita contagiarlo para que sea la primera persona en contraer *Virus*. Nuestro paciente cero realizará sus actividades con normalidad, por lo que interactuará con personas a su alrededor como vecinos, colegas de trabajo o personas con las que coincida durante sus traslados. A través de estas interacciones, *Virus* tendrá la oportunidad de propagarse. Cada vez que logre infectar a una persona nueva contabilizaremos un caso más. A continuación, cada persona contagiada desarrollará una respuesta inmune que le permita curarse de *Virus* después de algunos días.

Si llevamos la cuenta de las personas infectadas, las recuperadas y las que aún no han contraído a *Virus* (las que son susceptibles de contagiarse) cada día desde que

contagiamos a nuestro voluntario, hasta que *Virus* se extinga de la población, entonces podremos averiguar el total de personas que se contagiaron de *Virus* a partir del paciente cero y cómo fue ocurriendo su propagación. Podríamos después repetir el experimento con otro voluntario, distinto del primero, que inicie la propagación de *Virus*. Al finalizar este segundo experimento, podremos comparar el número de contagiados de los dos procesos. Seguramente, los dos pacientes cero tienen hábitos distintos y el número de personas con las que conviven es diferente; por lo que la propagación de *Virus* será distinta. A la luz de eso, nos preguntamos si convendrá a la propagación que los voluntarios interactúen con muchas personas o será mejor que sean tímidos e interactúen con pocas. A lo mejor, *Virus* alcanza más personas cuando está en contacto sólo con un número relativamente bajo de personas pero éstas conocen a personas *influyentes*.

En este capítulo definimos los algoritmos y propiedades de procesos de contagio, como el descrito anteriormente, que utilizaremos para simular y estudiar brotes epidémicos en la red de colaboración de actores que describimos en la Sección A.1. Cabe mencionar que este trabajo, y en particular, este capítulo, se basa en el artículo de (Kitsak et al., 2010), si bien ellos usan distintas redes para simular procesos epidémicos con algoritmos como los que describimos más adelante. Del análisis de sus resultados concluyen, entre otras cosas, que el *kshell* de los nodos en una red es la medida que mejor captura la importancia que tiene el nodo en la propagación de una enfermedad, superando como clasificador al grado y la cercanía. Con el fin de mantener este estudio como referencia a lo largo de los diferentes análisis, elegimos la red de colaboración de actores que también analizaron. Esta red representa un actor con un nodo, y tiene una arista entre cada dos actores distintos que actuaron en una misma película, se se compone de 47,719 nodos y 1,098,451 aristas.

## 4.1. Dinámicas del proceso epidemiológico

En esta sección diseñamos e implementamos un algoritmo que simula un proceso de contagio, el cual está inspirado en los modelos de campo medio descritos en el Capítulo 2, al tiempo que incorpora una característica de las poblaciones reales: cada individuo tiene contacto sólo con una fracción pequeña de la población total, por lo que puede contagiar a un número relativamente pequeño de personas.

Una vez asociados estos conceptos a un lenguaje de programación específico, la implementación de algoritmos como los que hemos descrito (*PageRank* y *K-shell*) o el que simula el proceso de contagio que describimos en este capítulo, también resulta fácil y transparente.

Vale la pena mencionar que el estudio de los procesos de contagio por medio del diseño e implementación de algoritmos lleva al estudio de una gran cantidad de temas secundarios. Por ejemplo, la eficiencia de *PageRank* implementado en *Python* se puede mejorar si se traduce a un lenguaje como *C* (o *Java*) y se hace buen uso de arreglos. Por otro lado, la utilización de estos lenguajes representaría una mejora en el manejo de la memoria principal, sacrificando abstracción y transparencia en su implementación. Sobre esta misma línea, si consideramos construir redes reales cuyo tamaño crece al orden de decenas de miles de nodos (o hasta millones), el cambio de paradigma de la herramienta en la que se implementen los algoritmos a uno que sea paralelo o concurrente incrementaría su eficiencia considerablemente.

Como el objetivo de este trabajo es estudiar los algoritmos descritos previamente y su relación con los procesos de contagio en un nivel abstracto, empleamos *Python* como lenguaje de programación principal, y lo complementamos con el uso del paquete *NetworkX* con el que manipulamos redes específicas. *NetworkX* incluye algoritmos para calcular distintas medidas de centralidad, como *PageRank* y *k-shell*. Finalmente, utilizamos el paquete *numpy* para mejorar el rendimiento del algoritmo que simula los

procesos de contagio en las redes, mismo que describimos más adelante.

Entre los aspectos que debemos definir al simular un proceso de propagación de la enfermedad están: primero, el nodo en que inicia la enfermedad, es decir, el primer nodo infectado, al que llamamos $v_0$; segundo, la probabilidad de que un nodo infectado contagie a un vecino susceptible; este parámetro lo denotamos $\beta$; y finalmente, la probabilidad de que un nodo infectado se cure o cambie su estado de *infectado* a *recuperado*, a la que denotamos $\gamma$. Puesto que la $\beta$ modela el contagio entre cada par *I-S*, esperamos que si el nodo en el que inicia la enfermedad está muy conectado, entonces la enfermedad afectará a más nodos y contagiará en el mediano o largo plazo una proporción relativamente alta de la población.

Una vez fijos la probabilidad de contagio $\beta$, la tasa de recuperación $\gamma$ y el nodo inicial $v_0$, simulamos la epidemia en la red de la siguiente manera: todos los individuos de la población (nodos de la red) inician en el estado *SUSCEPTIBLE*, salvo $v_0$, que comienza en estado *INFECTADO*. Ya que el nodo $v_0$ tiene estado *INFECTADO*, cada uno de sus vecinos puede contagiarse con probabilidad $\beta$. Cada vecino que se contagie cambiará su estado de *SUSCEPTIBLE* a *INFECTADO*. Enseguida, el nodo $v_0$ se recuperará con probabilidad $\gamma$, cambiando entonces su estado a *RECUPERADO* o manteniéndose como *INFECTADO* con probabilidad $1-\gamma$. En la siguiente etapa del proceso, nos fijamos en los nodos con estado *INFECTADO*, que son todos los vecinos de $v_0$ que contrajeron la enfermedad en el paso previo y $v_0$, si éste no tiene estado *RECUPERADO*. Cada uno de estos nodos infectados, intentará contagiar a cada uno de sus vecinos cuyo estado sea *SUSCEPTIBLE* con probabilidad $\beta$; enseguida se recuperará con probabilidad $\gamma$. El segundo paso termina después de definirse los contagios producto del contacto entre nodos susceptibles e infectados al inicio de esta etapa, y luego que se determine cuáles de los infectados se recuperaron. De esta manera, cada vez que se inicia una etapa nueva en el proceso, por cada nodo infectado $v_I$ se realizan los siguientes pasos: a) $v_I$ contagia a cada uno de sus vecinos en estado *SUSCEPTIBLE* con probabilidad $\beta$ b) $v_I$ cambia

al estado *RECUPERADO* con probabilidad $\gamma$. La simulación terminará una vez que no haya nodos infectados.

En el recuadro Algoritmo 2 damos el pseudocódigo del proceso de contagio descrito anteriormente.

**Algoritmo 2** Simula un proceso epidémico que inicia en el nodo *origen*, con probabilidad de contagio $\beta$ y recuperación $\gamma$. También recibe *states* un diccionario inicializado con todos los identificadores de los nodos como llaves y todos los valores como $SUSCEPTIBLE$. respectivamente. El Algoritmo 3 contiene detalles sobre las funciones utilizadas.

```
1: function EpidemicTransitions(G, origen, β, γ, states)
2:     Q1 ← [ ]                          ▷ Nodos infectados en el tiempo anterior
3:     Q2 ← [ ]                          ▷ Nodos infectados en el tiempo actual
4:     Q1 ← Add(Q1, origen)
5:     while HasElements(Q1) do
6:         for node in Q1 do
7:             for v in Neighbors(node) do
8:                 if random() ≤ β and IsSusceptible(v, states) then
9:                     Infect(v, states)
10:                    Add(Q2, v)
11:                end if
12:            end for
13:            if random() ≤ γ then
14:                Delete(Q1, node)
15:                Recover(node, states)
16:            end if
17:        end for
18:        Q1 ← Add(Q1, Q2)
19:        Q2 ← [ ]
20:    end while
21:    return CountRecoveries(states)
22: end function
```

**Algoritmo 3** Funciones que son utilizadas por el Algoritmo 2. El paramáetro *node* es un identificador a un nodo y *states* es un diccionario cuyas llaves son los identificadores de cada nodo de la red y los valores puede ser algún valores de: $SUSCEPTIBLE$, $RECOVERED$ o $INFECTED$.

```
1: function RECOVER(node, states)
2:     states[nodo] = RECOVERED
3: end function
1: function INFECT(node, satate)
2:     states[node] = INFECTED
3: end function
1: function ISSUSCEPTIBLE(node, satates)
2:     return states[node] == SUSCEPTIBLE
3: end function
1: function COUNTRECOVERIES(states)
2:     no_recov ← 0
3:     for node in KEYS(states) do
4:         if states[node] == RECOVERED then
5:             no_recov ← no_recov + 1
6:         end if
7:     end for
8:     return no_recov
9: end function
```

El Algoritmo 2 determina el número de nodos recuperados o afectados al finalizar la simulación de un proceso epidémico que se originó en el nodo *origen*. Como éstos son valores asociados a una simulación, y ésta es un proceso estocástico, es necesario, con el fin de obtener valores representativos, realizar muchas simulaciones. Esto permitirá aproximar numéricamente el estado final de la epidemia para las condiciones iniciales y parámetros utilizados. Por tal motivo, creamos una rutina que simula un brote $N$ veces con los mismos parámetros y devuelve una lista, $Rs$, tal que su $i$-ésimo elemento indica el número de recuperados que ocurrieron al finalizar la simulación del brote $i$. El pesudocódigo de esta rutina se muestra en el Algoritmo 4. El principal objetivo del Algoritmo 4 es ejecutar $N$ simulaciones para cada nodo y guardar los resultados a partir de los cuales realizaremos el análisis posterior. Este algoritmo tiene como segundo objetivo agregar una capa adicional en la que podamos organizar los datos generados de la forma que mejor nos convenga.

**Algoritmo 4** Realiza un proceso epidemiológico estocástico al simular $N$ brotes epidemiológicos independientes, cada uno utilizando los mismos parámetros nodo origen, $\beta$ y $\gamma$

```
1: function EpidemicalSimulationProcess(N, G, source, β, γ)
2:     Rs ← []
3:     for i ← to N do
4:         recuperados_i ← EpidemicalIteration(G, source, β, γ)
5:         Add(Rs, recuperados_i)
6:     end for
7:     return Rs
8: end function
```

.

El Algoritmo 2 es el más importante en esta investigación, ya que busca retratar las dinámicas propias de los procesos epidemiológicos. Por lo tanto, nos detendremos a

examinar algunas de sus características en los siguientes párrafos.

El primer nodo infectado en cada simulación lo denominamos nodo *origen* y el Algoritmo 2 lo recibe como condición inicial. Además, como la condición de paro es que no haya nodos *infectados*, al finalizar cualquier iteración o simulación epidémica no habrá nodos infectados y tendrá que existir, al menos, un individuo recuperado.

A pesar de que este algoritmo define con precisión un proceso de contagio inspirado en los descritos en el Capítulo 2, las implementaciones que utilizamos fueron evolucionando a fin de reducir el tiempo que demora cada simulación. A continuación describimos estas modificaciones.

Transformamos los nodos y aristas de un objeto tipo Gráfica (de *NetworkX*) en arreglos $n$-dimensionales de *numpy*. Los nodos de la red $G$ los re-etiquetamos usando números naturales (desde el 0), evitando que existan etiquetas duplicadas y que no se salte ningún número, pero sin importar el orden en que éstos se asignen a los nodos. Creamos un arreglo de tamaño igual al número de nodos y tal que la celda con el índice $i$ del arreglo almacene otro arreglo con las etiquetas de los vecinos del nodo $i$. De esta forma es fácil conocer los vecinos de un nodo y simular el contagio en un mismo bloque de código. Este algoritmo resulta ser más eficiente que el Algoritmo 2, ya que recuperamos a los vecinos mediante los índices, aprovechando de forma más eficiente el uso de la memoria principalmente a través de los arreglos de *numpy*.

