

Minería de Medios Sociales

Máster en Ciencia de Datos e Ingeniería de Computadores

Bloque I: Redes Sociales y Minería de Datos en Redes

Sesión I.5: Procesos en Redes

Oscar Cordón García

Dpto. Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada
ocordon@decsai.ugr.es

Difusión de información: Proceso mediante el cual una unidad de información (conocimiento) se difunde **en una red** y alcanza a los individuos mediante interacciones

Se estudia en una gran cantidad de disciplinas como la Sociología, la Epidemiología y la Etnografía, que se trasladan así a la Minería de Medios Sociales

Nos centraremos en técnicas que pueden modelar los procesos de difusión de información

EEUU, Febrero 2013, tercer cuarto de la *Super Bowl*: un apagón interrumpió el partido durante 34 minutos

Durante ese parón, la compañía **Oreo** twiteó y publicó en su cuenta de *Facebook* el mensaje: “*Power out? No Problem, You can still dunk it in the dark*”

La propagación fue desmesurada, alcanzando **más de 15,000 retweets y 20,000 likes en menos de 2 días**

Este sencillo tweet se difundió rápidamente en una gran población de individuos

Ayudó a la compañía a conseguir reconocimiento con una inversión mínima (*earned media*) en un entorno en el que las empresas gastan 4 millones de dólares en anuncios de 30 segundos



15,884 RETWEETS 6,488 FAVORITES

Torrejón, 28 Febrero 2017: una profesora de 3º de ESO diseña un experimento para concienciar de la rapidez de difusión de imágenes en redes

Dibujan un monigote con el texto “*Ayúdame a recorrer el mundo. Soy Nico*” y lo envían por sus perfiles (Instagram y Twitter) pidiendo su difusión. La profesora lo manda por Facebook y Whatsapp.

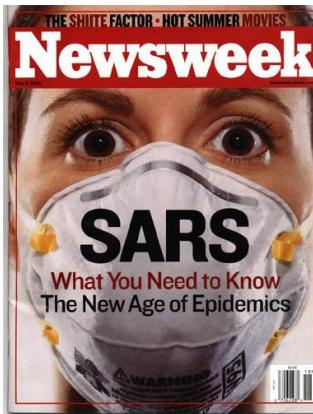
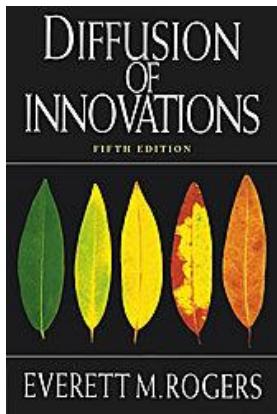
El dibujo lleva una maleta de una conocida marca deportiva. Muchos internautas pensaron que era una campaña de marketing

Nico se hizo viral en Whatsapp gracias a usuarios de 30 a 40 años. Llegó a los perfiles de Twitter de Guardia Civil y Policía Nacional. **Pasó por EEUU, Nicaragua, Venezuela, Honduras, Costa de Marfil, Italia, Francia, ...**



PROCESOS EPIDÉMICOS Y DE DIFUSIÓN

¿POR QUÉ ES TAN IMPORTANTE EL PROCESO DE PROPAGACIÓN?



***** SMART 11:37 AM

Search Twitter

#ALDubEBforLOVE 11.5M Tweets about this trend

#ShowtimeKapamilyaDay 363K Tweets about this trend

#NewAmericana 40.9K Tweets about this trend

#LarrysPureLove 180K Tweets about this trend

#2030NOW 3,760 Tweets about this trend

Epi + demos sobre pueblo



<http://es.wikipedia.org/wiki/Epidemia>

Biología:

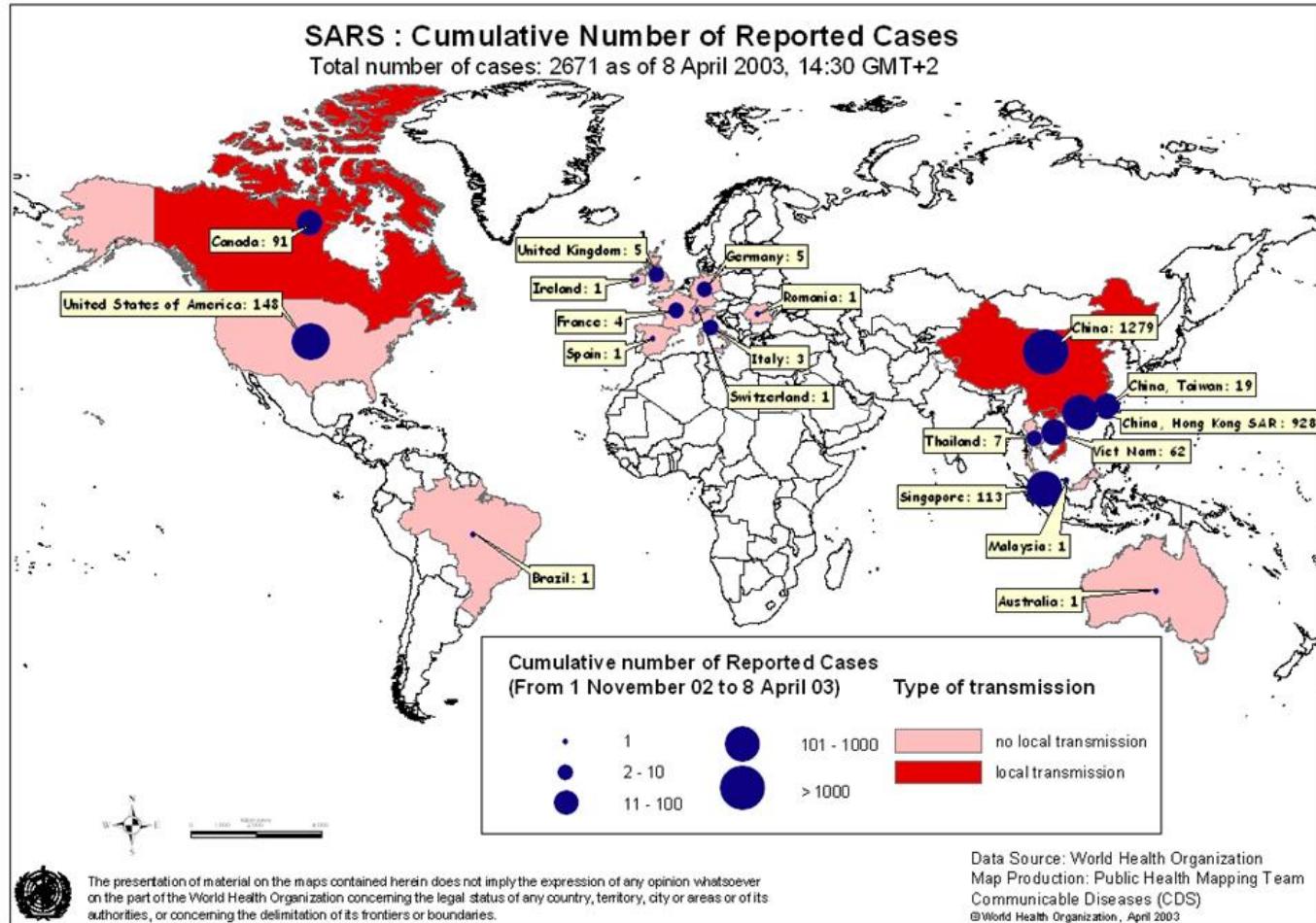
- Enfermedades transmitidas por el aire (gripe, gripe aviar, ...)
- Enfermedades venéreas (VIH, ...)
- Otras enfermedades infecciosas, incluidos algunos cánceres (VPH (virus del papiloma humano), ...)
- Parásitos (chinches, malaria, ...)

