Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Simulación de procesos de difusión de infecciones en grafos, y sus múltiples capas de análisis

Tesis de Licenciatura en Ciencias de la Computación

Carolina Lang

Director: Esteban Feuerstein
Codirector: Carlos Sarraute
Buenos Aires, 2022

Simulación de procesos de difusión de infecciones en grafos, y sus múltiples capas de análisis

En esta tesis se estudia el problema de simular la difusión de un agente en una población. Nos enfocamos principalmente en la difusión de infecciones biológicas entre personas. Este problema es complejo y hay varias capas de información involucradas en su simulación. En la tesis realizamos un recorrido teórico por algunas de ellas: los modelos epidemiológicos estándar para modelar procesos de difusión, la simulación del movimiento de personas en el espacio, y la creación sintética de grafos sociales. Luego implementamos un prototipo de simulador, incorporando varias ideas a nivel de implementación para apuntar a construir un software declarativo que sea fácil de mantener. También procesamos una base de datos de viajes en bicicleta de dominio público de los datos abiertos de la Ciudad de Buenos Aires para usar como datos de entrada para un modelo gravitacional de dos puntos para modelar movilidad humana. Finalmente realizamos varios experimentos sobre distintas partes del problema tratado: comparamos experimentalmente dos modelos de generación de grafos sociales, implementamos políticas complejas en dos módulos distintos del simulador y comparamos dos modelos de movilidad humana tomando como datos de entrada la base de viajes en bicicleta.

Palabras claves: Procesos de difusión, Grafos sociales, Epidemiología, Simulación, Movilidad humana.

SIMULATION OF SPREAD PROCESSES IN GRAPHS, AND ITS MULTIPLE LAYERS OF ANALYSIS

This thesis studies the problem of simulating a spreading process among a population. We mainly focused on biological infection spreading among people. This is a complex problem that has many layers of information that must be taken into account in its simulation. In this thesis we performed a survey of some of them: standard epidemics modelling techniques, the simulation of human mobility, and synthetic generation of social graphs. Then we developed a simulator prototype, and we intended to incorporate some ideas at the implementation level to build a declarative and easily mantainable software. We also processed a database of bycicle trips from the public data repository of the City of Buenos Aires as input for a two-point gravitational model for human mobility. Finally, we ran some experiments about different parts of the trated problem: we compared two models to generate social graphs, we implemented complex policies in two different models of the simulator, and we compared two human mobility models, taking the bycicle trips database as their input.

Palabras claves: Spread processes, Social graphs, Epidemiology, Simulation, Human mobility.

AGRADECIMIENTOS

Tengo mucho para agradecer, y además no tengo la habilidad de escribir textos cortos.

Esta tesis surgió bastante antes de su comienzo oficial (a principios de 2021). Mi primer agradecimiento es para mi director Charles, que junto con toda la gente del equipo de Labs de Grandata fueron mi primer contacto con el mundo científico. Gracias a esa experiencia aprendí muchas cosas, en particular de Chagas y de trabajo interdisciplinario. Hace muchos años que quiero que Charles me dirija una tesis y me pone muy feliz que, aunque cambiaron las circunstancias y las situaciones laborales, finalmente haya pasado. Es un privilegio haber podido hacer ciencia tan temprano en la carrera y lo valoro mucho.

También gracias a Esteban, mi otro director, que se involucró muchísimo y desde el principio. Le presentamos con Charles una idea poco ortodoxa, y el aporte de Esteban fue clave para que el trabajo tuviera forma de tesis de licenciatura. Me dio el espacio para que pudiera plantear todo tipo de dudas e ideas y me sentí en mucha confianza para trabajar. Gracias por acompañarme y orientarme, no me estaría recibiendo si no fuera por eso.

Les quería agradecer a Viviana y Rodrigo por aceptar ser mis jurados y por la predisposición para definir la fecha en medio de fin de año, fin de cuatrimestre y el mundial. Además, gracias a Rodrigo por los comentarios en la presentación de resultados intermedios que le hicimos.

Transité muchos roles en la facultad y todos son una parte fundamental de mi formación universitaria como estudiante, docente y militante. Les que dicen que "a la facultad se va solamente a estudiar", además de perjudicar al sistema, se pierden personalmente de aprender mucho. En particular le agradezco a todas las generaciones de militantes del FEM! y de activistas del CECEN y la ComCom en general con las que defendemos y repensamos la universidad pública; y a mis estudiantes y colegas docentes con quienes comparto una actividad que me apasiona. En esta nueva etapa las cosas pueden cambiar de forma pero las convicciones que me llevo son para toda la vida.

También quiero agradecer a mis socies en Eryx por el aguante y porque decidimos entre todes fomentar que la gente pueda estudiar, y no es algo común en un lugar de trabajo.

Por último me quiero agradecer a mi familia y a mis amigues, que me sostuvieron en todo el proceso de la tesis (y los altibajos de la carrera en general). Cerrar una etapa tan central de la vida lleva mucho trabajo en varios aspectos y tener una buena red de afectos hace la diferencia.

Índice general

1. INTRODUCCIÓN

El tema de esta tesis es el estudio de distintos aspectos que se ponen en juego a la hora de analizar un proceso de difusión de un agente en una población. Los procesos de difusión son de interés en distintas áreas, por ejemplo la epidemiología en casos de infecciones de distinto tipo en poblaciones de seres vivos, o del estudio de la transmisión de ideas, información u opiniones entre poblaciones de personas. En particular, se hizo foco en la transmisión de patógenos en personas. Este problema es complejo, en tanto tiene varias capas de información superpuestas que hay que entender para poder proponer modelos que las tengan en cuenta.

Este trabajo se da en una época particular de auge del *big data*, o el análisis de grandes volúmenes de datos, que aporta una gran cantidad de materia prima para incorporar información a las distintas capas de los modelos de estos procesos complejos. Estos datos son de diverso tipo: ubicación geográfica de las personas, relaciones o intensidad del contacto entre ellas, patrones de consumo cultural e intereses, o nivel socioeconómico, por dar algunos ejemplos. El aprovechamiento de estos datos, su organización y la extracción de información de ellos (el minado de datos o *data mining*) es un campo de estudio en sí mismo. En particular, en la tesis buscamos incorporar datos reales en la capa de movilidad del modelado. Cada vez más, se busca complementar las encuestas de movilidad con otras fuentes de datos que abarquen a una mayor parte de la población [Sar+17; Wes+15; LBH12]. En este trabajo se busca estudiar cómo se pueden inferir patrones de movilidad de personas a partir de sus datos *geolocalizados* y *ubicados temporalmente*[1] (en este caso, de viajes en bicicleta en la Ciudad de Buenos Aires, Argentina).

La primera parte de la tesis es un estudio bibliográfico de algunos de los aspectos que entran en juego a la hora de modelar el problema de difusión de infecciones en personas: la simulación de procesos de infección en general (modelo **SI** y derivados), la descripción y predicción de la movilidad humana, y la caracterización y generación sintética de grafos sociales. En cada parte del resumen une lectore familiarizade con algún tema podría saltear el capítulo correspondiente. En caso contrario, estos capítulos sirven como introducciones.

En una segunda parte se realizaron dos desarrollos: un simulador donde los distintos aspectos del problema fueron modelados como módulos configurables, para generar un software declarativo y extensible, y el procesamiento de un dataset real a partir del cual se infirió un proceso de movilidad que fue utilizado como insumo para realizar una simulación.

Luego, en la tercera parte se realizaron experimentos sobre los distintos aspectos tratados en la revisión bibliográfica, y casos de uso del simulador construido para poner en juego la expresividad de sus distintos aspectos.

Finalmente se sacaron conclusiones en varios ejes interesantes sobre los diferentes temas que se tocaron en la tesis.

En cuanto a la revisión bibliográfica, primero tratamos el tema del modelado y la simulación de procesos de difusión. Estudiamos distintos modelos de procesos de difusión derivados del modelo **SI** y presentamos las hipótesis de *compartimentarización* y *mezcla homogénea*, que funcionan como base para estos modelos más simples. Luego, a partir de distintos estudios que, de alguna manera u otra, rompen con alguna o ambas de ellas, ha-

[1] Un dato geolocalizado es aquel que tiene una o más ubicaciones geográficas asociadas, por ejemplo, pares latitud-longitud. Ubicado teporalmente es que tiene asociado uno o más momentos en el tiempo.

blamos de distintas líneas de desarrollo hacia modelos más complejos posteriores, también a partir de la incorporación de la simulación por computadora como herramienta.

Por otro lado estudiamos el modelado de la movilidad espacio-temporal de las personas como un proceso discreto, tanto para eventos particulares de algunas horas de duración como para procesos más largos. Partimos de modelos que consideran la movilidad humana como un proceso estocástico aleatorio, y luego desarrollamos la idea de la *regularidad del movimiento* de las personas, y de la importancia de los *puntos de interés* como caracterización de sus trayectorias. En este sentido, los movimientos de personas son periódicos (volviendo a los mismos lugares cada un día o una semana en la mayoría de los casos). Describimos con más detalle distintos casos de uso de la idea de puntos de interés, y en particular hablamos del modelo gravitacional de dos puntos usado en la tesis, que caracteriza la movilidad de una persona en función de dos puntos de interés: la ubicación de su casa y la de su trabajo.

Concluimos en este aspecto que, por un lado, existe un recorrido histórico desde modelos más simples con fórmulas matemáticas cerradas hacia otros más complejos que aprovechan el poder de procesamiento de las computadoras para simularlos explícitamente. Igualmente, distintos modelos resultan adecuados para distintos tipos de situaciones, en función de cuales son la duración y la escala en el espacio del fenómeno que se quiera modelar en general. Los modelos de caminante aleatorio sirven para procesos más acotados en el tiempo pero, para describir la movilidad humana a largo plazo, es mucho más efectivo trabajar sobre puntos de interés para calcular su trayectoria. Además, existen distintos casos de uso que requieren aplicar esta idea de diferentes formas.

Estudiamos la capa social del problema de difusión de infecciones analizando distintos modelos para expresar la topología de una red social, y, por lo tanto, también para generar redes sociales sintéticamente. Históricamente se introduce la idea de grafo generado aleatoriamente con el modelo de Erdös-Rényi (ER), como alternativa a modelos anteriores que utilizaban procesos determinísticos para crearlos (por ejemplo, ubicando a los nodos en una cuadrícula de casilleros y conectándolos con sus vecinos verticales y horizontales). Luego, el modelo de ER fue discutido al observar que la distribución de grados en redes sociales era distinta a la de un grafo de ER, entre otras diferencias. En particular se encontró que determinados grafos sociales estudiados tienen una distribución de nodos que cumple una *ley de potencia*, por lo que se introducen distintas características de estas distribuciones en función del valor de su parámetro γ. En particular presentamos el método de *preferential attachment*, que fue usado como base para uno de los algoritmos de generación de grafos sintéticos usados en la tesis. Este procedimiento genera una clase particular de grafo con una distribución de ley de potencia con $\gamma = 3$. Además, presenta la idea innovadora de hacer crecer un grafo orgánicamente en vez de partir de un conjunto fijo de nodos y unirlos con aristas de forma aleatoria.

Como parte de la tesis realizamos una comparación entre dos simulaciones similares, una con vínculos entre personas dados por un grafo de Erdös-Rényi y otra, por uno generado con *preferential attachment*, que se incluye en la sección de experimentación, y donde se comprueba experimentalmente que simular un proceso en un grafo generado con preferential attachment acelera la difusión, respecto de realizar el mismo experimento sobre un grafo generado usando Erdös-Rényi, aunque sólo se evidencia esta propiedad en grafos ralos (con relativamente pocas aristas). Esto es coherente con la redefinición que se da de la *propiedad del mundo pequeño* en los grafos independientes de escala respecto de los de Erdös-Rényi.

Concluimos que también existe en este caso un desarrollo histórico de distintos modelos, y las ideas innovadoras que rescatamos en este sentido son dos en especial. La primera es el cambio de paradigma de usar grafos definidos de forma determinística (por ejemplo cuadrículas bidimensionales) a no-determinística (por ejemplo el modelo de Erdös-Rényi) para representar grafos sociales. La segunda es la idea de *crecimiento* de un grafo, que expande las redes generadas mediante la adición de nodos en lugar de comenzar con una cantidad de nodos fija y definir sólo las aristas (y en particular la estrategia adoptada en *preferential attachment*, que además incorpora la idea de priorizar conexiones con nodos que ya son puntos nodales). Estas ideas van en la dirección de simular el proceso de creación de redes sociales orgánicamente.

En la parte de implementación, describimos la estructura del simulador desarrollado y sus distintas partes interesantes, entendiendo el desarrollo de un programa de software como un *modelo computable de un dominio de la realidad*. En este sentido presentamos los *módulos* que se implementaron espejando cada capa del problema: la capa social, la de movilidad, la de propagación de la infección y la de su desarrollo. Estos módulos no necesariamente son suficientes para representar cualquier proceso, pero la idea de módulo se puede extender a otros aspectos de la simulación que se quieran controlar u observar.

También se desarrolló la idea de *condiciones*, que aparece en distintos contextos, y que sirve para describir el comportamiento del modelo de forma explícita. Además, se resumen algunas decisiones de diseño que se consideraron no triviales, o tienen un marco teórico detrás: el uso de un *lenguaje ubicuo* entre el mundo del dominio de problema y el del desarrollo de software, la elección de una *estrategia* o *cosmovisión* de simulación y la *ubicación de la información* de cada nodo-persona en el software.

Mostramos, como casos de uso del software, el modelado de situaciones complejas expresadas usando el sistema de condiciones: en el primer caso se describió una forma compleja de propagación de la infección, donde parte de la población usa un método que reduce el contagio (uso de tapabocas). En el segundo, la complejidad se encontró en el desarrollo de la infección una vez que un nodo la contrae, ya que hay parte de los nodos que están inmunizados y cursan una versión más leve de la infección. En estos casos comprobamos que, al cambiar la proporción de nodos que adoptan cada medida, cambia la forma de las curvas características, con lo cual se evidencia el efecto de las medidas. El uso de tapabocas retrasa el pico del proceso, y en algunos casos también baja la proporción de nodos contagiados. El plan de vacunación puede bajar la altura del pico cuando el nivel de contagio de base es bajo, pero si es alto, no afecta el contagio sino que solamente acelera la recuperación de la proporción de nodos vacunada.

Desarrollamos conclusiones sobre la implementación del modelo. En general el desarrollo de un lenguaje ubicuo y el objetivo de mejorar la declaratividad del modelo nos parecen interesantes, si bien esto genera que la escalabilidad del prototipo sea menor. Este tipo de exploraciones pueden usarse para analizar situaciones complejas en una primera etapa de una investigación para pasar luego a una segunda etapa donde se piense en la escalabilidad. En particular el sistema de condiciones permite expresar explícitamente, mediante composición, condiciones arbitrariamente complejas sin necesidad de realizar traducciones entre el diseño del experimento y su programación. Existe en este aspecto un balance entre la declaratividad y la velocidad de corrida de una solución, por ejemplo, en que el prototipo desarrollado requiere computar cada condición cada vez que se ejecuta, lo cual es ineficiente, pero permite condiciones que cambien dinámicamente a medida que las condiciones de la simulación cambian. La cosmovisión elegida (basada en *escaneo de actividades*) también

va en este sentido, ya que es simple pero poco eficiente. Una cuestión pendiente para explorar es la incorporación de condiciones que cambian en el transcurso de la simulación (por ejemplo, un plan de vacunación que empieza cuando se registra cierta proporción de casos), ya que todos los ejemplos que se programaron son estáticos.

Se puso énfasis en lograr un nivel de acople bajo entre los componentes. En este sentido, cada módulo tiene separada la definición de su comportamiento de la del estado que necesita cada nodo-persona en el aspecto que el módulo representa. También se cuidó que las definiciones de formatos de salida de los experimentos se definieran con claridad en un sólo lugar en el código, para que el prototipo se pueda conectar fácilmente a distintas herramientas de visualización.

El segundo desarrollo fue de un modelo de movilidad humana a partir de una base de datos reales, de dominio público, publicados por el gobierno de la Ciudad de Buenos Aires (GCBA). Esta base contiene viajes registrados en un sistema de préstamo/alquiler de bicicletas que el gobierno licita a una empresa. En él damos una descripción de la base, que consiste en una serie de viajes geolocalizados y ubicados temporalmente. Luego explicamos cómo, a partir de estos datos, se infirieron la casa y el trabajo de les usuaries para incorporar en un modelo gravitacional de dos puntos, y correr una simulación de un proceso de difusión de una infección sobre él. Con estos datos realizamos un experimento de comparación de la dinámica del proceso gravitacional de dos puntos respecto de uno similar de un punto (donde sólo se considera la casa de cada persona como punto de interés). En este sentido pudimos observar que la movilidad de las personas entre sus dos puntos de interés es necesaria para que la infección alcance a toda la población.

De todas formas, hay que recordar que el dataset utilizado tiene limitaciones, ya que la proporción de personas que viajan en bicicleta como medio de transporte para trabajar únicamente es baja, y no se pueden observar viajes en otros medios de transporte e incluso otros en los que algunos trayectos son en bicicleta y otros no. Además, se consideraron ciertas simplificaciones (por ejemplo que los viajes de todas las personas ocurren a la misma hora) que quitan precisión al modelo. Nos encontramos con la dificultad de la falta de datos disponibles en el dominio público en este sentido.

1.1. Estructura de la tesis

A continuación se describe la estructura de la tesis.

En el capítulo 2 se trata el tema del modelado y la simulación de procesos de difusión. Se introducen el modelo **SI** y derivados, las hipótesis de *compartimentarización* y *mezcla homogénea*, y luego se plantean algunas líneas de desarrollo que complejizan dichas hipótesis.

En el capítulo 3 se presentan los distintos modelos de generación aleatoria de grafos (movimientos de caminante aleatorio con distintas distribuciones de saltos), y luego se introducen las nociones de *regularidad del movimiento* y de *puntos de interés*, y, con ellas, la idea de modelos gravitacionales.

En el capítulo 4 se encuentra el resumen del estudio sobre la capa social. Se presentan el modelo de Erdös-Rényi, y luego se da el análisis de la distribución de grados que muestra la necesidad de representar puntos nodales (*hubs*). Luego se presentan la idea de *ley de potencia* y, en particular, el método de *preferential attachment*.

En el capítulo 5 se describe la estructura del simulador desarrollado y sus distintas partes interesantes (módulos y el sistema de condiciones). También se dan tres discusiones

teóricas: las ideas de *lenguaje ubicuo* entre el mundo del dominio de problema y el del desarrollo, la elección de una *estrategia* o *cosmovisión* de simulación y la *ubicación de la información* de cada nodo-persona en el software.

El capítulo 6 describe la metodología usada para inferir la casa y el trabajo de cada usuarie de EcoBici, y la base de datos utilizada.

En el capítulo 7 se presentan todos los experimentos realizados: la comparación entre Erdös-Rényi y preferential attachment en cuanto a velocidad de los procesos de difusión, las simulaciones con distinta proporción de nodos implementando el uso de barbijo y la vacunación, y la comparación entre los modelos gravitacionales de un punto y de dos.

El capítulo 8 contiene conclusiones sobre los distintos temas abordados en el transcurso de la tesis, tanto del recorrido bibliográfico, como de diseño del prototipo de simulador y de los experimentos realizados.

2. MARCO DEL PROBLEMA DE DIFUSIÓN DE INFECCIONES

La difusión de agentes en redes es un fenómeno de interés en distintas áreas. En epidemiología es importante tener herramientas para entender cómo se difunden los vectores de distintas enfermedades, pero también, usando modelos similares, se puede modelar la difusión de información, o, en general, de todo lo que se pueda transmitir de una entidad a otra por algún medio[1]. Para infecciones en seres vivos también se usa el término "enfermedad comunicable"[2]. En adelante se va a hablar de transmisión de *infecciones* pero teniendo en cuenta la salvedad de que los mismos mecanismos son aplicables al resto de las áreas mencionadas.

El capítulo presenta un modelo básico usado en modelado matemático de difusión de infecciones (modelo **SI**) junto con variantes comunes. Estos modelos son simples, en el sentido de que no tienen en cuenta, por ejemplo, la distribución espacial de las personas ni las relaciones entre ellas, que pueden hacer que estén en contacto en mayor o menor medida. Tampoco tienen en cuenta la heterogeneidad de las personas en la población, que puede hacer que se comporten de diferente manera, ni la posibilidad de incorporar cambios en el proceso de difusión de forma dinámica. Existen alternativas para incorporar estos y otros elementos, y algunas de ellas se discuten también en este capítulo. La estructura es la siguiente:

La sección 2.1 está principalmente basada en partes de [BP16] y de [Bai+75] y es una introducción a algunos conceptos importantes en la teoría del mo-delo de difusión de infecciones más utilizado en los estudios matemáticos sobre epidemiología. En particular plantea el modelo **SI** y derivados y las hipótesis de compartimentarización y de mezcla homogénea, y explica la métrica R_0 o *número de reproducción básico*. En estos modelos, existen transmisiones entre cada persona infectada y otra(s) susceptible(s) de serlo con cierta probabilidad, que resume los distintos factores que influyen en transmisión. Por ejemplo, una gripe que se transmite entre personas que se encuentran cerca físicamente, o un virus informático se transmite entre dispositivos entre los cuales existió una comunicación.

La sección 2.2 plantea que en poblaciones que no respetan las hipótesis del modelo **SI**, sobre todo la hipótesis de mezcla homogénea, es más difícil obtener una solución analítica de un proceso de difusión. Por ejemplo, distintos estudios que incorporan la distribución espacial de las personas, o que las enriquecen con metadatos que cambian su comportamiento. Por lo tanto, presenta la simulación de estos procesos como una herramienta valiosa para atacar estos problemas con una perspectiva distinta. En particular, para que un modelo considere la estructura social o espacial, se modela a cada individuo de la población como un nodo en un grafo, que se conecta con otros a través de diferentes tipos de vínculos (aristas del grafo), a través de los cuales ocurren las transmisiones de la infección con cierta probabilidad.

2.1. Teoría del modelado de epidemias

La teoría que se presenta en esta sección está basada fundamentalmente en algunas secciones del el capítulo 10 del libro *Network Science* de Barabási ([BP16]), que da un

[1] Por ejemplo, en [PV01] se trabaja sobre la difusión de virus informáticos.
[2] Ver capítulo 3 de [Bai+75]

marco teórico del modelado de epidemias en general, para luego enfocarse en el problema de modelado de epidemias sobre grafos.

En particular en esta sección se presentan herramientas analíticas que no tienen que ver con grafos. En este sentido, se trabaja sobre dos hipótesis sobre el proceso de infección, con el objetivo de simplificar el análisis (por ejemplo ignorando distintas cargas virales o comorbilidades) y en particular para ignorar la topología del grafo que subyace a la epidemia particular. Estas hipótesis son:

1. **Compartimentarización:** se asume que una persona puede estar en uno (y sólo uno) de varios estados discretos. Los estados **S**usceptible (quien todavía no contrajo la infección) e **I**nfectade (quien ya la contrajo y tiene la capacidad de contagiar a otres) aparecen en todos los casos, y también se pueden agregar otros según las características de la infección particular (por ejemplo **R**emovide para quien ya no puede contagiar la infección, ya sea por fallecimiento, aislamiento o recuperación).

2. **Mezcla homogénea:** Se asume que cada persona tiene la misma probabilidad de encontrarse con una persona infectada (en particular siendo susceptible). Esta hipótesis es la que hace que se ignore la topología del grafo social y permite calcular la evolución del proceso únicamente a partir de las proporciones de personas en cada estado y en cada momento del tiempo.

Modelos epidemiológicos simples

En esta sección se presentan tres modelos epidemiológicos simples presentes en la literatura[3].

En el modelo **SI**, el más simple de todos, inicialmente todas las personas son **S**usceptibles al contagio, y cuando en un paso del proceso interactúan con una persona que tiene el virus, se **I**nfectan también, con cierta probabilidad.

También puede pasar que una persona infectada tenga cierta probabilidad de dejar de tener el virus y volver a ser susceptible (el modelo anterior sería el caso donde esta probabilidad es cero). Este modelo se denomina **SIS**.

Una complejización razonable es considerar que las personas que tienen una infección y se curan desarrollan anticuerpos (al menos por un tiempo), con lo que pasan a **R**ecuperarse, o removerse según la bibliografía (se puede identificar por separado a los **D**ecesos, o muertes). Este es un estado del cual no se es susceptible a la reinfección. En este caso el proceso termina cuando ya no hay gente en **I**, y la proporción de personas que quedaron en cada uno de los estados posibles (**S,R,D**) se pueden usar para caracterizar el proceso.

También existen otros estados posibles, como **L**atente (cuando una persona contrae la infección pero todavía no puede replicarla), o inmune cuando la persona inicialmente no es susceptible (por vacunaciones o por otras características). Por ejemplo, en diversos estudios sobre COVID-19 se estudió la división de la población entre distintos grupos etarios con distintas probabilidades de convertirse en pacientes hospitalarios, y se aplicaron probabilidades acordes al tipo de nodo del que se trataba, por ejemplo en un estudio ([Pai+21]) sobre la población de la localidad de Oro Verde, en Entre Ríos, Argentina. En

[3] En el capítulo 5 de [Bai+75] se trabaja sobre el proceso SI, en particular en la sección 5.3 se define como proceso estocástico con una variable aleatoria que indica la cantidad de susceptibles y otra de infectades, en un momento del tiempo. En la sección 10.2 de [BP16] se caracterizan los procesos SI, SIS y SIR. El paper [KM27] es uno de los trabajos fundacionales del modelo SIR.

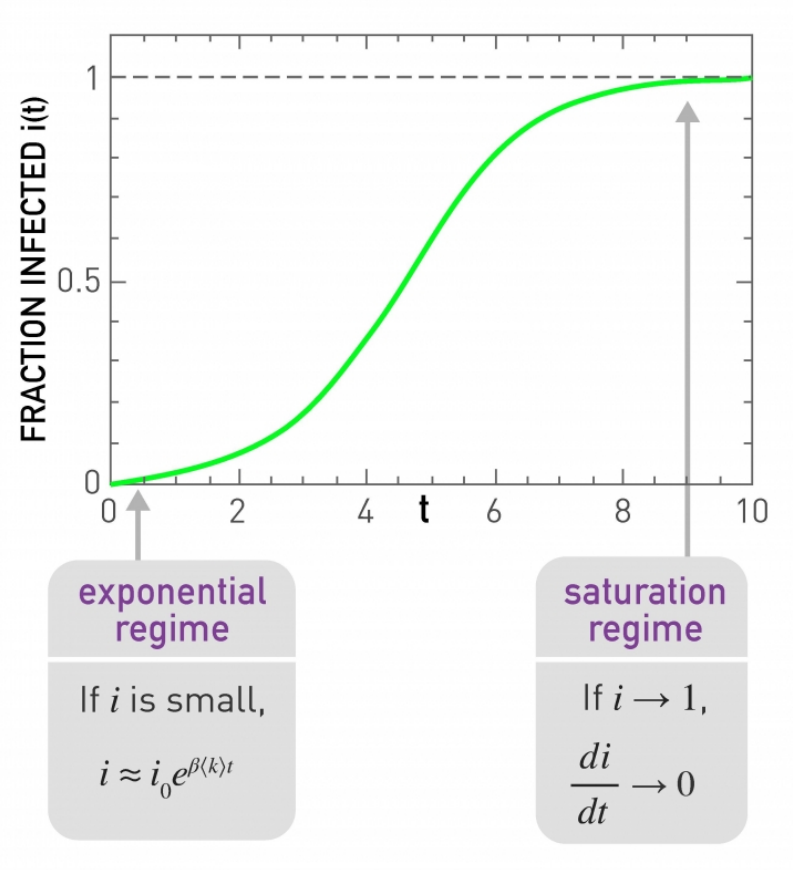


Fig. 2.1: Curva característica para el modelo SI. Se muestra la proporción de infectades (traducción del nombre del eje Y) a lo largo del tiempo. Los recuadros se traducen como: "régimen exponencial", "si i es pequeño" y "régimen de saturación". Fuente: [BP16] sección 10.2.

este estudio también se tuve en cuenta el estrés sobre el sistema de salud, y se hizo una diferenciación entre pacientes sintomátiques (**I**) y asintomátiques (**A**), también llamados portadores[4] en la bibliografía.

Curvas características y posibles formas de convergencia

En esta sección se presentan las formas de convergencia de los tres modelos más simples de difusión, siguiendo la hipótesis de mezcla homogénea. Se presentan también cuatro figuras de análisis, extraídas de la sección 10.2 de [BP16], donde se muestran gráficamente las curvas características para cada uno de los modelos anteriores. La figura 2.4 muestra las curvas características para el estado **I** en los tres modelos.

Modelo SI: En este caso, la curva característica de la proporción de infecciones a lo largo del tiempo se llama *curva logística*, caracterizada por un crecimiento exponencial inicial (porque casi todas las personas con las que alguien en **I** interactúa

[4] Ver capítulo 10 de [Bai+75].