Posteriormente implementamos la misma restricción impuesta por Kitsak et al. (2010): un nodo infectado se recupera con probabilidad 1 después de una ronda en que intenta infectar a sus vecinos. Esta condición en la dinámica de la simulación epidémica, nos permite realizar una última optimización en el tiempo de ejecución del algoritmo: dejamos de ejecutar la instrucción 13 del Algoritmo 2, ya que los nodos se recuperan inevitablemente tras terminar el ciclo de sus potenciales contagios. Si establecemos como unidad de tiempo una iteración, en la que todos los nodos infectados intentan contagiar a cada uno de sus vecinos, entonces por la restricción anterior, cada

nodo infectado durará exactamente una unidad de tiempo infeccioso, lo que equivale a tomar el lapso infeccioso como unidad temporal de las simulaciones. En nuestra implementación, después de recorrer todos los nodos de la lista de infectados actuales, registramos los nodos que se infectaron durante esa iteración y actualizamos la lista de los nodos infectados con esta lista. La lista anterior de infectados simplemente la desechamos, ya que trascurrió una unidad de tiempo en nuestra nueva escala temporal, con lo que todos estos nodos se habrán recuperado sin excepción.

Al finalizar cada simulación, obtenemos el número de nodos que se recuperaron, lo que, por la restricción explicada en el párrafo anterior, es igual al número de nodos infectados, en los distintos momentos de la simulación.

Al fijar los parámetros $N$, $\beta$ y $\gamma$ y aplicar el Algoritmo 4 obtenemos por cada nodo de la red una lista con $N$ valores independientes correspondientes al número de recuperados a que dio lugar esa simulación. A partir de esta lista realizamos los análisis posteriores.

## 4.2. Medidas de los brotes epidémicos

En esta sección definimos distintas medidas del alcance de una epidemia que consideran un aspecto particular del impacto global de la infección al finalizar su propagación en una población. La definición de estas métricas es esencial, ya que, por cada proceso epidémico simulado a partir de cada nodo, obtenemos un conjunto de magnitudes independientes que podremos comparar. La estrategia que adoptamos es la de calcular un único valor que represente a este conjunto de magnitudes y asignárselo al nodo origen. Este valor reflejará la influencia del nodo en un proceso de contagio, de tal manera que valores altos indicarán que los brotes originados en el nodo afectan a una fracción mayor de la población. Llamaremos "infectividad" del nodo a esta medida.

Así, la definición de estas métricas nos ayudará no sólo a ordenar el análisis, sino

también a formalizar el estudio del alcance de las epidemias en poblaciones heterogéneas modeladas en redes. A su vez, esto nos permite establecer un marco cuantitativo para comparar brotes epidémicos que inician en distintos nodos

En este trabajo utilizamos métricas que han sido empleadas en otras investigaciones (Zhang et al., 2016), (Chen et al., 2012) y desde luego (Kitsak et al., 2010).

Con el Algoritmo 2 ejecutamos $N$ simulaciones independientes con inicio en el mismo nodo que utilizan un mismo juego de parámetros. Para cada nodo consideramos entonces el conjunto de $N$ simulaciones.

El Algoritmo 4 es el que nos permite ejecutar repetidamente las $N$ simulaciones. Devuelve una lista de resultados de las $N$ simulaciones para cada nodo $v$. De esta manera, construimos al conjunto $R_v$ con $N$ elementos y cuyo i-ésimo elemento es el número de recuperados de la i-ésima simulación correspondiente al nodo $v$. El conjunto $R_v$ no considera al nodo que inició la epidemia (Ecuación 4.1), por tanto el menor número de individuos recuperados que reporta una simulación es cero, lo que ocurre cuando el infeccioso inicial no contagia a nadie.

Dada una red $G$ en la que simulamos la propagación de una enfermedad con parámetros $\beta$ y $\gamma$ dados, y $v$, el nodo origen, definimos los siguientes conjuntos:

$$E_{v,\mu}(R_v) = \left\{ \frac{r_i}{|V|} \geq \mu \middle| r_i \in R_v \right\} \tag{4.1}$$

$E_{v,\mu}$ considera sólo las simulaciones en las que el número de recuperados es mayor o igual a un cierto umbral $\mu$. Así, $E_{v,\mu}(R_v)$ es la fracción del total que resultó afectada en cada una de las simulaciones en las que se infectó una fracción mayor o igual a $\mu$. Se considera que las simulaciones que pertenecen a este conjunto son aquellas en las que sí hubo una epidemia.

Por lo anterior, tomamos la convención de llamar "simulaciones del brote" (o simplemente "brote") al conjunto de las simulaciones que iniciaron en $v$ y tienen el mismo juego de parámetros, o a una sola simulación, cuando se especifique. Cuando el bro-

te resulta en "muchos" contagios entonces diremos que estalló una epidemia. Esto se afirmará cuando la fracción de individuos contagiados rebase el umbral $\mu$.

Definimos las siguientes funciones que miden el potencial epidémico de un nodo $v_0$ en el proceso de propagación:

1. El promedio del número de nodos recuperados sobre todas las simulaciones que inician en el nodo $v_0$ para un juego de parámetros determinado:

$$f_1(R_{v_0}) = \frac{1}{N} \sum_{r \in R_{v_0}} r$$

   Notamos que $f_1$ toma valores en $[0, |V|]$. Aunque en gráficas reales, $f_1$ no devuelve valores cercanos a $|V|$, pues ningún nodo lleva al contagio de toda la red.

2. La fracción de simulaciones en las que sí hubo epidemia,

$$f_2(R_{v_0}, \mu) = \frac{\left| E_{v_0,\mu}(R_{v_0}) \right|}{N}$$

   que toma valores en el intervalo $[0, 1]$. $\left| \cdot \right|$ denota la cardinalidad del conjunto.

3. El promedio de individuos recuperados considerando sólo las simulaciones en que sí hubo epidemia:

$$f_3(R_{v_0}, \mu) = \frac{1}{N} \sum_{\varepsilon \in E_{v_0,\mu}} \varepsilon$$

Por último, definimos para cada $f_i$ las funciones $M_i(\sigma, k)$ como en Kitsak et al. (2010), que miden el grado de propagación de la enfermedad considerando dos medidas de centralidad de los nodos: el *kshell* $\sigma$ y el grado $k$:

$$M_i(\sigma, k) = \sum_{w \in \Upsilon(\sigma,k)} \frac{f_i(w)}{|\Upsilon(\sigma, k)|}$$

en donde $\Upsilon(\sigma, k) = \{v \in V | kshell(v) = \sigma \ \mathbf{y} \ grado(v) = k\}$

Cada función $M_i$ considera subconjuntos de nodos que tienen el mismo valor en dos medidas de centralidad distintas. Esta función obtiene los subconjuntos de nodos

que tienen el mismo grado y *kshell*, y promedia los valores que le asigna alguna de las funciones $f_1$, $f_2$ o $f_3$. Generalmente usaremos $f_1$ y denotaremos a $M_1$ como $M$, a menos que se indique otra cosa. Así, la función $M$ devuelve el promedio de individuos recuperados que iniciaron en nodos con un mismo grado y *kshell*. Análogamente, las funciones $M_2$ y $M_3$ calculan respectivamente el promedio de simulaciones en que hubo una epidemia y su tamaño promedio para conjuntos de nodos con el mismo grado y *kshell*, aunque, como se verá más adelante, será suficiente estudiar únicamente $M$.

En las siguientes secciones usamos las simulaciones efectuadas en la red de colaboración de actores para calcular el tamaño del brote epidémico y su relación con distintas medidas de centralidad del nodo origen, mediante el cálculo de las funciones $f_1$, $f_2$, $f_3$ y $M$, a las que nos referiremos como *funciones de infectividad*. Cada una de estas, por construcción, mide un aspecto diferente del alcance de la propagación de la enfermedad o del proceso de contagio.

## 4.3. Análisis de los brotes epidémicos

Estamos listos para combinar y aplicar los conceptos anteriores. A continuación describimos los resultados de las simulaciones del proceso de contagio que se llevaron a cabo en la red de colaboración de actores que describimos en la Sección A.1. Dichas simulaciones se ejecutaron mediante la implementación del Algoritmo 4 que corre un proceso epidémico de tipo $SIR$, según se describió en el Capítulo 2. Posteriormente realizamos el análisis del alcance de cada proceso de contagio utilizando las funciones de infectividad definidas en la sección anterior y concluimos discutiendo la relación entre el tamaño del brote que inicia en cada nodo y las medidas de centralidad que los caracterizan, específicamente el grado, *kshell*, y *PageRank*.

Para cada valor de $\beta$ realizamos $N = 100$ simulaciones empezando en cada uno de los nodos. Recordemos que, tras fijar la probabilidad de contagio $\beta$ y definir el nodo

infectado inicialmente, la probabilidad de recuperación de cualquier nodo infectado siempre será uno, de forma que cambiará de *INFECTADO* a *RECUPERADO* después de "intente" contagiar a cada uno de sus vecinos una vez, como se describió en la Sección 4.1.

Los resultados que describimos a continuación toman $\beta$ igual a 0.04, aunque también ejecutamos otro conjunto de simulaciones, con $\beta$ igual a 0.01. Uno y otro juegos de simulaciones se comparan en los aspectos sobresalientes y nuestras conclusiones de los resultados obtenidos para estos valores de $\beta$ se proporcionan en el resto del capítulo.

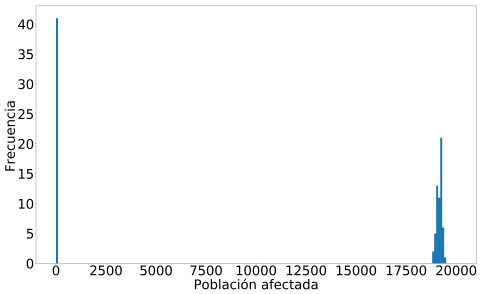
La Figura 4.1 muestra tres histogramas del número de individuos recuperados al finalizar cada simulación, considerando las 100 iteraciones del modelo que iniciaron en un mismo nodo. Seleccionamos tres nodos distintos para ilustrar tres comportamientos típicos de acuerdo con nuestras observaciones que resultan del proceso epidémico simulado.

En la Figura 4.1a se pueden apreciar dos picos, uno a la izquierda y el otro del lado derecho de la gráfica. En el primero se observa que en casi 40 simulaciones de un total de 100, el número de individuos *RECUPERADOS* fue cercano a cero [1]. En contraste, en el segundo pico el número de *RECUPERADOS* se encuentra entre 18,000 y 19,000 de un total de 47,719 nodos. Este segundo grupo identifica a las simulaciones en las que sí estalló una epidemia (aproximadamente el 60 %).

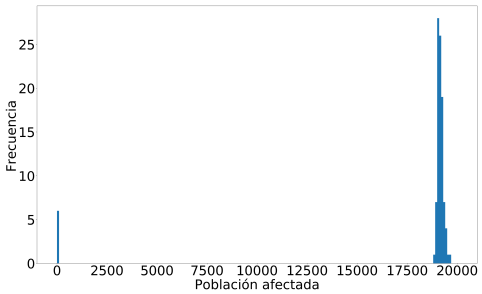
En contraste, la Figura 4.1b muestra el histograma del número de nodos recuperados para un nodo que en la mayoría de las simulaciones (más del 90 %) dio inicio a una epidemia que infectó entre 18,000 y 19,000 nodos. Sólo en 7 simulaciones la enfermedad no se difundió en forma epidémica.

Finalmente, la Figura 4.1c muestra el comportamiento de un nodo que siempre detona una epidemia. En este caso, el número de individuos alcanzados por la enfermedad

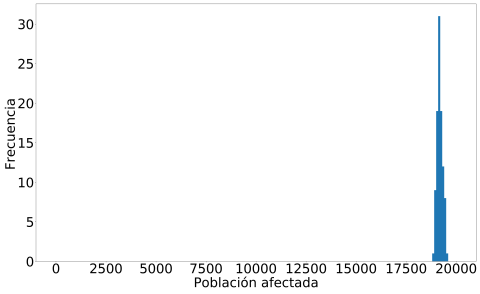
[1] Se obtuvieron valores entre 1 y 30. Puesto que éstos son muy pocos nodos para una red de 47,719, podemos decir que en esos casos no se observó epidemia.