TIC:

- Virus de ordenador, gusanos
- Virus de teléfonos móviles

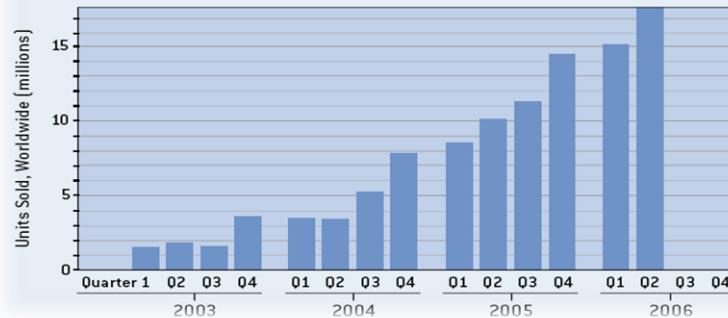
Conceptos/Aspectos intelectuales:

- **Difusión de información**
- **Difusión de innovaciones**
- **Rumores**
- **Memes (tweets, whatsapp, ...)**
- **Prácticas empresariales**



Fuente: Organización Mundial de la Salud

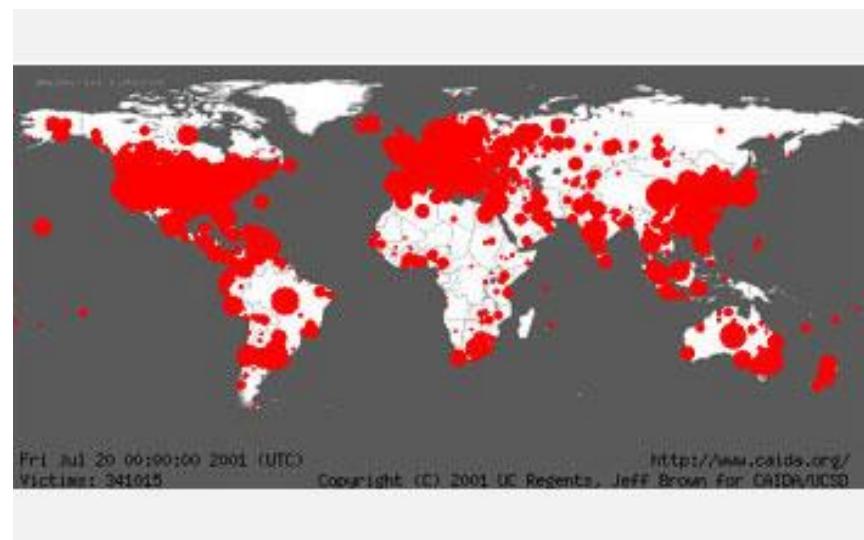
SMARTPHONES ON THE RISE



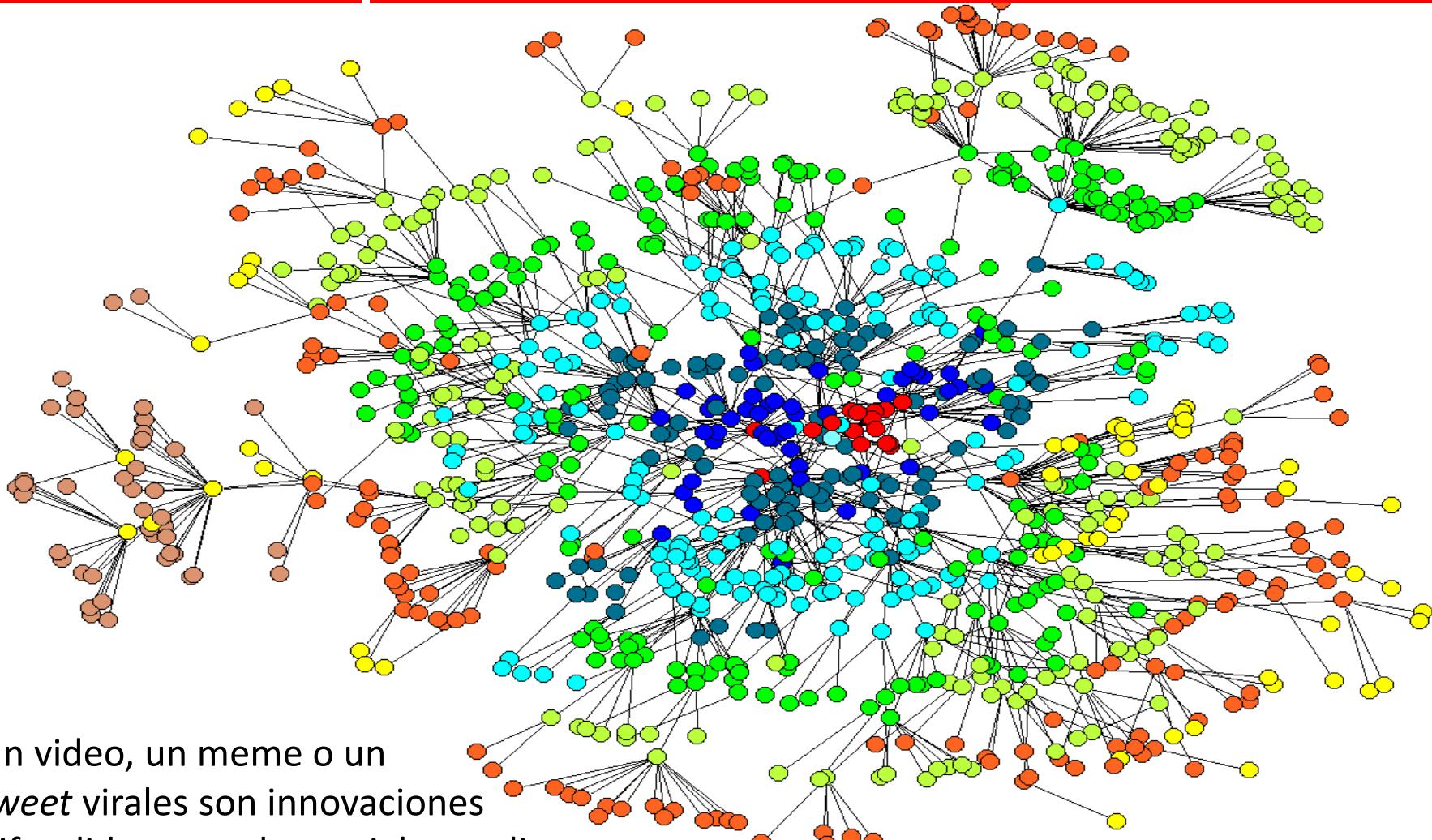
GROWTH IN MOBILE MALWARE



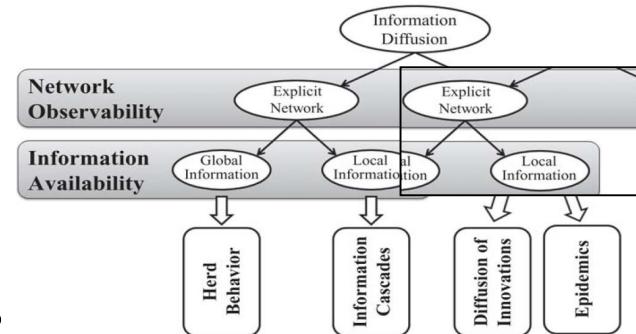
El código del *Red Worm* paralizó Internet en muchos países