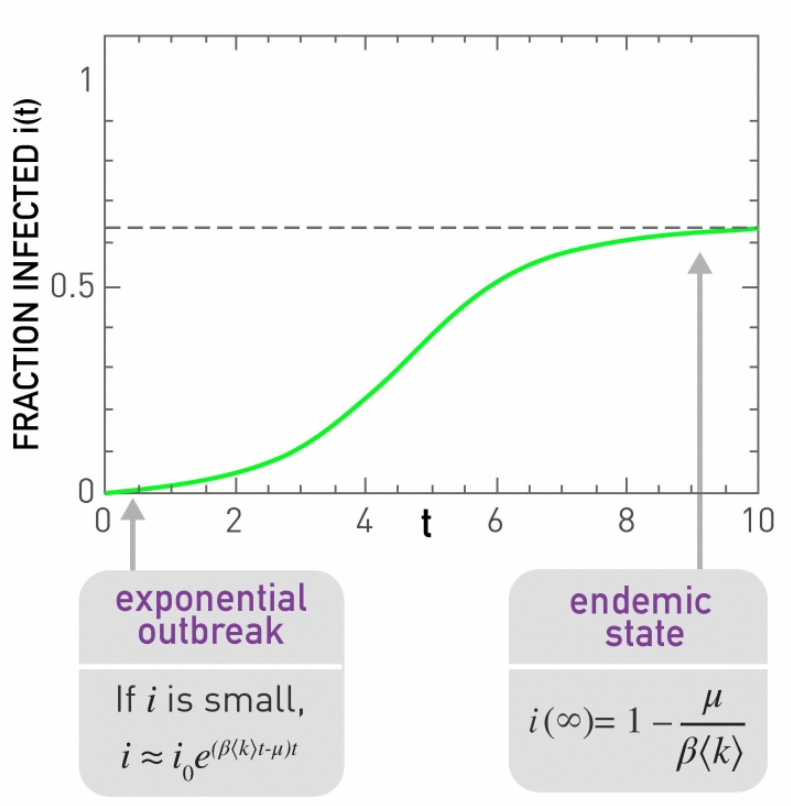


Fig. 2.2: Curva característica para el modelo SIS. Se muestra la proporción de infectades (traducción del nombre del eje Y) a lo largo del tiempo. Los recuadros se traducen como: "brote exponencial", "si i es pequeño" y "estado endémico". Fuente: [BP16] sección 10.2.

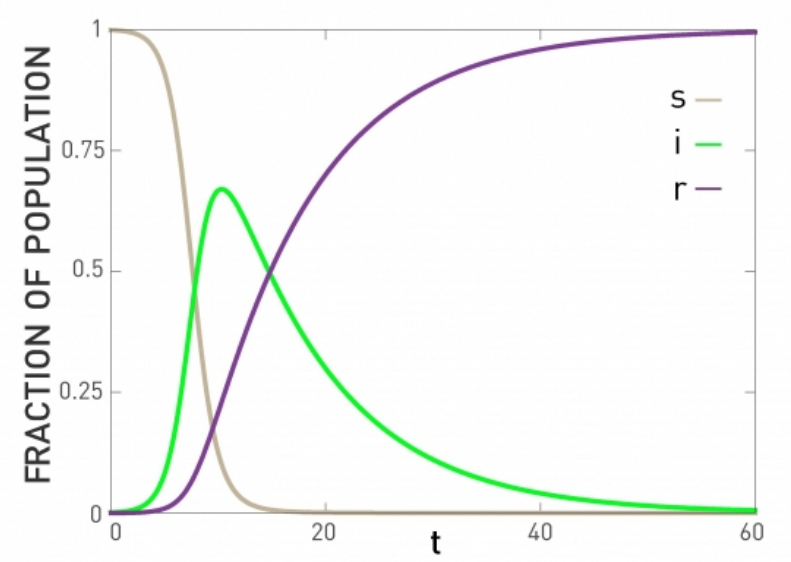


Fig. 2.3: Curvas características para el modelo SIR. Se muestran la proporción de susceptibles, infectades y removides a lo largo del tiempo. La traducción del nombre del eje Y es "fracción de la población". Fuente: [BP16] sección 10.2.

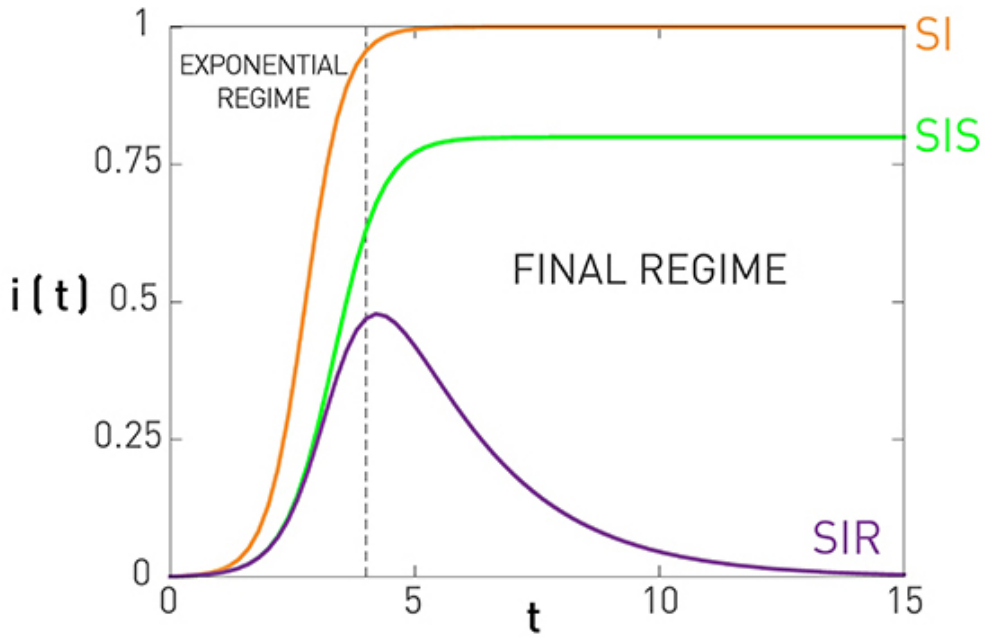


Fig. 2.4: Combinación en un mismo gráfico de las curvas características del estado **I** para los modelos **SI**, **SIS** y **SIR**. Las traducciones de las leyendas de izquierda a derecha son "régimen exponencial" y "régimen final". Fuente: sección 10.2 de [BP16].

están en **S**), seguido de una fase de saturación cuando quedan pocas personas en **S**. El proceso termina cuando el total de la población está en **I**. Ver figura 2.1.

Modelo SIS: Tiene la misma forma que el caso anterior, excepto que no tiende hacia el punto en el que toda la población está infectada, ya que las personas tienen la capacidad de recuperarse y volver a **S**. En este caso el sistema tiende al punto fijo en el que la tasa de contagio se equipara a la tasa de recuperación. En este punto se dice que el modelo llega a un estado endémico. Hay cierta probabilidad, si la tasa de recuperación es alta y la de infección baja, de que la infección se extinga también. Esto tiene que ver con cuál es el número de reproducción básico R_0 de la infección, que se detalla en la sección 2.1. Ver figura 2.2.

Modelo SIR: La infección siempre se extingue, ya que cada nodo sólo se puede infectar a lo sumo una vez, y todas las personas se remueven del sistema y mantienen inmunidad. En una etapa inicial, el crecimiento de la cantidad de infectades es exponencial (como en los otros casos), pero posteriormente alcanza un máximo local (un pico) para luego caer a cero. Ver figura 2.3.

Número de reproducción básico R_0

Es de interés intentar predecir la tendencia de una curva de infecciones, y para esto existen métricas para la rapidez con la que un patógeno se puede propagar, y cuánto va a persistir en la población. Una de las más usadas es el llamado *número de reproducción básico* o R_0, que es una medida de la cantidad de infecciones que cada persona infectada provoca.

En los procesos simples definidos anteriormente, se define como la esperanza de la cantidad de infecciones que provoca una persona en una población, bajo la hipótesis de que todos sus contactos son susceptibles[5]. Es decir,

$$R_0 = \frac{\beta\langle k\rangle}{\mu}$$

donde $\langle \mathrm{k} \rangle$ es la cantidad de contactos que tiene cada persona en promedio, β es la proporción de ellos que resultan en una propagación, y μ es la tasa de recuperación.

En estos modelos este número depende solamente de parámetros del agente que se transmite (es decir, de la probabilidad de contagio y del tiempo estimado de recuperación). En un caso real, en cambio, el cálculo puede ser más complejo y podría depender de condiciones variables en la simulación que afecten el proceso. Por ejemplo, es esperable observar un cambio en el valor de R_0 cuando se introduce en el sistema una medida para bajar la cantidad de contactos (reducir $\langle \mathrm{k} \rangle$) o prevenir la propagación con un método de barrera durante ellos (reducir β).

Por esta razón es una métrica útil para medir la efectividad de distintas políticas de salud pública que influyan sobre el modelo, comparando el efecto que tiene sobre el R_0 la aplicación de la medida vs. la misma situación pero sin ella. El número de reproducción aporta información sobre la velocidad de crecimiento de la población infectada en el modelo, y podemos distinguir tres situaciones en función de su valor:

- Si $R_0 > 1$, quiere decir que la cantidad de infectades tiende a crecer, y mientras más grande sea este valor, más rápido será este crecimiento.
- Si $R_0 = 1$ se llegó a un punto de estabilidad en la que la cantidad de infectades se mantiene constante.
- Por último, si $R_0 < 1$, la cantidad de infectades tiende a decrecer (mientras más pequeño, más rápido es este decrecimiento).

En la mayoría de los modelos (así como en la realidad), el R_0 no es constante a lo largo de una epidemia. Por ejemplo, en modelos más complejos como el SIR, donde la cantidad de nodos susceptibles se hace más pequeña con el tiempo, no tiene sentido mantener la hipótesis de que todos los contactos de las personas son con otras susceptibles. Como la cantidad de contactos con susceptibles baja a medida que hay más parte de la población en **R**, también tiende a hacerlo R_0.

2.2. Más allá de las hipótesis de los modelos de difusión

Las hipótesis presentadas en la sección 2.1 (compartimentarización y mezcla homogénea) son útiles para hacer una simplificación que permita el tratamiento analítico de un proceso de difusión. Sin embargo, resultan limitantes en el estudio de casos complejos sobre poblaciones grandes.

La hipótesis de compartimentarización plantea que los estados en los que se encuentran las personas son categorías internamente homogéneas, y esto es una simplificación. Por ejemplo, en el caso de infecciones que provocan enfermedades, se puede tener diferente carga viral de una infección, y esta condición se traduce en que no todas las personas propaguen

[5] Referencia: capítulo 10.2 de [BP16].

la infección de la misma manera. También, cierta proporción de la población puede tomar distintas medidas de cuidado para evitar contagios que modifiquen la probabilidad de que ocurra una transmisión (uso de barbijo para transmisiones aéreas, o de métodos de barrera para ITS[6]). Por ejemplo en [Pai+21], un estudio sobre COVID-19, se da una segmentación de la población en función de su edad y se simula un comportamiento distinto para distintos grupos etarios por separado (algunos grupos etarios realizan aislamiento y otros no). En los experimentos de las secciones 7.2 y 7.3 se muestran dos formas de incorporar heterogeneidad entre los nodos. En el primer caso se distingue entre personas que usan tapabocas vs. personas que no, y en el segundo, entre personas inmunizadas y no inmunizadas.

La hipótesis de mezcla homogénea, por otro lado, ignora la naturaleza heterogénea de las conexiones entre las personas. Existen varias razones por las cuales ciertos nodos pueden relacionarse de forma distinta con el resto del grafo: por cercanía física, porque los nodos tienen una relación social (de amistad o familiar, por ejemplo), porque comparten un espacio (trabajo, actividad, comunidad educativa), o por una combinación de estos factores. En un caso real existe una red social y/o espacial que subyace al proceso de difusión, que determina cuáles nodos en particular tienen más probabilidad de infectarse en un determinado momento. En particular sobre la distribución espacial, en la sección 2.6 de [Bai+75] se discuten líneas de investigación en modelado de epidemias, y en particular se mencionan trabajos soviéticos, a partir de [BR67], que analizan brotes de gripe en distintas ciudades distribuidas en la URSS, en la década del '70, y plantean modelos migratorios para explicar el contagio entre ciudades. En [Cac+20] se trabaja sobre el territorio de EEUU y se analiza cómo afecta la componente geográfica a dos olas distintas de COVID-19, en función del estado de origen de cada una de ellas. En [Kra+20], se analiza la correlación entre migraciones registradas desde Wuhan[7] hacia otras provincias de China con la aparición de brotes de COVID-19 en estas provincias. En cuanto a la estructura del grafo subyacente a un proceso en general, en [Fer+05] se estudia cómo mejora la contención de influenza la inmunización selectiva de ciertos nodos particulares en función de su centralidad entre la población.

Históricamente, existe un punto bisagra alrededor de las décadas del '60 y '70 en el que proliferaron los estudios de simulaciones computacionales en el campo de la epidemiología[8]. Hasta ese momento el enfoque principal fue el de encontrar y resolver modelos matemáticos, ya sean las ecuaciones diferenciales de un modelo determinístico o funciones de verosimilitud en un modelo estocástico. El surgimiento de herramientas computacionales suficientes para computar estas simulaciones en un tiempo razonable también favoreció el surgimiento del nuevo paradigma, ya que simular permite expresar condiciones complejas que no necesariamente tienen una resolución matemática e igualmente obtener resultados mediante el cómputo.

En este sentido se tomó la decisión en la tesis de complejizar el proceso de simulación para que el lenguaje en el que se expresen las condiciones del experimento en el prototipo se parezca lo más posible a los términos en los cuales se piensan los experimentos. En la sección 5.1 se hacen consideraciones implementativas para el prototipo desarrollado en particular.

[6] Infecciones de transmisión sexual.

[7] La provincia de Wuhan, China, fue donde se registraron los primeros casos de COVID-19 a nivel mundial.

[8] Ver recorrido histórico en el capítulo 2 de [Bai+75], en particular la sección 2.6.

3. MODELOS DE MOVILIDAD HUMANA

A lo largo de la historia se intentaron modelar los patrones de movimiento espacial de las personas. Existen distintos fenómenos de movilidad y cada uno, en principio, tiene características particulares. No es lo mismo, por ejemplo, el desplazamiento de pocas personas durante un evento particular (como un recital, una feria o un parque), respecto del flujo de personas a lo largo de una ciudad que se acercan al lugar de dicho evento [Pon+16], o de los desplazamientos entre barrios de una ciudad en un día determinado de la semana [Sar+17] o los flujos migratorios a lo largo de años [Mon+16].

En este capítulo se presentan distintos modelos de movilidad. Por un lado en la sección 3.1 se presenta un estudio que intenta explicar el movimiento humano como una trayectoria aleatoria, y que ataca el problema de predecir la distribución de la longitud de los saltos entre observaciones consecutivas de una persona, y otro que agrega la idea de la regularidad del movimiento humano para discutir la hipótesis del primer trabajo. En la sección 3.2 se habla sobre la idea de puntos de interés como predictores del movimiento humano. En la 3.3 se presenta la idea de movimiento gravitacional en función de los puntos de interés y, a grandes rasgos, cuál fue el modelo de dos puntos implementado en la tesis. Esta sección es general, para más detalle sobre la implementación del modelo sobre el dataset trabajado en concreto ver el capítulo 6. En particular las secciones 6.2 y 6.3 habla del uso de los puntos de partida y llegada de viajes en bicicleta para inferir los puntos de interés (*casa* y *trabajo*) usados en la experimentación de la tesis.

3.1. Trayectorias aleatorias y regulares

Un posible modelo nulo es asumir que cada persona es un punto en el plano que se mueve de forma aleatoria de un lugar a otro, sin priorizar algunas direcciones de desplazamiento sobre otras. Este modelo imita otros usados en física para describir movimientos de partículas en un fluido, y en particular describe el movimiento de un agente (en este caso una persona) como un proceso estocástico, definido como una secuencia de posiciones $p_i = (x_i, y_i)$ en instantes de tiempo sucesivos, donde

$$p_{i+1} = p_i + \Delta_i$$

Es más fácil definir la variación $\Delta_i = (r_i, \theta_i)$ en forma polar, de tal manera que $\theta_i \sim U[0, 2\pi)$, y r_i puede definirse de distintas maneras.

La forma más sencilla es fijar $r_i = k$ valor constante a lo largo de la simulación, pero también se podría definir tomando una muestra de cualquier distribución aleatoria continua. Por ejemplo, se podría fijar $r_i \sim \mathcal{N}(\mu, \sigma)$ y θ_i y obtener un caminante aleatorio con una distribución de saltos gaussiana. Estos son algunos intentos simples de describir la secuencia de longitudes de los saltos entre dos pasos sucesivos del proceso como un proceso aleatorio donde los r_i son independientes entre sí.

Una forma bastante estudiada de definir el proceso r_i fue la de *vuelos de Lévy*. La dirección en estos movimientos sigue eligiéndose uniformemente entre el giro completo. El libro *The Physics of Foraging* ([Vis+11]) hace un tratamiento formal de la distribución de Lévy en el capítulo 3. También recopila casos de estudio en distintos tipos de animales donde esta distribución aparece. En particular en el capítulo 7 habla sobre intentos de

aplicarla para explicar movimiento en poblaciones de personas, como algunos trabajos citados en esta tesis.

Un vuelo de Lévy es un proceso de caminante aleatorio en el que r_i sigue una distribución de Lévy[1]. Las características que la hacen de interés son que es una distribución de cola pesada, por lo cual da una proporción mayor de saltos largos entre posiciones consecutivas que en una gaussiana, por ejemplo, y que existe una cantidad considerable de bibliografía que usa esta distribución para explicar movimiento de agentes complejos (animales y personas en contraposición con partículas en suspensión).

La idea de aplicar una distribución de Lévy para explicar movimiento humano se basa en la observación de las trayectorias de animales en busca de comida. En biología se trabaja sobre la hipótesis de que los animales siguen patrones de recolección óptima de alimentos[2], y, si se modela la distribución de saltos como una gaussiana, se evidencia la sub-representación de saltos largos presentada en el párrafo anterior. Entre los intentos de extrapolar esta observación a poblaciones de personas se encuentra, por ejemplo, [BHG06], que sigue la trayectoria de billetes marcados y observa que se desplazan con un patrón más cercano a un vuelo de Lévy.

Otro estudio posterior a éste [GHB08] da una explicación alternativa que pone un matiz sobre esta afirmación. Por un lado este trabajo encuentra en una distribución de Lévy truncada una explicación de la distribución de saltos en los movimientos de las personas. Por otro lado, en este trabajo se muestra que las trayectorias observadas por [BHG06] provienen de una convolución entre dos movimientos: las trayectorias de las personas y el cambio de posesión del billete. En este sentido, la trayectoria resultante representa fracciones de las trayectorias completas de distintas personas, en sucesión. Entonces en [GHB08], al seguir movimientos de teléfonos celulares, que no cambian de dueñe, se muestra que la probabilidad de regresar a un sitio ya visitado es mucho más alta en la trayectoria de una misma persona, con picos de regularidad en múltiplos de 24 horas. En este sentido, determinar una dirección aleatoria para cada uno de los saltos parece ser un modelo insuficiente.

3.2. Regularidad del movimiento y puntos de interés

Los movimientos de personas, según lo discutido en las secciones anteriores de este capítulo, se pueden aproximar por procesos de camino aleatorio si se encuentran en un determinado lapso corto de tiempo (por ejemplo, una visita a un campus o una feria [Rhe+11]). Si, en cambio, se observa a una misma persona moverse todos los días durante un mes, por ejemplo, resulta claro que las trayectorias de cada día no son independientes entre sí. Esta persona hipotética frecuentará siempre los mismos lugares: tendrá un horario de dormir en el que va a su casa, tendrá lugares donde trabaja o estudia, lugares de esparcimiento y otros puntos de interés. Incluso en la sección 3.1 se presentó la aparición de períodos de 24 horas en el movimiento humano, lo que respalda esta hipótesis.

La forma de extraer puntos de interés a partir de una trayectoria observada y sostenida en el tiempo cambia mucho en función de con qué datos se cuente para observar el movimiento de la persona. Por ejemplo, si alguien participa de un estudio en el que se coloca un dispositivo que la geolocaliza a intervalos regulares, o si la base de datos contiene usos de antenas de telefonía celular o de redes WiFi, distinguir cuáles de las observaciones corresponden a puntos de interés efectivamente y cuáles corresponden a lugares de paso es

[1] Las ecuaciones de la distribución de Lévy se pueden ver en la sección 3.1 de [Vis+11].
[2] Ver sección 2.1 de [Vis+11].

un problema no trivial en sí mismo. También, si no se tiene cuidado en las observaciones se puede observar la convolución de varios movimientos, como es el caso de [BHG06]. Si la persona participa activamente de la recolección de datos puede declarar cuáles son los puntos de interés para ayudar a distinguirlos, pero en el análisis de grandes volúmenes de datos, y sobre todo al usar bases que no fueron generadas con un estudio de movilidad en mente, hay que tomar decisiones para definir y caracterizar estos puntos de interés. Estas decisiones, al mismo tiempo, cambian en función de si se observan apariciones puntuales o trayectorias completas. Por ejemplo, se puede inferir que si se observan varias apariciones puntuales en secuencia, en lugares muy cercanos, la persona observada está permaneciendo en un lugar y por lo tanto no está de paso, y este lugar es significativo de alguna forma.

Otro período natural para la repetición del movimiento humano, además del ya presentado de 24 horas, es el de una semana, ya que los días hábiles y no hábiles suelen tener patrones distintos entre ellos, pero que se repiten a través de las semanas. Existen trabajos previos que caracterizan los distintos días y horarios en los que las personas se mueven[3], y se pueden proponer distintos criterios en función de qué se esté buscando. Por ejemplo, si se quiere hacer un estudio sobre lugares de esparcimiento para caracterizar a las personas por sus perfiles de consumo cultural, se podría suponer que estos lugares se frecuentan los fines de semana principalmente, y quizás se deban eliminar los sitios de interés más frecuentes para no considerar casa y trabajo de las personas. Por otro lado, si se quiere encontrar un "día típico" laboral para las personas y usarlo para un plan urbano, se deben tomar días y franjas horarias laborales y tomar muy pocos puntos que sean frecuentes en esos horarios. Incluso se puede trabajar con outliers, y darle semántica a los puntos de interés que surgen los días feriados, por ejemplo, para predecir adónde se refugiará la gente en caso de un desastre natural (como es el caso de un estudio previo que predice dónde acudió la gente de Haití después del terremoto de enero de 2010 en función de puntos de interés etiquetados como vínculos sociales: por ejemplo, dónde habían pasado Navidad y año nuevo del año anterior [LBH12]).

3.3. Modelo gravitacional de movilidad humana

Los puntos de interés sirven, entre otras cosas, para caracterizar las trayectorias frecuentes de las personas. Si una persona es observada en un punto A un día determinado a la mañana, y en otro punto B ese mismo día a la tarde, durante varios días (o semanas) distintos, se puede considerar que en un día (o semana) típico, ocurre un desplazamiento en ese momento desde el punto A hacia el B. Por lo tanto, conociendo los puntos de interés de una persona y sabiendo en qué momento del día (o semana) ocurren frecuentemente, se puede modelar la *trayectoria típica* de esa persona mediante viajes en ciertos horarios entre los puntos de interés en orden. Esta es la idea de la generación de trayectorias sintéticas usando un modelo gravitacional: se considera que los puntos de interés atraen en ciertos horarios a la persona correspondiente y que por lo tanto viaja hacia estos lugares.

Hay varias formas de inferir esta trayectoria. La más simple es decir que en un horario determinado la persona deja de estar en el punto A y se mueve instantáneamente al B. Otra, apenas más compleja, es asumir que en cierto horario cada nodo-persona se mueve

[3] En `https://www.argentina.gob.ar/sites/default/files/enmodo.zip` se pueden encontrar los resultados de la ENMODO 2009-2010 (Encuesta de Movilidad Domiciliaria), donde se caracterizan los horarios pico para los distintos tipos de viaje, en particular los laborales. Además en `https://www.argentina.gob.ar/transporte/dgppse/publicaciones/encuestas` se pueden encontrar varias encuestas de transporte realizadas por el gobierno nacional argentino.

a través de un vector cuya magnitud es una velocidad determinada y su dirección es $\overline{AB}$ (es decir, el vector que va desde el punto A hacia el B). Esta segunda opción es la que se utilizó en la tesis. Se pueden generar algoritmos más sofisticados que se integren, por ejemplo, a la red de calles e incluso al transporte público de la ciudad. Este tipo de análisis quedó fuera del alcance del trabajo[4].

En particular se trabajó con un modelo gravitacional de dos puntos con una semántica particular: cada persona se desplaza entre dos puntos que corresponden con su hogar inferido y su trabajo inferido. Los horarios de estos desplazamientos ocurren durante los horarios pico establecidos por la encuesta de movilidad urbana ENMODO[5]. Se dejaron afuera del análisis todos los otros puntos de interés de cada persona y se consideró que una simulación consiste de una secuencia de días laborales consecutivos (es decir, no se simularon días que siguieran un patrón de movilidad de fin de semana).

Se trabajó con modelos simples (con relativamente pocos puntos de interés) para acotar la cantidad de información introducida en el modelo. En este sentido se eligió caracterizar el movimiento de cada persona usando hasta dos puntos. La implementación de los algoritmos de inferencia de hogar y trabajo se describe en la sección 6.3. Además, en el resto de la sección 6, se presentan los datos utilizados y se habla de los criterios concretos aplicados a la inferencia sobre ese caso particular, ya que, como se discute en 3.2, cada dataset y cada inferencia requieren hacer distintas preguntas a los datos.

3.4. Modelos de movilidad humana implementados en la tesis

Los modos de movilidad que se implementaron en la tesis son los siguientes:

- **Un proceso de caminante aleatorio**, en el cual todos los nodos comienzan a moverse siguiendo un proceso con dirección aleatoria y longitud de salto fija a partir de puntos elegidos dentro de un espacio determinado con una distribución uniforme.
- **Dos procesos gravitatorios** donde, para cada nodo, la información sobre su(s) ubicación(es) proviene(n) de una fuente externa al módulo.
  - **Uno de un solo punto** en el cual cada nodo se encuentra asociado a su casa y queda fijo en este punto.
  - **Uno de dos puntos** entre la casa y el trabajo de cada nodo-persona. Los movimientos se realizan a vuelo de pájaro, sin tener en cuenta la geografía subyacente.

Quedó fuera del alcance la implementación de los demás movimientos de caminante aleatorio definidos a lo largo del capítulo, manteniendo una dirección de movimiento elegida de forma uniforme pero variando la forma de definir r_i. En este sentido, se podrían programar dos variantes: un proceso tal que r_i siga una distribución gaussiana, y otra tal que siga una distribución de Lévy.

[4] Sin embargo, para realizar análisis en este sentido existe Open Street Map, una base de datos cartográfica colaborativa de software libre, que tiene su propia herramienta de ruteo, llamada OSRM (Open Source Routing Machine), disponible en `http://project-osrm.org/`

[5] Encuesta de Movilidad Domiciliaria 2009-2010. Ver nota al pie 3 para más detalles.

4. GENERACIÓN DE GRAFOS SINTÉTICOS

Este capítulo se trata de la capa social del problema de difusión de infecciones, para profundizar lo que adelantamos en la sección 2.2, en la que se tratan algunas líneas de desarrollo para complejizar los modelos epidemiológicos derivados del **SI**. Es importante tener la perspectiva de que muchos de estos modelos fueron desarrollados en la primera mitad del SXX, antes de la masificación de la computadora como herramienta de simulación.

Estos modelos de procesos de difusión más simples asumen la hipótesis de mezcla homogénea de la población a estudiar[1] a la hora de calcular números de contagios. Es decir, no tienen en cuenta la componente espacial ni asumen que algunas conexiones entre personas pueden ser más fuertes que otras.

También por las limitaciones en las herramientas de cada período histórico, los intentos de tener en cuenta la topología de la red subyacente fueron por el lado de definir matemáticamente las propiedades espaciales de las poblaciones estudiadas[2]. Por ejemplo, se buscó expresar distribuciones de personas en el espacio usando funciones de densidad en el plano. Otro enfoque fue modelar poblaciones distribuidas en grafos regulares, por ejemplo grillas bidimensionales[3].

En general, los grafos sociales no tienen una estructura regular, como puede ser una cuadrícula de calles o la estructura de un cristal, por ejemplo. Entonces, para tratar el problema de difusión en redes de personas se necesita un modelo alternativo. También, en este sentido, poder usar una computadora para representar un grafo con nodos y conexiones arbitrarias de forma enumerativa, y computar procesos sobre estos grafos, amplía el universo de grafos sociales "analizables", porque ya no hace falta restringir el modelado a estructuras regulares, que son más fáciles de definir matemáticamente pero comparten menos propiedades con una red social real.

La historia de la teoría de grafos aleatorios[4] comenzó con el trabajo de Paul Erdös y Alfréd Rényi entre los años '50 y '60, en particular con la presentación del modelo que lleva su nombre. Más adelante, en la década del '90 se dieron avances en distintos frentes, a partir de los trabajos en el modelo de mundos pequeños de Watts y Strogatz ([WS98]) y el modelo de *preferential attachment* de Barabási y Albert (1999, en [BP16]). Estos avances provocaron a su vez una gran proliferación de trabajos en diversas áreas de la ciencia (y no solamente en la matemática). En la opinión de le autore, también esta explosión de trabajos sobre el tema se posibilitó por la disponibilidad de computadoras, y por el acceso a estudiar redes de datos relacionadas con Internet. Por ejemplo, en 1998, se realizó en la Universidad de Notre Dame un estudio ([AJB99]) de la topología de la WWW[5]. Actualmente, por

[1] Ver sección 2.1 de esta tesis para referencias sobre la hipótesis de la mezcla homogénea.

[2] Ver discusiones dadas en el capítulo 9 de [Bai+75].

[3] Se puede tomar por ejemplo un paper ([HW65]), mencionado en el capítulo 9 de [Bai+75], en el que se define un modelo para trabajar con procesos de difusión aleatorios en grafos de forma general pero que se centra en el caso de estudio en el que el grafo subyacente es una grilla en dos dimensiones. El paper trabaja sobre el ejemplo de un campo de árboles plantados en una cuadrícula que contraen una infección de sus vecinos.

[4] Reseña extraída del capítulo 1.1 de [Dur07].