(a) Nodo 5972




(b) Nodo 6551



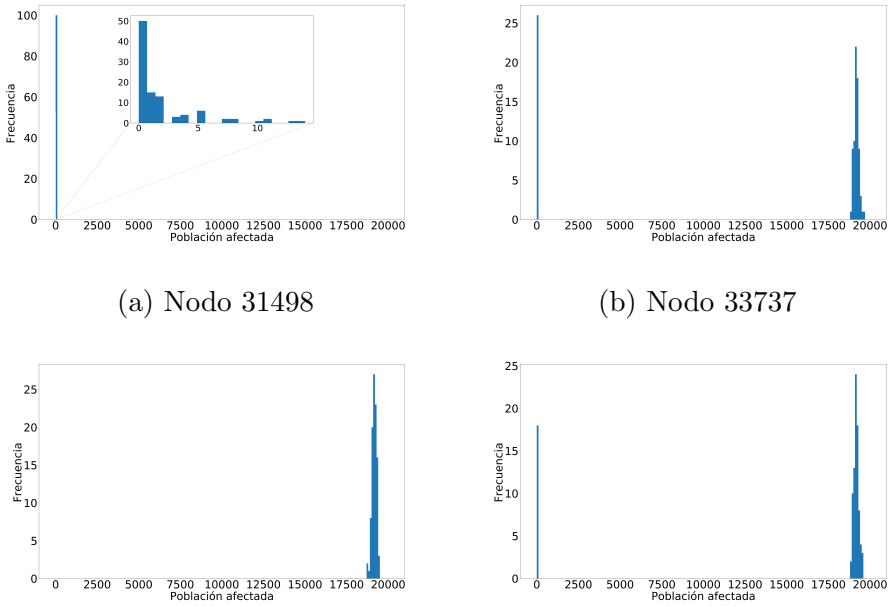

(c) Nodo 5166

Figura 4.1: Histogramas del número de recuperados al término de cada una de las 100 iteraciones. Los parámetros epidémicos son $\beta = 0.04$ y $\gamma = 1$, después de que cada nodo intenta contagiar a todos sus vecinos

se encuentra entre 18,900 y 19,600 nodos para todos las iteraciones.

Otra observación, que más adelante retomaremos con mayor detalle, es que el número de individuos afectados en una epidemia tiene siempre valores muy parecidos. Por ejemplo, en los tres histogramas de Figura 4.1, el número de individuos recuperados del grupo de la derecha, que representan a las simulaciones en las que estalló epidemia, se encuentra entre 18,811 y 19,647. El valor más pequeño y más grande del porcentaje de recuperados en simulaciones en que estalló una epidemia iniciada en el nodo 5972 es de 39.48 y 40.70 % respectivamente. Para el nodo 6,551 de 39.42 y 41.72 %, y para el nodo 5,166, 39.56 y 41.033 %.

En la figura Figura 4.2 graficamos otros nodos con el fin de ilustrar la diversidad de estados finales de las simulaciones epidémicas originadas en nodos determinados.




(a) Nodo 31498

(b) Nodo 33737

(c) Nodo 31051

(d) Nodo 9866

Figura 4.2: Comportamiento del número de individuos recuperados tras las epidemias que iniciaron en los nodos identificados como 9866, 31051, 33737 y 33747.

La Figura 4.2a muestra un ejemplo de nodo que es mal propagador, ya que nunca da lugar a brotes epidémicos y en Figura 4.2c vemos un nodo que siempre produce una epidemia. Diremos que estos últimos tienen una *infectividad* alta, mientras que la *infectividad* de los primeros es baja.

De este primer vistazo a los datos generados por las simulaciones que iniciaron en distintos nodos, observamos la diversidad en el número de individuos recuperados al finalizar la propagación para un mismo conjunto de condiciones iniciales y de parámetros (nodo origen, $\beta$, $\gamma$). También nos permite resaltar la importancia de cuantificar el alcance de la infección. Utilizaremos las funciones $f_1$, $f_2$, $f_3$, y $M$, descritas previamente, para conocer el alcance de la infección para los distintos nodos. En primer lugar,

vemos un nodo como un vector cuyas entradas son los resultados de estas funciones, y para compararlos, lo hacemos a través de comparar una a una sus entradas. También realizaremos un análisis para averiguar las correlaciones entras las funciones y poder seleccionar una única función. En última instancia la *infectividad* de un nodo estará dada por la magnitud de estos valores asignados.

Iniciamos calculando la función cuya construcción es las más simple, $f_1$. Recordemos que esta función tiene como base a $R_v$ (el conjunto del número de individuos recuperados al final de cada una de las simulaciones de la epidemia originada en $v$). En nuestro caso, el tamaño de estos conjuntos es de 100. $f_1(R_v)$ devuelve el promedio de individuos recuperados en los brotes que iniciaron en $v$ (ya sea que hayan generado epidemia o no). La función $f_2$ calcula la fracción de simulaciones en las que estalló una epidemia. Intuitivamente, $f_2$ mide qué fracción del total representa el grupo de simulaciones que se ubican a la derecha de cada histograma en las Figuras 4.1 y 4.2. El segundo parámetro, $\mu$, de $f_2$ es el umbral que define si hubo una epidemia: si en una simulación la fracción de recuperados rebasa $\mu$ decimos que en esta simulación estalló una epidemia. En este trabajo utilizamos un valor de $\mu$ de 5 % de la población total.

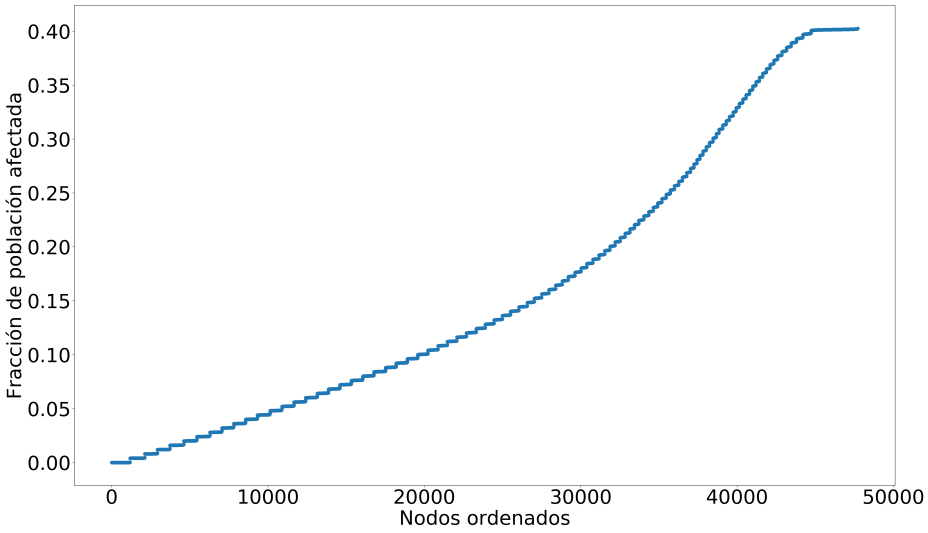
Calculamos las funciones $f_1$ y $f_2$ para cada nodo de la red de colaboración de actores. Los resultados se muestran en la Figura 4.3.

Los valores devueltos por $f_1$ (antes de la normalización) van desde 0 hasta el 19,217. Vale la pena observar que en la Figura 4.1c las simulaciones con mayor número de individuos recuperados para el mismo nodo reportaron un valor de casi 19,600, que es ligeramente mayor que el valor máximo para $f_1$. Esto se debe a que el promedio que devuelve $f_1$ se encuentra generalmente por debajo del valor máximo reportado por las simulaciones.

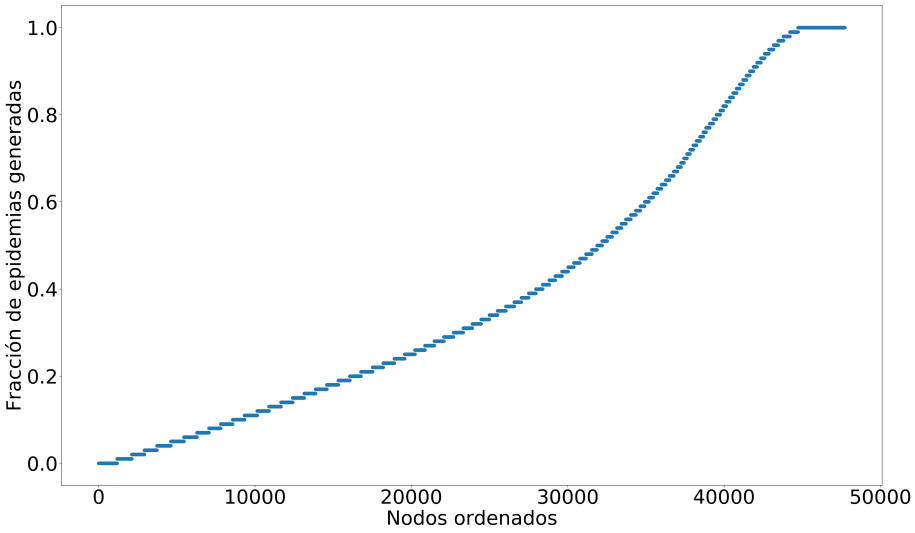
Los valores devueltos por $f_2$ van de 0 a 1 por su definición, de manera que no hay necesidad de normalizar sus valores. Cuando la función alcanza el valor máximo de 1, significa que todas las simulaciones que iniciaron en ese nodo originaron una epidemia,

aunque no se toma en cuanta el número de afectados preciso.

Finalmente, revisamos la función $f_3$, que calcula el promedio de individuos infectados pero únicamente del conjunto de simulaciones que originaron una epidemia (grupo de la derecha en los histogramas iniciales). Esta función mide el impacto del brote en sí mismo, ignorando la frecuencia con que sucedió éste. Así, si un nodo fue el lugar de origen de una epidemia, aunque sea en pocas simulaciones, sólo éstas se toman en cuenta y se omiten el resto. Después de calcular la función $f_3$ en todos los nodos de la red de colaboración de actores, se encontró que 46,544 de los 47,719 nodos de la red devolvieron un valor mayor a 0. Es decir, menos del 3 % de los nodos no dieron lugar a una epidemia. A continuación mostramos la gráfica de los valores de $f_3$, de la que eliminamos aquellos valores que son iguales a 0, que son minoría.




(a) $f_1$




(b) $f_2$

Figura 4.3: Lista ordenada de los valores devueltos por $f_1$ y $f_2$ en cada nodo de la red de colaboración de actores. El eje $x$ indica el rankeo del nodo, por lo que para un mismo lugar $x$ 4.3a y 4.3b no necesariamente les corresponde el mismo nodo. Los valores de $f_1$ están normalizados con respecto al total de la población 47,719 individuos

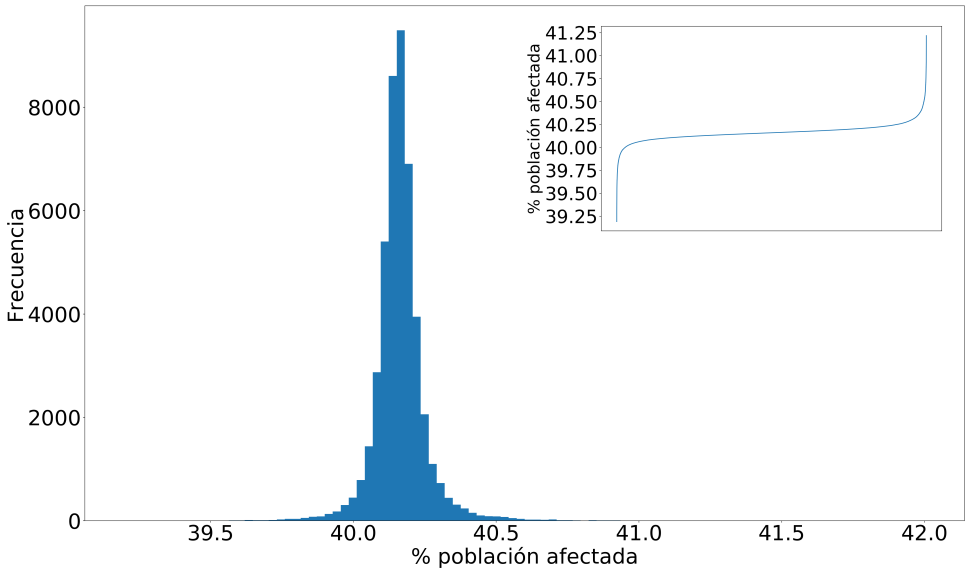
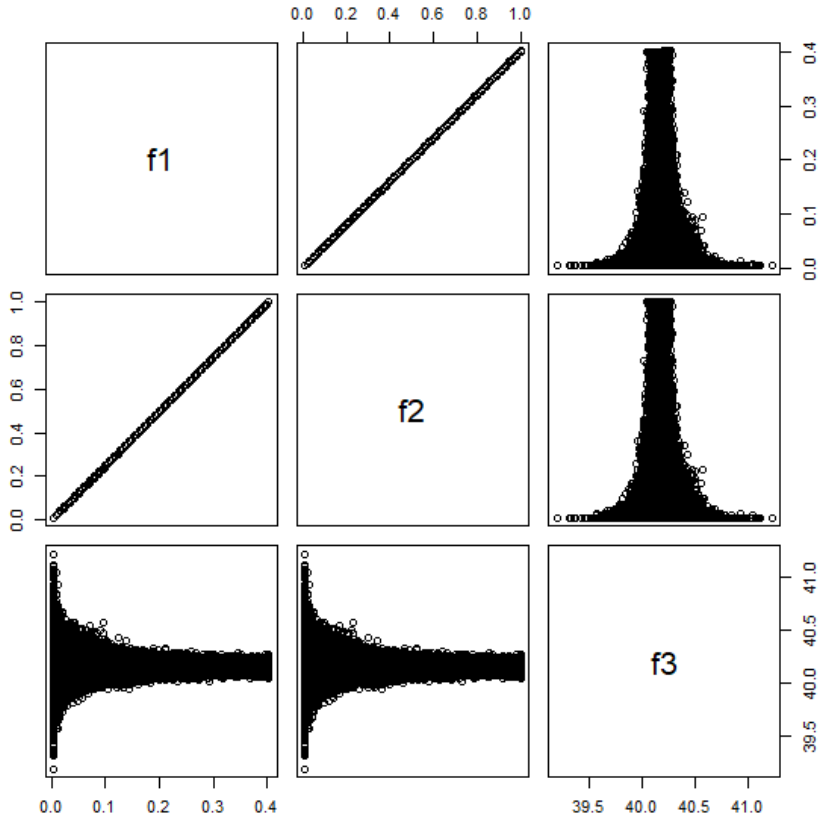



Figura 4.4: Gráfica de la lista ordenada de los valores devueltos por $f_3$ e histograma de la frecuencia con que ocurren. En ambas omitimos los valores igual a 0. La fracción de población afectada claramente sigue una distribución normal.