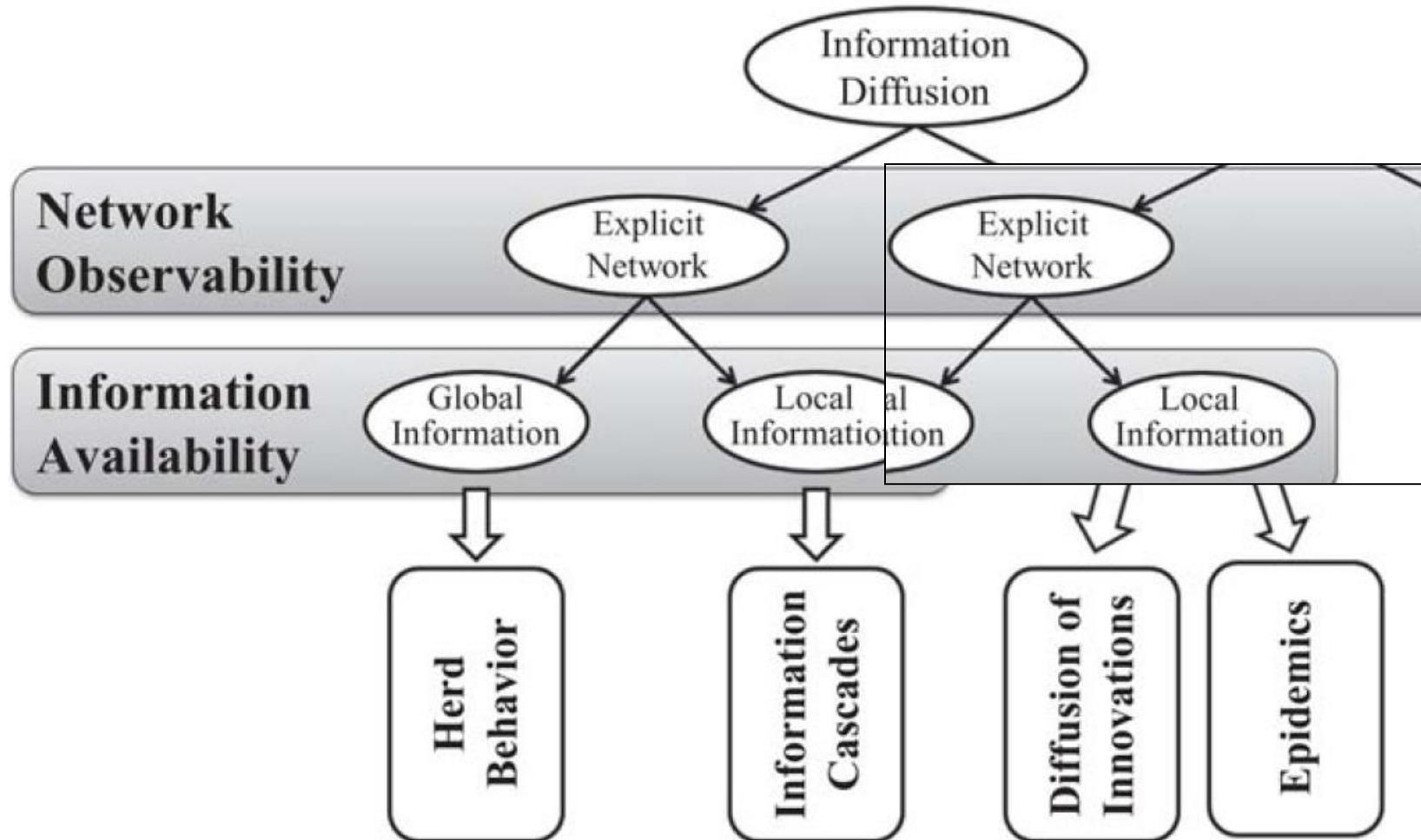


<http://www.caida.org/publications/visualizations/>



- Los modelos de difusión de información existentes permiten modelar distintas situaciones:
 - **Cascadas de Información y Modelos Epidémicos en redes**: la difusión se produce sólo vía los amigos (**contagio/decisión con información local**)
 - **Difusión de innovaciones en redes**: Tres variantes con **información global y local**: sólo innovación (global), sólo imitación (local) y mixto (global y local)
- En los modelos epidémicos y de cascada *centrados en el emisor*, los individuos no toman la decisión por si mismos. En el resto sí
- Los **contagios** pueden ser **simples** (individuales por probabilidad) y **complejos** (umbrales)





MODELOS CLÁSICOS DE PROPAGACIÓN DE EPIDEMIAS

[http://es.wikipedia.org/wiki/Modelaje_matemático_de_epidemias](http://es.wikipedia.org/wiki/Modelaje_matem%C3%A1tico_de_epidemias)

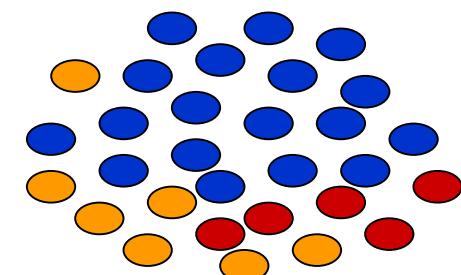
http://www.uni-tuebingen.de/modeling/Mod_Pub_Software_SIR_en.html

La **Epidemiología** describe el proceso mediante el que se difunden las enfermedades. Se basa en:

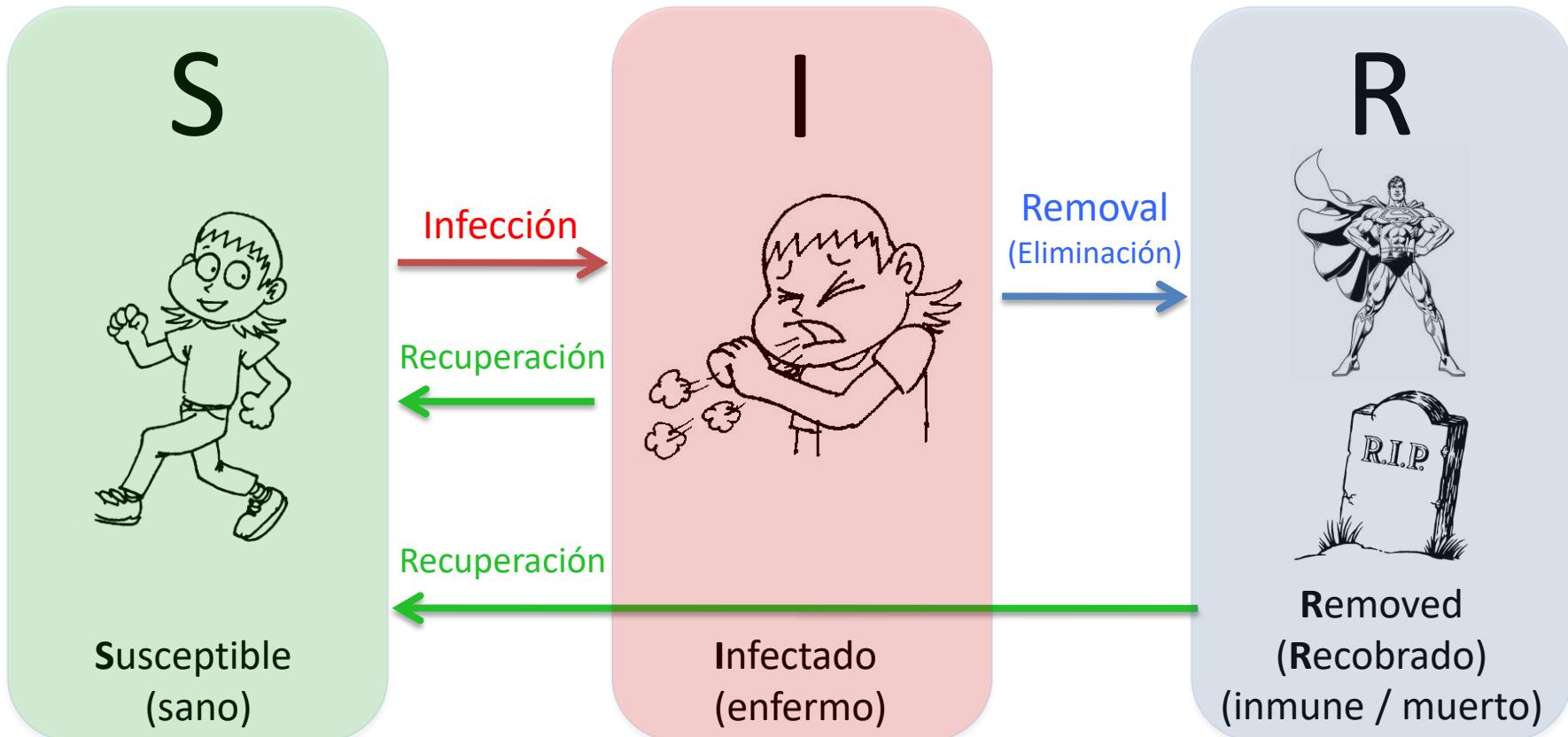
- Un **patógeno** (la enfermedad que se está transmitiendo),
- Una **población de huéspedes** (humanos, animales, plantas, etc.) en **distintos estados**
- Un **mecanismo de transmisión** (respiración, bebida, actividad sexual, etc.)