[5] La WWW, si bien no es una red social, es una red de información creada por personas, en la que los nodos son páginas. Existe una arista dirigida entre la página A y la B si hay un hipervínculo en A que se dirige a B.

otro lado, se da un proceso que suma a esta explosión: estamos viviendo en una era de generación continua de una inmensa cantidad de datos. Esto pone a disposición (con el procesamiento adecuado) todo tipo de insumos para generar redes sociales enriquecidas con todo tipo de información (localización, intereses, edades, por ejemplo), y estudiar sus características y su estructura subyacente. Nuevamente, esta información es de interés en áreas muy diversas del conocimiento: economía, epidemiología, publicidad, planificación de campañas y planificación urbana, por ejemplo.

El capítulo estudia dos aproximaciones al problema: por un lado el modelo de Erdös-Rényi, y por otro la idea de redes libres de escala y, en particular, el modelo de *preferential attachment* de Barabási-Albert. Presentamos el resumen de varias secciones de cada capítulo como forma de introducir estos modelos y, si bien no pretendemos que el capítulo sea autocontenido, sí quisimos hacer un resumen lo más completo posible de las características básicas que presenta el libro de cada modelo. Por esta razón, no todas las caracterizaciones que presentamos en el capítulo son estrictamente necesarias para comprender la tesis, aunque sirven para desarrollar intuiciones respecto de los dos modelos en un sentido más general.

El capítulo está estructurado de la siguiente manera:

En la sección 4.1 se define el modelo de Erdös-Rényi y se analizan distintas propiedades de los grafos generados con este método: la esperanza de la cantidad de aristas y la función de distribución de los grados de sus nodos. En la parte final de la sección se presenta la propiedad del mundo pequeño (relacionada con la hipótesis de los 6 grados de separación), que es una propiedad importante para esta tesis ya que habla de a qué distancia están los nodos de un grafo, lo que tiene una relación con qué tan rápido se difunde un agente a través de él.

Al estudiar la topología de distintas redes sociales, y otras redes reales experimentalmente, se descubrieron diferencias entre propiedades observadas respecto del modelo de ER, en particular en relación con la distribución de grados y la propiedad del mundo pequeño. Por ejemplo, el estudio realizado en 1998 de la topología de la WWW antes mencionado [AJB99]. En este estudio se observó que la distribución de distancias en esta red era fundamentalmente distinta al modelo generado por ER para condiciones similares. Además, se sub-representa la cantidad de puntos nodales (*hubs*) del grafo en relación a la red real.

A partir de esta observación se desarrollaron otros modelos que replican la distribución de grados observada en redes sociales reales, además de la densidad esperada de la red. En la sección 4.2 de este trabajo se estudia el modelo de *redes independientes de la escala*, definiendo la función de distribución de los grados para este caso también. Una red se dice independiente de la escala si los grados de sus nodos siguen una *distribución de ley de potencia*. Luego se presentan los distintos regímenes de comportamiento de esta familia de redes en función de los valores del parámetro γ de esta distribución. Finalmente se presenta el método de *preferential attachment* como una forma de implementar fácilmente un caso particular de generador de redes independientes de la escala en el régimen de $\gamma = 3$.

El contenido de este capítulo en general, salvo que se indique lo contrario, forma parte de un resumen de algunas secciones de los capítulos 3, 4 y 5 de [BP16]. El capítulo 3 corresponde con la sección 4.1 y los otros dos con la 4.2.

4.1. Redes aleatorias de Erdös-Rényi

La primera aproximación a estimar la topología de las redes sociales con un modelo de red aleatoria fue mediante el modelo en el cual cada arista tiene la misma probabilidad de existir. Este modelo fue desarrollado por Paul Erdös y Alfréd Rényi a partir del año 1959 [ER59]. En este modelo, el parámetro es esta probabilidad, que da como resultado una densidad esperada del grafo (es decir, dada una cantidad de nodos fija, una cantidad de aristas esperada).

Si une se imagina una fiesta en la que cada persona inicialmente no conoce a nadie, y deja que las personas hablen entre sí y se conozcan, la estructura resultante probablemente parezca irregular y azarosa[6]. Esto motiva a que, a la hora de intentar reproducir la estructura de grafos sociales, se piense en algoritmos aleatorios.

Supongamos que queremos modelar una situación social, como la fiesta descrita anteriormente, como un grafo en la que los nodos son personas y se relacionan mediante una arista si hablaron entre ellas en la fiesta. El modelo más simple que se puede definir para generar un grafo de conversaciones es asumir que no tenemos información sobre cuáles vínculos son más probables que otros y que, por lo tanto, todas las aristas deberían ser igualmente probables. En este contexto surgen los primeros algoritmos de generación de grafos aleatorios.

Hay dos formas de definir estos grafos en los que todas las aristas son igualmente probables. Siempre, primero se fija la cantidad de nodos N. Luego, las formas de unir los nodos con aristas son:

- Elegir de forma equiprobable un conjunto de L aristas entre las $\frac{N(N-1)}{2}$ posibles
- Fijar una probabilidad p como medida de la densidad del grafo, y, de forma independiente, elegir con probabilidad p si cada arista se agrega al grafo o no

Se suele usar el segundo método porque es más cómodo para hacer cuentas que el primero, renunciando a fijar exactamente la cantidad de aristas del grafo[7]. Además, es más fácil de programar ya que las probabilidades para cada arista son independientes, mientras que trabajar con subconjuntos de aristas es más complicado.

Una red obtenida mediante alguno de estos métodos se denomina *grafo aleatorio* o bien *red aleatoria*. También se los denomina grafos de Erdős-Rényi.

En las siguientes secciones se describen diferentes aspectos de estos grafos que serán relevantes en el resto de la tesis.

En la sección 4.1 se calculan la cantidad de aristas total esperadas, y el grado esperado de un nodo elegido al azar.

En la sección 4.1 se muestra que la fórmula de probabilidad puntual $p(k)$ (la probabilidad de que un nodo elegido al azar tenga grado k) sigue una distribución binomial. Además, que en el caso de grafos poco densos esta probabilidad está bien aproximada por una distribución de Poisson. Se usa esta distribución en vez de la binomial porque facilita las cuentas.

La sección 4.1 describe las limitaciones del modelo para representar redes sociales reales. Este modelo fue el primero en encarar el problema de representar un grafo social de forma

[6] Este ejemplo está extraído del capítulo 3 de [BP16], que además se utilizó como base para escribir el resto de esta sección.

[7] Ver capítulo 1.2 de [Dur07].

aleatoria. Analizando sus limitaciones fue como se construyeron otros modelos posteriores que las corrigen.

En la sección 4.1 se introduce la idea de la propiedad del mundo pequeño en grafos, que habla sobre la cercanía de los nodos entre sí, en grafos sociales. Si se define la distancia entre dos nodos de un grafo como la cantidad de nodos intermedios que tiene un camino más corto entre ellos, la propiedad del mundo pequeño da un orden de crecimiento de esta distancia en función de la cantidad de nodos del grafo. Se muestra que para el caso de grafos de Erdös-Rényi este crecimiento es $O(ln(N))$ siendo N la cantidad de nodos del grafo.

Cantidad esperada de aristas y grado esperado

Si fijamos la cantidad de nodos N y la probabilidad p de que cada par de nodos resulte en una arista:

- La cantidad de ejes máxima del grafo es de

$$\frac{N(N-1)}{2}$$

- la probabilidad puntual de que el grafo tenga exactamente L aristas es

$$p(L) = p^L(1-p)^{\frac{N(N-1)}{2}-L}\binom{N}{L}$$

- Como esta es una distribución binomial, la esperanza de L es:

$$< L >= \frac{pN(N-1)}{2}$$

- El grado promedio $< k >$ es:

$$< k >= \frac{2 < L >}{N} = p(N-1)$$

Distribución de grados

Para un nodo se puede realizar un cálculo similar para encontrar su distribución de grados, es decir, caracterizar la probabilidad de que, eligiendo un nodo con una probabilidad uniforme, éste tenga grado k. La probabilidad resulta del producto de 3 factores:

- La probabilidad de que k aristas estén presentes
- La probabilidad de que $N-k-1$ aristas no estén presentes
- El combinatorio de los diferentes órdenes en los que estas aristas pueden aparecer

La cuenta queda:

$$p(k) = p^k(1-p)^{N-k-1}\binom{N-1}{k} \tag{4.1}$$

Esta ecuación corresponde con una distribución binomial en función de N (cantidad de nodos totales) y p (probabilidad de aparecer para cada arista).

Cuando la cantidad de nodos crece, y p es relativamente bajo (lo que da una relación $k << N$), la distribución binomial es bien aproximada por una distribución de Poisson, que tiene la siguente función de densidad:

$$p(k) = e^{-\mu}\frac{\mu^k}{k!} \tag{4.2}$$

donde $\mu = (N-1)\cdot p$ es la esperanza de la distribución binomial calculada anteriormente.

Las ventajas de usar la distribución de Poisson en vez de la binomial son que la ecuación es más simple y, sobre todo, que depende de un sólo parámetro μ en vez de dos (N y p).

Diferencias entre redes sociales y el modelo de Erdös-Rényi

El modelo de Erdös-Rényi fue, históricamente, la primera forma de encarar el problema de caracterizar/crear sintéticamente grafos que modelaran redes sociales. Sin embargo (como todo modelo) no refleja exactamente la realidad en varios aspectos. Por ejemplo, la función de densidad del grado de los nodos es distinta a la observada en redes sociales reales. Por esta razón, sobre este modelo se construyeron otros posteriores, para reflejar más fielmente ésta y otras características de las observaciones reales.

A continuación se presenta un ejemplo que permite generar una intuición sobre la diferencia entre las distribuciones de los grados de los nodos. Supongamos que una red social mundial sigue un proceso de Erdös-Rényi. Se supone que 1000 personas es una buena estimación del grado promedio de un nodo en esta red[8], por lo cual vamos a fijar $\mu = 1000$ y asumir que la red sigue una distribución de grados Poisson. Siendo $N = 7 \times 10^9$ podemos extraer conclusiones del modelo:

- El nodo más conectado tiene aproximadamente 1185 conexiones, y el menos conectado aproximadamente 816.
- el desvío estándar σ de la distribución es de $\sqrt{\mu} = 31{,}62$, con lo cual una gran proporción de la población está en el rango

$$[\mu - \sigma, \mu + \sigma] = [968{,}38, 1031{,}62]$$

En estos grafos, la gran mayoría de los nodos tienen una cantidad de conocides muy cercana al promedio. No existen prácticamente en el modelo personas muy famosas, relacionadas con muchas más de 1000 personas, ni personas con una red mucho más chica. Observando redes reales (por ejemplo la red social subyacente a Facebook) vemos que estas personas existen, y que cuando se intentan replicar estas redes usando el algoritmo de Ërdos-Rényi, no se pueden generar con una probabilidad suficiente nodos muy grandes ni muy pequeños.

Esta diferencia entre el modelo de ER y los grafos reales es relevante para el estudio de procesos de difusión porque, como se discute luego en la sección 4.2, la distribución de grados de los nodos tiene un impacto en la forma de conexión de los nodos del grafo, y por lo tanto en la velocidad de los procesos de difusión sobre ellos. Este impacto está relacionado con cómo opera en cada modelo la propiedad del mundo pequeño presentada a continuación.

[8] en [dK78] se habla de algunos estudios que se hicieron para estimar la cantidad de conocides de una persona.

La propiedad del mundo pequeño

La propiedad del mundo pequeño, también conocida como la propiedad de los 6 grados de separación, establece que si se eligen dos personas al azar entre todas las personas del mundo, se puede encontrar un camino de 6 personas o menos que los une. El nombre proviene de un cuento del escritor húngaro Frigyes Karinthy, del año 1929.

Más allá del número 6 en particular, la idea detrás de esta propiedad es que, en un grafo social, la distancia entre dos nodos (entendida como cantidad de aristas en el camino más corto entre ellos) elegidos de forma aleatoria tiende a ser corta. Esto ocurre no sólo entre nodos/personas que estén cerca físicamente (por ejemplo que vivan en la misma localidad) sino también entre personas de distintas partes del mundo.

Une podría preguntarse qué tan pequeño resulta el mundo en este sentido, y qué es lo que genera que las distancias sean tan cortas. Una intuición detrás de este fenómeno puede ser que, si nos olvidamos de la repetición de nodos por un momento y tratamos de estimar de forma gruesa la cantidad de vecinos a distancia d de un nodo particular, siendo k la esperanza del grado de un nodo, podemos realizar el siguiente cálculo rápido:

- Si $d = 1$, la cantidad de nodos vecinos es aproximadamente k
- Si $d = 2$, es aproximadamente k^2
- En general, para distancia d la cantidad de nodos es aproximadamente k^d

Es decir que la cantidad de nodos a distancia menor o igual a d es:

$$N(d) \sim k + k^2 + \cdots + k^d = \frac{k^{d+1} - 1}{k - 1}$$

Sea d_{max} el diámetro del grafo. Esto quiere decir que la cantidad de nodos a distancia d_{max} o menos de cualquier nodo origen debería ser todo el grafo. Es decir:

$$N(d_{max}) \sim N \tag{4.3}$$

$$\frac{k^{d_{max}+1} - 1}{k - 1} \sim N \tag{4.4}$$

$$k^{d_{max}+1} - 1 \sim N(k - 1) \tag{4.5}$$

Para N y k suficientemente grande los -1 son despreciables:

$$k^{d_{max}+1} \sim Nk \tag{4.6}$$

$$k^{d_{max}} \sim N \tag{4.7}$$

$$d_{max} \sim log_k(N) = \frac{ln(N)}{ln(k)} \tag{4.8}$$

Este cálculo tiene problemas, por ejemplo que la cantidad de nodos está acotada por N, pero sirve para tener una intuición: si la cantidad de nodos a distancia menor o igual que d crece de forma exponencial respecto de la distancia, en aproximadamente $ln_k(d)$ iteraciones esperamos cubrir todos los nodos del grafo.

Es decir que esta estimación gruesa nos da una intuición de que el crecimiento del diámetro[9] de un grafo aleatorio crece con orden $O(ln(N))$.

[9] El diámetro de un grafo es la máxima distancia entre dos de sus nodos cualesquiera.

La formulación, en resumen, tiene la siguiente forma:

$$d_{max} = O\left(\frac{ln(N)}{ln(k)}\right)$$

Sin embargo, la evidencia empírica muestra que esta intuición no funciona exactamente para el camino más largo pero sí da una buena aproximación de la longitud promedio de un camino. Intuitivamente esto es porque el camino más largo puede ser particularmente largo mientras que promediar los caminos elimina fluctuaciones estadísticas. Por lo tanto la formulación típica de la propiedad del mundo pequeño es la siguiente:

$$< d >= O\left(\frac{ln(N)}{ln(k)}\right) \tag{4.9}$$

La ecuación 4.9 indica que la distancia promedio crece logarítmicamente respecto al tamaño del sistema, y también que depende de forma inversa de la densidad del grafo[10] (es decir, las distancias se acortan a medida que el grafo es más denso).

La propiedad del mundo pequeño es antiintuitiva porque muchos grafos más regulares que se suelen utilizar no la cumplen. Por ejemplo, las grillas N-dimensionales, que fueron usadas como modelos de difusión antes de la generación de grafos aleatorios[11], se comportan de la siguiente manera:

- En las grillas de una dimensión (es decir, una línea) la distancia máxima es $O(N)$
- En una grilla bidimensional, con $N = n^2$ nodos la distancia máxima es $2n$, con lo cual la distancia máxima es $O(N^{1/2})$
- En general, en una grilla k-dimensional la distancia máxima es $O(N^{1/k})$

En todos los casos el crecimiento es una función polinomial, que es mayor asintóticamente al crecimiento logarítmico que se da en un grafo aleatorio. Esto hace que, sobre todo para redes grandes (como es el caso de la mayoría de redes sociales que se estudian actualmente), los nodos de una red aleatoria estén más cerca entre sí que los de una red determinística regular como las que se muestran en esta sección.

En el contexto del análisis de procesos de difusión de infecciones entre agentes, esto quiere decir que, intuitivamente, un agente en un grafo aleatorio se debería transmitir con mayor rapidez en un grafo de Ërdos-Rényi que en una grilla con una cantidad de nodos y una densidad de aristas similares. En un sentido más general, Manfred Kochen e Ithiel de Sola Pool discuten distintas implicaciones de la estructura de este tipo de grafos en [dK78], un paper fundacional e interdisciplinario[12] publicado en la década del '70, que ataca el problema de la estructura de los grafos sociales y que inspira el desarrollo de la hipótesis del mundo pequeño. También habla de las distintas implicancias sociales de esta idea, y de cómo impactan los estratos sociales en la formación de redes (tema que queda fuera del alcance de esta tesis).

La relación logarítmica entre d y N debe ser ajustada para adaptarse a redes reales. Como ya se dijo anteriormente en esta sección, el modelo de Ërdos-Rényi tiene limitaciones

[10] Recordemos que esta fórmula es una estimación. Existen correcciones sobre ella que se discuten con más detalle en el anexo 3.F de [BP16].

[11] Recordar ver por ejemplo [HW65] para grillas bidimensionales.

[12] Es interesante que se trata de un trabajo interdisciplinario porque Kochen fue matemático e informático y de Sola Pool, científico social.

a la hora de explicar ciertas características de los grafos sociales, y esta relación es una de ellas. En la sección siguiente se discute una revisión de la hipótesis del mundo pequeño en el contexto de nuevos modelos, que dan distancias aún más cortas que éste.

4.2. Redes independientes de la escala y *preferential attachment*

Si bien el desarrollo de las redes de Erdös-Rényi (ER) sirvió para expresar lazos sociales en situaciones puntuales y acotadas en el tiempo, se encontró que varias redes sociales y otras redes de conocimiento (como por ejemplo la jerarquía de páginas web con sus hipervínculos) presentan características distintas a este modelo. En la introducción de esta sección se presentan y se comentan brevemente a modo de ejemplo algunos trabajos de la década del '90 que van en este sentido.

Todos estos trabajos apuntan a la idea de que, por un lado, las redes sociales y de información no son completamente regulares, es decir que tienen un componente aleatorio, pero por otro lado las aristas no son todas equiprobables como en el modelo de ER. En particular se observó experimentalmente una presencia mayor de nodos de grados más extremos que los que da una distribución de Poisson, que es la función de distribución para el modelo de ER[13]. Es decir, esta distribución tiene una acumulación de nodos con un grado cercano a la esperanza de la distribución, mientras que los grafos reales muestran una mayor proporción de nodos tanto con un grado bastante mayor como menor. Sobre todo, es de interés el estudio de los nodos con grado bastante mayor (puntos nodales o *hubs*) porque su existencia afecta los procesos de difusión sobre estos grafos.

En particular, se vio que la WWW es un grafo *independiente de la escala*[AJB99]. En lo que sigue de la sección se presenta este modelo y distintas propiedades que presentan los grafos aleatorios que siguen esta distribución.

Se define entonces una *red (o grafo) libre de escala* como aquella en la que su p_k está bien aproximado por una ley de potencia.

En este contexto se consideraron solamente grafos no dirigidos, pero hay grafos sociales con aristas dirigidas en las que los grados de entrada y de salida siguen distribuciones de independientes de escala con valores de γ diferentes.

El resto de la sección se estructura de la siguiente manera: en la sección 4.2 se presentan las versiones discretas y continuas de la función de distribución derivada de una ley de potencia, $p(k)$. La sección 4.2 habla de la diferencia entre la cantidad y el tamaño de los puntos nodales (*hubs*) entre el modelo de Erdös-Rényi y el de leyes de potencia. Se distinguen los distintos regímenes de comportamiento en función de ambas funciones para los distintos rangos de valores de k. La sección 4.2 habla de los diferentes tipos de grafos que se pueden generar usando una distribución independiente de la escala para distintos valores de γ. Se habla de que, en general, las redes sociales tienden a tener valores de γ entre 2 y 3. La sección 4.2 revisita la idea de *mundo pequeño* introducida en la sección 4.1 y discute cómo cambian las distancias entre nodos del grafo cuando sus nodos siguen una ley de potencia. Finalmente la sección 4.2 presenta el método de *preferential attachment* como una forma de generar grafos que sigan una ley de potencia de forma orgánica, para $\gamma = 3$, y la sección 4.2 habla sobre el algoritmo construido a partir del método de PA en el contexto de la tesis.

[13] Ver ecuación 4.2 en la sección 4.1.

Función de distribución de grados p_k

En esta sección se describe la distribución de los grados de los nodos en un grafo independiente de la escala. Se usan dos formalismos distintos, uno discreto (una función de probabilidad puntual) y otro continuo (una función de distribución).

Se muestra a continuación la derivación de p_k en el caso discreto por su simplicidad. Al ser una distribución independiente de la escala, dado γ fijo se puede definir p_k:

$$p_k = Ck^{-\gamma}$$

donde C es una constante de normalización para que la función de probabilidad puntual sume 1 sobre todo el dominio. Es decir:

$$\sum_{k=1}^{\infty} p_k = 1$$
$$\sum_{k=1}^{\infty} Ck^{-\gamma} = 1$$
$$C\sum_{k=1}^{\infty} k^{-\gamma} = 1$$
$$C\zeta(\gamma) = 1$$
$$C = \frac{1}{\zeta(\gamma)}$$

donde ζ es la función zeta de Riemann. Entonces la distribución de probabilidad puntual es:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

La versión continua asume que cada nodo puede tener como grado cualquier número real, que es una suposición más conveniente para realizar cálculos. En este caso planteamos:

$$p(k) = Ck^{-\gamma}$$

donde C es tal que $p(k)$ integra a 1. Es decir:

$$\int_{k_{min}}^{\infty} p(k) = 1$$
$$\int_{k_{min}}^{\infty} Ck^{-\gamma} = 1$$
$$C = \frac{1}{\int_{k_{min}}^{\infty} k^{-\gamma}}$$
$$C = \frac{-\gamma + 1}{-k_{min}^{-\gamma+1}}$$
$$C = (\gamma - 1)k_{min}^{\gamma-1}$$

Por lo tanto reemplazando C se obtiene la expresión para p(k):

$$p(k) = (\gamma - 1)k_{min}^{\gamma-1} \cdot k^{-\gamma} \tag{4.10}$$

donde k_{min} es el mínimo valor posible de k. En el contexto de la tesis se trabaja sobre el formalismo continuo.

Puntos nodales (hubs) de la red

La principal diferencia entre una red aleatoria de Erdös-Rényi (ER) y una red independiente de la escala está en su función de distribución p_k. En particular, estas últimas son *de cola pesada*. Esto quiere decir que es más frecuente encontrar nodos con grado alto (puntos nodales) que en una red de ER.

Si comparamos ambas distribuciones a lo largo del dominio se distinguen tres regiones[14] con comportamiento distinto:

- Para valores pequeños de k, la distribución independiente de escala siempre es mayor que la de Poisson, es decir que en un grafo independiente de la escala se esperan más nodos con grado bajo que en una red de ER.
- Para k cercanos a la esperanza de la Poisson ocurre lo contrario, es decir que en una red de ER hay un exceso de nodos de grado cercano a k respecto de un grafo independiente de la escala.
- Para valores grandes de k la distribución independiente de escala vuelve a superar a la Poisson, por lo cual también en un grafo independiente de la escala se esperan más nodos con grado alto que en una red de ER.

También se observan diferencias en cómo cambia el tamaño de los puntos nodales de la red en función de su cantidad de nodos totales N. Es decir, qué tan grandes son los *nodos grandes* en cada una de las distribuciones. Para esto se puede observar la distribución de probabilidad de k_{max}, el nodo con mayor grado de la red, en cada uno de los casos.

En el caso de un grafo independiente de la escala se observa un crecimiento polinomial de k_{max} en función de N, mientras que para una red aleatoria este crecimiento es subpolinomial (con un ritmo $O(ln(N))$)[15].

La fórmula específica para la esperanza del tamaño del nodo máximo para una distribución independiente de escala es:

$$k_{max} = k_{min} N^{\frac{1}{\gamma - 1}} \tag{4.11}$$

Esto quiere decir que si se generan dos redes usando cada uno de los métodos, con distribuciones con esperanzas similares, el valor esperado del nodo más grande es órdenes de magnitud mayor para el grafo independiente de la escala que para el grafo de ER.

Caracterización de momentos de p_k

Una forma analítica de observar estas diferencias entre los dos tipos de redes es observando los momentos de la función de distribución de una ley de potencias y cómo se

[14] Hay un gráfico que ilustra las tres regiones mencionadas en la sección 4.3 de [BP16].

[15] El análisis para un grafo independiente de la escala se puede encontrar en la sección 4.3 de [BP16]. El correspondiente a una red de ER no se reproduce en dicha fuente, pero se puede hallar usando la acumulada de la p_k, y que la probabilidad de que el máximo sea menor o igual a un valor x es el producto de las N acumuladas hasta x.

traducen en características del grafo subyacente. La fórmula para el enésimo momento (utilizando el formalismo continuo presentado en 4.2) es:

$$< k^n >= \int_{k_{min}}^{k_{max}} k^n \cdot p(k)dk = C \cdot \frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1}$$

El valor de k_{min} está fijo, mientras que k_{max} aumenta en función de N. Vamos a analizar el comportamiento de cada momento al hacer tender N a infinito (y por lo tanto k_{max} a infinito también) para entender su comportamiento en redes de gran tamaño. Este comportamiento cambia en función de la relación entre n y γ:

- Si el exponente $n-\gamma+1$ es menor que cero, $k_{max}^{n-\gamma+1} \to 0$ y por lo tanto los momentos que satisfacen $n < \gamma - 1$ son finitos. Si es igual, $k_{max}^{n-\gamma+1} = 1$ y el momento también es finito.
- Cuando el exponente es mayor que 0, $k_{max}^{n-\gamma+1} \to \infty$, y por lo tanto estos momentos divergen.

La mayoría de las redes sociales tienen un γ entre 2 y 3, por lo que su esperanza (primer momento) converge mientras que su varianza (relacionada con el segundo momento) diverge[16]. En particular, el comportamiento de σ^2 es el que le da el nombre a las redes independientes de la escala.

- Se dice que las redes aleatorias tienen una escala ya que, independientemente de la cantidad de nodos N, tanto la esperanza como la varianza de una distribución de Poisson valen $< k >$, entonces el desvío estándar σ vale $\sqrt{< k >} =< k >^{\frac{1}{2}}$, y la mayoría de los nodos tendrán un grado en el rango $[\mu-\sigma, \mu+\sigma] = [< k > - < k >^{\frac{1}{2}}, < k > + < k >^{\frac{1}{2}}]$.
- En cambio en el caso de un grafo independiente de la escala, en cambio, no se puede acotar el desvío estándar independientemente de N y por lo tanto se dice que no tiene una escala (en particular, no se puede afirmar que la probabilidad de tener un nodo con grado grande tiende a 0 a medida que $N \to \infty$).

Revisión de la propiedad del mundo pequeño

Cabe preguntarse si la presencia de puntos nodales con un grado muy alto afecta la propiedad del mundo pequeño discutida en la sección 4.1, porque la caracterización de los diámetros de los grafos estudiados está directamente relacionada con la difusión de un agente a través de sus aristas.

El efecto sobre esta propiedad depende del valor de γ y además está relacionado con la discusión anterior sobre la convergencia de los momentos de p_k al divergir el tamaño del grafo.

En el análisis se tendrán en cuenta solamente valores de γ que sean mayores o iguales a 2. Esto es porque, cuando $\gamma < 2$, mirando la ecuación 4.11, el exponente $\frac{1}{\gamma-1}$ es mayor que 1, por lo que el grado del nodo más alto crece más rápidamente que la cantidad de nodos del grafo. Con un N suficientemente grande un grafo de estas características es imposible de crear.

[16] Tabla 4.1 de [BP16]

Efectivamente existe una diferencia entre los grafos de Erdös-Rényi y los indepdndientes de escala: en [BP16] se enumeran los distintos regímenes de dependencia del diámetro del grafo $< d >$ como una función de su cantidad de nodos N, para distintos rangos de γ:

- Cuando $\gamma = 2$, el exponente en la ecuación 4.11 vale 1, y por lo tanto $< d >$ crece linealmente con N. En este régimen el diámetro esperado es constante (no depende de N).
- Cuando $2 < \gamma < 3$, el régimen se denomina *ultra pequeño* ya que el diámetro es asintóticamente distinto al caso de una red aleatoria común. La dependencia en este régimen es:
$$< d >= O(lnln(N)) \tag{4.12}$$
- Cuando $\gamma = 3$ se encuentra el *punto crítico*, en el que la varianza de la distribución de grados deja de divergir. En este punto se cumple:
$$< d >= O\left(\frac{ln(N)}{lnln(N)}\right) \tag{4.13}$$
.
- Cuando $\gamma > 3$ los puntos nodales (hubs) no resultan suficientemente grandes como para achicar el diámetro del grafo con respecto a la ley de *mundo pequeño* que se encuentra en las redes aleatorias, y por lo tanto la dependencia es:
$$< d >= O(ln(N)) \tag{4.14}$$

Esta es la razón por la cual es de interés trabajar sobre grafos que tengan un γ entre 2 y 3, y sobre grafos con $\gamma > 3$ por separado a la hora de realizar comparaciones entre distintos algoritmos de generación de grafos aleatorios, ya que cada rango genera grafos con catacterísticas distintas, que afectarán de diferente manera el proceso de simulación.

En particular se espera que, a medida que crece el valor de γ, los grafos se parezcan cada vez más a los de Erdös-Rényi, ya que baja la presencia de puntos nodales, es decir, de nodos con grado mucho más alto que la media.

La distribución de grados de un grafo que cumple una ley de potencia es, entonces, cualitativamente distinta a la de un grafo creado por el método de Erdös-Rényi (que resulta similar a la del régimen $\gamma > 3$ para grafos que siguen una ley de potencia). Además, esto parece implicar diferencias en la distribución de las distancias entre nodos del grafo, lo que a su vez podría cambiar la velocidad con la que se da un proceso de difusión en este mismo grafo.

Método de *preferential attachment* para generación de grafos que cumplan una ley de potencia

En esta tesis se implementó un caso particular de método de generación de grafos independientes de la escala: el de construcción por *preferential attachment*, presentado en la bibliografía[BP16].