De lo anterior concluimos que prácticamente todos los nodos de la red dan lugar a una o más epidemias en cien simulaciones. También observamos que en todas las simulaciones que presentaron una epidemia, ésta tiene un tamaño similar. La magnitud de la epidemia está acotada entre 39.2 % y 41.3 %. De forma intuitiva, la frecuencia con que un nodo produce una epidemia está relacionada con su importancia en la propagación de infecciones. Los nodos que son malos puntos de inicio en la propagación de la infección casi nunca detonan epidemias, pero cuando lo hacen, éstas tienen el mismo tamaño que aquellas que inician en otros nodos.

Continuamos con el análisis observando la siguiente laFigura 4.5.




(a) $f_3$

Figura 4.5: Comparativo de las funciones $f_1$, $f_2$ y $f_3$. Eliminamos los nodos con $f_3 = 0$.

Mediante la aplicación de las funciones $f_1$, $f_2$ y $f_3$ logramos reducir los $N$ valores independientes asociados a cada proceso epidemiológico originado en cada nodo a tres únicos valores, uno por cada función $f_i$. Resumimos algunos de los puntos que más resaltan en este análisis:

- El tamaño de un brote que inicia en un nodo está reflejado por dos componentes: población afectada globalmente ($f_1$) y tamaño final (total de afectados) de las epidemias producidas ($f_3$).

- $f_1$ y $f_2$ están relacionadas linealmente, por lo que emplearemos $f_1$, que es el modelo

más sencilla de definir y calcular.

- La fracción total de recuperados durante una epidemia ($f_3$) está acotada en un intervalo pequeño en todos los casos estudiados. En la red de colaboración de actores y para los parámetros estudiados el tamaño de la epidemia alcanza entre 39.2 % y 41.3 % de la población.

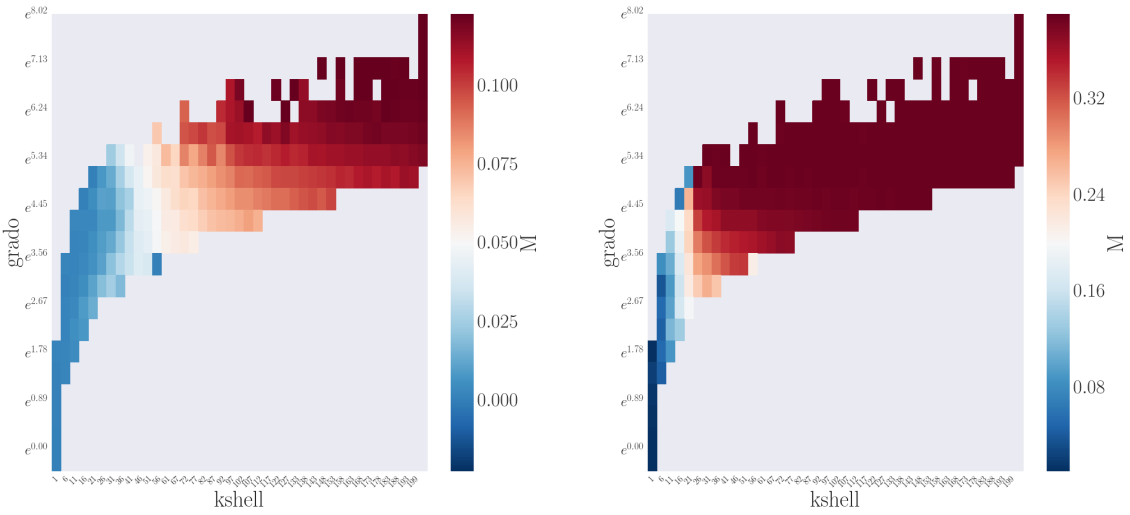
| Valor( %) | Min | 1er Q | Mediana | Media | 3er Q | Max |
|---|---|---|---|---|---|---|
| $f_1$ | 0.0 | 5.6 | 12.5 | 16.1 | 25.3 | 40.3 |
| $f_2$ | 0.0 | 14.0 | 31.0 | 40.1 | 63.0 | 1.000 |
| $f_3$ | 39.2 | 40.1 | 40.2 | 40.2 | 40.2 | 41.2 |

Figura 4.6: Tabla con valores estadísticos básicos de las tres funciones. En $f_3$ omitimos los valores igual a 0. Los valores están en formato del porcentaje respecto a la población.

La cuarta y última función de infectividad que estudiamos es $M_1$, definida originalmente en el trabajo de (Kitsak et al., 2010). Recordamos que esta función promedia las $f_1$ de los nodos que tienen un mismo grado y *kshell*.

Observemos la Figura 4.7 que grafica los valores de $M_1$ para dos conjuntos de simulaciones independientes en los que $\beta$ toma los valores de 0.01 y 0.04 respectivamente.

En esta figura, el eje de las abscisas contiene los valores de *kshell* y el eje de las ordenadas el grado del nodo. El gradiente de color representa los valores de $f_1$ promediados sobre todos los nodos con la misma combinación de *kshell* y grado. Un corte horizontal es equivalente a fijar un valor del grado y variar su *kshell*, mientras que en un corte vertical se fija un valor de *kshell* y se varía el grado. Las dos imágenes se obtienen de 100 simulaciones por nodo realizadas en la red de colaboración de actores. La diferencia es que en la Figura 4.7b, el parámetro $\beta$ que se utilizó es de 0.04, que es el valor con el que hemos trabajado hasta este momento y en la Figura 4.7a, se realizaron las simulaciones con $\beta$ 0.01.




(a) $\beta = 0.01$ (b) $\beta = 0.04$

Figura 4.7: Comparativo

La Figura 4.7a reproduce los resultados de (Kitsak et al., 2010), mientras que la Figura 4.7b permite observar el efecto que tiene sobre $M$ una tasa de transmisión más alta.

En las dos imágenes observamos que generalmente en los cortes horizontales, el color cambia gradualmente de azul a rojo cuando avanzamos de izquierda a derecha. Es decir, la infectividad de los nodos aumenta conforme aumenta el *kshell* si dejamos fijo el grado. Por otro lado, si realizamos un corte vertical, es decir, si fijamos un valor de *kshell*, el incremento del grado no suele acompañarse de un incremento en la infectividad de los nodos conforme nos desplazamos de abajo hacia arriba. Además, para un valor de *kshell*, la variación de colores es menor que la mostrada al fijar un valor del grado. Este hecho indica que valores similares de *kshell* se asocian a infectividades similares, a diferencia del grado, en el que un mismo grado puede vincularse con una variedad de infectividades. Por lo tanto, concluimos que existe cierta independencia de la infectividad del nodo con respecto al grado, pero ésta está fuertemente ligada al

*kshell*.

Además, en la Figura 4.7b, la región en rojo es bastante grande: a partir de un valor de *kshell* y grado todo tienen valores similares que se encuentra alrededor del punto medio de los dos ejes. Es decir, los nodos cuyos grado y *kshell* sean al menos los señalados por esa celda en la gráfica alcanzan el valor máximo de infectividad o uno muy cercano. En contraste, para $\beta = 0.01$ (como muestra la imagen Figura 4.7a), la celda donde inician los nodos de máxima infectividad se ubica hacia arriba y derecha de donde empezaba para $\beta = 0.04$.

El contraste de las imágenes en la Figura 4.7 también sugiere que cuando aumentamos la probabilidad de contagio $\beta$ y mantenemos iguales los otros parámetros, aumenta la infectividad de todos los nodos de la red de manera uniforme, y no se modifica la infectividad relativa de los nodos.

Finalmente, definimos la función "de imprecisión" $\xi_{f_i}$ propuesta originalmente en (Kitsak et al., 2010), que calcula la variación de infectividad de subconjuntos rankeados a partir del grado, *PageRank* y *kshell* respecto a la infectividad máxima; teniendo de esta manera una suerte de función de error, o mejor dicho, de una función que calcula la imprecisión de la infectividad de las medidades de centralidad de subconjuntos de nodos. A continuación la describimos.

Primero, construimos cuatro listas de tamaño $N$, cada una ordenada de manera descendente con base a los valores de los nodos de $f_1$, *kshell*, *PageRank* y grado. Luego, establecemos lo siguiente: uno, la infectividad de un subconjunto de nodos de cualquier tamaño está dada por el promedio de las infectividades ($f_1$) de sus nodos. Dos, para subconjuntos del mismo tamaño construidos con los primeros nodos de las cuatros listas, la infectividad del construido de la lista ordenada por $f_1$ será la infectividad máxima.

Denotamos con $\Gamma_{ef}$ a la lista de los nodos ordenados por sus valores $f_1$, y con $\Gamma_{ks}$, $\Gamma_{gr}$ y $\Gamma_{pr}$ a las listas ordenados respecto a *kshell*, grado y *PageRank* respectivamente. Y definimos la función $F_\Gamma$ que calcula la infectividad de un subconjunto de elementos de

alguna de las listas ordenadas como $F_\Gamma = \frac{1}{|V|\cdot p}\sum_{v\in\Gamma}^{p} f_1(v)$, con $p \in |[0,1]$ y $\Gamma$ representa alguna de las listas ordenadas.

Finalmente, definimos las funciones de imprecisión para cada medida de centralidad:

$$\xi_{\text{kshell}}(p) = 1 - \frac{F_{\Gamma_{ks}}(p)}{F_{\Gamma_{ef}}(p)} \qquad \xi_{\text{grado}}(p) = 1 - \frac{F_{\Gamma_{gr}}(p)}{F_{\Gamma_{ef}}(p)} \qquad \xi_{\text{pagerank}}(p) = 1 - \frac{F_{\Gamma_{pr}}(p)}{F_{\Gamma_{ef}}(p)}$$

Notamos que el parámetro $p$ indica la fracción de nodos que se tomará en las listas ordenadas $\Gamma$ y el denominador es el mismo en las tres funciones, debido a que $F_{\Gamma_{ef}}$ está ordenada respecto a las infectividades de los nodos $f_1$, con lo que se garantiza que $F_\Gamma(p) \leq F_{\Gamma_{ef}}(p)$, $\forall p \in [0,1]$ independientemente del criterio de orden, $\Gamma$, utilizado. Por tanto las funciones de imprecisión $\xi_\alpha$ toman valores entre 0 y 1.

Así, valores cercanos a 0 indican que la medida de centralidad es un buen clasificador de importancia epidemiológica, ya que el promedio de infectividad de la fracción de nodos seleccionados con esta medida es muy cercano al promedio de infectividad sobre los nodos de infectividad más alta.