Los individuos se clasifican en varios estados según la etapa de desarrollo de la enfermedad (**compartimentación**):

- *Susceptible (S)*: Individuos sanos que no han contraído la enfermedad
- *Infectado (I)*: Individuos contagiados que la han contraído y pueden infectar a otros
- *Recobrado (R)*: Individuos previamente infectados que se han recuperado de la enfermedad y ya no son infecciosos



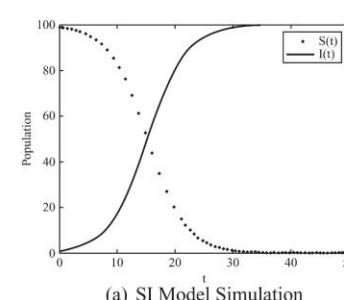
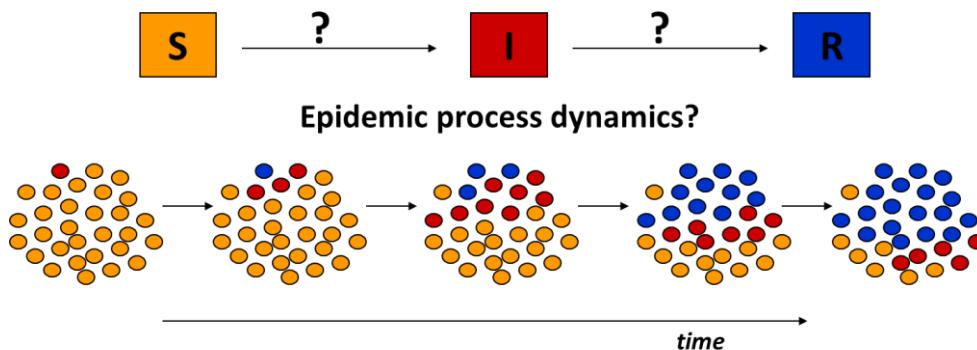
ESTADOS BÁSICOS Y TRANSICIONES DEL MODELO CLÁSICO SIR



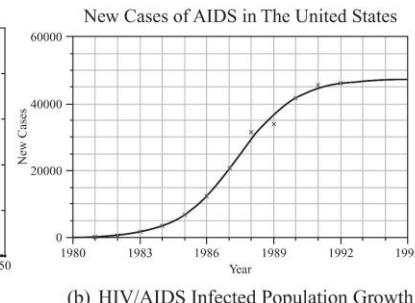
Los modelos clásicos (1927) asumen una red de contactos implícita y que no se conocen las conexiones entre los individuos → **No hay red = mezclado homogéneo (cada individuo puede infectar a cualquier otro en cualquier momento):**

- Soluciones analíticas basadas en ecuaciones diferenciales
- Sólo interesa la obtención de **patrones globales** (tendencias y ratios de población infectada) y no quién infecta a quién
- Estimación errónea de la dinámica de contagio (velocidad, picos, ...)

$$i(t) = \frac{i_0 \exp(\beta \cdot \langle k \rangle \cdot t)}{1 - i_0 + i_0 \exp(\beta \cdot \langle k \rangle \cdot t)}$$

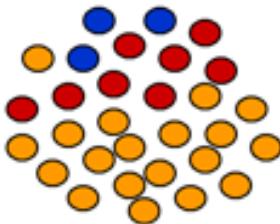


(a) SI Model Simulation



(b) HIV/AIDS Infected Population Growth

Logistic growth function compared to the HIV/AIDS growth in the United States



Mezclado Homogéneo: En cada unidad de tiempo, cada individuo tiene $\langle k \rangle$ contactos con otros individuos de la población escogidos aleatoriamente

La probabilidad de que quede infectado por esos contactos es $\beta \in [0,1]$ (**CONTAGIO SIMPLE**). El ratio de transmisión de la enfermedad es $\beta \cdot \langle k \rangle$ y determina el contagio

Los infectados se recuperan, volviéndose inmunes, o mueren con una probabilidad μ

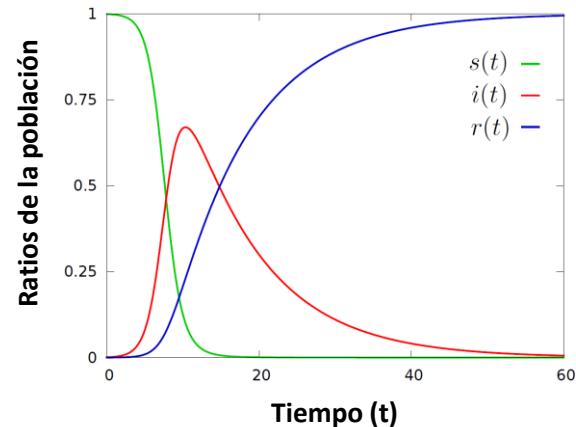
En tiempo t , la fracción de individuos infectados en una población de tamaño N es $i = I(t)/N$, la de individuos sanos $s = S(t)/N$ y la de individuos recobrados es $r = R(t)/N$. Lógicamente, $N = i + s + r$

Al haber i infectados, el **ratio medio de nuevas infecciones** en tiempo t es $\beta \cdot \langle k \rangle \cdot s \cdot i$: y el **ratio medio de recuperaciones/fallecimientos** es $\mu \cdot i$

$$\frac{ds(t)}{dt} = -\beta \langle k \rangle i(t) [1 - r(t) - i(t)]$$

$$\frac{di(t)}{dt} = -\mu i(t) + \beta \langle k \rangle i(t) [1 - r(t) - i(t)]$$

$$\frac{dr(t)}{dt} = \mu i(t).$$



Comportamiento temprano: Patrón de comportamiento de la epidemia en las fases iniciales. Es importante porque:

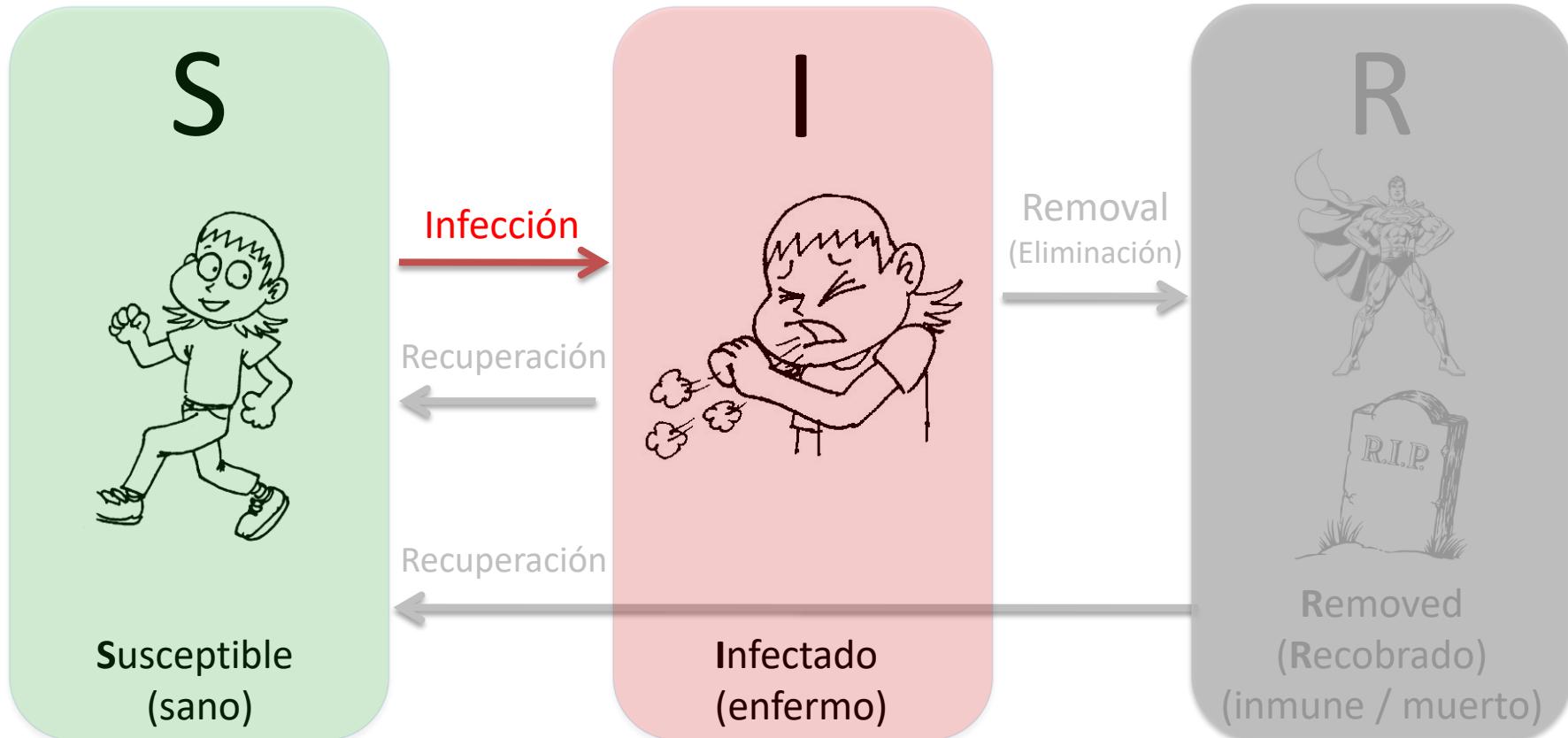
- Se necesita tiempo para desarrollar las vacunas y las intervenciones médicas
- La mejor forma de detener/contener la epidemia es la **cuarentena y/o la vacunación tempranas**

Un parámetro de estudio importante es el **tiempo característico τ , que indica el tiempo necesario para que se contagie una fracción $1/e$ de la población** (aprox. un 36%):
$$\tau = \frac{1}{\beta \cdot \langle k \rangle}$$

Es la inversa de la velocidad con la que el patógeno se extiende por la población. Aumenta con el número de contactos o la probabilidad de contagio

Comportamiento tardío: Patrón de comportamiento de la epidemia en las fases finales (cuando $t \rightarrow \infty$). Es importante porque:

- Permite medir (y por tanto predecir) el alcance de la epidemia, el número de individuos afectados (su pico), etc.



MODELO SI

Comportamiento



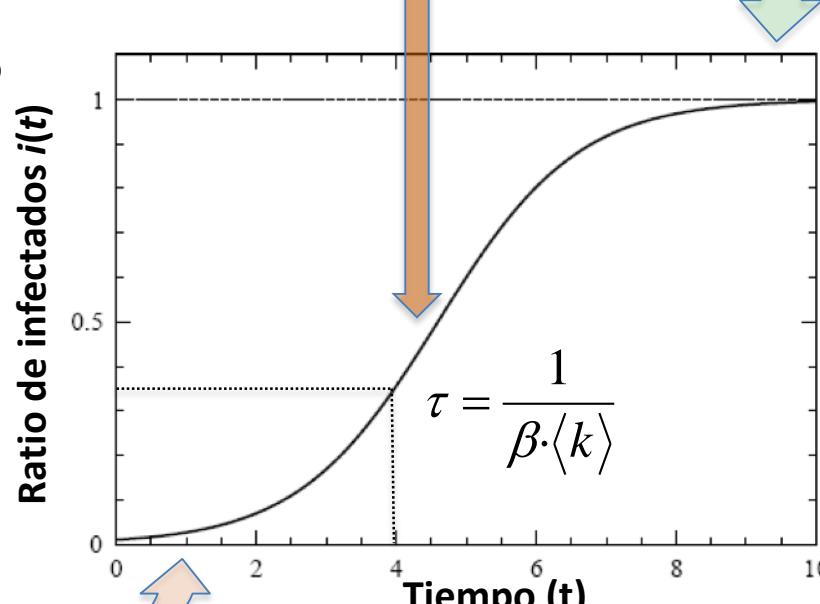
$$i(t) = \frac{i_0 \exp(\beta \cdot \langle k \rangle \cdot t)}{1 - i_0 + i_0 \exp(\beta \cdot \langle k \rangle \cdot t)}$$

Ecuación logística: modelo básico de crecimiento de poblaciones

Al principio, sólo hay uno o unos pocos individuos infectados

brote exponencial

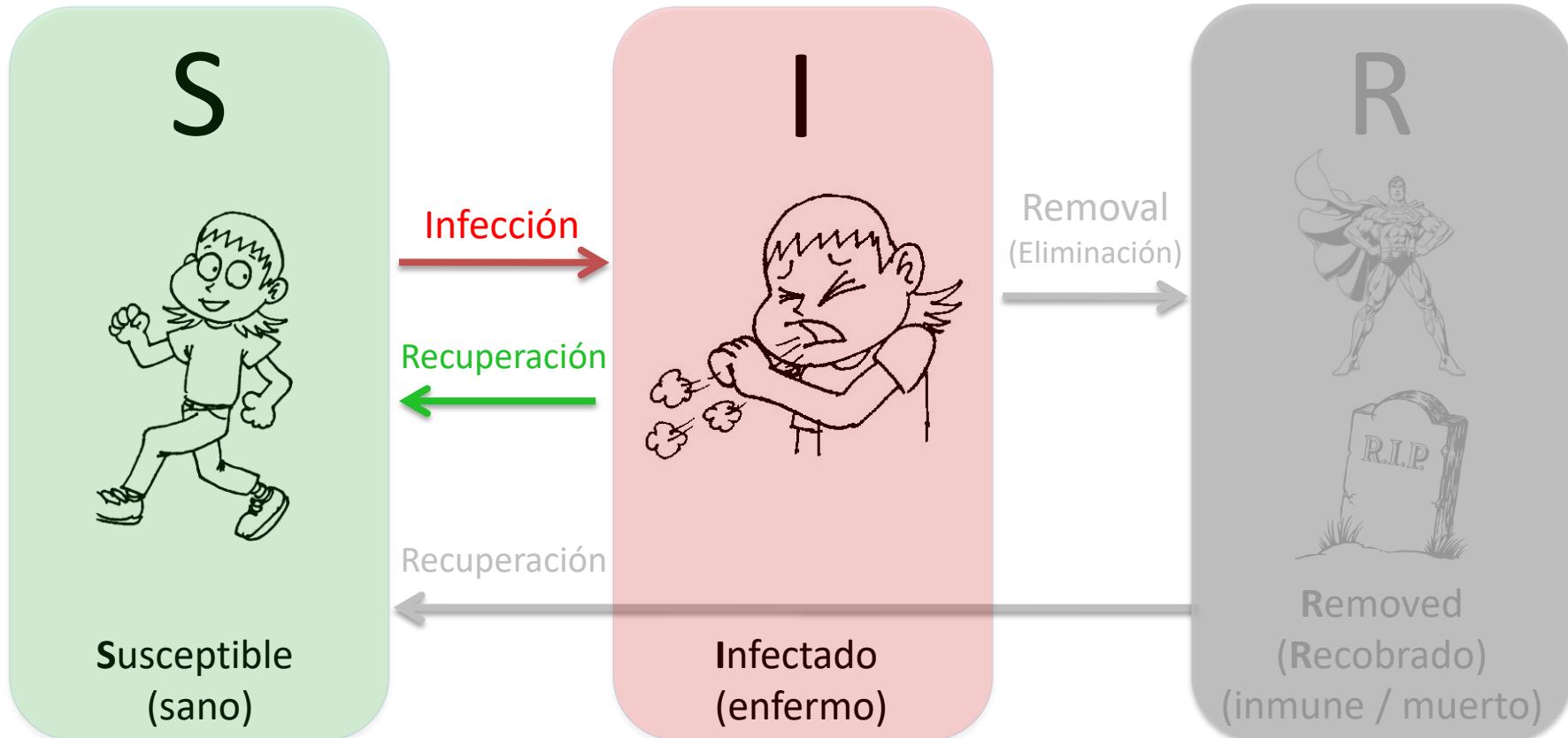
La pendiente de la curva depende del ratio de transmisión $\beta \cdot \langle k \rangle$