El algoritmo genera un grafo independiente de la escala con $\gamma = 3$, por lo cual no se puede aplicar para generar una red en el caso general, pero sí para crear fácilmente grafos para comparar con los de Erdös-Rényi en el contexto de procesos de difusión. La justificación de estas propiedades están en el anexo A (secciones A.1 y A.2).

Este método fue elegido por dos razones. La primera es que es más fácil de programar que otras alternativas. En la bibliografía, la otra forma propuesta en [BP16] de construir un grafo que siga una ley de potencia con γ arbitrario es la siguiente:

1. se fija el grado de cada nodo tomando una muestra de la distribución deseada[17],
2. se conectan los nodos entre sí eligiendo pares de vértices a quienes les falten vecinos y uniéndolos, hasta que no queden más vértices por unir,
3. se corrigen autoejes y multiejes haciendo intercambios de extremos con alguna otra arista elegida aleatoriamente, conservando los grados de todos los nodos involucrados.

Esta forma de generar grafos aleatorios es bastante más compleja que el algoritmo utilizado para *preferential attachment*.

La segunda razón para elegir *preferential attachment* es que explora una idea interesante, que es la de *hacer crecer* una red mediante la adición iterativa de nodos a un grafo, en lugar de partir de un conjunto de nodos fijo e intentar unirlos con aristas de alguna forma que tenga sentido. Es una forma orgánica de construir grafos que pretende simular el surgimiento de una red social. Esto induce una nueva forma de generar grafos de forma aleatoria, esencialmente distinta a la forma propuesta por Erdös y Rényi.

La forma propuesta de hacer crecer un grafo orgánicamente es mediante dos propiedades[18] de los sistemas reales:

- **Crecimiento**: se supone que las redes comienzan teniendo una cantidad de nodos fija N sin ninguna arista que los conecta, pero las redes se expanden mediante la adición de nodos nuevos.
- **Conexión preferencial**: existe la hipótesis de que cada vínculo puede generarse o no con una probabilidad aleatoria (en el caso de Erdös-Rényi, incluso, con una probabilidad independiente de los otros ejes), pero los nodos en redes reales prefieren conectarse con los nodos más conectados. Es más probable que se citen papers más populares o que se contraten actores más famoses a que estas probabilidades se repartan de forma uniforme.

Algoritmo de generación del grafo

A continuación se presenta la metodología propuesta por [BP16] para generar un grafo por *preferential attachment*. No se lo presenta como algoritmo porque el paso 2 no es un paso mecánico.

1. Se crean m_0 nodos vacíos
2. Se conectan estos nodos de forma arbitraria tal que constituyan una única componente conexa
3. Para cada instante de tiempo entre 0 y $N - m_0$ (donde N es la cantidad de nodos deseada)

[17] Esto incluye dos correcciones: la primera es un ajuste en algún nodo para que la suma de grados sea par, y la segunda es no permitir nodos de grado 0 para que no queden vértices aislados.

[18] Se puede encontrar una justificación de por qué ambas propiedades son necesarias en el anexo A.3

a) Se crea un nuevo nodo n

b) Se conecta a n con $m < m_0$ otros nodos de acuerdo a la siguiente distribución de probabilidad. Sea $\Pi(k_i)$ la probabilidad de que el nuevo nodo se una al nodo preexistente k_i.

$$\Pi(k_i) = \frac{\text{grado}(k_i)}{\sum_j \text{grado}(k_j)}$$

Dadas una cantidad de nodos iniciales m_0 y una de ejes e_0, luego de t pasos se genera un grafo que tiene $m_0 + t$ nodos y $e_0 + tm$ ejes.

Esta definición del modelo deja cuestiones abiertas que deben decidirse al momento de una demostración formal o de una implementación de un algoritmo de generación de grafos a partir del modelo de Barabási-Albert. En [BP16], sección 5.3, se presenta el formalismo de Bollobás et al. para las demostraciones que se hacen a lo largo del capítulo, que asume que los grafos son creados de esta manera, pero se decidió hacerle modificaciones para la tesis.

En general se tomó la decisión de no permitir autoejes ni multiejes en los grafos generados. El formalismo de Bollobás et al. no sólo los permite sino que siempre produce un autoeje en su primer paso, por lo que se optó por usar una variación que sigue la premisa original pero no este formalismo exactamente. El algoritmo usado consiste en comenzar a partir de una clique (grafo completo) de tamaño m, y, al momento de agregar un nodo nuevo, tomar una muestra sin repetición de m nodos preexistentes para que sean sus vecinos, usando la distribución de probabilidades propuesta (en función del grado previo de cada nodo). Esto además asegura que el grafo resultante es conexo.

5. IMPLEMENTACIÓN Y ESTRUCTURA DEL PROTOTIPO

En este capítulo se describe el funcionamiento del prototipo construido, así como distintas consideraciones de diseño del software y decisiones de implementación que se exploraron.

El software simula una población de personas a través del tiempo, dividido en unidades de tiempo discretas. Dado el sistema en el tiempo t, a partir de su estado interno, calcula su estado en el tiempo $t + 1$. Este proceso en el caso general no es determinístico, ya que las decisiones tomadas en esta transición pueden depender de variables aleatorias. Esto corresponde a una estrategia de simulación basada en *escaneo de actividades*, o *activity-scanning*.

En el prototipo, el estado de la población es el conjunto de los estados para cada nodo-persona que pertenece a ella, y los vínculos entre personas (aristas del grafo social). El estado de cada persona está conformado a su vez por su estado de salud/infección, y por su estado de movilidad.

Como grafo social se usaron como entrada grafos generados sintéticamente usando las técnicas descriptas en la sección 4, o bien un grafo sin aristas en los casos en los que no se quiso considerar la componente social por algún motivo (por ejemplo si el contagio únicamente ocurre por el aire entre personas que están cerca físicamente).

Se pueden colocar etiquetas arbitrarias a las personas para usar en conjunto con las reglas (por ejemplo, para expresar si la persona se vacunó, su edad o su pertenencia a algún grupo de riesgo).

Los distintos aspectos de la simulación fueron encapsulados en módulos, excepto por el grafo social y las etiquetas sobre las personas. Esta decisión solamente se tomó por una cuestión de alcance, ya que tiene sentido considerar si estos dos aspectos se pueden pensar también como módulos.

El resto de la sección presenta distintos aspectos de implementación tenidos en cuenta en la tesis. La sección 5.1 introduce la idea de *lenguaje ubicuo*[1], un lenguaje común entre el dominio que se desea modelar en una simulación y el software desarrollado. En la tesis se tomó la decisión de explorar el concepto de *lenguaje ubicuo* y la idea de generar una biyección entre el dominio de problema y el programa de software que lo modela, usando este lenguaje, para reducir la carga cognitiva al modelar y expresar requerimientos.

En la sección 5.2 se describen los distintos módulos que fueron implementados. En la 5.3 se presenta el modelado de *Condiciones* en varios puntos de la implementación. La sección 5.4 trata la decisión de diseño de que la *Persona* sea la entidad responsable de conocer la información sobre un nodo particular en el sistema, pero que los módulos definan el formato de los distintos aspectos de esta información. La sección 5.5, por último, trata la elección de una estrategia o cosmovisión de simulación para el prototipo, entre varias opciones posibles.

5.1. Uso de lenguaje ubicuo para modelar la realidad

La simulación es la creación de un programa que reproduzca un sistema que se quiere modelar, para poder extraer información sobre cómo se da el desarrollo de este sistema

[1] Ver capítulo 2 de [EE04].

a lo largo del tiempo. Entonces, una simulación es en particular un programa, entendido como un modelo computable de un dominio de problema de la realidad.

Es deseable que un programa de software represente el dominio de problema de una forma lo más fiel posible, entre otras razones porque eso facilita que sea usado como herramienta para entender un sistema. Por eso, un indicador de calidad del software es el grado de correspondencia entre las entidades presentes en el dominio y los objetos del programa. Esta correspondencia crece cuando el lenguaje[2] en el que está escrito el programa se acerca al que se usa para expresar el dominio, y se desarrolla un lenguaje común entre el mundo del dominio (en este caso la epidemiología) y el mundo del desarrollo de software. Este lenguaje se denomina *lenguaje ubicuo*[3].

En este caso, el dominio a representar (uno de los dominios centrales que entran en juego en el software desarrollado) es el de la difusión de infecciones entre personas. Es entonces valioso hacer una exploración de cuáles son los conceptos que existen en este dominio y tratar de modelarlos lo más explícitamente posible y usando el lenguaje ubicuo.

En el momento en que alguien plantea un experimento sobre un proceso, piensa los datos de entrada del modelo como una enumeración de condiciones y decisiones. Por ejemplo, a la hora de realizar una simulación de contagio de una infección respiratoria, una pregunta posible es ¿cómo afecta el proceso que cierta proporción de los nodos use tapabocas? El planteo explícito de esta pregunta se hace definiendo cuál es la probabilidad de contagio entre dos personas que no usan tapabocas, entre dos que lo usan, de una que sí hacia otra que no, y viceversa. Por eso, si el modelo permite dar como datos de entrada estos valores, resulta explícito en este sentido. Si solamente permite introducir una única probabilidad de contagio, se debe hacer un análisis intermedio a partir de estas probabilidades y de la proporción de nodos que usan tapabocas para obtener una probabilidad ponderada tal que el proceso resulte similar. Este es un modelo que embebe la complejidad del proceso de contagio en un único valor y por lo tanto representa implícitamente esta situación. En este caso se pierde la declaratividad del modelo ya que existe una traducción o adaptación entre el dominio del problema y el programa.

Esta tesis se propuso generar una prueba de concepto de un sistema explícito, donde se pudieran expresar condiciones complejas sobre el modelo en lenguaje ubicuo, y que por lo tanto eliminaran la necesidad de resumirlas en parámetros más puntuales. En particular, se trabajó sobre los procesos de decisión de cuándo ocurre un contagio entre dos agentes dados, y, una vez que un nodo se encuentra infectado, de cuál es el desarrollo interno de la infección dentro de él.

No hay un límite teórico de la complejidad que pueden tener las condiciones/predicados (sí existe, obviamente, un límite práctico que proviene de la capacidad de cómputo). En principio son estructuras definidas de forma recursiva y cada parte puede ejecutar código arbitrario por lo que las condiciones se pueden volver arbitrariamente grandes. Se eligió priorizar la declaratividad sin buscar optimizar los tiempos de corrida, ya que, más allá de que el prototipo se pueda usar en su estado actual para producir experimentos, y de hecho se usó en el marco de la tesis, aporta el valor adicional de explorar distintos conceptos o ideas de diseño que son adaptables individualmente a otros sistemas.

En particular, la definición de condiciones arbitrarias ayuda a la hora de definir simulaciones para condiciones más complejas, quitando la traducción del camino para poder

[2] No en el sentido de lenguaje de programación sino en el sentido de las palabras que se usan para nombrar conceptos en el programa.

[3] Ver capítulo 2 de [EE04].

enfocarse en el modelado del dominio.

A modo de ejemplo, algunas posibles extensiones sobre el modelo:

- Procesos de aislamiento social intermitente (por períodos de tiempo predefinidos, o dependiendo de variables del mismo sistema, por ejemplo ante cierto nivel de ocupación de los hospitales). Implementación intermitente de otras medidas de cuidado, como el uso de alcohol en gel o tapabocas.
- Planes progresivos de vacunación que ocurren simultáneamente con el proceso de contagio.
- Comportamiento diferenciado de los agentes en función de la información que tienen (por ejemplo, si una persona necesita un test para saber si está infectada o no, y a partir de esta información decide si seguir trabajando o aislarse).
- Comportamiento diferenciado en función de la edad o posibles factores de riesgo de cada agente en particular.

Esta tesis pretende dar un marco para expresar estas situaciones complejas de forma explícita. Para esto separamos distintas capas de análisis dentro del modelo. En particular en este trabajo nos centramos en la difusión de patógenos físicos entre personas. Por eso consideramos en particular *la capa social*, que modela las relaciones sociales entre personas, y *la capa de movilidad espacial*, que habla sobre su posición geográfica. Estas capas corresponden a distintos módulos del sistema de software desarrollado, con el espíritu de generar una relación uno a uno entre las partes que tiene el problema de difusión en grafos y las de la solución de software producida.

El valor de esta arquitectura modular es que admite la adición de nuevas capas en caso de ser necesario. Por ejemplo, en el prototipo que se generó en la tesis no existe un módulo encargado de monitorear el sistema e implementar medidas dinámicamente (por ejemplo, un plan de vacunación progresiva de la población o aislamiento social intermitente) pero el sistema soporta este tipo de extensiones, mediante la creación de nuevos módulos que tengan la capacidad de cambiar el comportamiento de otros módulos, o actualizar la información que cada persona sabe de sí misma. Por ejemplo, para simular aislamiento intermitente, se puede implementar una entidad que, al ver que la cantidad de contagios supera cierto umbral, se comunica con el módulo de movilidad y le informa que tiene que cambiar los patrones de movilidad de cada nodo, y éste actúa en consecuencia, afectando la forma en la que los nodos se van a mover en el paso siguiente de la simulación.

5.2. Módulos

Las acciones que los nodos realizan se programan en distintos módulos que controlan los aspectos en los que el proceso de cada persona avanza. Esta implementación del sistema tiene tres:

- Movilidad de las personas en el espacio (sección 5.2.1).
- Lógica de contagios/propagación del agente (sección 5.2.2).
- Progresión de la infección dentro de las personas (sección 5.2.3).

El módulo de movilidad inyecta en la persona la forma de moverse en el espacio. El de propagación usa reglas para determinar, para cada nodo infectado, si ocurre un contagio hacia otros nodos. La condición de contagio se evalúa a través de reglas, un objeto que tiene la capacidad de predicar una condición arbitraria sobre el modelo, lo cual le da flexibilidad al sistema. El de progresión usa una configuración de infección para simular su progreso. La infección funciona como una máquina de estados no determinística y en cada instante discreto de tiempo se simula. Cada transición tiene asociada una regla.

Todos los módulos se inicializan guardando el modelo al que refieren porque necesitan vincularlo a las condiciones que crean (porque como las condiciones son arbitrarias necesitan estar acopladas al modelo entero para poder hacerle consultas de forma general). Esto ocurre porque, por ahora, el modelo no está suficientemente maduro como para determinar si hay ciertas partes en particular sobre las cuales las condiciones siempre van a predicar, y por lo tanto, si se puede reducir o acotar el acople entre las condiciones y el modelo.

5.2.1. Módulo de movilidad espacial

El modelo de movilidad espacial configura, para cada persona, la forma en la que se mueve. Consiste en dos partes, una que se inyecta en el modelo y describe el comportamiento del módulo, y otra que contiene el estado de cada nodo, y de la que cada uno recibe su propia copia. En todos los casos contiene su posición, pero dependiendo del modelo puede contener su velocidad u otros datos relevantes para decidir cómo se mueve.

El protocolo de la información de movilidad tiene dos métodos: `move()` que es llamado en cada paso para que cada nodo avance un paso, y `position()`, que informa la posición del nodo.

En cada instante de tiempo, el nodo realiza la acción de moverse en el espacio mediante el método `node.move()`, lo que modifica su posición. Esta posición puede ser consultada por otros módulos en sus condiciones mediante la clase `Persona`.

El módulo de movilidad, por otro lado, puede crear una información de movilidad para asignar a un nodo mediante el método `new_mobility_information(n)` (donde `n` es un nodo), y responder si dos nodos están *cerca* mediante el método `nodes_are_in_close_contact(n1, n2)`, donde `n1` y `n2` son dos nodos. Es decir, contiene la definición de cercanía para el modelo. Este último método es usado por el módulo de propagación para saber si existe un contacto entre determinado par de nodos.

Existe un proceso de movilidad nulo, que se utiliza cuando no se quiere incluir el factor de movilidad en el modelo (siempre responde que un par de nodos está suficientemente cerca para un contacto, todos los nodos están en la posición $(0, 0)$ y no hacen nada cuando se les indica que se muevan).

La figura 5.1 es un diagrama de secuencia que ilustra la activación del módulo en un ciclo particular para cada uno de los nodos.

5.2.2. Módulo de propagación de la infección

El módulo de propagación se llama en cada instante de tiempo, para cada par de nodos del sistema, en el método del simulador `run_simulation_step()`, para determinar si entre este par de nodos existe posibilidad de propagación o no.

No se puede acotar esta operación solamente a los nodos vecinos de cada persona en el grafo (lo cual aceleraría el cálculo) social porque para las propagaciones por aire la condición de cercanía física alcanza para generar un contagio. Igualmente existe la opción

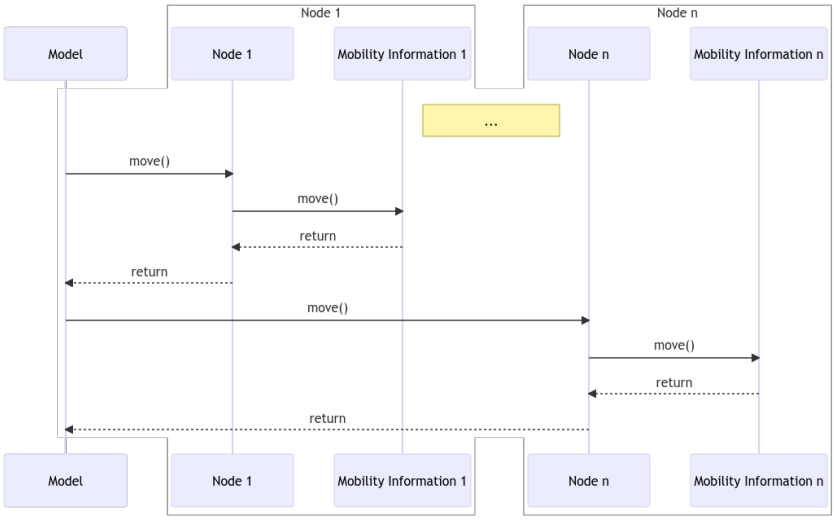


Fig. 5.1: Diagrama de secuencia para el módulo de movilidad espacial

de subclasificar el método `node_neighbors_of(n)` (que lista los vecinos válidos de un nodo `n` dado) en caso de que se requiera una optimización y tenga sentido aplicarla.

El método `propagate_infection_from_node(n)` (donde `n` es un nodo) es el método principal del simulador que llama al módulo, que recorre los vecinos de `n` y, para cada uno de ellos, llama al método `spread(n2)` (donde `n2` es un nodo vecino) para evaluar si la infección se propaga. Esta evaluación se da mediante la evaluación (`value()`) de una *condición de propagación*[4] que se configura en el módulo en su creación.

La figura 5.2 es un diagrama de secuencia que ilustra la activación del módulo en un ciclo particular para el j-ésimo nodo. El resto de los nodos tienen un comportamiento similar.

5.2.3. Módulo de progresión de la infección

La forma en la que la infección se desarrolla se inyecta en el modelo mediante la clase `InfectionConfiguration`. Este módulo puede crear una nueva infección para asignar a un nodo, y conoce los posibles estados de salud de cada persona.

El método `new_infection` genera una instancia nueva de infección para un nodo en particular. Esta infección comenzará en el estado **I** y tendrá un mapa de posibles resoluciones de la forma `{estado: condición}`. Por ejemplo, en el caso del modelo $SIRD$, un mapa podría ser:

```
b = Bernoulli(p=0.001)
c = CountDown(n=15)
possible_resolutions = {
  Recovered(): IndicatorDevelopmentCondition(b),
  Deceased(): IndicatorDevelopmentCondition(c),
}
```

[4] Una *condición de propagación* es una condición que se evalúa sobre un par de nodos. Ver más sobre condiciones en la sección 5.3.

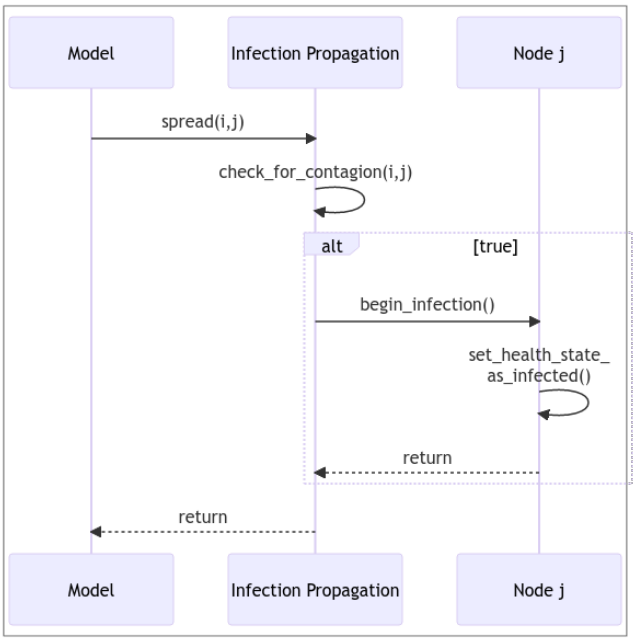


Fig. 5.2: Diagrama de secuencia para el módulo de propagación de la infección

Este mapa se lee como: "mientras la persona está infectada, cada instante de tiempo tiene una probabilidad de $0,1\,\%$ de morir en ese instante, y luego de 15 instantes se recupera".

Esta es una primera iteración que, por ejemplo, no tiene la capacidad de incorporar un estado latente, porque cuando un nodo se infecta pasa a estar en un estado fijo, y luego esa situación se resuelve hacia **R** o **D**, pero estos estados ya no son consultados para seguir transicionando. Para que el modelo pueda tener esa capacidad se requiere una nueva iteración.

El módulo además conoce una *condición de parada* que corresponde a la infección se pasa por parámetro, ya que las simulaciones de los procesos **SI** se detienen idealmente cuando toda (o una gran proporción) de la población está infectada, ya que el estado absorbente es **I**, mientras que en los procesos **SIR** o **SIRD** en los que la infección siempre se resuelve de alguna manera, se puede detener cuando la cuenta de infecciones llega a cero (o baja lo suficiente) luego de subir inicialmente.

En la tesis se implementaron 3 subclases que corresponden a los modelos **SI**, **SIR** y **SIRD**. Cada una de ellas conoce cuáles son las resoluciones posibles del estado infeccioso y la condición que se debe cumplir para realizar esa transición.

Cada estado de la infección tiene su propia clase. Las infecciones son binarias (es decir, se está infectade o no: no existe noción de gravedad, síntomas o carga de la infección). Los estados tienen preguntas básicas, como por ejemplo si corresponde a un estado contagioso, si es susceptible, o si le huésped está vive (importante, por ejemplo, a fines de calcular el desplazamiento espacial).

Se podría complejizar la infección introduciendo nuevos estados infecciosos. Para esto último hay que definir las transiciones entre estados de la forma que corresponda: el método `state_on_infection_spread` del estado **S** tiene que resolver cuál de los estados infecciosos se aplica en cada caso. También se puede agregar estado a **I** para dar una noción de carga

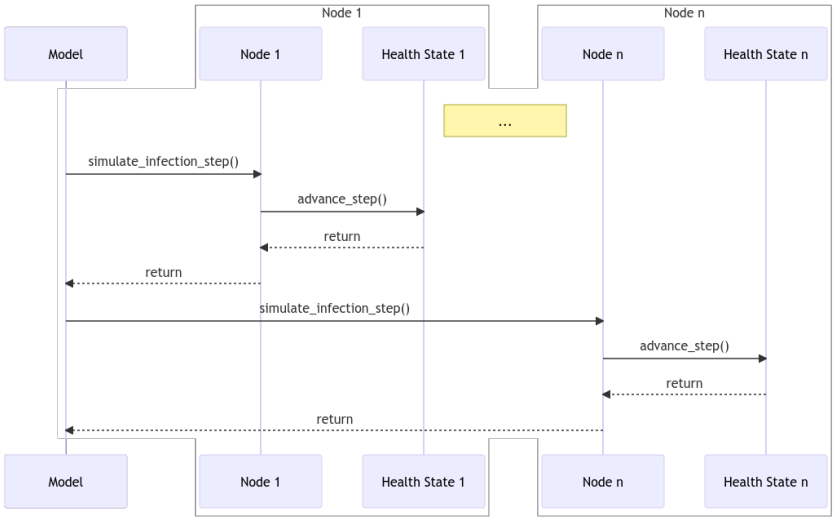


Fig. 5.3: Diagrama de secuencia para el módulo de progresión de la infección

viral que complejice la interacción del contagio.

La figura 5.3 es un diagrama de secuencia que ilustra la activación del módulo en un ciclo particular para cada uno de los nodos.

5.3. Condiciones sobre el modelo

En el modelo existe la idea de *Condiciones* sobre el modelo, que representan predicados sobre éste. En el modelo hay jerarquías de clases distintas, ya que algunas condiciones deben ser evaluadas sobre un nodo solamente (en el caso del avance de la infección de un nodo particular), sobre dos (en el caso de las condiciones de propagación a través de la red) o sobre ningún nodo en particular (en el caso de las condiciones de parada de la simulación).

En todos los casos el protocolo es que existe un método `value` que tiene un tipo de retorno booleano.

Cada jerarquía de condiciones sigue el patrón de diseño *Composite*[5], en el que la intención es generar una jerarquía de objetos en forma de árbol, tales que algunos son hojas y resuelven su comportamiento independientemente, y otras están compuestas por otras entidades y resuelven su comportamiento a partir de componer el de sus distintos objetos hijos. En este caso las condiciones compuestas son condiciones *algebraicas*, es decir, operaciones booleanas sobre otras condiciones (en el prototipo se implementaron la negación, la conjunción y la disyunción).

Algunos ejemplos de condiciones hojas `C` implementadas, para distintas aridades son:

- `C1.value(n1, n2)` si `n1` es vecino de `n2` en el grafo social.
- `C2.value(n)` si el nodo-persona `n` se encuentra en estado **S**.
- `C3.value()` si la cantidad de personas en **I** es cero.

[5] Ver sección 4.3 de [Gam+95].

5.4. Ubicación de la información sobre los nodos

Un problema atacado en el desarrollo de la tesis es cómo manejar la información de cada nodo-persona dentro del modelo. Como la estructura del prototipo es modular, en principio la información que constituye el *estado* de cada nodo en el sistema es variable. Por otro lado, es una característica deseable del modelo que cada instancia de la clase *Persona* pueda responder mensajes acerca de cuál es su estado, de tal forma que dé una interfaz única por la cual el sistema puede acceder a los distintos aspectos de la información de una persona, ya que agrupar esta información es la responsabilidad de las instancias de Persona.

Une podría solucionar este problema haciendo que cada Persona contenga una colección de datos no jerarquizados que informen todo sobre ella. Por ejemplo, que Persona tenga como colaboradores internos datos tan variados como su velocidad, su edad, su posición y en qué estado de la infección está. El problema de que esos campos estén definidos en ese lugar es que los módulos no tienen coherencia interna. Es decir, el comportamiento del módulo se encuentra en éste, pero los datos que la persona necesita saber de sí para que este módulo funcione se encuentran fuera, en la instancia de Persona. Esto hace que si se quiere modificar el módulo, o cambiar por otro, haya que redefinir estos datos en la clase Persona también, y esto constituye un acoplamiento.

Si por otro lado hacemos que cada módulo sepa su información correspondiente, accedida por ejemplo mediante una clave que corresponda al ID de cada nodo, surge un problema distinto: a la hora de consultar datos sobre las personas hay que referirse a distintas partes del sistema en función de cuál dato se necesite. Por ejemplo, si el módulo de movilidad establece que las personas fallecidas no se mueven (una condición bastante razonable), debe consultar al módulo de avance de la infección para conocer ese dato, lo cual saca responsabilidad a la clase Persona. Esta clase pierde su cohesión y de hecho se vuelve inútil al no centralizar la información de cada persona.

La solución generada en la tesis es hacer que cada módulo del sistema encapsule la información que aporta a cada nodo en un *aspecto*[6], que es un objeto que los módulos inyectan en cada instancia de Persona al inicio de la simulación. Esto hace que la Persona, al conocer todos sus aspectos (los tiene como colaboradores internos), pueda contestar preguntas acerca de su propio estado, con relación a todos los módulos. Así, la forma de los datos que cada módulo tiene es responsabilidad del módulo, y el conocer todos los datos para centralizar las consultas, de Persona. Además, estos módulos ofrecen protocolo para cambiar este estado (por ejemplo, el aspecto de movilidad de una persona tiene un método `move` que cambia su posición, pero no puede contestar preguntas o realizar acciones que también dependan de otros nodos).

Esta estrategia se acerca al paradigma de tener una persona "configurable", pero sigue existiendo un acople, ya que los métodos para hacerle preguntas a las Personas siguen teniendo que ser modificados en función de cuáles datos constituyen su estado (que a su vez depende de los módulos). Sin embargo, las instancias de Persona son un punto donde se puede interceptar cualquier pregunta que se le haga a la Persona y responderla como resulte conveniente, por lo cual el sistema resultante es más flexible, y además oculta detrás de la Persona las interfaces que ofrecen los aspectos, funcionando de adaptador entre estos aspectos y el sistema, y confinando el acople a las implementaciones de los métodos de

[6] Pensamos que el mejor nombre es el de *aspectos*, pero en el código se llaman *módulos de información*, que es un mal nombre porque no da cuenta de que estos objetos también codifican comportamiento.

Persona.

5.5. Cosmovisiones en simulación

A la hora de programar un mecanismo de simulación en el tiempo, existen tres *estrategias*, llamadas también *cosmovisiones*[7], en particular para solucionar el problema del paso del tiempo.

Estas son:

- *activity-scanning* (o de escaneo de actividades): divide la simulación en pasos de tiempo discreto y, en cada uno de ellos, actualiza el estado y desencadena las acciones de cada una de las entidades del sistema si corresponde.
- *event-scheduling* (o de agenda de eventos): encola eventos en una cola global, a ocurrir en un momento particular en el futuro, y la simulación avanza sobre esta cola resolviendo los eventos presentes en orden de tiempo creciente.
- *process-interaction* (o de interacción entre procesos): resuelve interacciones entre procesos que se suspenden y reanudan en función de distintas condiciones del sistema.