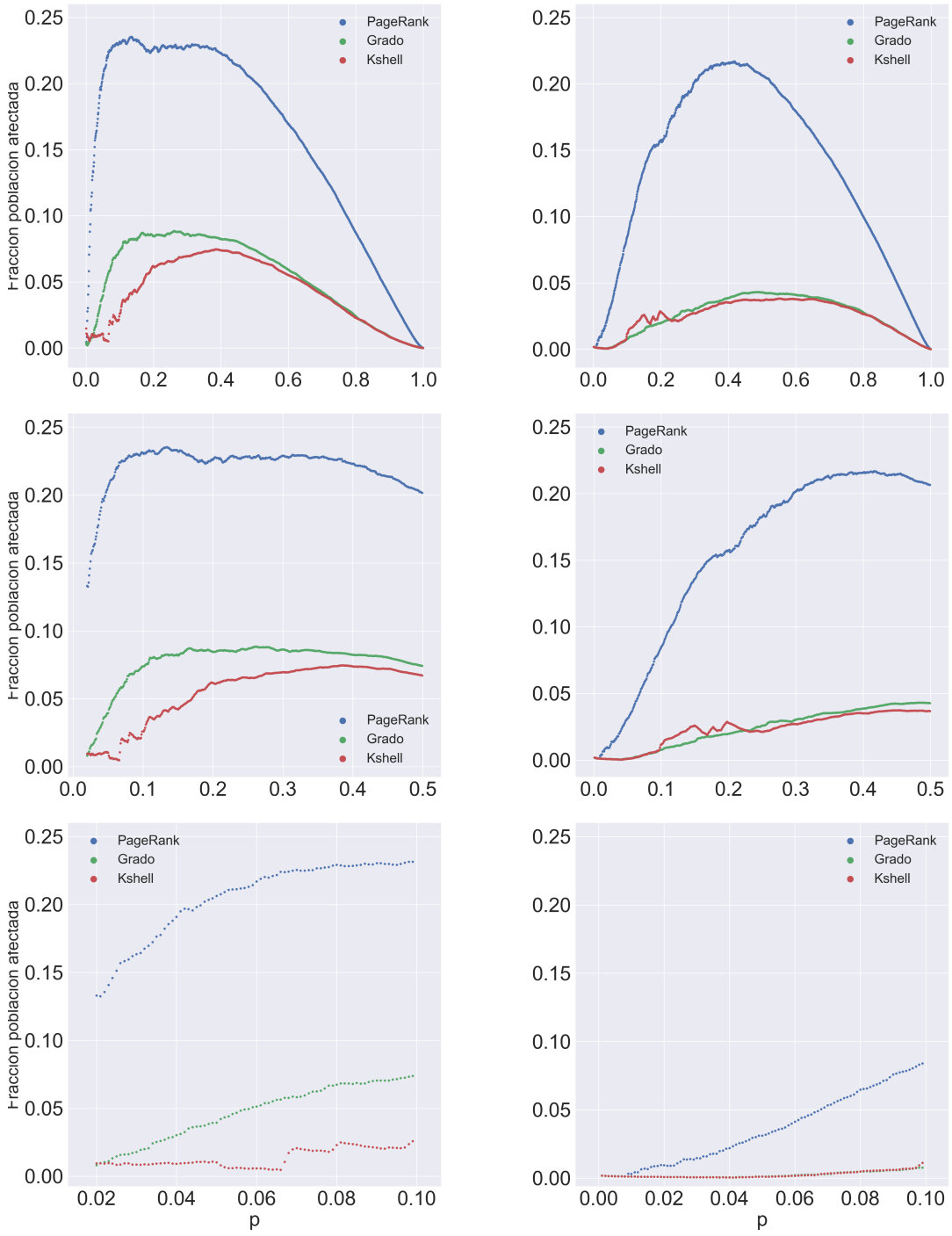
A continuación mostramos las gráficas de las funciones de imprecisión para *kshell*, grado y *PageRank* tomando valores de $p \in [0,1]$ para los dos valores de $\beta$ utilizados [2]. Mostramos tres gráficas distintas por cada valor de $\beta$: la primera presenta las curvas completas y las otras dejan ver detalles en acercamiento. De las primeras dos imágenes podemos observar que la imprecisión del *PageRank* es siempre mucho mayor que la del grado y el *kshell*. Además, para la $\beta$ de 0.04 (Figura 4.8b), la imprecisión de *kshell* y el grado es muy parecida; mientras que para el valor menor de $\beta$, el *kshell* supera inicialmente al grado hasta aproximadamente el punto $p = 0.7$, en el que las curvas se vuelven casi indistinguibles.

Concluimos que cuando la tasa de transmisión $\beta$ es baja, el *kshell* identifica mejor que el grado a los nodos de mayor infectividad y que cuando la tasa de transmisión es

[2]La obtención de la gráfica asociada al valor de $\beta = 0.01$ en la Figura 4.8 culmina, junto con la Figura 4.7a, la reproducción de una parte del trabajo en (Kitsak et al., 2010). Otros de los resultados que alcanzamos se inspiran en las preguntas y lecturas derivadas de éste.

alta (valores altos de $\beta$) el detalle capturado por el grado es suficiente y el desempeño del *kshell* y el grado es similar.

En resumen, el *kshell* hace una mejor labor en la identificación de los nodos que tienen mayor influencia en procesos de contagio. Sin embargo, el *PageRank* no debe descartarse por completo, ya que puede usarse para estudiar otras facetas del fenómeno epidemiológico, como veremos en el siguiente capítulo. El grado tampoco debe desecharse, pues a pesar de que es la medida más sencilla, también logra capturar información útil, en nuestro caso, la infectividad asociada a los nodos.




(a) $\beta = 0.01$ (b) $\beta = 0.04$

Figura 4.8: Comparativo de la impresición, $\xi_{f_i}$, reflejada por *kshell*, *PageRank* y grado obtenidas a partir de dos tasas de transmisión $\beta$ 0.01 (columna izquierda) y 0.04 (columna derecha). Además, los renglones capturan diferentes niveles de acercamiento.

# Capítulo 5

# Estrategias de vacunación

En este capítulo culminamos nuestra investigación con la evaluación de estrategias de vacunación basadas en el rankeo de nodos que dan las medidas utilizadas. Las estrategias de vacunación que proponemos consisten en elegir un conjunto de individuos usando información de las distintas medidas de centralidad. Este conjunto se "inmuniza" dándoles como estado inicial el de recuperado, que mantendrán durante el resto del proceso epidémico. De esta manera, un nodo vacunado no podrá adquirir la enfermedad y tampoco podrá transmitirla. El objetivo de este capítulo es estudiar cuáles medidas de centralidad maximizan el efecto de la vacuna sobre la magnitud de la epidemia.

Es claro que podemos erradicar cualquier enfermedad si vacunamos a toda la población, pero esto no aporta ningún conocimiento y dista mucho de ser un escenario realista, ya que durante el control y prevención de enfermedades, solamente existe una cantidad limitada de recursos y vacunas.

## 5.1. Procedimiento

En cada campaña de vacunación seleccionamos a un conjunto de indiviudos a los que hacemos inmunes, de forma que no pueden contraer ni propagar la enfermedad.

Una vez que inmunizamos a un subconjunto de individuos, damos inicio a la simulación del proceso epidémico con origen en cada uno de los nodos de la red usando los parámetros antes utilizados. Finalmente, calculamos las medidas de infectividad de cada nodo mediante las funciones $f_1$ y $f_3$.

Tomaremos como referencia los resultados de las simulaciones que estudiamos en Capítulo 4 en las que no se vacunó a ningún individuo y que reflejan el alcance *máximo* de la epidemia en la red y la infectividad de sus nodos. Los grupos de individuos que seleccionamos para vacunar son 2 % y 15 % del total, elegidos:

- en forma aleatoria (control),
- con el *kshell* más alto,
- con el *PageRank* más alto.

A nivel de código, para simular los procesos de contagio en la población con un subconjunto de individuos vacunados, hacemos lo siguiente: cambiamos el estado de los nodos seleccionados de susceptible a recuperado entre las líneas 3 y 4 del Algoritmo 2, y ejecutamos el resto del código normalmente. De esta manera, los nodos vacunados no cumplirán la condición de la línea 8 (ser susceptible), por lo que nunca pasarán al estado infectado y no podrán transmitir la enfermedad.

Otras consideraciones que debemos tener al modelar la vacunación de esta manera son: cuando un nodo vacunado es donde comienza el proceso de contagio, consideraremos que la vacuna falló y permitiremos que sea el primer nodo en contraer y transmitir a sus vecinos la enfermedad (líneas 7 y 8 de Algoritmo 2). Debido a que simulamos un proceso de contagio estocástico y vacunamos $c$ nodos por cada nodo de la red, entonces, de los 47,719 brotes simulados, la vacuna habrá fallado $c$ veces [1] que son los casos en los que el brote se origina en cualquiera de los nodos vacunados y, en cada uno de estos brotes, el número efectivo de nodos inmunizados es $c - 1$.

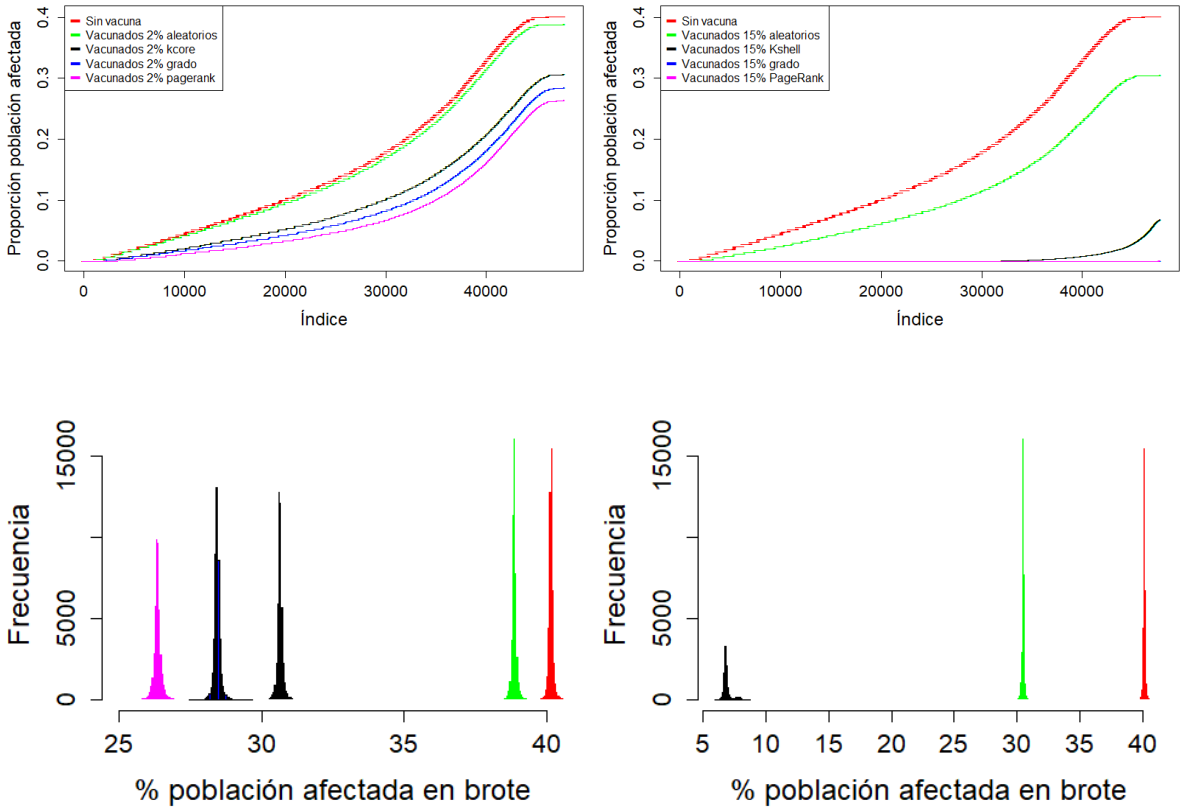
[1] En este capítulo nos referimos como un "brote simulado" a las $N$ simulaciones.

Estas características pueden ser implementadas con las siguientes instrucciones: a) inmunizar nodos (cambiar estado susceptible a recuperado de los $k$ nodos seleccionados), b) hacer infeccioso al nodo origen y agregarlo a la lista de nodos infectados, c) continuar normalmente el contagio (ejecutar a partir de la línea 5 del Algoritmo 2).

## 5.2. Inmunizando la red de colaboración de actores

En este apartado describimos los resultados de las distintas estrategias de vacunación implementadas en la red de colaboración que describimos en la Sección A.1.

La Figura 5.1 muestra el efecto de la vacuna en los dos aspectos generales que cuantifican la magnitud de la epidemia: el número de individuos alcanzados por la enfermedad (capturada por la función de infectividad $f_1$). Y el tamaño de los brotes cuando la propagación alcanza magnitud epidémica (función $f_3$). La gravedad de un brote epidémico es una ponderación de estas dos medidas. Por este motivo analizamos los valores devueltos por $f_1$ y $f_3$.