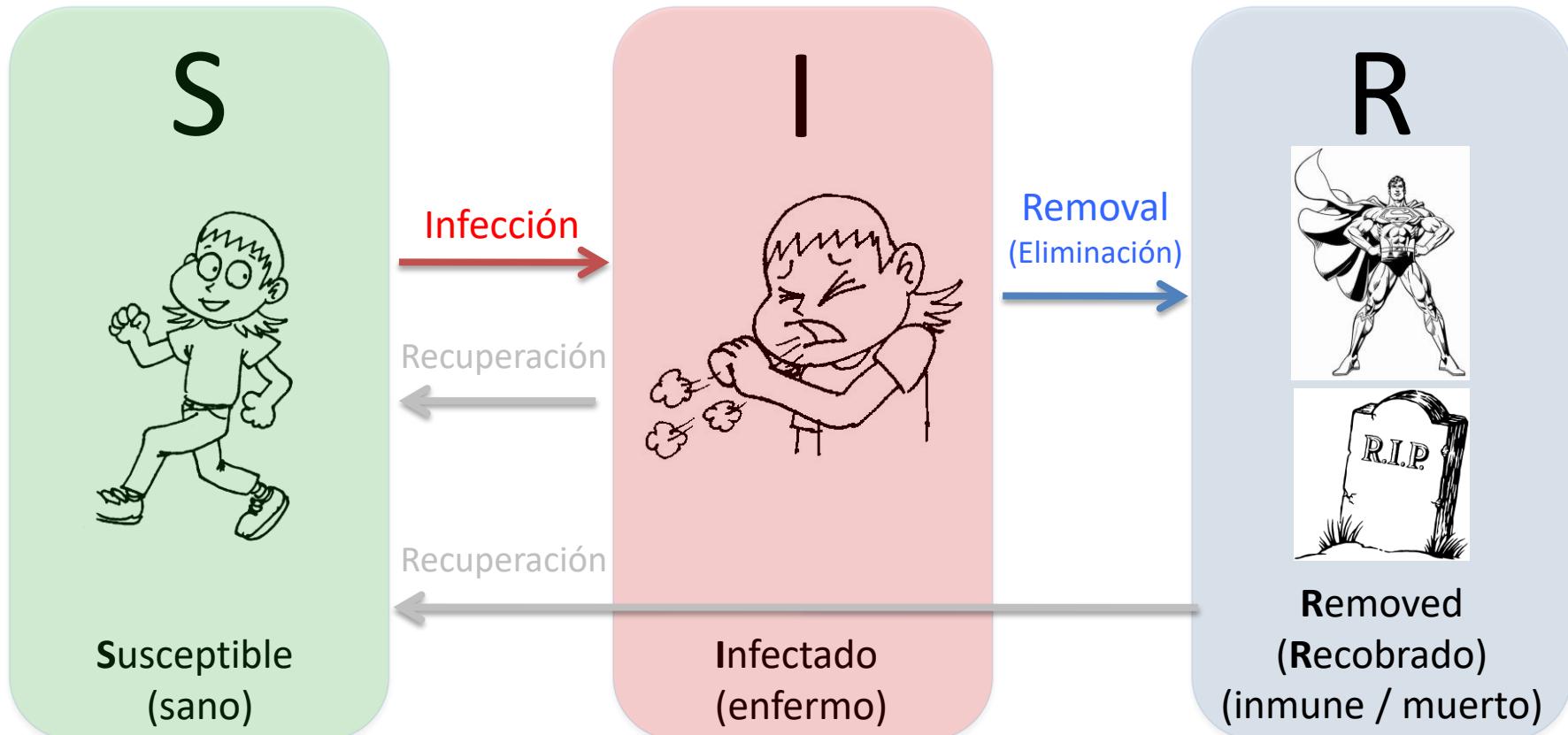


Al final, toda la población se infecta, aunque en cada paso hay menos nuevos infectados
saturación

http://es.wikipedia.org/wiki/Función_logística
<http://mathworld.wolfram.com/LogisticEquation.html>

Modelo SI: el ratio de infectados aumenta hasta que todo el mundo queda infectado

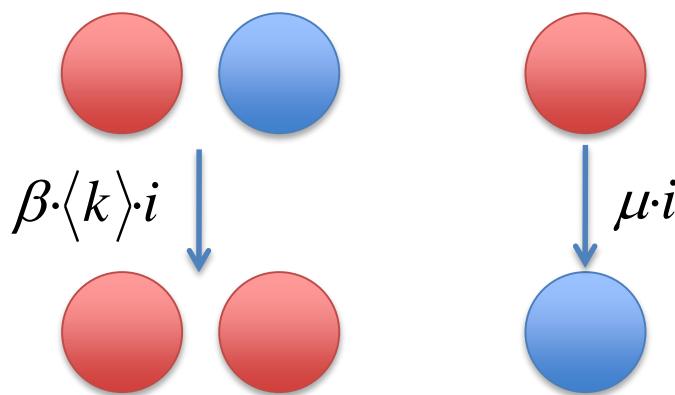




$$\frac{di}{dt} = \beta \cdot \langle k \rangle \cdot \underbrace{i}_{I} \cdot \underbrace{(1-i)}_{S} - \underbrace{\mu \cdot i}_{I \rightarrow R}$$

Si $\mu \approx \beta \cdot \langle k \rangle$, $i \rightarrow 0$

“Umbral epidemiológico”
 R_{0c} , Transición de fase



$$I = I + 1$$

$$I = I - 1$$

$$R_0 \equiv \frac{\beta \cdot \langle k \rangle}{\mu}$$

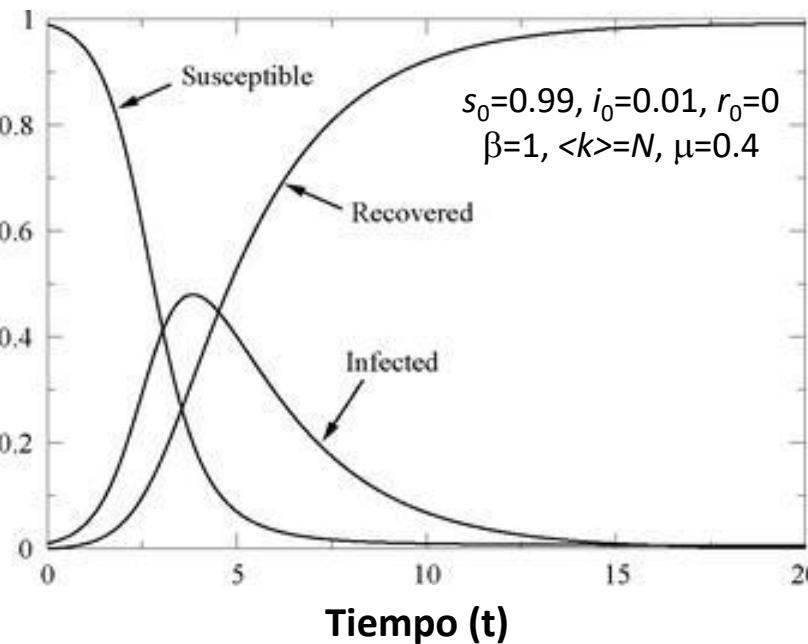
“Número reproductivo básico”