En la tesis usamos una estrategia de *escaneo de actividades*, principalmente por una cuestión de alcance, ya que esta estrategia resulta más simple de programar, y permitió aumentar la complejidad del prototipo en otras partes del modelo.

Si comparamos con la cosmovisión de agenda de eventos, puede haber una pérdida de flexibilidad en algunos casos. Por ejemplo, si se usa una estrategia por eventos y se agenda que una persona se recuperará de su infección en un día determinado, quizás muere antes y ese evento queda sin efecto, o quizás recibe atención médica en un momento determinado por una campaña que el sistema decide después de haber determinado cuándo se recuperará. Se puede salvar esta dificultad descartando eventos que resultan no ser relevantes pero lidiar con esos casos es una complicación en el código.

La simulación por interacción entre procesos parece una estrategia que puede mantener esta flexibilidad ya que no se agendan eventos fijos en el futuro. En este caso se debe resolver el problema de que los procesos dormidos no puedan generar cambios en sí mismos (por ejemplo, una persona que está suspendida esperando el estímulo de un sistema externo debe curarse de su infección).

Cada una de las estrategias tiene sus propias limitaciones y sus consideraciones a la hora de implementarse. En particular la estrategia elegida trabaja sobre todos los pares de nodos posibles para determinar si existe un contagio, en cada paso de la simulación. Es decir, cada paso es $O(N^2)$ (suponiendo que la cantidad de operaciones que se hace para cada par de nodos es $O(1)$). Esto resulta prohibitivo para trabajar con datasets de órdenes de nodos más grandes que miles en una computadora de uso personal.

Es una continuación interesante modificar el prototipo para poder comparar tiempos de corrida entre las distintas estrategias, y sería una forma de extender este trabajo. Quedó fuera del alcance de la tesis ya que optamos por implementar únicamente *escaneo de actividades*, por tiempo y por simplicidad, pero para que el prototipo pueda realmente atacar problemas con una gran cantidad de nodos hace falta explorar las otras cosmovisiones.

[7] Ver sección 1.2 de [Ban+14].

6. USO DE DATOS DE ECOBICI COMO INPUT DEL MODELO GRAVITACIONAL

Además del estudio de distintas formas de generación de datos sintéticos, en el contexto de la tesis se decidió poner a prueba el modelo con un dataset real, para mostrar un caso de uso posible para el sistema. En esta sección se describen el dataset utilizado, el preprocesamiento que se le dio y la forma en la que estos datos fueron incorporados al modelo.

El dataset corresponde, a grandes rasgos, a un conjunto de viajes en bicicleta, diferenciados por usuarie y geolocalizados en distintos puntos de la Ciudad de Buenos Aires, Argentina, a lo largo del año 2019. Los trayectos registrados fueron realizados en bicicletas de la empresa TemBici, en la Ciudad de Buenos Aires.

Esta empresa tiene actualmente un convenio con el GCBA (Gobierno de la Ciudad de Buenos Aires) y son el único sistema de bicicletas integrado que tiene la Ciudad. Existen distintos puntos de retiro/entrega de bicicletas distribuidos por la ciudad, y les usuaries pueden, logueándose en la aplicación de TemBici, usar las bicicletas para viajar entre dos estaciones (pueden ser distintas, o puede ser la misma en el caso de que la bicicleta se use para ir y volver al mismo lugar).

El uso que se le dio a estos datos es inferir, para las personas que utilizaron el servicio durante este año, cuáles son las geolocalizaciones de su *casa* y su *trabajo* (en una definición de estos dos términos que es especifica en la sección 6.2). Luego, estos puntos son incorporados al modelo para realizar una simulación de un proceso de difusión de una infección. En este proceso, los nodos se mueven siguiendo un proceso gravitacional de dos puntos entre los puntos definidos como su casa y su trabajo según el dataset de viajes en bicicleta.

El recorrido de las bicicletas no se encuentra geolocalizado, pero sí las estaciones de inicio y fin. Además el viaje tiene como datos los momentos de retiro y devolución y la identificación de le usuarie que efectuó el viaje.

En la sección 6.1 se explica cuál es la relevancia del modelo gravitacional de dos puntos[1] en este caso, y por qué la casa y el trabajo de una persona son puntos razonables de input para este modelo de movilidad.

En la sección 6.2 se define qué se considera la *casa* y el *trabajo* de une usuarie, y ciertas salvedades que se deben tomar a la hora de trabajar con este tipo de datos geolocalizados, y en la 6.3 se describe cómo se utilizó el conjunto de viajes de una persona dada para inferir estos puntos.

En la sección 6.4 se describe el contenido de las columnas del dataset que fueron utilizadas en la tesis, así como métricas importantes de las distintas tablas de la base (cantidad de líneas, proporción de líneas válidas o cantidad de usuaries y estaciones de bicicleta, por ejemplo).

6.1. Consideraciones sobre el modelo gravitacional de dos puntos

En esta sección se discute la relevancia del modelo gravitacional de dos puntos para caracterizar los movimientos entre personas para este dataset en particular, y las limitaciones que tiene resumir las trayectorias típicas de personas al usarlo.

[1] En la sección 3.3 se explica qué es un modelo gravitacional en el marco de la tesis.

El uso del modelo se basa en la idea de que la mayoría de la gente, en un día laboral (de lunes a jueves, ya que el viernes tiene un patrón ligeramente distinto) gravita principalmente entre su casa (durante la noche) y su trabajo (durante el día) (en [Sar+17]).

Por esta razón, se puede conseguir una buena aproximación del movimiento de un agente infiriendo estos dos puntos de interés y pautando horarios (momentos de la simulación) en los que se dará el movimiento entre su casa inferida y su trabajo inferido. En este sentido hay consideraciones que se deben tomar:

Por un lado, se debe elegir **en qué momento se realizan los viajes**. En este caso, por simplicidad, se eligió que todos los nodos se movieran en los mismos momentos del día, pero agregando más información al modelo esto se puede complejizar haciendo que, por ejemplo, cada persona tenga su propio horario de salida al trabajo, observando sus horarios de movimiento y construyendo a partir de ellos un día "típico" para esa persona. En esta tesis se utilizaron horarios estándar guiados por las franjas horarias propuestas en [Sar+17][2].

Por otro lado, se debe tener en cuenta **la diferencia entre la casa y el trabajo efectivos de una persona y los puntos inferidos por el modelo**, y entre el viaje efectivamente realizado y el que podemos inferir solamente sabiendo el origen y el destino. Respecto de esto se plantean varias cuestiones.

Primero, no se pueden observar viajes en otros medios de transporte sin cruzar datos con otra fuente. Eso hace que si una persona utiliza la bicicleta en combinación con otro medio de transporte sea indistinguible de una persona que se mueve en bicicleta solamente. Además, obviamente, otra limitación del dataset usado es que toda persona que no se mueva por el sistema de bicicletas es invisible para el modelo. En este sentido no es el objetivo de este caso de uso dar un análisis exhaustivo de patrones de movilidad de la Ciudad de Buenos Aires, aunque si se realiza un análisis similar sobre una base de datos más completa sí se podría hacer.

Segundo, incluso asumiendo que se está observando la totalidad del viaje, los puntos observados no son su casa y su trabajo reales. Por ejemplo, en un viaje de ida al trabajo, la persona realiza siempre un trayecto de caminata entre su casa real y la estación donde la observamos partir, y entre la estación a la que llega y su trabajo real. Al tomar la geolocalización de las estaciones como los puntos de caracterización de su movimiento estamos obviando la dispersión que se da después de que las personas llegan a una estación de destino. Por eso, el modelo puede afirmar que dos personas estuvieron en contacto durante mucho tiempo cuando en realidad se separaron de la estación al llegar.

Sobre este punto, si se quiere modelar de alguna forma esta dispersión, surgen por lo menos dos maneras. La primera es que el modelo de movilidad no haga que las personas se queden quietas en la estación al llegar, sino que se dirijan a algún punto cercano al azar a distancia caminable (habiendo definido esta distancia para cada persona). La segunda es decrementar artificialmente la probabilidad de contagio entre dos personas para tener en cuenta que existe una probabilidad de que estas personas ya no estén en contacto. Estas opciones no fueron implementadas pero podrían serlo: el ruido en la posición se puede implementar en el módulo de movilidad y el decremento de la probabilidad, directamente configurando el módulo de contagio en ese sentido.

Tercero, existe una limitación del modelo gravitacional de dos puntos, que es que por construcción no permite explicar trayectorias que no estén bien aproximadas por dos pun-

[2] Estas franjas horarias a su vez están basadas en los horarios pico y valle detectados por la Encuesta de Movilidad Doméstica (ENMODO) 2009-2010.

tos. Esto puede parecer una obviedad, pero esta suposición trabaja sobre la hipótesis de que un día típico de una persona transcurre entre una única casa y un único trabajo (o escuela). Existen muchas realidades que no son bien representadas por esta hipótesis: jóvenes con xadres separades que tienen dos casas, personas con niñeces a cargo que les llevan cotidianamente al colegio, estudiantes adultes de cualquier nivel educativo, o personas que tienen más de un trabajo en ubicaciones distintas, por ejemplo. En particular, al hablar de tareas de cuidado, se introduce un sesgo de géneros en el que el modelo explica con más facilidad trayectorias de varones que de mujeres. No se cuantificó sobre el dataset una caracterización de las trayectorias en este sentido pero es importante ser conscientes de los sesgos introducidos en los análisis, al menos cualitativamente.

En cuarto lugar, no se realizó una integración con el mapa de calles de la Ciudad de Buenos Aires, sino que se asumió que todos los viajes se realizan de forma instantánea o a vuelo de pájaro. En este sentido están disponibles de forma pública el mapa de Open Street Map, una plataforma de mapeo colaborativa, y una herramienta de ruteo a través del sistema de calles (OSRM). Ambas herramientas son de software libre.

En la sección siguiente discutiremos cómo usamos los datos de viajes en bicicleta para inferir estos dos puntos para los usuarios de las bicicletas del GCBA.

6.2. Patrones de movilidad en distintos días de la semana

En esta sección proponemos una simplificación de la idea de "día típico" para una persona para poder combinar los viajes realizados en días distintos y utilizar esta información para definir de qué manera se resumirá el conjunto de viajes de una persona en su *casa* y su *trabajo*.

El modelo propuesto en [Sar+17], separa:

- Los días en 4 patrones diferentes: lunes a jueves (día laboral típico), viernes (que tiene un patrón disitnto por la noche), sábado y domingo. Se considera que la gente en su mayoría trabaja de lunes a jueves por lo que se utilizaron sólo viajes ocurridos en esos días como un indicador de viajes entre casa y trabajo.

- Los horarios en 4 franjas: la mañana (5:00 a 11:00), el mediodía (11:00 a 15:00), la tarde (15:00 a 20:00) y la noche (10:00 a 5:00 del día siguiente).

En [Mon+16], un estudio posterior que usa la misma división de días y horarios[3], el dataset no registra viajes sino que, mediante llamados telefónicos, se registran apariciones puntuales de cada persona. Por lo tanto, si se divide un día hábil en las 4 franjas horarias propuestas, considera que la persona se encuentra en el trabajo en la franja del mediodía, y en su casa en la franja de la noche, y toma el conjunto de estas apariciones respectivamente como muestra para estimar estas ubicaciones.

En cambio, en este estudio, el dataset tiene viajes, por lo cual se utilizó una aproximación distinta: se consideró que los viajes son de ida al trabajo en la franja de la mañana, y de vuelta del trabajo durante la tarde. Además, como cada viaje tiene dos momentos asociados (el inicio y el fin), se tomó otra definición: la de considerar que un viaje *ocurrió durante una franja horaria* cuando se encontró completamente contenido en ella.

[3] Este estudio usa la misma división salvo porque establece el principio de la mañana y el fin de la noche a las 6:00 en vez de a las 5:00

6.3. Cálculo de H y W

Se define una ruta en bicicleta como un objeto con los siguientes campos:

- Le *usuarie* que realizó el viaje (campo accedido con el mensaje .usr)
- La estación de *origen* (campo accedido con el mensaje .org)
- La estación de *destino* (campo accedido con el mensaje .dst)
- El instante de tiempo de inicio del viaje
- El instante de tiempo de fin del viaje

Usando estas rutas, para cada usuarie n se realizó un análisis que tiene como resultado estimaciones de la geolocalización de su hogar h_n y de su trabajo t_n.

Se filtraron todos los viajes que constituían un ciclo (es decir, que empiezan y terminan en el mismo lugar) para que las ubicaciones de casa y trabajo sean distintas.

Se utilizaron solamente viajes registrados entre los días lunes y jueves inclusive, para obtener patrones típicos de días laborales, de acuerdo con la sección 6.2.

Sean R_m las rutas que ocurren por la mañana y R_t, a la tarde. Si un viaje ocurre a la mañana, se toma a su origen como evidencia de la ubicación de h_n, y su destino, de t_n. En cambio si una ruta ocurre a la tarde se considera un viaje en sentido contrario.

Es decir, se definen las muestras que se usarán para calcular h_n y t_n de la siguiente manera:

- $H_n = \{r.org | r \in R_m \wedge r.usr = n\} \cup \{r.dst | r \in R_t \wedge r.usr = n\}$
- $T_n = \{r.dst | r \in R_m \wedge r.usr = n\} \cup \{r.org | r \in R_t \wedge r.usr = n\}$

Luego se definieron las ubicaciones inferidas de hogar y trabajo como el elemento que más frecuentemente aparece en el conjunto correspondiente[4]. La cantidad de usuaries que contaban con suficiente información en las franjas horarias definidas es de 205954, que es un $48,53\,\%$ de les usuaries totales.

6.4. Estructura del dataset

El dataset utilizado son las listas de estaciones[5] por un lado, y el registro de viajes durante 2019[6] por otro. Todos los archivos se obtuvieron del portal de datos abiertos del GCBA.

Existen dos listas de estaciones porque conviven dos sistemas de nomenclatura distintas, y existen estaciones que sólo están en el primer sistema y otras que sólo están en el segundo. Se tomó la decisión de usar las estaciones del sistema nuevo, y complementar con el sistema viejo cuando no existiera en el nuevo una estación con este código.

A continuación se muestran los esquemas de datos, teniendo en cuenta que solamente se mencionan las columnas relevantes para la experimentación:

[4] El manejo de cuál valor de usa en caso de empate corresponde a la biblioteca *collections* de Python y no está especificada en la documentación de dicha biblioteca.

[5] Disponibles en https://data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas

[6] Disponibles en https://data.buenosaires.gob.ar/dataset/bicicletas-publicas

Para el sistema viejo de geolocalización de estaciones:

id_estacion	Número real, siempre entero. Se lo interpretó como un entero
nombre_estacion	Texto que se usó para corroborar manualmente la correspondencia de ids
long_estacion	Número real, en grados
lat_estacion	Número real, en grados

Para el sistema nuevo geolocalización de estaciones:

codigo	Número entero.
nombre	Texto que se usó para corroborar manualmente la correspondencia de ids
WKT	Punto en formato WKT, con la longitud en la primera coordenada y la latitud en la segunda (números reales, en grados)

El sistema viejo tiene 6 líneas inválidas y 198 válidas. El nuevo tiene sólo 229 estaciones válidas. En total la unión de ambas listas tiene 282 estaciones únicas geolocalizadas.

Para el listado de recorridos realizados:

id_usuario	Número entero.
fecha_estacion_origen	Instante de tiempo (fecha y hora) en la que comenzó el viaje.
id_estacion_origen	Número entero que identifica a la estación origen del viaje.
fecha_estacion_destino	Instante de tiempo (fecha y hora) en la que terminó el viaje.
id_estacion_destino	Número entero que identifica a la estación destino del viaje.

El listado tiene en total 6367314 líneas, de las cuales 953421 ($14,97\,\%$) corresponden a data mal formada y 5413893 ($85,03\,\%$) a recorridos válidos.

La cantidad de usuarios totales es de 424383.

7. EXPERIMENTOS

En esta sección se presentan los distintos experimentos que se realizaron en el transcurso de la tesis.

El primer experimento es una comparación entre el modelo de Erdös-Rényi de generación de grafos aleatorios, y el algoritmo de preferential attachment presentado por Barabási. Simulamos el mismo proceso de difusión sobre grafos de tamaño y densidad similar generados por cada algoritmo, y comparamos las curvas características de la simulación para cada método. Este experimento se encuentra en la sección 7.1.

Luego se introducen dos experimentos que constituyen variaciones sobre las hipótesis de mezcla homogénea y compartimentarización, donde el cálculo analítico es difícil. Ambos experimentos están inspirados en el caso de uso de una infección entre seres humanos.

En el primero se plantea la situación de que existe un virus que se transmite por el aire, y que cierta proporción de personas en el sistema usan tapabocas y otras no. La probabilidad de contraer la infección varía en función de si alguna o ambas personas de la interacción tienen un tapabocas puesto. Se observa cómo varían las curvas características de una simulación para distintas proporciones de la población que adoptan la medida. Para conseguir modelar esta medida se modifica el módulo de propagación de la infección con condiciones en función de metadatos introducidos en cada nodo. Este experimento se presenta en la sección 7.2.

En la segunda variación se asume que existe una vacuna para la infección, que, en particular, acorta el tiempo de recuperación. Hay cierta proporción fija de la población que se encuentra vacunada durante toda la simulación y otra que no. Observamos cómo varían las curvas características en función de cuál es la proporción de la población que se encuentra vacunada. No se realizó un experimento en el que la vacunación fuera progresiva a lo largo del tiempo pero es un siguiente paso natural a partir de este experimento. Este experimento se presenta en la sección 7.3.

Para todas las pruebas de los experimentos anteriores se realizaron varias repeticiones del mismo experimento para poder generar una distribución de resultados, porque tanto la generación de grafos como el proceso de simulación son procesos aleatorios, entonces no alcanza con una sola repetición. Se modificó la herramienta de generación de gráficos para procesar una muestra de varias repeticiones y considerarlas una misma serie. Para cada una, graficamos la mediana con una línea, y el área entre los percentiles 25 y 75 de la muestra correspondiente.

En un último experimento modelamos un proceso de movilidad humana gravitacional de dos puntos usando el módulo correspondiente del software producido. Analizamos el comportamiento de un proceso de contagio sobre un sistema de movilidad humana real, extraído de la base de datos pública del gobierno de la Ciudad de Buenos Aires (GCBA), correspondiente a los viajes en bicicleta realizados por usuaries del programa de bicicletas del GCBA. Los datos se utilizan para inferir, para cada persona, los dos puntos de su propio sistema gravitacional (que corresponden con su casa y su trabajo), y luego se realiza una simulación de un proceso de contagio sobre los nodos, mientras se mueven por el espacio entre estos dos puntos. El experimento y sus resultados se presentan en la sección 7.4. Hay distintas referencias a lo largo de la tesis para más información sobre el experimento: en la sección 3.3 definimos qué se entiende por modelo gravitacional, y a lo largo del capítulo 6

describimos con más detalle tanto la estructura del dataset como el proceso de inferencia de los puntos de los sistemas gravitacionales de los nodos.

7.1. Comparación de grafos sociales sintéticos

El primer experimento fue una comparación entre los distintos modelos de grafos sociales, para observar de qué manera la topología de las redes generadas sintéticamente afecta el proceso de difusión.

En el capítulo 4 se caracterizaron dos distintas maneras de generar grafos sociales sintéticos que emularan las propiedades de los grafos reales. En particular, el desarrollo del modelo independiente de escala es una forma de corregir la diferencia entre la distribución de grados observada en los grafos sociales reales y la fabricada por el modelo de Erdös-Rényi, en particular generando una cola más pesada.

Esto tiene varias consecuencias, y en particular cambia el orden de magnitud de la distancia esperada entre dos nodos, en función del tamaño del grafo: en el caso del modelo de ER, la distancia esperada es $O(\ln(N))$ mientras que en un grafo generado usando preferential attachment, la distancia corresponde con una función de ley de potencia con $\gamma = 3$. Esto entra dentro del régimen de punto crítico para γ y da una complejidad de $O\left(\frac{\ln(N)}{\ln\ln(N)}\right)$[1].

En el contexto de un proceso de difusión, se esperaría intuitivamente que la velocidad de transmisión del agente sea mayor mientras menores son las distancias esperadas entre distintos nodos. Para comprobar esto se generaron dos grafos con la misma cantidad de nodos N, y densidades p similares. Uno de ellos fue generado usando un proceso de Erdös-Rényi, y el otro uno de preferential attachment.

Para el grafo de preferential attachment se eligieron valores de N (cantidad de nodos totales del grafo) y m (cantidad de nodos de la clique inicial) tales que $N >> m$. Se generó una clique inicial de m nodos y luego se agregaron los $N - m$ restantes con $k = m$ (es decir uniéndolos, sin autoejes y sin multiejes, con m de los anteriores siguiendo la distribución de probabilidad propuesta por Barabási).

En este grafo, la cantidad de aristas obtenidas de esta forma es

$$\frac{m(m-1)}{2} + (N-m)m$$
$$m\left(\frac{m-1}{2} + N - m\right)$$
$$m\left(\frac{1-m}{2} + N\right)$$

Para el grafo de ER se tomó el mismo N que en el caso anterior y se tomó p de tal manera de tener la misma cantidad de aristas esperadas que para el otro grafo (en este caso no se puede asegurar exactamente la igualdad de la cantidad de aristas porque es aleatoria). La cantidad esperada de aristas es de $\left(\frac{N(N-1)}{2}\right)p$, y este valor tiene que ser igual a las aristas obtenidas en el modelo de preferential attachment.

Sea E el conjunto de aristas del modelo de preferential attachment. Entonces tiene que ocurrir:

[1] Estas fórmulas provienen del capítulo 4 de [BP16] y se introducen en la sección 4.2.

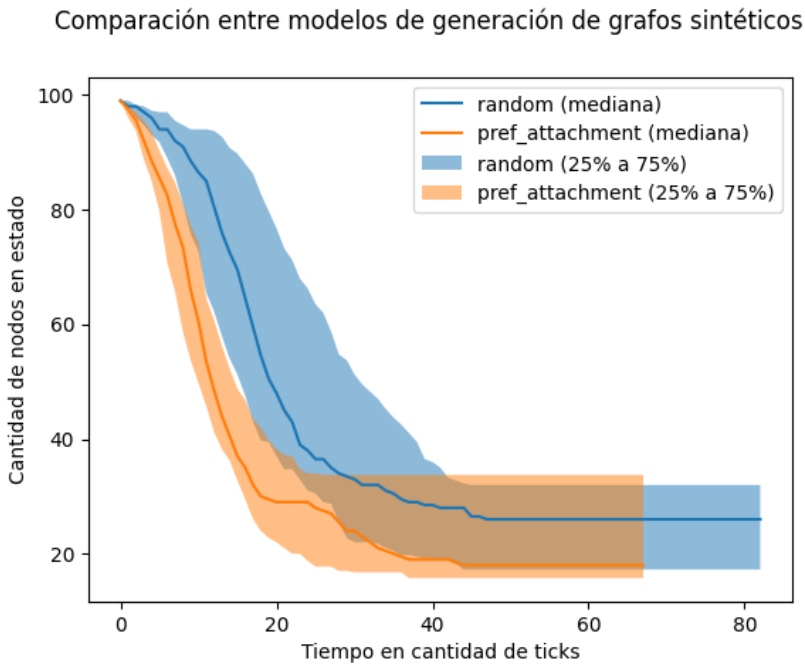


Fig. 7.1: Cantidad de nodos susceptibles en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=100, m=2, p=3.98 %.

$$|E| = \frac{N(N-1)}{2}p$$
$$p = \frac{2|E|}{N(N-1)}$$

Se presentan los resultados para distintas combinaciones de los parámetros N y m, tomando muestras de 20 corridas de cada serie. En todos los casos se generó un grafo nuevo para cada una de las repeticiones.

En las figuras 7.1, 7.2 y 7.3 se muestran los percentiles correspondientes a correr el experimento con valores pequeños. Para la generación del grafo con preferential attachment se utilizaron los parámetros $N = 100$, $m = 2$. Esto da una densidad esperada de 3,98 %, que fue usada como p para la generación de grafos de Erdös-Rényi.

En este caso se observa experimentalmente la diferencia entre ambas curvas, y se puede observar que existe una aceleración del proceso de difusión en el caso del grafo que usa preferential attachment.

Además, como se puede observar en la figuras 7.4, 7.5 y 7.6, que muestra el resultado del mismo experimento para valores más grandes, la tendencia de los grafos que se crean con preferential attachment de acelerar el proceso de difusión se mantiene.

Esta diferencia constituye una prueba empírica de la intuición desarrollada en la sección 4.2 de que los grafos independientes de la escala tienen distancias entre nodos más pequeñas que los generados de manera aleatoria, y que esta diferencia acelera el proceso de difusión.

La diferencia es observable incluso para los grafos generados usando preferential attachment, que no muestran la distribución de distancias $O(\ln \ln(N))$ (ecuación 4.12) de grafos

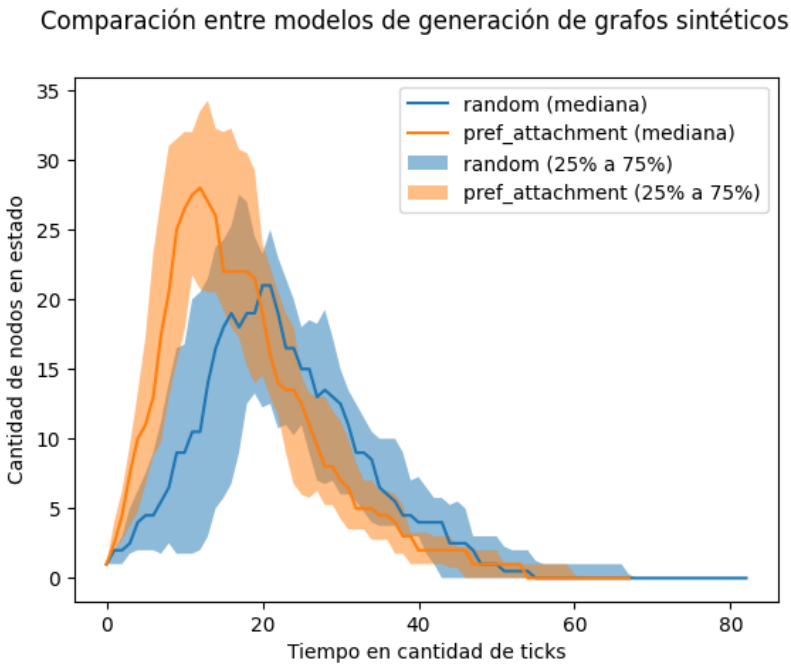


Fig. 7.2: Cantidad de nodos infectados en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=100, m=2, p=3.98 %.

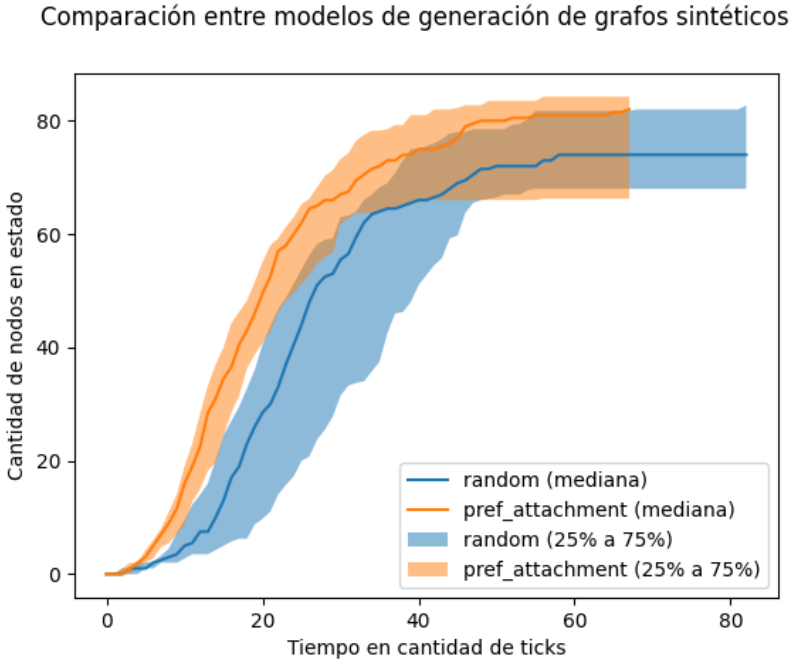


Fig. 7.3: Cantidad de nodos recuperados en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=100, m=2, p=3.98 %.

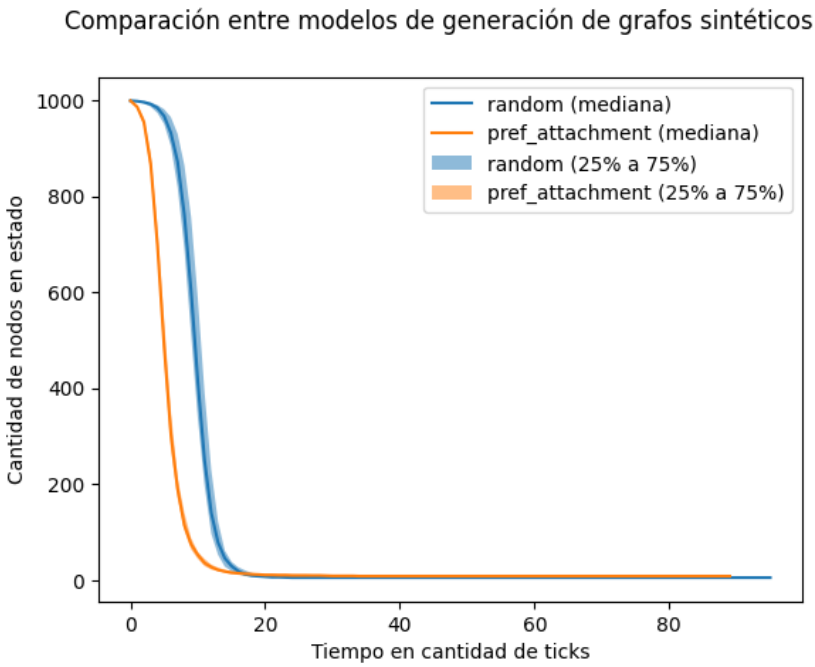


Fig. 7.4: Cantidad de nodos susceptibles en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=1000, m=5, p=0.98 %.