(a) Vacuna aplicada al 2 % de la población (b) Vacuna aplicada al 15 % de la población

Figura 5.1: Efecto en la propagación de la enfermedad de vacunar al dos y quince porciento de la población seleccionada de cuatro formas distintas: al azar, por su *kshell*, su grado o su *PageRank*. Arriba graficamos los valores ordenados de $f_1$ y abajo los valores $f_3$ (eliminando los valores igual a 0)

En la Figura 5.1a mostramos los valores de $f_1$ (arriba) y $f_3$ (abajo) obtenidos de simulaciones de la propagación de la infección como lo hicimos en el capítulo anterior, pero vacunando al 2 % de los individuos elegidos conforme el orden (de mayor a menor) que reciben según su: a) grado, b) *kshell* y c) *PageRank*. d) al azar. De manera consistente, el *PageRank* minimiza $f_1$ y $f_3$. Después del *PageRank*, la estrategia de vacunación que reduce más el tamaño de la epidemia es el grado, le sigue el *kshell* y al último está

| | 2 % vacunados | | | | 15 % vacunados | | | |
|---|---|---|---|---|---|---|---|---|
| | f1( %) | | f3( %) | | f1( %) | | f3( %) | |
| Vacunación | Media | Max | Media | Max | Media | Max | Media | Max |
| Sin Vacuna | 16.1 | 40.3 | 40.2 | 41.2 | 16.1 | 40.3 | 40.2 | 41.2 |
| Aleatorio | 15.4 | 39.0 | 38.8 | 40.0 | 10.9 | 30.6 | 30.5 | 31.4 |
| *Kshell* | 10.0 | 30.8 | 30.6 | 31.9 | 0.0 | 6.9 | 6.9 | 8.7 |
| Grado | 8.6 | 28.6 | 28.4 | 29.6 | 0.0 | 0.1 | - | - |
| *PageRank* | 7.4 | 26.5 | 26.3 | 27.9 | 0.0 | 0.0 | - | - |

Figura 5.2: Tabla que contiene la media y el máximo de los valores devueltos por las funciones $f_1$ y $f_3$ de las poblaciones al aplicar la vacuna a al 2 % y 15 % de nodos con las mayores medidas de centralidad. Para la $f_3$ eliminamos los valores igual a 0. En las dos ultimas columnas, todos los valores fueron 0.

la de elegir los nodos de manera aleatoria.

Cuando aumentamos la fracción de individuos que recibe la vacuna, (ver la Figura 5.1b), los resultados coinciden con lo que observamos antes (ver la Figura 5.1a): el *PageRank* tiene el mayor impacto sobre el tamaño de la epidemia, seguido por el grado, el *kshell* y la selección aleatoria.

Estos resultados sugieren que la estrategia de vacunación basada en el *PageRank* es la mejor. Si la comparamos con la estrategia que utiliza el *kshell*, reduce casi un 30 % la $f_1$ promedio y un 13 % la $f_3$ promedio. Aún así, la vacunación del 2 % de los *kshell* más altos, tiene un efecto unas 8 veces más efectivo que elegir a nodos de manera aleatoria, como muestran los resultados de la vacunación aleatoria del 15 % de los nodos.

Finalmente, cabe recalcar que al elegir el 15 % de los nodos con mayores valores de *PageRank* y del grado, se logra evitar el brote independientemente de dónde estuviera el caso inicial.

A continuación damos otros detalles del efecto que tienen las diferentes estrategias

de vacunación en la red, a fin de conocer qué medidas de centralidad son mejores guías de nuestra estrategias.

En la Figura 5.3 se muestra el $f_1$ de cada nodo después de haber sido vacunado contra su $f_1$ sin vacuna. Además, graficamos la identidad (línea discontinua) como referencia y la recta que resulta de calcular una regresión lineal sobre dicha relación(línea continua). De esta manera, un punto (círculo verde) en la gráfica, corresponde a un individuo, cuya coordenada $x$ es su valor $f_1$ que ocurre en la población vacunada, y su coordenada $y$ su valor $f_1$ en una población que no recibió ningún tipo de vacuna. Las cruces rojas indican nodos que recibieron la vacuna, en este caso, son los del *kshell* más alto. Los puntos que se encuentran por arriba de la identidad corresponden así a los nodos que tuvieron menor infectividad en la población vacunada que en la que no recibió vacuna. Por su parte, los nodos que se encuentran por debajo de la identidad tuvieron mayor infectividad. Aunque no para todos los nodos se redujo el tamaño de la epidemia, las epidemias que iniciaron en los nodos más peligrosos (nodos con valores mayores en la coordenada $y$) sí fueron consistentemente de menor tamaño.

Comparamos el efecto de la vacuna aplicada con el rankeo del *kshell* contra el efecto de aplicar la vacuna en los subconjuntos seleccionados por su *PageRank* (Figura 5.4)y su grado (Figura 5.5).

La estrategia de vacunación basada en el *PageRank* resultó en la mayor reducción de la propagación de la infección. Los nodos que tenían el *PageRank* más alto, aunque se acumulan en la región superior de la gráfica, están dispersos en el eje vertical en mayor medida que los de *kshell* y grado más altos. Los nodos con el mayor grado (Figura 5.5), muestran una dispersión en el eje $y$ similar a la del *kshell*, aunque no se distribuyen a la largo de la franja superior como lo hacen los nodos para el *kshell*.

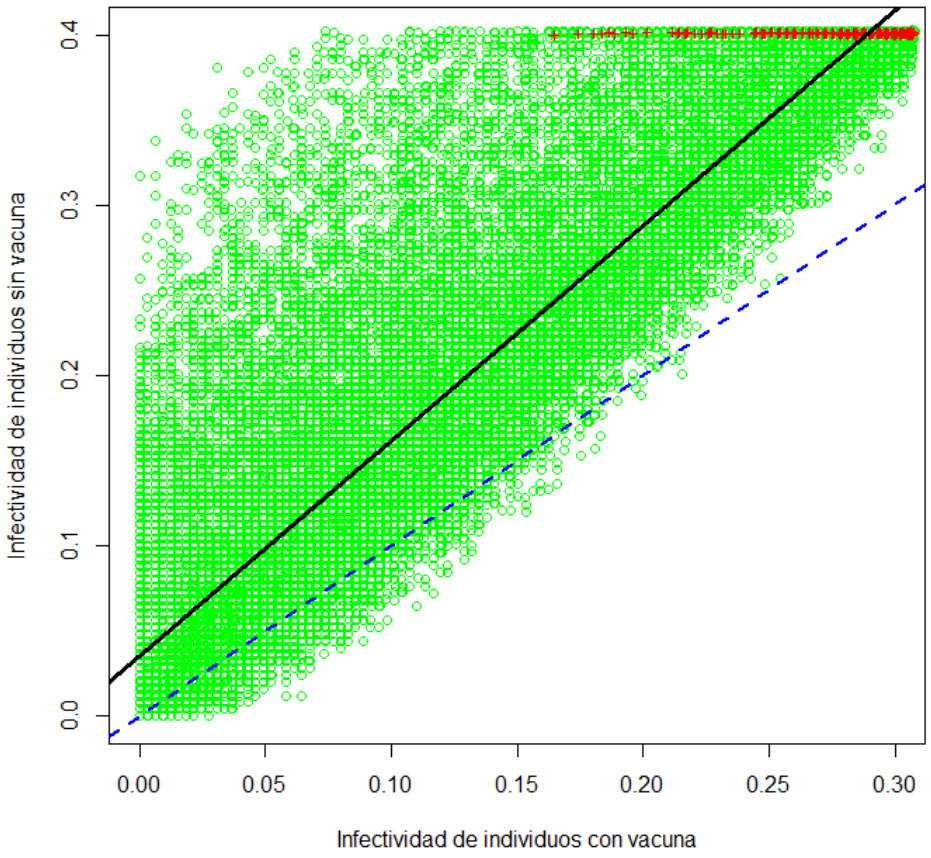



Figura 5.3: Gráficas de los valores $f_1$ en la población que recibió la vacuna contra esto mismos valores en la población que no recibió la vacuna. La línea discontinua es la identidad. La línea sólida es la recta que resulta de realizar una regresión lineal entre estos dos conjuntos. Las cruces rojas son los nodos que recibieron la vacuna.

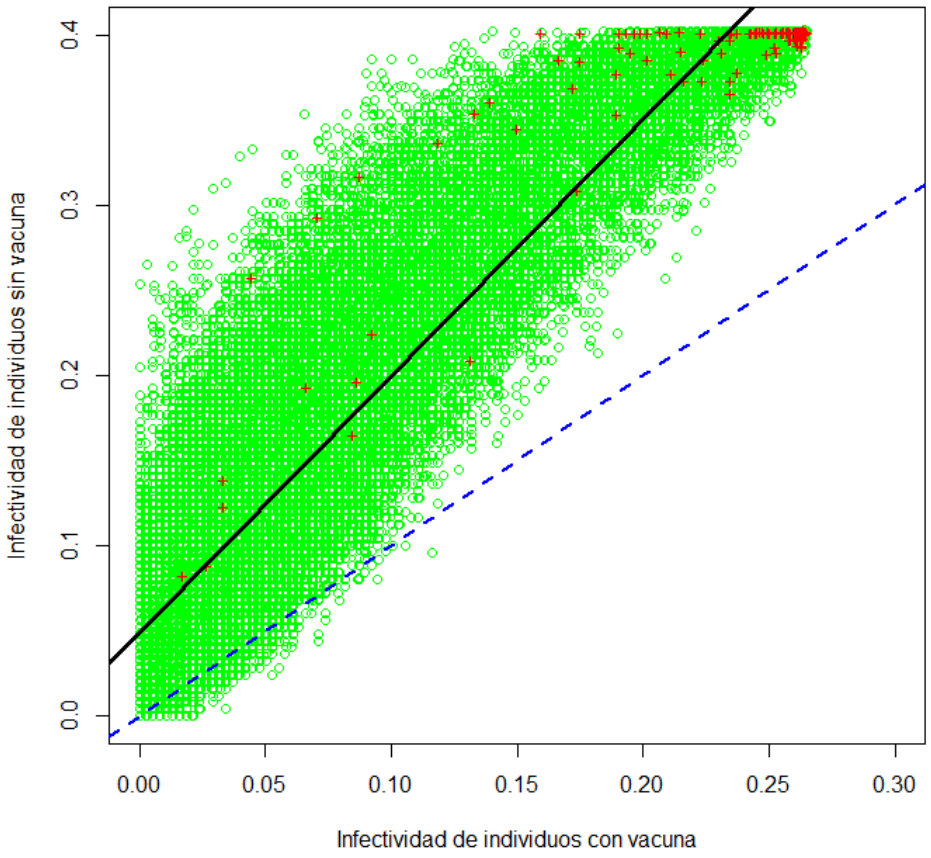



Figura 5.4: Valores de $f_1$ en la población en que se vacunó al 2 % de *PageRank* más alto contra los mismos valores en la población sin vacuna. Así como en la Figura 5.3 , la identidad y la recta de la regresión lineal están indicadas por una línea discontinua y una continua respectivamente.

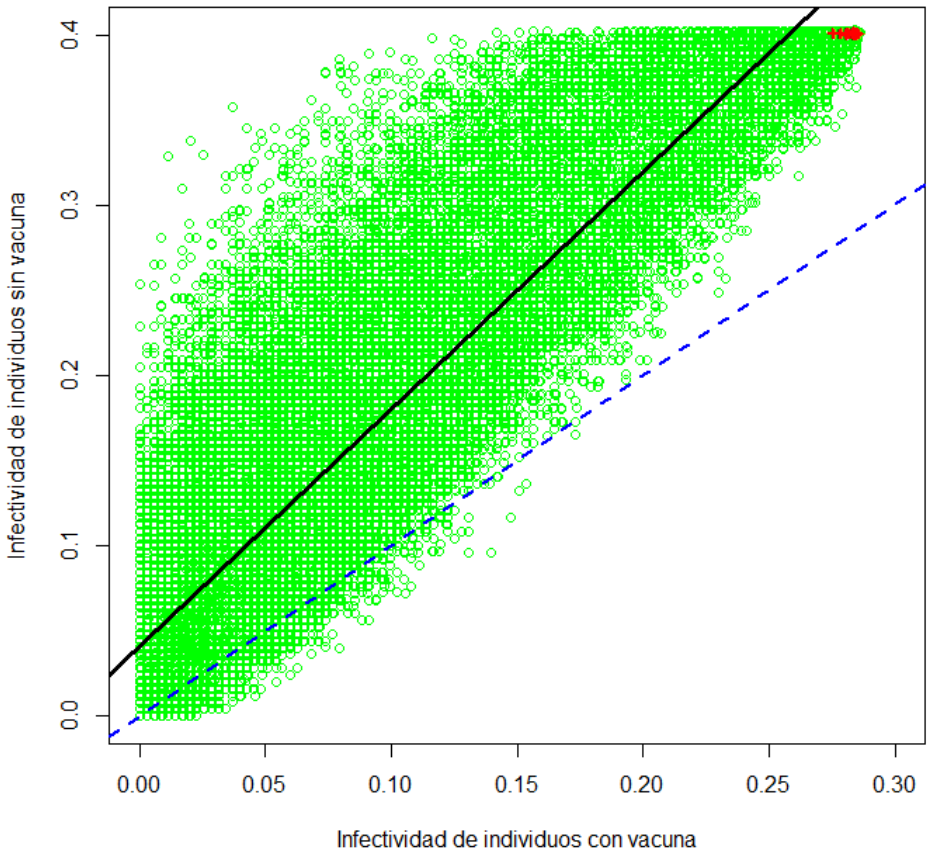



Figura 5.5: Valores de $f_1$ en la población donde se vacunó al 2 % de mayor grado contra los mismos valores en la población sin vacuna. Así como en la Figura 5.3 y la Figura 5.4, la identidad y la recta de la regresión lineal están indicadas por una línea discontinua y una continua respectivamente.