En promedio, indica cuántos individuos serán infectados por cada individuo infectado

$R_0 > 1$: Estado endémico (epidemia)

$R_0 < 1$: Estado sano (extinción del brote)

Ratios de la población



Condiciones iniciales más habituales:
o bien $I=1$ o bien $I=c$, un número pequeño

$$i_0 = \frac{c}{N} ; s_0 = 1 - i_0 = 1 - \frac{c}{N} ; r_0 = 0$$

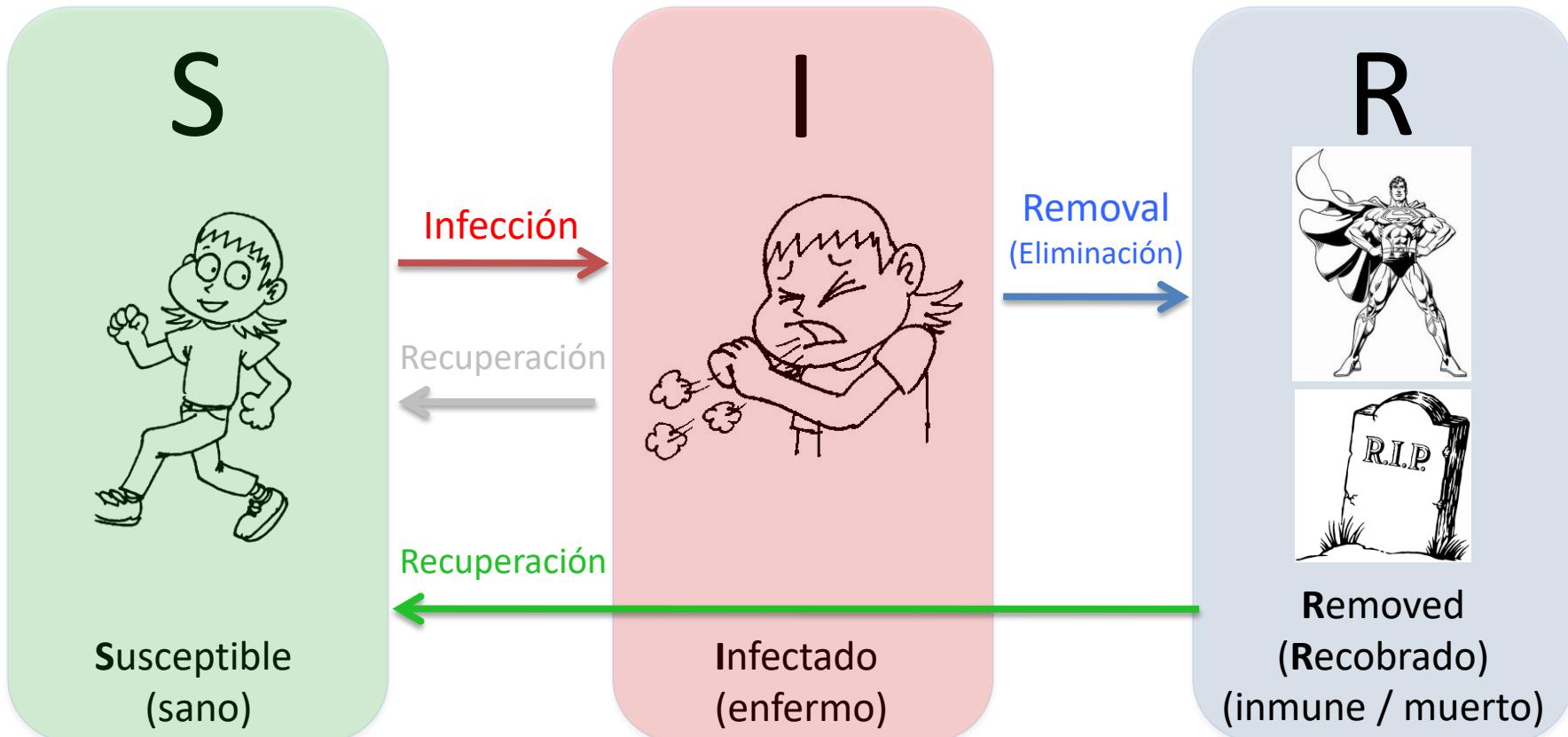
Modelo SIR: Cuando $\beta \cdot \langle k \rangle > \mu$ (estado endémico del modelo SIS), i crece hasta un **pico máximo y luego decrece hasta valer 0**

El ratio de susceptibles decrece de forma monótona y el de recobrados crece igual

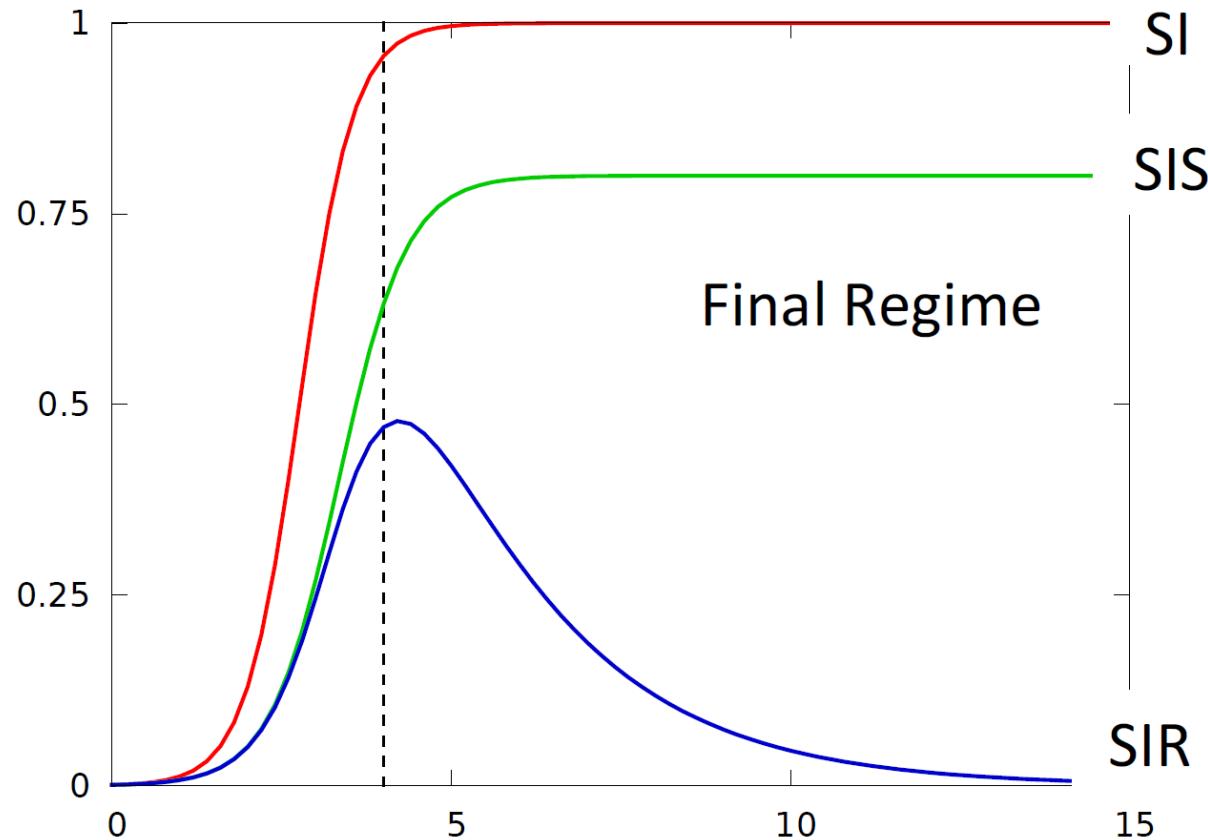
s satura pero no llega a cero porque con $i \rightarrow 0$ ya no hay individuos que puedan infectar

Los individuos que consiguen mantenerse sanos hasta fases avanzadas pueden no infectarse nunca