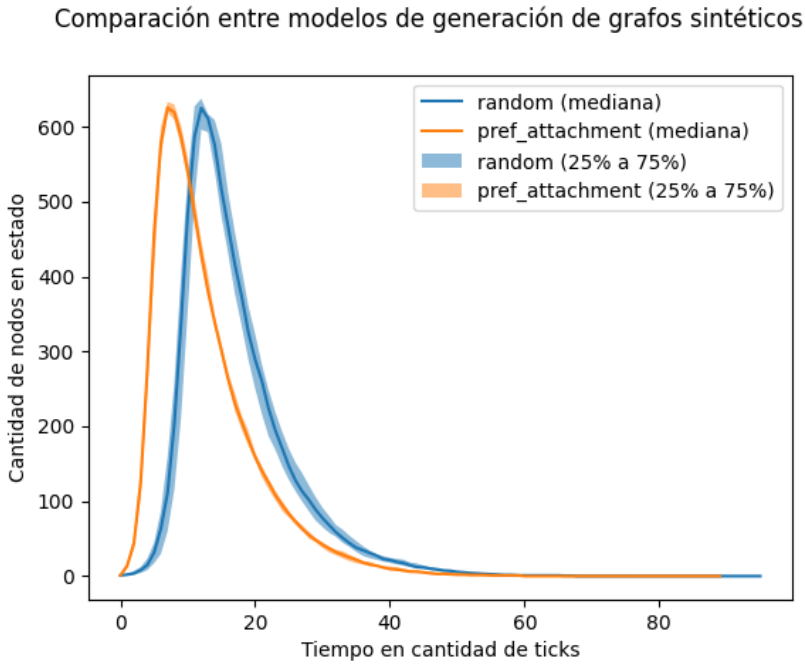


Fig. 7.5: Cantidad de nodos infectados en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=1000, m=5, p=0.98 %.

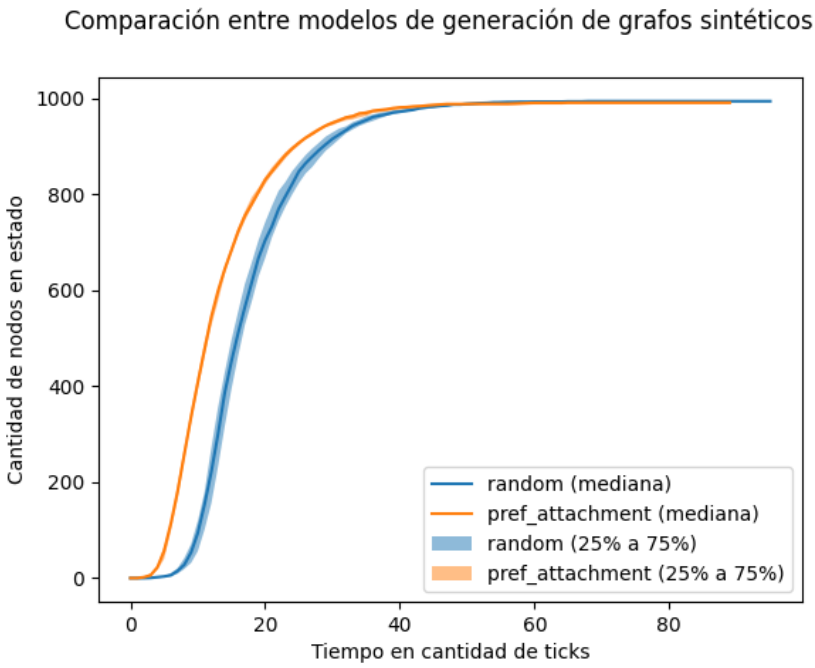


Fig. 7.6: Cantidad de nodos recuperados en función del tiempo. Grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=1000, m=5, p=0.98 %.

con γ más pequeño sino la distribución logarítmica con corrección $\langle d \rangle = O\left(\frac{\ln(N)}{\ln\ln(N)}\right)$ (ecuación 4.13). Intuitivamente, con un grafo que tenga un γ menor el proceso de debería acelerar aún más pero esta comprobación no se realizó experimentalmente.

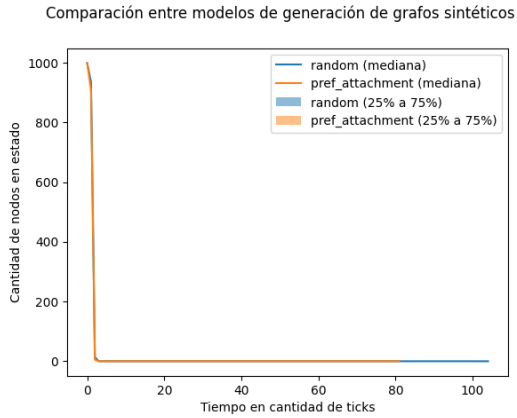
Por último se muestra una última corrida en la que, a diferencia de las anteriores, no ocurre que $N >> k$, para ver si la diferencia también se observa en este caso. Se estableció que en las redes sociales reales suele ocurrir que el grado (cantidad de personas conocidas) por cada nodo del grafo es mucho menor que la cantidad total de nodos, y que la presencia de puntos nodales afecta la distribución de las distancias entre nodos bajo el supuesto de que $N >> k$.

Como se observa en la figura 7.7, que da una densidad esperada del 51.02 %, las curvas son indistinguibles en la versión lineal del gráfico, y apenas se observa una diferencia al graficar en escala logarítmica. Esto apoya la idea de que la condición $N >> k$ es necesaria para observar una diferencia entre los modelos.

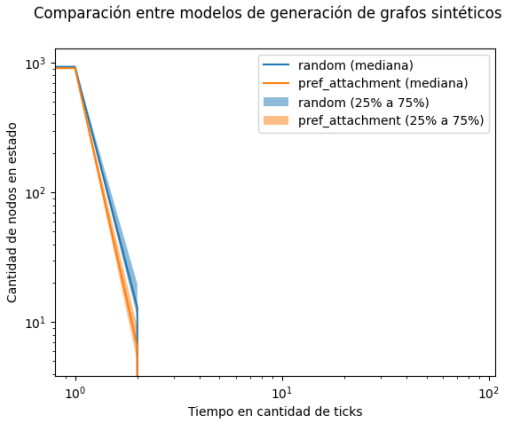
7.2. Variación en el uso de tapabocas

Este experimento trabaja sobre el uso de tapabocas en una infección humana transmitida a través del aire, pero puede aplicarse a toda situación en la que haya factores que modifiquen, para ciertos nodos, la probabilidad de transmisión de un agente en un grafo social. Por ejemplo, situaciones similares pueden ser:

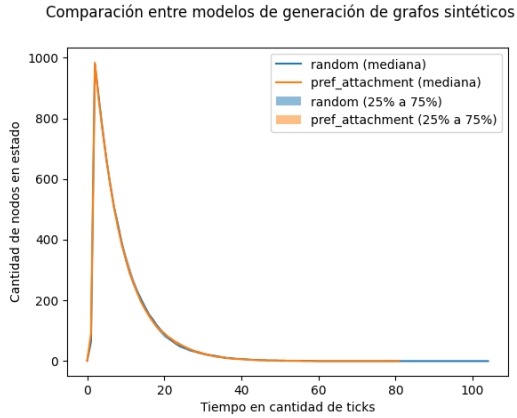
- El uso de antivirus en el caso de un virus informático
- El uso de profilaxis en una infección de transmisión sexual



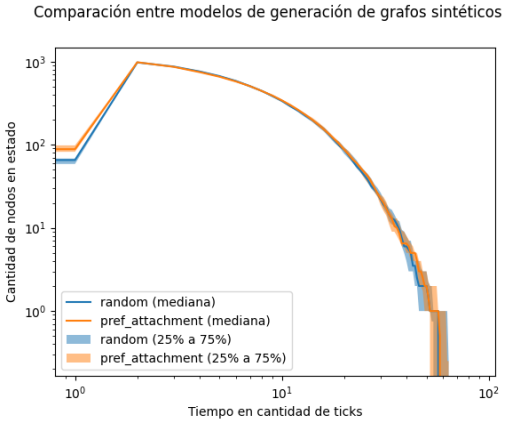
(a) Cantidad de nodos susceptibles en función del tiempo.



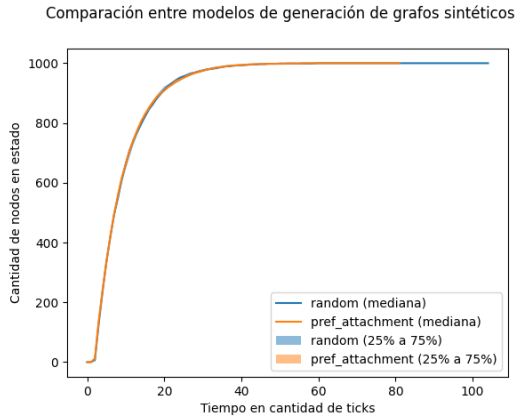
(b) Cantidad de nodos susceptibles en función del tiempo. Escala logarítmica.



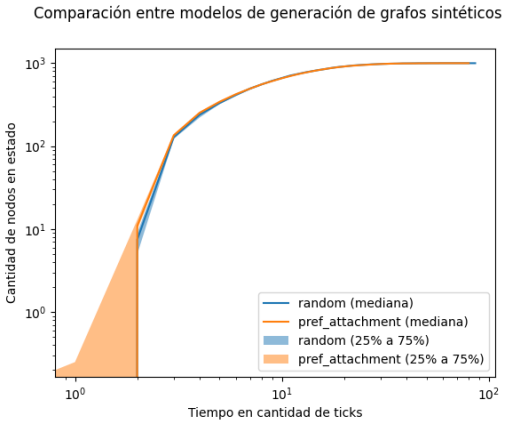
(c) Cantidad de nodos infectados en función del tiempo.



(d) Cantidad de nodos infectados en función del tiempo. Escala logarítmica.



(e) Cantidad de nodos recuperados en función del tiempo.



(f) Cantidad de nodos recuperados en función del tiempo. Escala logarítmica.

Fig. 7.7: Curvas características del proceso de difusión de la infección para grafos generados con el modelo de Erdös-Renyi (azul) y preferential attachment (naranja) para N=1000, m=300, p=51.02 %. Se muestran en escala lineal (izquierda) y en logarítmica (derecha).

- Alguna medida de carga viral que afecte la capacidad de un nodo de transmitir una infección

Se trabajó sobre un grafo conectado únicamente por cercanía, ya que representa un escenario sintético similar al proceso de transmisión de una infección por vías respiratorias. Entonces, se ignoró la capa social en este caso y se usó como indicador de vínculo solamente si las personas se encontraban cerca. La definición de "cerca" se corresponde con la configuración del módulo de movilidad.

Este módulo se configuró con un movimiento aleatorio de los nodos siguiendo un proceso de caminata aleatoria de velocidad 1. Esto quiere decir que, en cada paso de la simulación, cada nodo se mueve a través de un vector de longitud exactamente 1 en una dirección elegida uniformemente entre todas las posibles. Inicialmente todos los nodos se ubicaron usando una distribución uniforme en un cuadrado de S unidades de lado. Como se comentó en el párrafo anterior, en este módulo se definió que dos nodos se encuentran "cerca" (y que por lo tanto es posible el contagio) cuando se encuentran a C unidades o menos de distancia.

Se simuló para $p \in [0,0{,}2, 0{,}4, 0{,}6, 0{,}8, 1]$ un proceso en el que Np nodos elegidos al azar se etiquetaron como usuaries de tapabocas. Para que la cualidad de usar tapabocas afectara la probabilidad de propagación de la infección, se redefinieron las probabilidades según la siguiente condición:

```
CasesInfectionPropagationCondition({
  self.both_wear_mask_condition():
    BernoulliRandomVariablePropagationCondition(0.005),
  self.exactly_one_wears_mask_condition():
    BernoulliRandomVariablePropagationCondition(0.02),
  self.both_do_not_wear_mask_condition():
    BernoulliRandomVariablePropagationCondition(0.4)
})
```

Es decir:

- Si ambos nodos usan tapaboca, la probabilidad de contagio es de 0,5 %
- Si exactamente uno lo usa, es de 2 %
- Si ningún nodo lo usa, es de 40 %

Esta configuración hace que, para cada evento de contagio, la variable aleatoria que se utilice para decidir si el contagio ocurre dependa de si los nodos involucrados usan tapabocas o no. El comportamiento esperado a priori del sistema es que a medida que la proporción de nodos que adopta la medida crece, el proceso de contagio se vea ralentizado.

Lo primero que se puede observar es que efectivamente existe una ralentización del proceso de contagio para proporciones más altas de uso de barbijo. Esto se ve en los gráficos 7.8, 7.9 y 7.10, en un corrimiento hacia la derecha del pico de la infección a medida que la probabilidad del uso de tapabocas crece.

Además, en este mismo sentido, para probabilidades de uso de 0.6 en adelante, de observa además una disminución progresiva de la altura de este pico. Esto indica que además de retrasar el contagio, la aplicación de medidas que bajan la proporción de contagio funciona además para que una proporción menor de la población se contagie. No es posible en

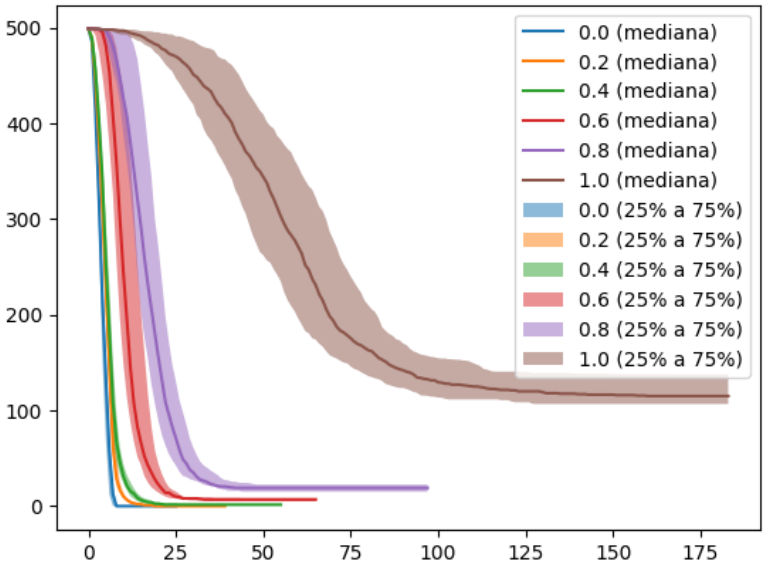


Fig. 7.8: Cantidad de nodos susceptibles en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0).

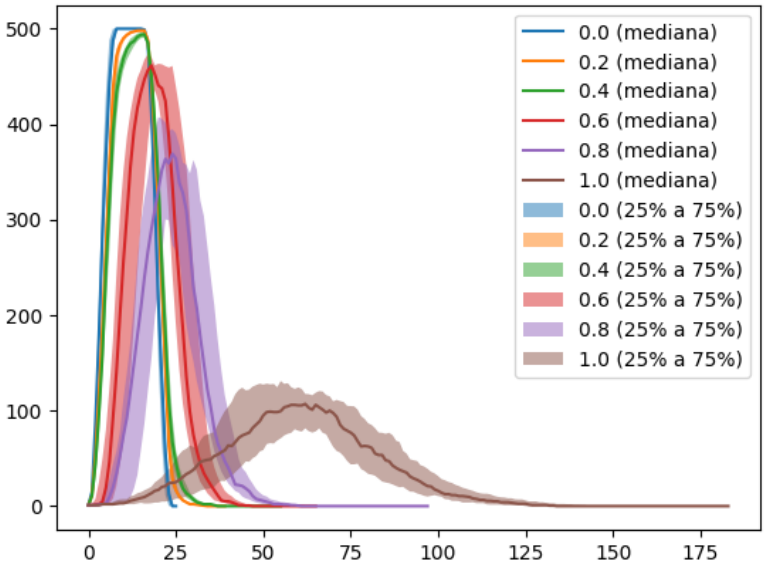


Fig. 7.9: Cantidad de nodos infectados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0).

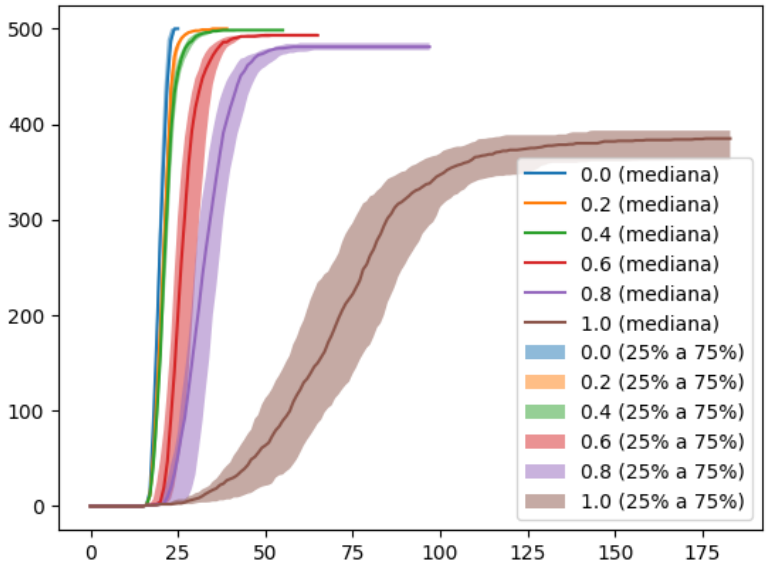


Fig. 7.10: Cantidad de nodos recuperados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0).

este caso hablar de inmunidad de rebaño porque el uso de barbijo no es una inmunización, pero sí se puede hablar de un efecto de contención del proceso de infección que proviene de disminuir la cantidad de contagios totales que cada persona individualmente provoca (es decir, de la disminución del parámetro R de las curvas de contagio características del proceso de infección).

En las figuras 7.11, 7.12 y 7.13 se muestra una simulación similar para un umbral de cercanía menor (2 unidades en vez de 8). Se muestran en el gráfico sólo las proporciones de uso de barbijo entre 0.0 y 0.6 inclusive ya que los valores mayores siguen la tendencia de la serie de valor 0.6 de contención del proceso de infección sin masificación.

En este proceso, al tener menor umbral de contagios, incluso sin aplicar la política de uso de barbijo no se contagia el 100 % de la población. Para este caso menos denso se observa un corrimiento a la derecha del pico de la infección solamente en los casos con proporciones más bajas de uso de barbijo ya que en las más altas el proceso en la mayoría de los casos es contenido antes de que la infección se masifique en la población. Esto tiene el efecto de acortar el proceso, ya que la población total afectada por la infección resulta cada vez menor. En un caso extremo, no existe ningún contagio y se termina el proceso cuando el nodo semilla pasa a estado R.

7.3. Variación de la proporción de nodos vacunados

En este experimento se puede observar cómo cambian las curvas características del proceso de contagio si distinta proporción de la población está vacunada. Existen varios efectos posibles de una vacuna, pero en particular se eligió suponer que el único efecto que

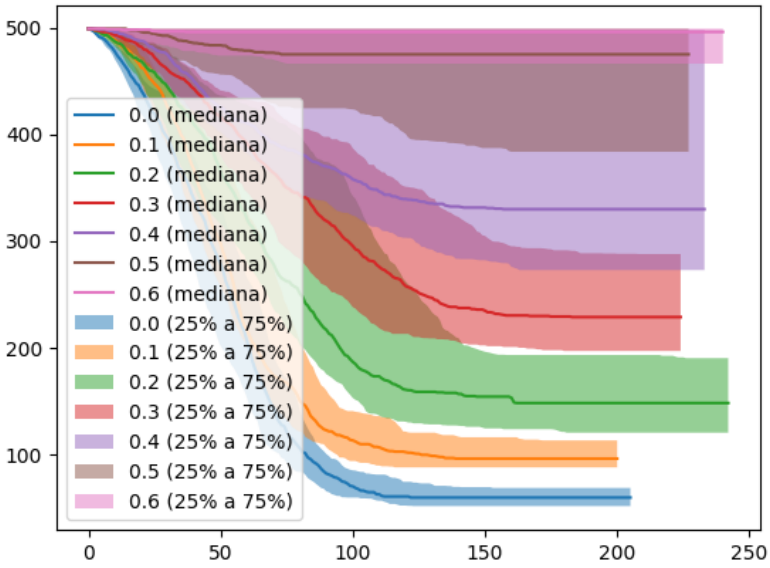


Fig. 7.11: Cantidad de nodos susceptibles en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0).

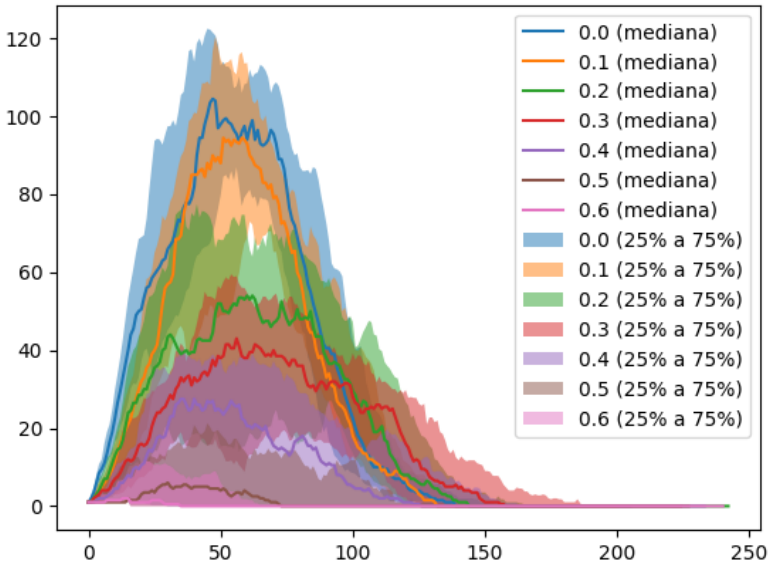


Fig. 7.12: Cantidad de nodos infectados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0).

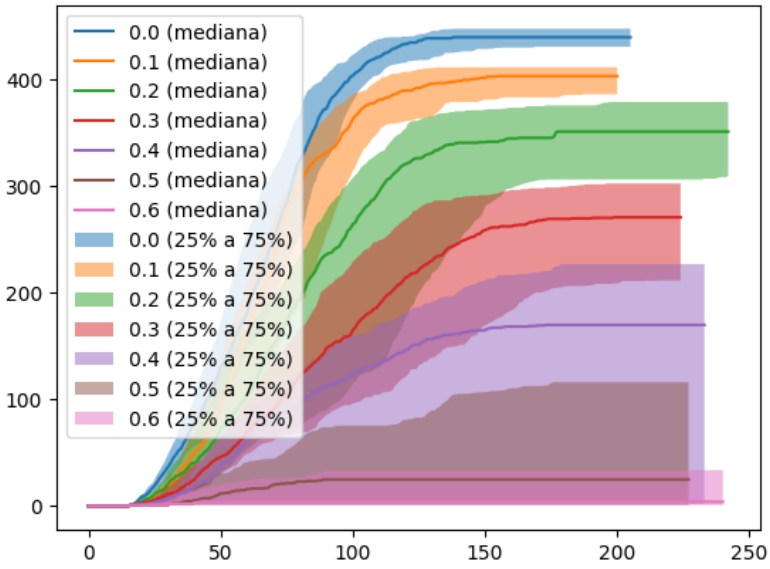


Fig. 7.13: Cantidad de nodos recuperados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0).

tiene es generar una versión más leve de la enfermedad para las personas que contraen la infección.

En este ejemplo, además, por una cuestión de simplicidad modelamos el transcurso de la infección usando un contador con una cantidad fija de días, sin un estadío latente previo.

Es decir, en código, se usó la siguiente indicadora de recuperación:

```
CountdownIndicatorFunction(X)
```

Esta función es determinística, y vale *falso* las primeras X veces que se la invoca, y *verdadero* desde la vez número $X + 1$ en adelante.

La condición de estar vacunade afecta el transcurso de la enfermedad para cada persona: si lo está, tiene una versión menos severa de la enfermedad, lo que en este modelo particular implementamos como que tarda menos días en recuperarse.

Existen formas más complejas de implementar las consecuencias de la vacunación, por ejemplo generando distintas máquinas de estado para la versión leve y la versión grave de la enfermedad, y/o teniendo en cuenta el efecto del estrés sobre el sistema de salud que el aumento de la cantidad de casos graves provoca. Elegimos esta manera por ser simple, al no requerir la implementación del sistema de salud ni de distintas versiones de la infección (al menos de forma explícita), pero las otras implementaciones de las consecuencias de la medida pueden hacerse.

Cada serie representa una proporción p distinta de nodos vacunados al azar, con $p \in [0,0{,}2, 0{,}4, 0{,}6, 0{,}8, 1]$.

Se redefinieron las funciones de avance de la infección como:

```
vaccinated_condition =
```

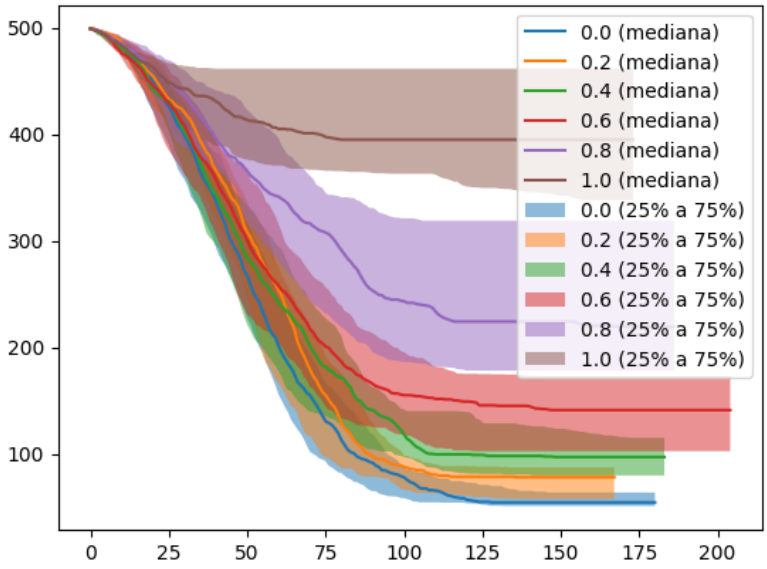


Fig. 7.14: Cantidad de nodos susceptibles en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 40 %.

```
  NodeIsTaggedAsDevelopmentCondition(TagNames.VACCINATED)

indicator_with_vaccine = CountdownIndicatorFunction(5)
indicator_without_vaccine = CountdownIndicatorFunction(15)

MatchesAnyDevelopmentConditions([
  MatchesAllDevelopmentConditions([
    vaccinated_condition,
    IndicatorDevelopmentCondition(indicator_with_vaccine)
  ]),
  MatchesAllDevelopmentConditions([
    NotDevelopmentCondition(vaccinated_condition),
    IndicatorDevelopmentCondition(indicator_without_vaccine)
  ])
])
```

Es decir, que para los nodos vacunados la infección se mantiene por 5 días mientras que para los nodos no vacunados, por 15.

En el gráfico se puede ver cómo cambian las curvas de contagio en función de la proporción de nodos vacunados.

En la situación que se grafica en las figuras 7.14, 7.15 y 7.16, incluso sin que nadie en la población esté vacunade, el proceso no alcanza a la totalidad de los nodos. En este caso, vemos que a medida que una mayor parte de la población tiene una versión de la enfermedad que dura menos tiempo, baja la proporción total de personas que llega a

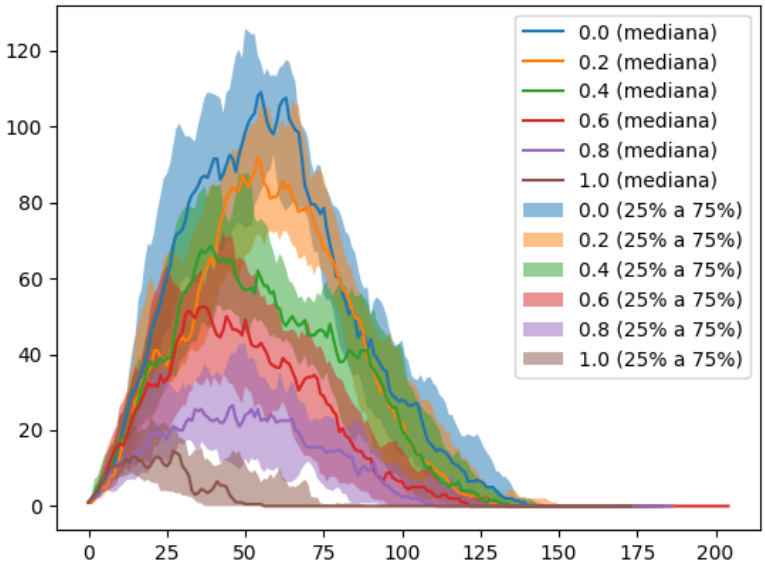


Fig. 7.15: Cantidad de nodos infectados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 40 %.