Por último, mostramos el efecto de la vacunación de 5 % y 15 % de nodos seleccionados aleatoriamente en la Figura 5.6. En la gráfica de la derecha se puede apreciar que los puntos rojos están distribuidos uniformemente a lo largo de los ejes $x$ y $y$, precisamente por haberse elegido a los puntos de manera uniforme.

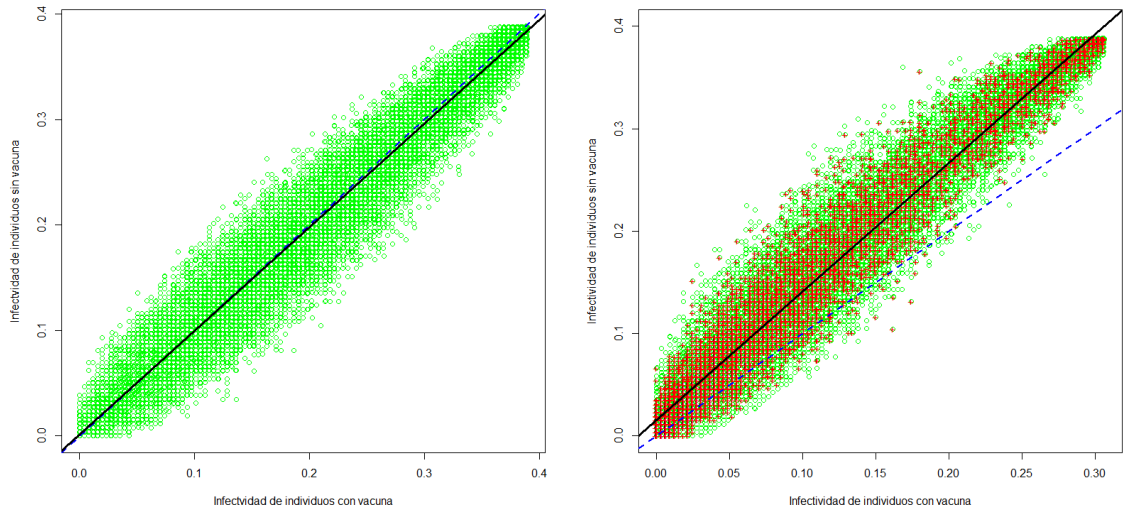



Figura 5.6: Gráficas de la infectividad de la epidemia devuelta por $f_1$ en la población con vacuna aplicada a un subconjunto elegido aleatoriamente contra la población sin vacuna. La línea discontinua muestra la identidad. La recta sólida es la regresión lineal de estos dos conjuntos.

# Capítulo 6

# Conclusiones

Estudiamos una estrategia que determina la influencia que tienen los nodos en un proceso de contagio regido por dinámicas epidemiológicas, que bajo el enfoque de redes complejas, puede pensarse como la definición de una medida de centralidad. Esta medida refleja la capacidad que tiene un nodo para inducir un brote epidémico. Una vez establecida la medida de infectividad, exploramos sus relaciones con otras medidas de centralidad. Finalmente contrastamos estas relaciones con estrategias de vacunación, que es una prueba crítica para de su utilidad para la epidemiología.

Al comparar el grado, el *PageRank* y el *kshel*l, esta última es la medida que mejor identifica los nodos de mayor infectividad, pues nodos con valores altos de *kshell* inducen brotes epidémicos de mayor magnitud. El grado resultó estar por debajo del *kshell* y en último lugar quedó el *PageRank*. Curiosamente estos resultados se invirtieron al estudiar el efecto de estrategias de vacunación en la propagación de la enfermedad, basadas en el rankeo que da cada una de estas medidas. El *PageRank* llevó a la mayor reducción en el tamaño de los brotes. Le siguió el grado y por el último el *kshell*, aunque cualquiera de estas estrategias de vacunación es mejor que vacunar de forma aleatoria.

Estos dos resultados sugieren que determinar a los nodos que inducen brotes epidémicos de mayor magnitud no es equivalente a identificar el conjunto de nodos que reducen

la propagación de la infección con mayor eficacia. Cuando menos, el grado, *kshell* y *PageRank* no capturan ambas propiedades simultáneamente en la red de colaboración de actores que estudiamos. Los nodos con *kshell* alto pertenecen a *kcores* de la red con valores de $k > 0$ altos. Éstas son subestructuras de la red fuertemente conectadas, mismas que facilitan la propagación de una infección dentro de ellas, y que a su vez aumentan la posibilidad de que la enfermedad se propague hacia el exterior de la red, incrementando con ello la magnitud total del brote. Pero al elegir nodos con mayor *kshell* estos constituirán el mismo kcore o se eligirán de un número reducido de kcores diferentes, por lo que al momento de vacunar estos nodos, en realidad uno o pocos kcores están recibiendo la vacuna, lo que disminuye su efecto global sobre la propagación de la enfermedad. Tal vez para aumentar el efecto de la vacunación, baste con elegir a un representante de cada kcore.

La medida de infectividad que estudiamos se basa en simulaciones de brotes epidémicos, que a su vez están inspiradas en los procesos de contagio de los modelos clásicos de la epidemiología teórica, en este caso en el modelo $SIR$. Estas simulaciones tienen diferentes restricciones, como el supuesto de que el tiempo que toman los individuos en recuperarse de la infección es constante. Esta restricción del tiempo de infecciosidad difiere de la del modelo $SIR$, ya que en éste, su distribución sigue una función exponencial. Pero ambos casos fallan en capturar lo que se ha observado en estudios médicos(Keeling y Rohani, 2011), en los que el tiempo que toma a los individuos recuperarse de la infección se distribuye en forma normal alrededor de un valor promedio. Por lo anterior, los resultados obtenidos en este trabajo se pueden complementar, y su caso refinar, si incorporamos a las simulaciones otras características propias de los fenómenos epidemiológicos; por ejemplo, las reglas de transición entre compartimientos que capturan los modelos $SIS$ o $SEIR$. O incluir una distribución más realista del periodo de infecciosidad.

Futuras investigaciones podrían enfocarse en generalizar los resultados obtenidos

en este trabajo. Esta generalización podría beneficiarse de realizar simulaciones que consideren otros aspectos de los procesos epidemiológicos y de comparar simulaciones de tiempo continuo contra simulaciones de eventos discretos. Lo anterior podría combinarse con los resultados que se obtengan de reproducir el análisis que describimos en este trabajo en otras redes específicas. Seguramente estos resultados traerían avances en nuestra comprensión de procesos epidemiológicas y generalización de redes complejas.

No debemos perder de vista que en este trabajo no exploramos las relaciones de las medidas de centralidad aprovechando los resultados teóricos que se conocen en redes. Por ejemplo, podríamos analizar las relaciones del *kshell*, *PageRank* e infectividad a partir de obtener y comparar las funciones que mejor las aproximan sus distribuciones. Pero existen otras caracterizaciones de redes que también podrían emplearse, tales como la distribución condicional del grado, el factor de *clustering* de la red y la modelación de la población con redes multicapa.

Finalmente, debemos reconocer la importancia que toman las herramientas computacionales y los beneficios que nos ofrecen en el entendimientos de problemas que derivan en la generación de nuevos resultados. En el caso de nuestra investigación, fue imprescindible preguntarnos qué tipo de lenguaje de programación nos ayudaba más, así como buscar mayor eficiencia en los cálculos a través del uso de diferentes estructuras de datos. Por último, la elección de lenguaje y decisiones en la estructura y diseño del código tratarse con recelo debido a la diversidad enorme de herramientas computacionales que existen hoy en día.

# Apéndice A

# Descripción de la red de colaboración de actores

## A.1. Descripción de la red

Construimos y estudiamos una red de colaboración de actores de películas a partir de la información que se obtuvo de *Research on Social Networks: IMDB networks* (2010). Esta red representa un actor con un nodo, y tiene una arista entre cada dos actores distintos que actuaron en una misma película.

La red de actores se compone de 47,719 nodos y 1,098,451 aristas. Los datos se descargaron de `http://www-levich.engr.ccny.cuny.edu/webpage/hmakse/software-and-data/` y, de acuerdo con el mismo sitio, ésta es una red de "conexiones entre actores que co-actuaron en películas, tales que el género está etiquetado como 'adulto' por el portal *Internet Movie Database*, y con fechas de estreno de las últimas décadas".

Una propiedad que se ha visto en redes complejas, es la llamada *mundo pequeño*. Este nombre viene del hallazgo realizado por el psicólogo Stanley Milgram a través de una serie de experimentos que iniciaron en los años 1960. Milgram eligió al azar (M. Newman, 2010, págs. 54-58) un número de personas de una ciudad en Estados

Unidos y les mandó un paquete, pidiéndoles que se lo hicieran llegar a una persona específica con residencia en otra ciudad a miles de kilómetros de distancia; pero con la restricción de que no se lo enviaran directamente, sino que se lo enviaran a alguna persona que supieran que lo conocería directamente o a alguna persona que tuviera más posibilidades de conocerlo. Milgram, y otros investigadores, repitieron este experimento con variaciones como el número de paquetes y los lugares de origen y salida. En el inicial, Milgram, envió 96 paquetes y 18 lograron llegar a la persona indicada. Los que iniciaron el envío residían en Omaha, Nebraska y la persona objetivo, en Boston, Massachussetts ((M. Newman, 2010, págs. 54-58)). Del experimento incial, se observó que de los paquetes que sí llegaron al destino, el promedio de pasos que les tomó en llegar fue de 5.9. De aquí se popularizó el dicho de hay 6 grados de separación entre cualesquiera dos personas en el mundo (M. Newman, 2010, págs. 54-58).

Aunque este experimento se realizó en una red específica y bajo condiciones muy particulares, estos resultados también se han presentado en muchas otras redes reales. Este efecto depende en realidad de distintas propiedades de la red, en particular de su topología, según se aprecia en la importancia que tiene la distribución de grado (libre de escala).

Debido a que el objetivo de este trabajo es estudiar principalmente las relaciones de las medidas de centralidad de los nodos con la relevancia que tienen durante un proceso de contagio, a continuación damos una caracterización de la red de *IMDB* usando las medidas de centralidad presentadas anteriormente.

| Centralidad | Media | Dev Std | Min | 1er Cuartil | Mediana | 3er Cuartil | Max |
|---|---|---|---|---|---|---|---|
| Grado | 46.0 | 124.3 | 1 | 9 | 15 | 33 | 3031 |
| *Kshell* | 25.5 | 35.2 | 1 | 8 | 13 | 26 | 199 |
| *PageRank* ($\times 10^{-6}$) | 21 | 35 | 3 | 8 | 13 | 19 | 10.23 |

Tabla A.1: Valores estadísticos descriptivos de los valores crudos del grado *kshell* y *PageRank* para la red actores.

Primero, obtenemos la Tabla A.1 que contiene distintas medidas estadísticas descriptivas del grado, *kshell* y *PageRank* de los 47,719 nodos que componen la red de colaboración de actores obtenidos de IMDB.

A continuación damos simplemente una descripción de los valores de las medidas...

El valor del grado mínimo que puede tener un nodo es 0, y en redes conexas como ésta, 1. El valor mínimo del *kshell* en esta red también es 1. Por su parte, el *kshell* más alto es unas 15 veces más chico que el máximo grado. Así, tanto el *kshell* como el grado toman valores enteros no negativos, y, en el *kshell*, el rango es menor. En cambio, los valores de *PageRank* están en el intervalo [0,1].

A continuación exploramos las relaciones de las frecuencias entre las diferentes medidas que son las que nos interesa interpretar en un contexto epidemiológico: grado, *kshell* y *PageRank*. Las gráficas del renglón de arriba de la **??** muestran los valores ordenados de cada medida, y las de abajo, los histogramas de los valores normalizados. Estas imágenes sugieren que la distribución de valores de *kshell*, comparadas con las otras dos, tiene un mejor ajuste lineal. Por otro lado, presenta un pico, lo cual, lo hace inusual.

Para cada nodo se calculan estos tres valores. En la Figura A.1 gráfica los contrastamos con el fin de visualizar posibles correlaciones entre medidas.