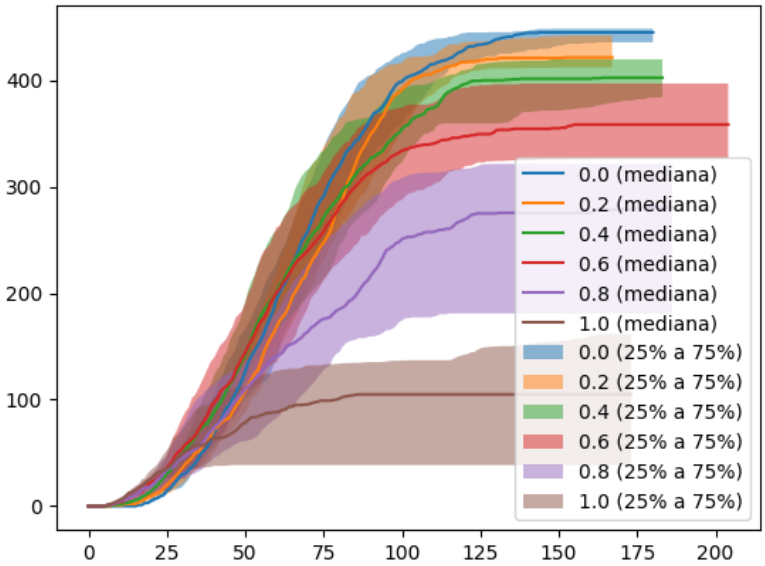


Fig. 7.16: Cantidad de nodos recuperados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 40 %.

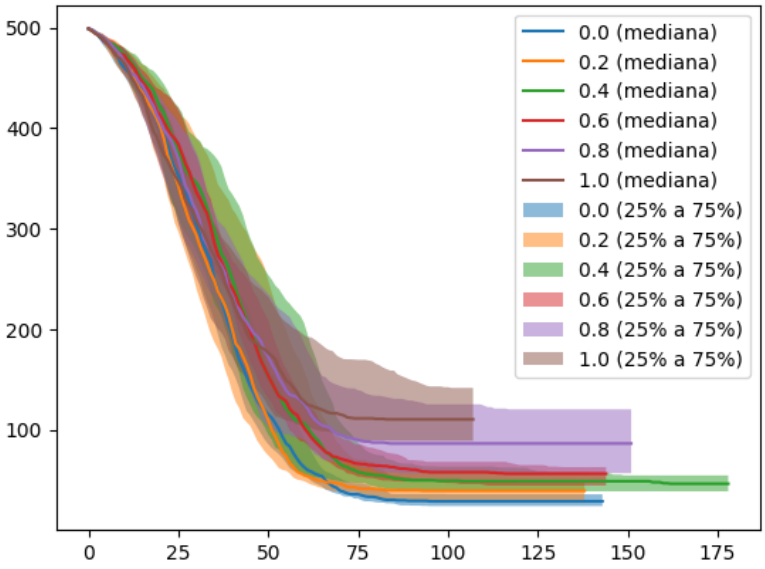


Fig. 7.17: Cantidad de nodos susceptibles en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 80 %.

contraer la infección. Esto ocurre porque, como cada persona vacunada que se infecta tiene menos días para propagar la infección, baja la cantidad de contagios por persona en total, con lo cual el índice R de propagación es menor, lo que aplana la curva de contagios.

Además se observa un ligero corrimiento del pico de contagios hacia la izquierda a medida que la proporción de la población vacunada sube. Este efecto se corrobora también en las curvas de los gráficos 7.17, 7.18 y 7.19, donde se mantuvieron las condiciones pero se aumentó la probabilidad de contagio al 80 %.

Además, con una probabilidad más alta de contagio, las curvas de los estados se parecen más entre las distintas proporciones de nodos vacunados. Esto se explica porque, si rápidamente se infecta una proporción alta de los nodos del grafo, la gran mayoría de los nodos alcanzan a contagiarse antes de que las personas que tienen una versión leve se recuperen.

Los gráficos de las figuras 7.20, 7.21 y 7.22 muestran una situación similar a las 7.14, 7.15 y 7.16 pero aumentando el umbral de cercanía a 8 unidades. En este caso se puede ver que las curvas del estado **S** son similares para cualquier proporción de nodos vacunados. Esto es porque la cantidad de contactos es tan grande que se alcanza rápidamente a contagiar a toda la población.

Sin embargo las curvas de los estados **I** y **R** son distintas: baja el pico de casos más rápidamente, y la cantidad de recuperaciones crece más rápidamente dada más gente vacunada. Esto es consecuencia de que existe una proporción de la población que, al estar vacunada, simplemente se recupera más rápidamente entonces sale antes del estado **I** hacia el **R**, y esto hace que se acumule menor parte de la población en **I**. Este efecto no afecta a la población en **S** por un efecto del que se habló anteriormente: el umbral elegido hace

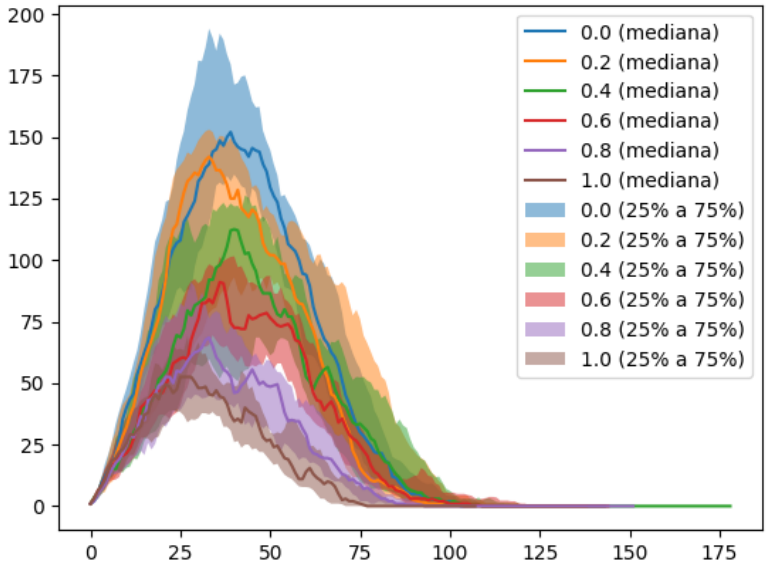


Fig. 7.18: Cantidad de nodos infectados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 80 %.

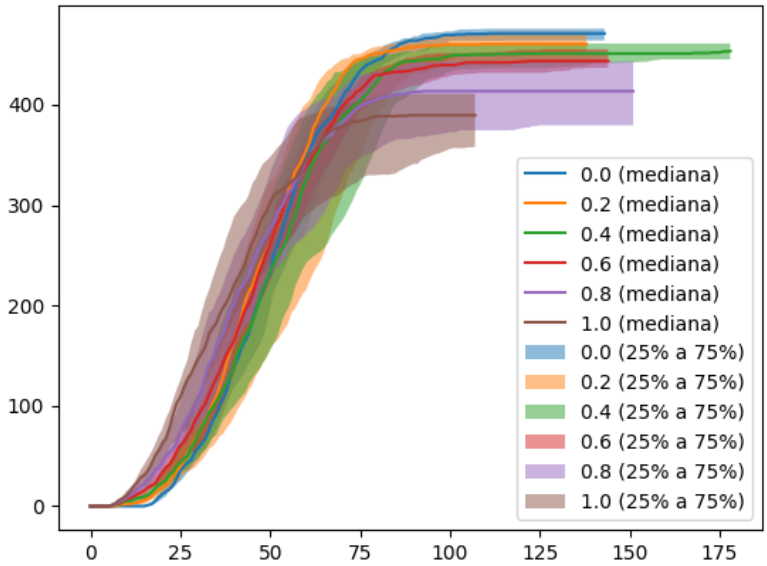


Fig. 7.19: Cantidad de nodos recuperados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 2 unidades (C=2.0). La probabilidad de contagio ante un contacto es del 80 %.

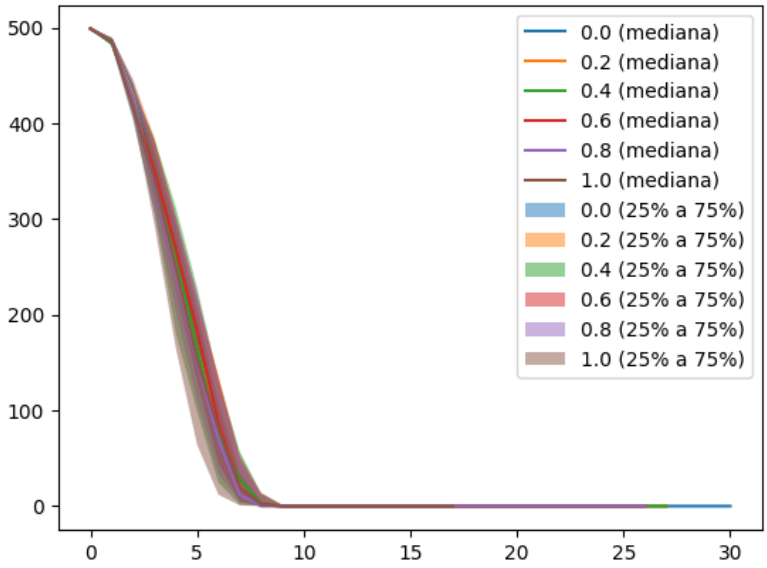


Fig. 7.20: Cantidad de nodos susceptibles en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0). La probabilidad de contagio ante un contacto es del 40 %.

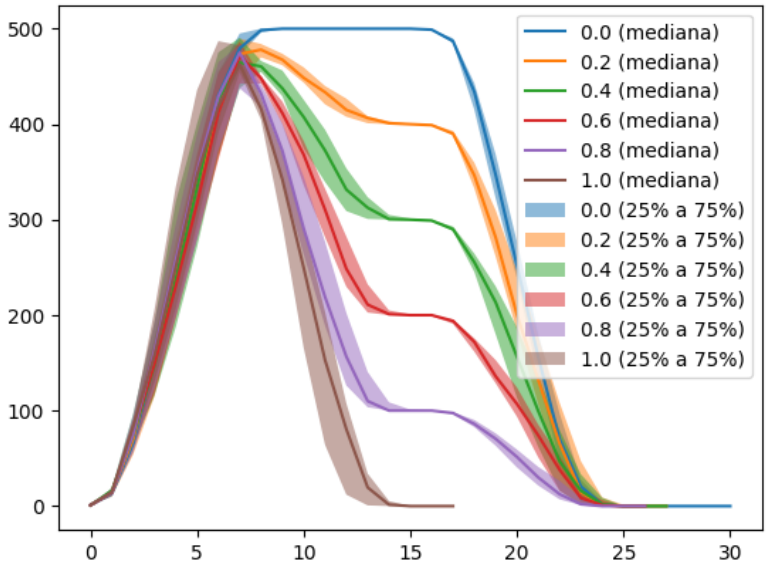


Fig. 7.21: Cantidad de nodos infectados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0). La probabilidad de contagio ante un contacto es del 40 %.

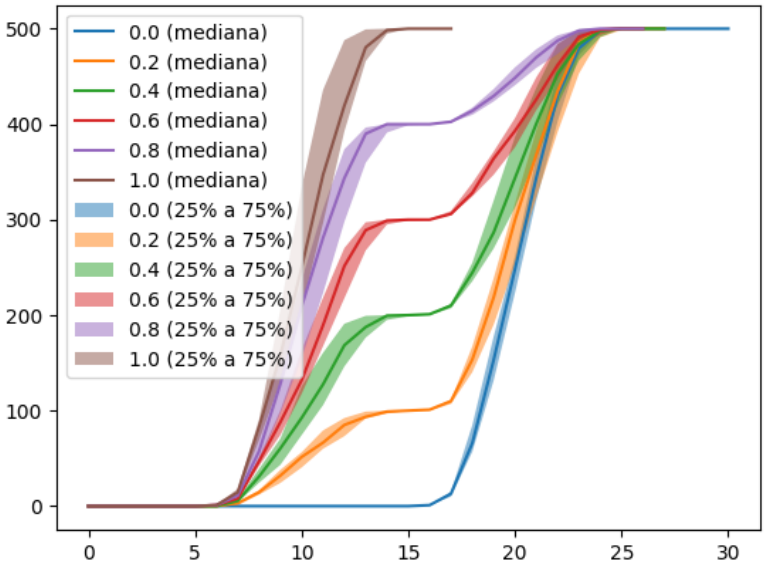


Fig. 7.22: Cantidad de nodos recuperados en función del tiempo. Grafos de 500 nodos (N=500) en un espacio inicial de 50x50 (S=50.0), con un umbral de cercanía de 8 unidades (C=8.0). La probabilidad de contagio ante un contacto es del 40 %.

que, en todos los casos, se contagie rápidamente el total de la población, en la mayoría de los casos antes de que las personas vacunadas comiencen a recuperarse en una proporción significativa.

Se observan en estos gráficos mesetas a diferentes alturas que muestran la recuperación de las personas vacunadas en una primera etapa y de personas no vacunadas en una segunda. Sin embargo en los casos anteriores (gráficos 7.16 y 7.19) no se observa este efecto tan notoriamente como en el último gráfico, donde, al contagiarse todos los nodos prácticamente al mismo tiempo, se diferencian claramente las dos etapas de recuperación.

7.4. Análisis de datos de Ecobici

En este experimento se simuló de un proceso de infección **SIRD** para nodos que se desplazan por el espacio de acuerdo con un modelo gravitacional de dos puntos. La capa social no se encuentra presente, sino que un nodo infectado contagia a uno susceptible con cierta probabilidad solamente cuando se encuentran cerca en el espacio.

Para elegir el conjunto de nodos se cargó la lista completa y se tomó una muestra de nodos distintos de tamaño N, para un valor fijo de N fijado de antemano. Se tomó como semilla (nodos infectados iniciales) a 10 personas elegidas aleatoriamente de la muestra.

Cada instante de tiempo representa una hora. La simulación empieza a las 0h, y es una sucesión de días laborales en función de las predicciones de hogar y trabajo de cada nodo, sin considerar fines de semana. A las 7h cada día todos los nodos simultáneamente se mueven a cierta velocidad hacia su trabajo, y cuando llegan se quedan en ese lugar. A las 17h ocurren los viajes inversos entre el trabajo y la casa de cada nodo, a la misma

velocidad. Cuando cada nodo llega a su casa permanece ahí hasta el día siguiente. Todos los movimientos se realizan a vuelo de pájaro.

Una infección, desde el momento en el que se contrae, tarda una cierta cantidad de días fijos (es decir `24*DIAS` instantes de tiempo) en pasar de estado **I** al estado final **R**. Cada día a las 0h, además, con cierta probabilidad la persona pasa al estado final **D**.

Se considera que dos personas están en contacto estrecho cuando se encuentran a **2m** de distancia o menos, y que se desplazan a **10km/h** ya que es una velocidad estimada para viajes en bicicleta.

Se comparó la forma del proceso de movilidad gravitacional respecto de otro en el que la información del trabajo no se tuvo en cuenta. En este segundo proceso, se implementó cada nodo como un caminante aleatorio siguiendo un proceso gaussiano, a la misma velocidad que el proceso anterior, pero en cada paso en una dirección aleatoria, y con un proceso en el que cada persona permaneció en su nodo hogar sin moverse. La posición inicial de cada nodo fue la geolocalización inferida como su casa.

En los gráficos de las figuras 7.23, 7.24, 7.25 y 7.26 se muestra una comparación entre las curvas de contagio para los procesos de movilidad entre dos puntos, y el proceso sin movimiento que ubica a cada punto en su hogar. El proceso aleatorio donde cada nodo se movió con la velocidad de un viaje en bicicleta a partir de su hogar inferido no se muestra en los gráficos porque para los parámetros introducidos en la simulación no generó contagios en ninguna corrida de la simulación.

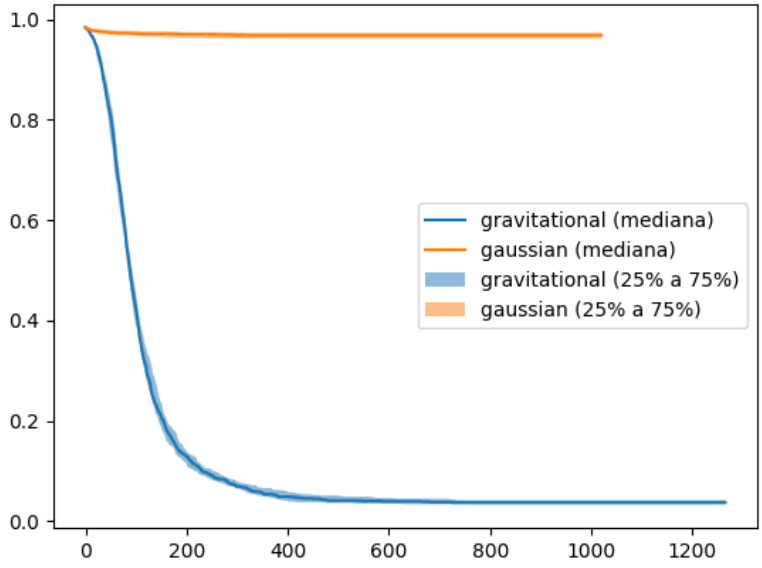


Fig. 7.23: Proporción de nodos susceptibles (nodos en este estado sobre nodos totales) en función del tiempo en horas. Los grafos tienen 600 nodos (N=600) tomados aleatoriamente de la población de usuaries de EcoBici, con un umbral de cercanía de 2 metros (C=2.0). La probabilidad de contagio ante un contacto es del 2 %.

Se puede observar una gran diferencia en la proporción de nodos contagiados en cada caso. Esto es porque, cuando la gente se queda fija en una ubicación, se confina la propa-

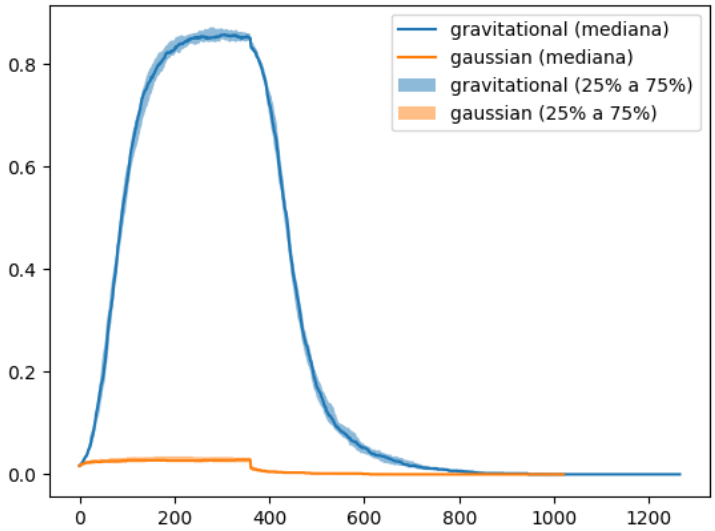


Fig. 7.24: Proporción de nodos infectados (nodos en este estado sobre nodos totales) en función del tiempo en horas. Los grafos tienen 600 nodos (N=600) tomados aleatoriamente de la población de usuaries de EcoBici, con un umbral de cercanía de 2 metros (C=2.0). La probabilidad de contagio ante un contacto es del 2 %.

gación de la infección a las (a lo sumo) 10 ubicaciones de las personas que están infectadas inicialmente (entre las 282 que tiene en total el dataset). En el momento en el que estas personas ya se encuentran contagiadas el proceso se satura. Para la simulación en la que las personas se mueven en una dirección aleatoria a velocidad bicicleta, la falta de contagios se atribuye a que la dirección aleatoria de los nodos hace que la probabilidad de encuentro entre dos de ellos sea muy baja y que por lo tanto no se den suficientes situaciones de posible contagio. Para el proceso de movilidad de dos puntos, por otra parte, se observa que la infección alcanza a gran parte de la población. Esto quiere decir que el hecho de que cada nodo pase tiempo en su casa y en su trabajo alcanza para conectar casi todas las ubicaciones posibles y propagar la infección hacia ellas.

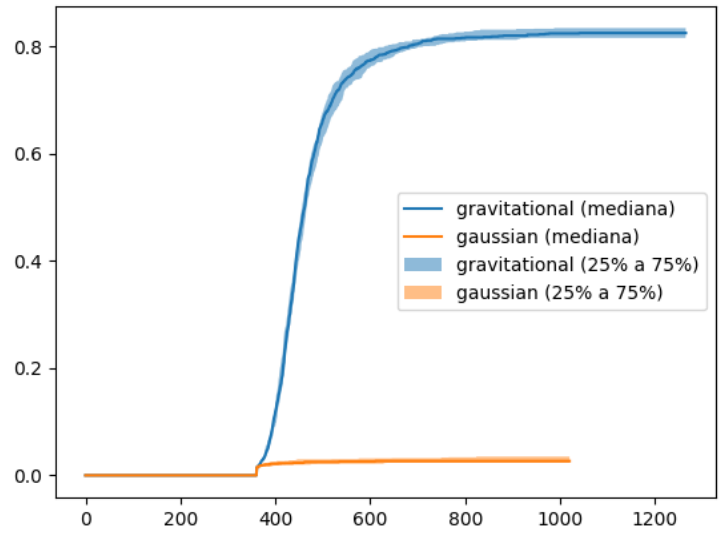


Fig. 7.25: Proporción de nodos recuperados (nodos en este estado sobre nodos totales) en función del tiempo en horas. Los grafos tienen 600 nodos (N=600) tomados aleatoriamente de la población de usuaries de EcoBici, con un umbral de cercanía de 2 metros (C=2.0). La probabilidad de contagio ante un contacto es del 2 %.

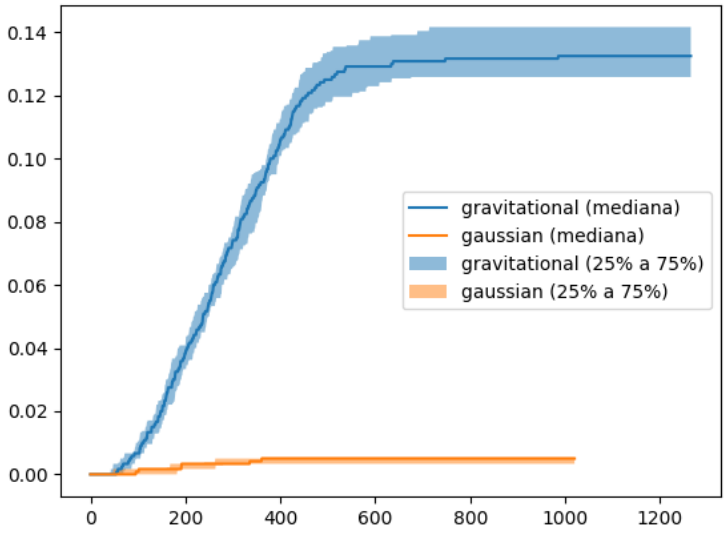


Fig. 7.26: Proporción de nodos fallecidos (nodos en este estado sobre nodos totales) en función del tiempo en horas. Los grafos tienen 600 nodos (N=600) tomados aleatoriamente de la población de usuaries de EcoBici, con un umbral de cercanía de 2 metros (C=2.0). La probabilidad de contagio ante un contacto es del 2 %.

8. CONCLUSIONES

En esta sección se presentan conclusiones y líneas de trabajo futuro sobre los distintos temas explorados en el transcurso de la tesis. En la sección 8.1 se trata el tema de la simulación de procesos de difusión, y de la decisión de realizar una simulación explícita como prueba de concepto. En la sección 8.2 se recogen consideraciones y aprendizajes respecto del diseño del prototipo como software, y se discuten los problemas y soluciones en el plano de la arquitectura del software de simulación. En la sección 8.3 se resume el aprendizaje sobre el estudio histórico de distintos modelos de movilidad humana y de generación sintética de grafos sociales, y sobre en qué contexto cada modelo resulta más conveniente. En la sección 8.4 se revisan los experimentos realizados utilizando el software, entre ellos una comparación entre procesos de difusión sobre grafos generados con distintos algoritmos, modificaciones hipotéticas en dos capas distintas de la definición de un proceso de difusión sobre grafos con características similares, y una prueba de concepto sobre datos reales de movilidad en la Ciudad de Buenos Aires, que infiere un modelo de movilidad gravitatorio en torno a puntos de interés.

8.1. Sobre la simulación de procesos de difusión

El análisis de procesos de difusión se puede atacar de forma analítica o mediante la simulación. Además, dentro de la categoría de simulación, existe un espectro de posibilidades de qué tan explícitas se hacen las condiciones y características del modelo, contra qué tanto se embeben de forma implícita en otros parámetros. Por ejemplo, a la hora de realizar una simulación de contagio de una infección respiratoria, y proponer que existen nodos que usan tapabocas y otros que no, un modelo puede declarar que ciertas entidades "nodo" poseen la característica de tener tapaboca y predican sobre qué implica esto en el modelo. Otro modelo distinto podría, por ejemplo, reducir la probabilidad general de contagio para representar que hay ciertas personas que son menos susceptibles, modelando implícitamente que esto es a causa de que ciertas personas se están comportando de forma diferente.

Para este ejemplo resulta sencillo encontrar una representación implícita posible del efecto que se quiere simular, pero para comportamientos locales más complejos, que resultan de la composición de varias características, no queda tan claro que al simplificarlas y embeberlas en cambios de parámetros no se pierdan efectos colaterales que emerjan de la interrelación de las distintas condiciones que ocurren en el modelo simultáneamente. En este sentido, pensar en términos del modelado de sistemas complejos permite predicar sobre el comportamiento micro de forma explícita, y observar el comportamiento macro que emerge de realizar las simulaciones.

En particular el problema de la simulación de procesos de difusión de infecciones es muy complejo. Existe una superposición de distintas capas de información que se suelen colapsar en la definición de una probabilidad puntual de contagio para cada par de personas (lo que resulta en un modelo de representación implícita de estas capas). Esta tesis explora la posibilidad de explicitar algunas de estas capas como una prueba de concepto. Los módulos elegidos provienen de un análisis de distintos casos de uso posibles presentes en trabajos relacionados citados a lo largo del trabajo. No se pretende afirmar que son

suficientes y necesarias para representar cualquier proceso. Al contrario, la idea de tener un sistema modular es que sea fácilmente expansible según la necesidad de cada fenómeno en particular. Por ejemplo, hay casos en los que se puede querer representar el estrés sobre el sistema de salud de forma directa para que el sistema pueda reaccionar a ello.

Una línea de trabajo es realizar una comparación más exhaustiva entre los resultados que da el framework propuesto y los resultados de otras estrategias de planteo del problema (tanto soluciones analíticas, como simulaciones que modelan condiciones complejas implícitamente y otros frameworks de modelado de sistemas complejos), para ver de qué manera afectan los resultados la cantidad de información que se le da al modelo y qué tan explícito es.

También es interesante comparar estos distintos resultados con ejemplos de fenómenos que hayan ocurrido efectivamente. A fines de esta tesis se realizó una comparación cualitativa en el caso del experimento de la sección 7.4, a partir de datos reales, con los modelos analíticos presentados en el libro de Barabási [BP16], que indican que el modelo es capaz de representar procesos de contagio fidedignamente. Esto es una primera iteración en ese sentido, aunque tiene varias limitaciones en el alcance de las conclusiones (discutidas en 6.2).

Otra línea para explorar es realizar experimentos con condiciones del modelo que cambien a lo largo del tiempo. Por ejemplo, personas que cambien su comportamiento en general en función de su estado de infección (por ejemplo que se muevan distinto o que decidan ponerse tapabocas sólo si tienen síntomas). Una estrategia posible para modelar este tipo de situaciones es asociar distintos modelos de movilidad o contagio a etiquetas sobre los nodos en una primera iteración, o directamente a condiciones en una iteración más avanzada. Esto requeriría una complejización, en la que por ejemplo el módulo de movilidad tendría que ser compuesto de otros módulos más pequeños y definir en función del estado del modelo cuál se aplica. Esto trae distintos problemas de implementación que escapan al alcance de la tesis pero que son continuaciones interesantes. El modelado de un álgebra de condiciones en esa situación puede ser un camino para lograr la expresión de estas complejidades, ya que, como pueden hacerle preguntas generales al modelo en cada paso, y las respuestas a estas preguntas potencialmente pueden cambiar, el comportamiento puede cambiar dinámicamente. Por ejemplo, si hay una condición que predica sobre la cantidad de infectades en un momento dado y se basa en esa pregunta para tomar una decisión, naturalmente el comportamiento cambia cuando lo hace esta cantidad, de un paso a otro de la simulación.

8.2. Aporte y alcance del prototipo propuesto

El modelo explora distintas ideas de diseño con el objetivo de acercarse a un framework general que pueda abarcar la simulación de procesos de difusión de disintas características. La idea de modelos de movilidad puede significar distintas cosas en función de la escala temporal y espacial en la que se esté simulando. Por ejemplo, si hay un proceso que ocurre en algunas horas, los módulos serán distintos que si el proceso ocurre en el transcurso de meses, pero en ambos casos el prototipo resuelve problemas comunes a estos distintos casos de uso.

Uno de estos problemas comunes es la presentación y el formato de salida del resultado de los experimentos. El prototipo separa la capa de presentación de la información de la de simulación mediante objetos presentadores, lo que permite fácilmente que se integre

a distintas herramientas de generación de gráficos, o de visualización en general. Aparece implícitamente la idea de *Resultado* de un experimento como una serie de fotos sucesivas de un estado definido en el presentador, aunque en el prototipo este concepto no se explicitó sino que se representa como una lista de diccionarios de Python con información. En una iteración futura sería interesante *reificar*[1] el resultado para explicitarlo.

Otro problema común que se atacó es la localización de la información de cada nodo dentro del modelo. La estructura en módulos permite separar el problema de modelado (multicapa y complejo) en sus componentes, definiendo el comportamiento en cada dimensión por separado, pero esta separación trae consigo una pérdida de claridad de dónde se encuentra la información de cada nodo: si la información se ubica dentro de los módulos, no resulta tan fácilmente accesible por el resto de las partes del sistema, y si se hardcodea dentro del objeto `Persona`, este objeto queda como una colección no jerarquizada de datos sobre la persona que resulta poco clara. La solución propuesta en el prototipo es la separación de la información de la persona en *aspectos*. Cada módulo inyecta el aspecto correspondiente en cada persona al comienzo de la simulación. Esto hace que la definición de cuál es el estado inicial quede dentro de cada módulo, pero que el acceso a la información de un nodo y su comportamiento básico pasen a través del objeto `Persona`.

Un tercer problema común atacado es cuál es la cosmovisión usada por el simulador (en la sección 5.5). Existen tres estrategias: basada en escaneo de actividades, en interacción de procesos y de agenda de eventos. Se eligió implementar una simulación basada en escaneo de actividades, por una cuestión de simplicidad. Esto tiene la contracara de que esta estrategia no es eficiente, en particular a la hora de propagar contagios, ya que en este caso cada paso de la simulación corre en $O(N^2)$ sobre un grafo de N personas. Es una continuación interesante modificar el modelo para cambiar la estrategia usada y ensayar su performance en relación con el camino tomado. Intuitivamente, parece que esto mejoraría el tiempo de corrida del modelo y permitiría que pueda manejar volúmenes más grandes de datos, a costa de complejizar la interacción entre las distintas partes del sistema.

En el prototipo se tomó la decisión de modelar cada persona individualmente, en contraposición con un modelo en el que, por ejemplo, cada nodo sea un lugar y haya aristas que representan el nivel de tráfico entre estos lugares. Se eligió esta estrategia como una forma de explorar la declaratividad que se podía alcanzar siendo más explícites. Sin embargo, si se quiere que la clase Persona pase a ser una población, y que los distintos módulos reflejen ese comportamiento, se podría hacer redefiniéndolos de una forma que tenga sentido para este nuevo problema (por ejemplo, en una población existen personas en distintos estados que conviven, en vez de tener un único estado para cada nodo).

Otro tema de exploración durante la tesis fue el modelado de las *Condiciones*. Esta es otra prueba de concepto de declaratividad, ya que modela explícitamente la configuración de condiciones que llevan a que los eventos ocurran de una manera u otra en el modelo. Cuando se plantea un experimento, inicialmente estas condiciones se plantean informalmente como una serie de decisiones encadenadas, que luego se deben procesar y resumir para condensarlas en una probabilidad que refleje un resultado similar al fenómeno que se quería expresar inicialmente. El prototipo explora la idea de generar un lenguaje que haga innecesaria esa transformación, y que permita la expresión de una forma declarativa de condiciones complejas y compuestas. Esto constituye una biyección más fiel entre el lenguaje original en el que se piensa la configuración de un experimento y el correspondiente

[1] Reificar, en la jerga de la programación orientada a objetos, significa modelar un concepto con un objeto. La palabra "reificar" literalmente significa "convertir en un objeto".

en el contexto del programa. Las condiciones se encuentran implementadas en dos lugares: en el módulo de configuración de la infección y en el módulo de propagación, pero la idea de condición es un concepto abstracto que puede aplicarse a cualquier proceso complejo de decisión. Existe un compromiso que se hace entre declaratividad y eficiencia, y en este caso la solución propuesta permite la expresión de condiciones arbitrarias combinadas de cualquier forma, pagando el costo de evaluar de toda la condición para llegar a una conclusión sobre un nodo cada vez que se necesita. En este sentido se ensayó con condiciones probabilísticas, contadores con estado, y condiciones que se evalúan realizando preguntas al estado de la simulación en cada momento. Esta exploración respecto de la declaratividad del modelo apunta a un objetivo más general: el de establecer un *lenguaje ubicuo*, común entre el dominio estudiado (en este caso la difusión de infecciones en poblaciones) y el software producido, lo que contribuye a acortar la brecha entre la definición de los experimentos y su configuración/implementación en el prototipo.

8.3. Marcos teóricos estudiados

En el marco de la tesis se exploraron distintos modelos para distintas partes del problema de la simulación de procesos de difusión de infecciones, principalmente en grupos humanos. Se dio principal atención a los modelos de movilidad humana, y a la generación aleatoria de grafos sociales.

Como conclusión general, existen distintos modelos que son más adecuados para representar distintas situaciones. Por un lado existe un recorrido histórico en el que ciertos modelos más modernos incorporan nuevos elementos y son capaces de representar más fielmente los procesos, emulando características que los modelos previos no. Este es el caso de los vuelos de Lévy corrigiendo la sub-representación de saltos largos respecto del modelo de ruido gaussiano en los procesos de movilidad humana, o la corrección que hacen modelos como el de leyes de potencia sobre la distribución de grados y de distancias entre nodos sobre el modelo de Erdös-Rényi en la generación sintética de grafos sociales. Por otro lado, distintas dimensiones espacio-temporales requieren aproximaciones distintas al modelado, porque el comportamiento que emerge de cada simulación es esencialmente distinto.

La diferencia en comportamiento de procesos a diferente escala, por ejemplo, se ve al comparar la forma de aplicar el concepto de puntos de interés a distintos procesos de movilidad humana. Los puntos de interés ejercen una fuerza de atracción sobre las personas, que hace que los patrones de sus recorridos sean más predecibles que una trayectoria puramente aleatoria. Esto funciona tanto para movimientos de agentes individuales como para poblaciones pero en el marco de la tesis se trabajó sobre el movimiento de personas.

En este sentido, se puede comparar un proceso de movilidad en en transcurso de horas en un evento acotado respecto de otro en una ciudad entera durante varios meses. En ambos casos la idea de puntos de interés es relevante pero de una forma distinta: probablemente en una ciudad cada persona tenga sus propios puntos de interés (ya que tienen diferentes hogares, ocupaciones e intereses) y en un evento los atractores son compartidos entre muchas de las personas presentes. En una ciudad probablemente las personas vuelvan con una frecuencia diaria o semanal a distintos puntos de interés, lo cual indica su importancia o el tipo de actividad que la persona desarrolla en ese lugar, y en un evento particular quizás no hay regresos a un mismo punto de interés pero sí se puede medir la cantidad de tiempo que se pasa en ese lugar o las personas con las que se comparte para extraer información. Es un desafío interesante para un modelo general poder representar esta diversidad de escalas

y permitir expresar comportamiento y preguntas al modelo en estos distintos marcos.

Esta diferencia también aplica para el modelo de contagio que se elija: en una difusión sostenida en el tiempo hay un margen para que la gente se recupere, que tenga un período de latencia de algunos días y que su comportamiento cambie en función de cómo se siente o qué información tiene sobre su estado de infección. En cambio en una situación puntual no tiene sentido hacer muchas de estas distinciones y un modelo más simple (por ejemplo un **SI**) es más adecuado. En estos casos, la complejidad puede venir del análisis de medidas de prevención u otras variaciones más locales del comportamiento de las personas.

En la generación de grafos sociales, primero la idea de decidir la existencia del vínculo entre personas como una variable aleatoria, y luego la incorporación de la idea de leyes de potencia a los procesos humanos fueron cambios de paradigma importantes. Las distribuciones de cola pesada representan más fielmente vínculos entre personas ya que las distribuciones normales sub-representan las personas con un círculo social mucho más grande que la esperanza para el grafo en general.

El algoritmo de Barabási-Albert para generar grafos sociales sintéticos por adición sucesiva de nodos es interesante porque toma en cuenta que las redes sociales son cambiantes en el tiempo, algo que los modelos previos no consideran tan explícitamente, y que los nodos más centrales funcionan como atractores (en este caso no de movimiento sino de vínculos sociales) en la generación de nuevas conexiones. Además resulta fácil de implementar y barato de ejecutar en relación con otros algoritmos que fabrican grafos independientes de la escala. Tiene la limitación, sin embargo, de que únicamente puede generar redes con un coeficiente $\gamma = 3$.

Se tiene que hacer la salvedad de que los modelos de generación de grafos sociales aplican con diferente grado de efectividad a distintas situaciones, y un modelo de Erdös-Rényi puede ser adecuado para registrar interacciones en un evento particular en un espacio acotado y con una cantidad de personas predefinidas que no se conocen entre sí previamente. Para este tema particular, sin embargo, es difícil encontrar una situación tan ideal en el mundo real, ya que incluso en eventos acotados existen relaciones previas entre las personas que asisten, y referentes dentro de las comunidades que funcionan como atractores más fuertes de nuevas conexiones que una persona promedio.

8.4. Experimentos y trabajo sobre datos reales

En el contexto de la tesis se realizaron distintos experimentos. En primer lugar se hizo una comparación entre dos algoritmos de generación de grafos sintéticos para entender cómo se comportaba un proceso de difusión en grafos producidos por cada uno de ellos. En segundo lugar se generaron dos complejizaciones sobre los módulos de propagación y de desarrollo de la infección respectivamente, por un lado para observar de qué manera estas modificaciones afectan las curvas características de los procesos de difusión, y también como forma de poner a prueba el poder expresivo del sistema de condiciones y de la definición de experimentos, como casos de uso del prototipo programado. En tercer lugar, se realizó una prueba de concepto sobre una base de datos real, del sistema de bicicletas de la Ciudad de Buenos Aires (CABA), que consistió de la inferencia de un modelo de movilidad para un conjunto de nodos, a partir de los viajes observados en el dataset, y de la posterior simulación de un proceso de difusión sobre ellos.

En la comparación de algoritmos de generación de grafos sintéticos, para grafos ralos, se ve cualitativamente un cambio en la rapidez del proceso de difusión: para el grafo generado

con *preferential attachment* el proceso resulta más rápido que para el grafo generado por el algoritmo de Erdös-Rényi. Esto ocurre tanto para la ubicación temporal del pico de personas en estado **I** como para el fin del proceso. Este fenómeno se alinea con la intuición de que, como las distancias entre nodos para el primer grafo son menores que para el segundo, el proceso de difusión llega desde el nodo semilla hasta cualquier otro con mayor velocidad. Para grafos más densos este efecto deja de observarse. Este último resultado verifica experimentalmente el planteo que hace la bibliografía consultada ([BP16]) en este sentido.

Las dos medidas implementadas (uso de barbijo para cambiar en ciertos nodos la probabilidad de contagio, y vacunación para generar una versión más corta de la infección) tuvieron efecto sobre las curvas características del proceso de difusión. En el caso del tapabocas se dan dos fenómenos distintos: cuando la infección alcanza a todos los nodos para las distintas curvas, el pico igualmente se retrasa cuando una mayor parte de la población usa tapabocas. A medida que la proporción sube más, se logra además que no se propague la infección hacia la población completa, por lo que este método funcionaría como contención del avance de la infección, ya que se reduce el R_0 (la cantidad de contagios que cada persona ocasiona en el transcurso de su infección). En el caso de la vacunación de personas, para una distancia de contagio baja (es decir que la cantidad de situaciones de contagio es escasa) se observa también un descenso de la altura del pico (es decir, que se contagia menos gente a medida que la proporción de nodos vacunados sube). En un proceso con una distancia de contagio más alta, en la que toda la población se contagia con una proporción nula de personas vacunadas, la curva del estado **S** es prácticamente similar para todas las proporciones (es decir que la vacunación en este caso no funciona para evitar el contagio) pero se distinguen claramente dos fases para las curvas de los estados **I** y **R**, que corresponden respectivamente a las recuperaciones de las personas que cursan una versión leve y grave de la infección. Vale la pena recordar que estos dos últimos experimentos son corridas sobre grafos sintéticos de un proceso de difusión con parámetros no extraídos de una situación real, por lo cual el alcance de las conclusiones extraídas no puede trasladarse directamente a un proceso real.

En el experimento presentado en 7.4, al menos a nivel cualitativo, se observa que el modelo presentado reproduce los resultados esperados a nivel teórico: las curvas de contagio sobre un dataset real (a pesar de las limitaciones de alcance de los datos) tienen una forma esperada para curvas de procesos SIR. Por otro lado, el prototipo permite declarar experimentos comparativos de forma flexible.

El trabajo con datos reales requiere un procesamiento previo realizado por fuera del prototipo, de conversión de los datos en la información que el modelo requiere para funcionar. En cuanto a modelos de movilidad en particular, se trabajó con una base de datos de trayectorias con un origen y un destino geolocalizadas (es decir, sin información sobre el recorrido entre estos dos puntos) que se usó como entrada para definir puntos de interés de características distintas: la casa de la persona y su trabajo. Esta inferencia se hizo en función del momento de la semana en el que los viajes transcurrían.

Es importante tener en cuenta las limitaciones de los datasets reales que se usan para hacer experimentos. En este caso el prototipo se pudo probar con datos reales, pero las conclusiones tienen un alcance bastante limitado por la calidad de los datos y por el modelo utilizado. Se habla en la sección 6.1 de algunas limitaciones del modelo: la incapacidad de representar a la población que no tiene exactamente una casa y un trabajo, con que se debe asumir que la gente vive exactamente en las estaciones de bicicleta o aplicar una corrección,

y que se asume que todas las personas viajan a la vez. En cuanto a las limitaciones del dataset, la gran mayoría de las personas no se mueven en el sistema de bicicletas de EcoBici en la CABA, o lo usan en combinación con otros métodos de transporte que no nos resultan observables. No existe en este momento una base de datos de dominio público a la que se pueda tener acceso que haga un estudio integral del uso de transporte en la ciudad de forma masiva, pero aplicar el mismo análisis a una base de datos de mayor calidad permitirá aumentar el alcance de las conclusiones.

Apéndice

A. PROPIEDADES DE GRAFOS CREADOS CON *PREFERENTIAL ATTACHMENT*

En este anexo se discuten con más detalle algunas propiedades que justifican la pertinencia del algoritmo de preferential attachment (PA) para resolver el problema de generación de un grafo independiente de escala (en particular con $\gamma = 3$), y la justificación de la idea detrás del algoritmo. Para esto se analizan algunas propiedades de los grafos generados con el algoritmo de PA.

En la sección A.1 se analiza cómo se comporta el grado de un nodo en función del tiempo (tanto en función del tiempo actual del sistema como del instante de tiempo en el que el nodo fue agregado a él). Esta función (ver ecuación A.1 de la sección A.1) resulta ser una ley de potencia efectivamente, con un exponente $\beta = \frac{1}{2}$ para todos los nodos. Sin embargo, como es inversamente dependiente del tiempo t_i en el que el nodo se agrega al sistema, mientras más antiguo es este nodo tiene mayor probabilidad de tener un grado grande.

En la sección A.2 se analiza la distribución de grados de un grafo generado con PA en función del tiempo. Para esto se caracteriza la probabilidad acumulada de elegir un nodo de un grafo generado por este proceso con grado menor o igual a k, para un instante de tiempo t fijo. Esta sección concluye que un grafo generado por PA sigue una ley de potencia con $\gamma = 3$.

Finalmente, en la sección A.3 se discute brevemente cuál es la intuición de haber elegido las propiedades de crecimiento y conexión preferencial como base del algoritmo, y qué pasa cuando se usa solamente una de ellas para intentar construir un grafo.

A.1. Evolución del grado de un nodo en función del tiempo

Para comprender el comportamiento del modelo de PA, vamos a analizar la evolución del grado de un nodo en el tiempo. Este grado cambia conforme se agregan nuevos nodos en el grafo en cada instante de tiempo $t \geq t_i$, donde t_i es el momento de tiempo en el que el i-ésimo nodo fue agregado al grafo.

Sea i un nodo. En cada momento de tiempo luego de t_i, se agrega un nuevo nodo al grafo, con m aristas que se unen a nodos preexistentes. El nodo i tiene cierta probabilidad de ser elegido en función de su grado k_i y los grados del resto de los nodos preexistentes. Sea $\Pi(k_i)$ la probabilidad de elegir a un nodo con grado k_i. Esta probabilidad es:

$$\Pi(k_i) = \frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

Para seguir las cuentas supongamos que existe la posibilidad de multiejes, por una cuestión de simplicidad. La cantidad esperada de aristas nuevas del nodo en función de t es el resultado de realizar m veces el proceso de elegir una arista, es decir:

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

Para eliminar la sumatoria del denominador calculamos de otra manera la cantidad de aristas totales preexistentes, como $m(t-1)$. Estamos tomando como despreciables las

aristas de la configuración inicial porque N es órdenes de magnitud más grande. Entonces la suma de grados es $2m(t-1)$ y la cuenta queda[1]:

$$m\Pi(k_i) = m\frac{k_i}{2mt-2m} = \frac{k_i}{2t-2}$$

Si consideramos el -2 como despreciable queda:

$$\frac{dk_i}{dt} = \frac{k_i}{2t}$$
$$\frac{dk_i}{k_i} = \frac{dt}{2t}$$
$$\int_{k_i(t_i)}^{k_i(t)} \frac{dk_i}{k_i} = \frac{1}{2}\int_{t_i}^{t} \frac{dt}{t}$$

Usamos que $k_i(t_i) = m$ porque en el instante de tiempo i agregamos el nodo i con m aristas.

$$\int_{m}^{k_i(t)} \frac{dk_i}{k_i} = \frac{1}{2}\int_{t_i}^{t} \frac{dt}{t}$$
$$ln(k_i(t)) - ln(m) = \frac{ln(t) - ln(t_i)}{2}$$
$$e^{ln\left(\frac{k_i(t)}{m}\right)} = e^{\frac{ln(\frac{t}{t_i})}{2}}$$
$$\frac{k_i(t)}{m} = \left(\frac{t}{t_i}\right)^{\frac{1}{2}}$$
$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}$$

Llamamos $\beta = \frac{1}{2}$ al exponente de la ley de potencia. Reemplazando en la ecuación nos queda que el grado del nodo i en función del instante de tiempo es:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \tag{A.1}$$

Mirando la ecuación A.1 se pueden extraer varias conclusiones sobre el comportamiento del modelo:

- El grado de cada nodo sigue un proceso de ley de potencia con el mismo exponente $\beta = \frac{1}{2}$.
- Como $\beta < 1$, el crecimiento es sub-linear. Esto tiene sentido ya que, con cada nodo que se agrega a la red, se tienen que conectar las m aristas eligiendo cada vez entre un conjunto más grande de nodos. Esto hace que en cada paso la probabilidad de que un nodo preexistente particular sea elegido es menor que en el paso anterior.

[1] En la fuente está escrito como $2mt - m$ pero ésta es la forma correcta. El hecho de que haya un 1 o un 2 es despreciable más adelante así que esto es un detalle.

- La diferencia es el momento en el que cada nodo es agregado, lo cual le da da mayor probabilidad de ser puntos nodales a los nodos que fueron agregados primero al grafo.
- El crecimiento de k_i está dado por la derivada de la ecuación A.1:

$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}$$

Esto nos indica dos cosas. La primera es que como $\sqrt{t_i}$ está en el denominador, los nodos más antiguos tienen más probabilidad de incorporar nuevas aristas. La otra es que, como $\sqrt{t}$ también está en el denominador, para un nodo fijo, a medida que pasa el tiempo, es cada vez menos probable que incorpore nuevos vecinos.

A.2. Distribución de grados

En esta segunda parte, el objetivo es usar la función de grados de los nodos en función del tiempo $k_i(t)$ para encontrar la función de distribución de la totalidad de los nodos del grafo. Para esto, es más sencillo calcular la función de distribución acumulada.

Partimos del formalismo continuo para describir una distribución de grados de un grafo, presentado en la sección 4.2 de [BP16], ya que es más simple de trabajar que el formalismo discreto. La distribución de grados para un grafo que sigue una ley de potencia es[2]:

$$p(k) = (\gamma - 1)k_{min}^{\gamma-1} \cdot k^{-\gamma}$$

Al calcular esta función para los grafos generados con PA, deberíamos encontrar una expresión similar, con un valor de γ. En el resto de la sección se hace este cálculo para llegar encontrar $\gamma = 3$.

Empezamos por encontrar una fórmula que exprese la cantidad de nodos que, en un instante de tiempo dado, tienen un grado menor a k, para cada valor de k. Es decir, queremos calcular la cantidad de nodos que cumplen:

$$k_i(t) < k$$

La ecuación A.1 nos da el valor esperado para el grado de un nodo que se agrega al grafo en tiempo t_i. Si reemplazamos $k_i(t)$ (que es el resultado del cálculo de la sección A.1) por su definición en esta expresión queda:

$$\begin{aligned} m\left(\frac{t}{t_i}\right)^{\beta} &< k \\ \left(\frac{t}{t_i}\right)^{\beta} &< \frac{k}{m} \\ \frac{t}{t_i} &< \left(\frac{k}{m}\right)^{\frac{1}{\beta}} \\ t\left(\frac{m}{k}\right)^{\frac{1}{\beta}} &< t_i \end{aligned}$$

[2] Retomamos la ecuación 4.10, en la sección 4.2 que también refiere a este tema.

Intuitivamente, esta inecuación nos dice que, dado un tiempo t fijo y posterior a t_i, los nodos que cumplen que su t_i es mayor que la expresión de la izquierda son los que tienden a tener grado menor que k^3.

Entonces esta ecuación nos permite, si fijamos un instante de tiempo t, separar los nodos en *grandes* y *pequeños* en función del momento t_i en el que fueron agregados al grafo. Los nodos cuyo t_i cumple la inecuación son los pequeños y los que no la cumplen son los grandes. Así, la inecuación funciona como una indicadora de que el nodo es pequeño.

Según la inecuación, los primeros $t\left(\frac{m}{k}\right)^{\frac{1}{\beta}}$ nodos son grandes y el resto pequeños. Por lo tanto la cantidad de nodos grandes es:

$$t\left(\frac{m}{k}\right)^{\frac{1}{\beta}}$$

y la de nodos pequeños es:

$$N - t\left(\frac{m}{k}\right)^{\frac{1}{\beta}}$$

Sabemos que en el instante de tiempo t la cantidad de nodos M es $m_0 + t$, pero cuando t crece podemos decir que $M \sim t$. Con esta transformación la cantidad de nodos pequeños es:

$$t - t\left(\frac{m}{k}\right)^{\frac{1}{\beta}} = t\left(1 - \left(\frac{m}{k}\right)^{\frac{1}{\beta}}\right)$$

Supongamos que fijamos el instante de tiempo t. La probabilidad de elegir un nodo con grado menor a k es la proporción de nodos pequeños respecto del total (es decir, tenemos que dividir la expresión por t). Esta, además, es la definición de densidad acumulada hasta k, que llamamos $P(k)$. Por lo tanto escribimos:

$$P(k) = 1 - \left(\frac{m}{k}\right)^{\frac{1}{\beta}}$$

La densidad es la derivada de la acumulada, por lo tanto es:

$$p(k) = \frac{1}{\beta}\left(\frac{m^{\frac{1}{\beta}}}{k^{1+\frac{1}{\beta}}}\right)$$

Recordando que $\beta = \frac{1}{2}$ tenemos:

$$p(k) = 2\left(\frac{m^2}{k^3}\right) = 2m^2k^{-3}$$

Esta expresión tiene una interpretación como una ley de potencia respecto del parámetro k, retomando la ecuación 4.10 y reemplazando $\gamma = 3$ y $k_{min} = m$ (porque cuando se agrega una arista al grafo tiene m ejes, por lo tanto es el mínimo grado que puede tener un nodo) queda:

$$p(k) = (3-1)m^{3-1} \cdot k^{-3} = 2m^2k^{-3}$$

que es exactamente la ecuación anterior, por lo que podemos decir que la distribución de grados de los nodos en un proceso de BA corresponde a una ley de potencia con $\gamma = 3$. Por lo tanto el algoritmo de BA es un procedimiento efectivo para generar sintéticamente un grafo que cumple una ley de potencia, en particular con $\gamma = 3$.

[3] En la fuente (sección 5.5 de [BP16]) el signo de esta inecuación aparece al revés pero esto es un error.

A.3. Necesidad de crecimiento y de conexión preferencial

Se puede comprobar experimentalmente la premisa de que ambas propiedades son necesarias (es decir, no alcanza solamente con una) para lograr un grafo cuyos nodos tengan grados que sigan una distribución independiente de escala. En el primer experimento se crea un grafo a partir solamente de la premisa de crecimiento, y en el segundo solamente de conexión preferencial. Se observa en ambos casos que los grafos que se obtienen no siguen la distribución objetivo.

- **Modelo que utiliza crecimiento únicamente**: para eliminar la propie-dad de conexión preferencial se cambió la función de probabilidad puntual de elegir m nodos con los cuales unir cada nodo nuevo en el paso t. En vez de que la probabilidad crezca en función del tamaño del nodo se distribuye uniformemente entre todas las posibilidades. En este caso la distribución de grados se ajusta a una exponencial. Esta distribución tiene una cola más fina que la independiente de escala, y por lo tanto tiene menor presencia de outliers (hubs).
- **Modelo que utiliza conexión preferencial únicamente**: para eliminar el crecimiento se comienza con N nodos (una cantidad fija) y en cada unidad de tiempo se elige un nodo al azar, que se une a otro siguiendo la distribución de conexión preferencial que se usó para el modelo original (probabilidad proporcional al grado de cada nodo). En este caso el grafo tiende a convertirse en una N-clique, por lo que la distribución de grados tampoco sigue una ley de potencia.

BIBLIOGRAFíA

[KM27] William Ogilvy Kermack y Anderson G McKendrick. «A contribution to the mathematical theory of epidemics». En: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), págs. 700-721.

[ER59] Paul Erdős y Alfréd Rényi. «On random graphs I». En: *Publicationes mathematicae* 6.1 (1959), págs. 290-297.

[HW65] John M Hammersley y Dominic JA Welsh. «First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory». En: *Bernoulli 1713, Bayes 1763, Laplace 1813*. Springer, 1965, págs. 61-110.

[BR67] OV Baroyan y LA Rvachev. «Deterministic models of epidemics for a territory with a transport network». En: *Cybernetics* 3.3 (1967), págs. 55-61.

[Bai+75] Norman TJ Bailey y col. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[dK78] Ithiel de Sola Pool y Manfred Kochen. «Contacts and influence». En: *Social Networks* 1.1 (1978), págs. 5-51. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90011-4. URL: https://www.sciencedirect.com/science/article/pii/0378873378900114.

[Gam+95] Erich Gamma, Richard Helm, Ralph Johnson, Ralph E Johnson, John Vlissides y col. *Design patterns: elements of reusable object-oriented software*. Pearson Deutschland GmbH, 1995.

[WS98] Duncan J Watts y Steven H Strogatz. «Collective dynamics of 'small-world' networks». En: *nature* 393.6684 (1998), págs. 440-442.

[AJB99] Réka Albert, Hawoong Jeong y Albert-László Barabási. «Diameter of the world-wide web». En: *nature* 401.6749 (1999), págs. 130-131.

[PV01] Romualdo Pastor-Satorras y Alessandro Vespignani. «Epidemic spreading in scale-free networks». En: *Physical review letters* 86.14 (2001), pág. 3200.

[EE04] Eric Evans y Eric J Evans. *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley Professional, 2004.

[Fer+05] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn y Donald S Burke. «Strategies for containing an emerging influenza pandemic in Southeast Asia». En: *Nature* 437.7056 (2005), págs. 209-214.

[BHG06] Dirk Brockmann, Lars Hufnagel y Theo Geisel. «The scaling laws of human travel». En: *Nature* 439.7075 (2006), págs. 462-465.

[Dur07] Richard Durrett. *Random graph dynamics*. Vol. 200. 7. Citeseer, 2007.

[GHB08] Marta C. Gonzalez, Cesar A. Hidalgo y Albert-Laszlo Barabasi. «Understanding individual human mobility patterns». En: *nature* 453.7196 (2008), págs. 779-782.

[Rhe+11] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim y Song Chong. «On the levy-walk nature of human mobility». En: *IEEE/ACM transactions on networking* 19.3 (2011), págs. 630-643.

[Vis+11] Gandhimohan M Viswanathan, Marcos GE Da Luz, Ernesto P Raposo y H Eugene Stanley. *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, 2011.

[LBH12] Xin Lu, Linus Bengtsson y Petter Holme. «Predictability of population displacement after the 2010 Haiti earthquake». En: *Proceedings of the National Academy of Sciences* 109.29 (2012), págs. 11576-11581. DOI: 10.1073/pnas.1203882109. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1203882109. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1203882109.

[Ban+14] Jerry Banks, Johan S Carson II, David M Nicol y Barry L Nelson. *Discrete event system simulation*. 2014.

[Wes+15] Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen y Caroline O Buckee. «Impact of human mobility on the emergence of dengue epidemics in Pakistan». En: *Proceedings of the National Academy of Sciences* 112.38 (2015), págs. 11887-11892.

[BP16] Albert-László Barabási y Márton Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016. ISBN: 9781107076266. URL: http://barabasi.com/networksciencebook/.

[Mon+16] Juan de Monasterio, Alejo Salles, Carolina Lang, Diego Weinberg, Martin Minnoni, Matias Travizano y Carlos Sarraute. «Analyzing the spread of chagas disease with mobile phone data». En: (2016), págs. 607-612. DOI: 10.1109/ASONAM.2016.7752298.

[Pon+16] Nicolas B Ponieman, Carlos Sarraute, Martin Minnoni, Matias Travizano, Pablo Rodriguez Zivic y Alejo Salles. «Mobility and sociocultural events in mobile phone data records». En: *Ai Communications* 29.1 (2016), págs. 77-86.

[Sar+17] Carlos Sarraute, Carolina Lang, Nicolas B Ponieman y Sebastian Anapolsky. «The city pulse of Buenos Aires». En: *arXiv preprint arXiv:1707.01032* (2017).

[Cac+20] Giacomo Cacciapaglia, Corentin Cot, Anna Sigridur Islind, María Óskarsdóttir y Francesco Sannino. «You better watch out: US COVID-19 wave dynamics versus vaccination strategy». En: *arXiv preprint arXiv:2012.12004* (2020).

[Kra+20] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Open COVID-19 Data Working Group†, Louis Du Plessis, Nuno R Faria, Ruoran Li y col. «The effect of human mobility and control measures on the COVID-19 epidemic in China». En: *Science* 368.6490 (2020), págs. 493-497.

[Pai+21] Carlos Pais, José Alberto Biurrun Manresa, Abelardo Del Prado, Hugo Leonardo Rufiner y col. «Modelo a escala ciudad de la epidemiología de COVID-19 con movilidad de personas y sus actividades representadas por un conjunto de Modelos Ocultos de Márkov». En: (2021).