A simple vista el *PageRank* y el grado están más correlacionados entre ellos que con el *kshell*, aunque en todos los casos se logran apreciar algunos patrones que relacionan

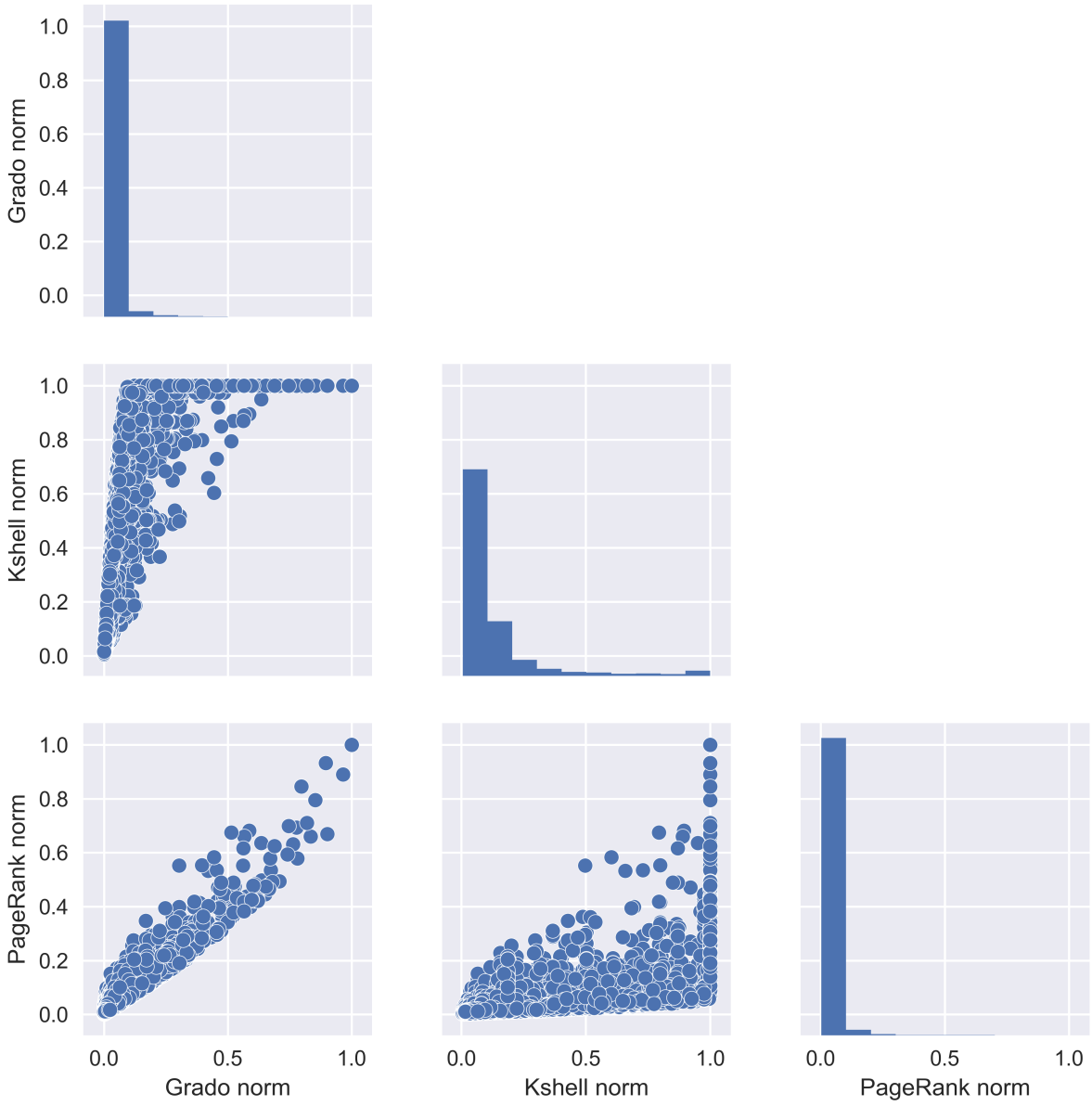



Figura A.1: Comparativo de los valores normalizados del grado, *kshell* y *PageRank* de los nodos.

distintos segmentos de los valores.

Debido al tipo de distribuciones que corresponden a las medidas, nos planteamos si para un nodo que se encuentra al principio de la distribución, su grado, *PageRank* y *kshell* se relacionan de la misma manera que para nodos tomados en otras regiones de la distribución. Con el fin de encontrar relaciones específicas de nodos tomados en distintas regiones, generamos la Figura A.2, que trabaja solo con los valores normalizados de cada medida. Estas imágenes organizan los nodos de acuerdo con: el orden que les da el *kshell* (columna izquierda) y el del *PageRank* (columna derecha). Arriba tomamos mil nodos de forma aleatoria de la lista ordenada de acuerdo con cada medida. Para el *kshell*, observamos que, aunque se encuentra sistemáticamente por arriba, esto no es una regla, ya que el *PageRank* de algunos nodos rebasa su *kshell*, como muestran los puntos verdes que están por arriba de la línea roja (aunque estos nodos representan una minoría, también se presentan al tomar otros conjuntos aleatorios). Además, esta imagen también sugiere que de acuerdo con *PageRank*, los nodos de valores más altos están por debajo del grado, pero conforme avanzamos en el rankeo, se invierte esta situación, quedando el *PageRank* por arriba del grado. Esta hipótesis la confirma la imagen de abajo que grafica *PageRank* contra grado, mostrando en distintos colores grupos de mil nodos tomados cada 10 mil en la lista rankeada por el *kshell*. En ésta, observamos que los mil nodos con *kshell* más alto (en azul) en general se encuentran por debajo de la identidad. Para el siguiente grupo (púrpura) formado por los nodos rankeados del diez mil al once mil, los puntos se encuentran por arriba de la identidad o encima de ella. Para los siguientes grupos, observamos que estos se ubican invariablemente arriba de la identidad.

Al ordenar respecto al *PageRank*, obervamos que el grado se distribuye alrededor de éste de manera casi uniforme (imagen arriba). El *kshell*, aunque sigue estando por arriba de las otras dos medidas, presenta mayor dispersión de valores, de forma que algunos

nodos pueden estar ligeramente por arriba de la línea verde (*PageRank*), aunque otros estén muy por encima de la línea verde. La segunda imagen (abajo) toma también 6 grupos de mil nodos cada diez mil lugares y grafica su *kshell* contra su grado. En el grupo de los mil nodos con *PageRank* mayor (azules), el *kshell* no muestra una relación lineal con el grado. En cambio, se observan una variedad de patrones. Por ejemplo, para los nodos con *kshell* igual a uno, sus grados toman casi cualquier valor posible. Aunque también podemos distinguir dos conjuntos de nodos que se agrupan alrededor de dos líneas distintas, una que acota por debajo y otra por arriba a todos los nodos azules.

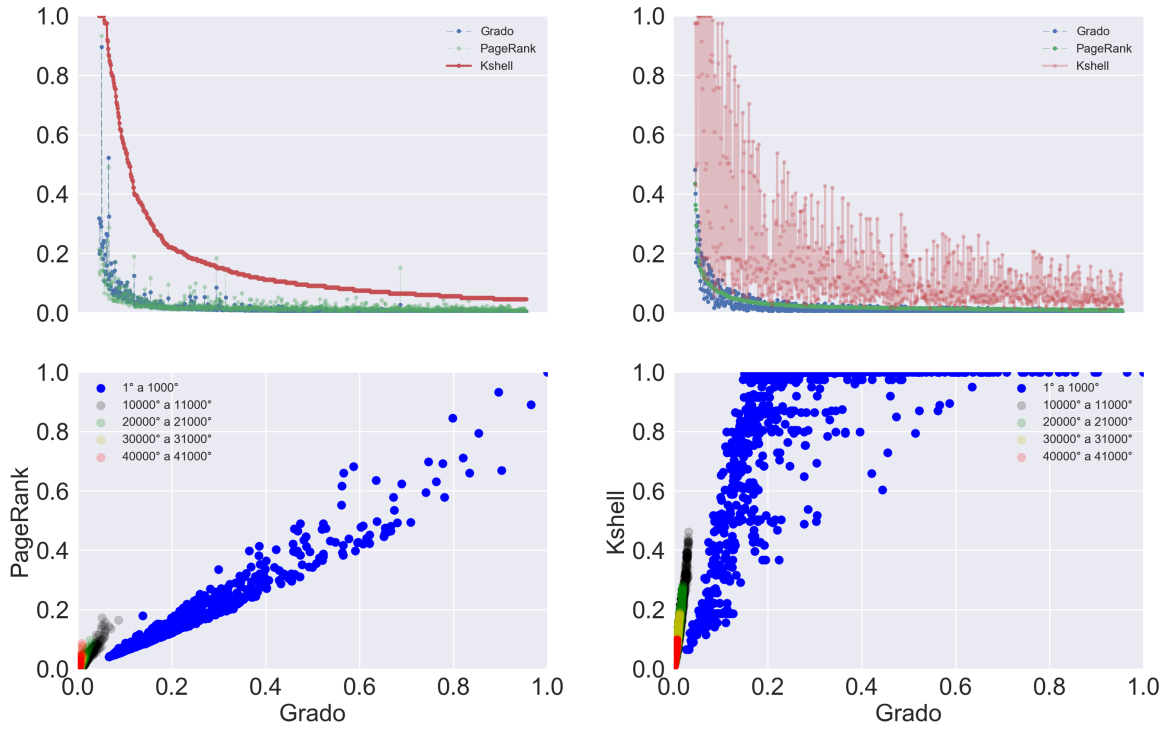



Figura A.2: Comparativo entre las medidas de centralidad en dos formas distintas. Arriba, se eligieron mil nodos aleatoriamente y se grafican su grado, *PageRank* y *kshell* ordenados por su *kshell* a la izquierda y por su *PageRank* a la derecha. Abajo a la izquierda se grafica el grado contra el *kshell* en seis grupos distintos de mil nodos elegidos cada diez mil lugares. El color azul corresponde a los primeros mil nodos, el púrpura a los nodos entre los 10mil y 11mil nodos con *kshell* más alto. El negro entre los veinte y veintiún mil, etc. Abajo a la derecha se contrastan el grado y *kshell* para los seis grupos elegidos similarmente. En ambas imágenes de arriba, los colores rojo, verde y azul, corresponden al *kshell*, *PageRank* y grado respectivamente.

# Referencias


Barabási, A.-L., y cols. (2016). *Network science*. Cambridge University Press.

Bernoulli, D. (1760). *Réflexions sur les avantages de l'inoculation," mercure de france (1760) 173–190. english translation by r. pulskamp, department of mathematics and computer science, xavier university, cincinnati (2009).*

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., y Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical mechanics and its applications*, *391*(4), 1777–1787.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., y Stein, C. (2009). *Introduction to algorithms*. MIT Press.

Esteva, L. G. G. H. J. y. Z. M. (1991). Matemáticas y epidemiología. *Ciencias*, *24 octubre-diciembre*, 57-63.

Gephi. (2019). *Gephi - the open graph viz platform.* Descargado de `https://github.com/gephi/gephi/wiki/PageRank`

Jaramillo Arango, C. J., Martínez Maya, J., y cols. (2010). *Epidemiología veterinaria.* Editorial El Manual Moderno.

Keeling, M. J., y Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermark, M., y Mckendrick, A. (1927). Contributions to the mathematical theory of epidemics. part i. *Proc. R. Soc. A*, *115*(5), 700–721.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., y Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, *6*(11), 888.

Nava-Frías, M., Searcy-Pavía, R. E., Juárez-Contreras, C. A., y Valencia-Bautista, A. (2016). Enfermedad por virus de chikungunya: Actualidad en méxico. *Boletín Médico del Hospital Infantil de México*, *73*(2), 67–74.

NetworkX. (2019). *Networkx.* Descargado de https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

Newman, M. (2010). *Networks: An introduction.* Oxford University Press.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, *66*(1), 016128.

OMS. (2018). *Organización mundial de salud.* Descargado de http://www.who.int/es

OPS. (2018). *Organización panamerica de la salud.* Descargado de https://www.paho.org/hq/index.php?lang=es

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., y Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925.

*Research on social networks: Imdb networks.* (2010). Descargado de http://www-levich.engr.ccny.cuny.edu/webpage/hmakse/software-and-data/

Rodrigues, F. A. (2019). Network centrality: An introduction. En *A mathematical modeling approach from nonlinear dynamics to complex systems* (pp. 177–196). Springer.

Rousseau, C., Saint-Aubin, Y., Antaya, H., Ascah-Coallier, I., y Hamilton, C. (2008). *Mathematics and technology.* Springer.

Seidman, S. B. (1983). Network structure and minimum degree. *Social networks*, *5*(3), 269–287.

Wearing, H. J., Rohani, P., y Keeling, M. J. (2005). Appropriate models for the management of infectious diseases. *PLoS Medicine*, *2*(7), e174.

Weng, J., Lim, E.-P., Jiang, J., y He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. En *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270).

Zhang, J.-X., Chen, D.-B., Dong, Q., y Zhao, Z.-D. (2016). Identifying a set of influential

spreaders in complex networks. *Scientific reports*, *6*, 27823.