Igualmente, r nunca llega a valer 1



CARACTERÍSTICAS BÁSICAS DE LOS MODELOS EPIDÉMICOS (2)



Los tres modelos tienen un comportamiento temprano similar pero son muy distintos en el comportamiento tardío

Aplicación de los modelos epidemiológicos para predecir el ciclo de vida (la **dinámica de adopción y abandono de usuarios**) en redes sociales on-line (RSO) como **Facebook**

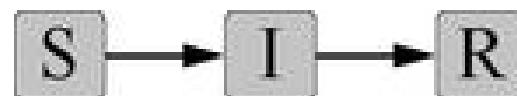
Se basa en la idea de que dicha dinámica se asemeja a los procesos epidémicos porque los usuarios se unen y abandonan la RSO porque sus amigos lo hacen



Se cambia ligeramente el modelo SIR tradicional y se usan datos de la única RSO conocida que “desapareció”, **MySpace**, para predecir el comportamiento futuro de Facebook

El nuevo modelo se denomina **infectious recovery SIR** (irSIR). Se modifica la recuperación porque los usuarios no piensan abandonar la RSO en principio. En teoría, los usuarios que se unen antes tienen más interés y se espera que permanezcan más tiempo

El proceso contagioso de abandono está controlado por los usuarios que dejan la RSO. Sus contactos consideran también abandonarla, como en el **churning** en compañías de móviles



S= individuos susceptibles de unirse a la RSO vía contagio. I= usuarios actuales de la RSO. R= individuos contrarios a unirse a la RSO (los que nunca se han unido y los que la han dejado)

Para incluir la dinámica de recuperación de la infección **se modifica su ratio para que sea proporcional a la población de recuperados**, los que provocan el abandono de la RSO:

Modelo SIR: $\frac{di}{dt} = \beta \cdot s \cdot i - \mu i ; \quad \frac{ds}{dt} = -\beta \cdot s \cdot i ; \quad \frac{dr}{dt} = \mu i ; \quad s + i + r = 1$

Modelo irSIR: $\frac{di}{dt} = \beta \cdot s \cdot i - \nu \cdot i \cdot r ; \quad \frac{ds}{dt} = -\beta \cdot s \cdot i ; \quad \frac{dr}{dt} = \nu \cdot i \cdot r ; \quad s + i + r = 1$

La probabilidad de recuperación se nota por ν para indicar el cambio (μ en el modelo SIR)

Symbol	Units	Disease Model Parameter	Equivalent OSN Model Parameter
S	People	Susceptible	Potential OSN users
I	People	Infected	OSN users
R	People	Recovered/Immune	Population opposed to OSN use
β	Time ⁻¹	Infection rate	Rate at which potential users join OSN
γ	Time ⁻¹	recovery rate	-
ν	Time ⁻¹	-	OSN abandonment rate

El ratio inicial de recuperados r_0 (individuos que se resisten inicialmente a unirse a la RSO) es un parámetro importante en el modelo irSIR . **Si $r_0=0$, no hay abandonos**

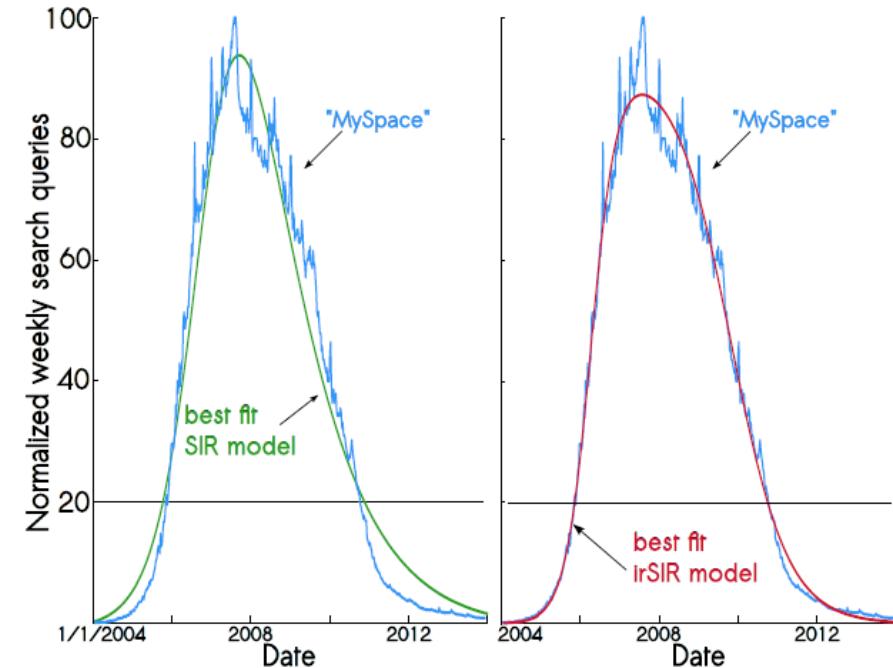
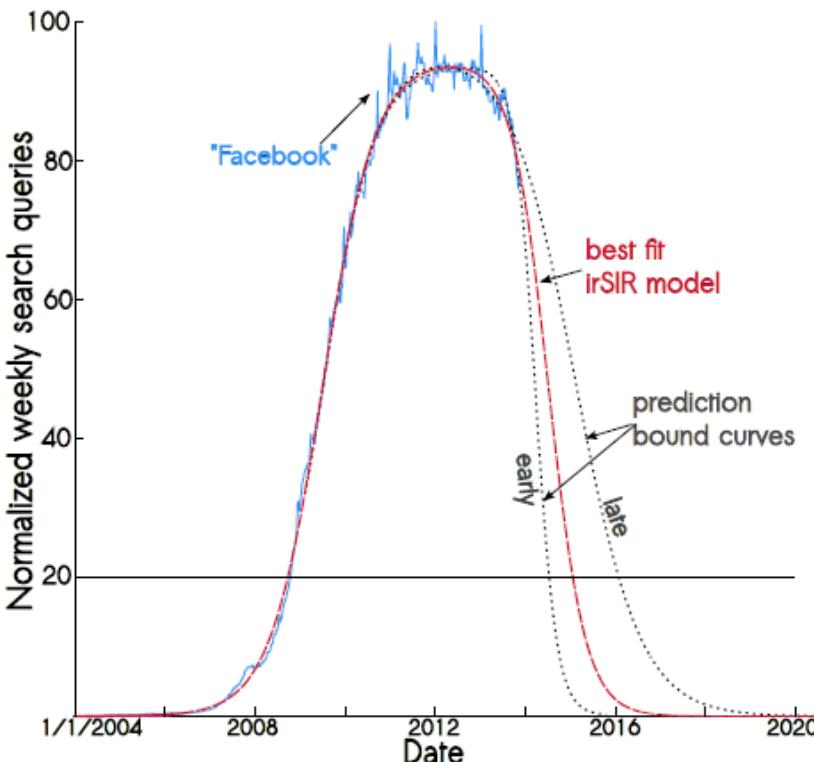
El umbral epidemiológico depende de ν , β , s_0 y r_0 . Se considera al revés que en SIR, se produce infección de recuperación (el proceso de contagio de abandono provoca que se reduzca progresivamente el número de usuarios de la RSO) si:

$$\frac{\nu}{\beta} > \frac{s_0}{r_0}$$

Salvo que $r_0=0$, esa es la dinámica habitual. Se espera que, **tarde o temprano, las RSOs siempre acaben perdiendo usuarios**

En los experimentos realizados, se ajustan los parámetros de los modelos SIR e irSIR con datos de Google Trends conocidos para MySpace.

El ajuste del modelo irSIR es mucho mejor



Para Facebook, se ajusta el modelo irSIR con datos 2004-2013 y se predice el comportamiento futuro

Se muestra también un intervalo de confianza con los tiempos de declive más cercano y más lejano

MODELOS DE PROPAGACIÓN DE EPIDEMIAS BASADOS EN REDES

REDES COMPLEJAS Y MODELADO DE EPIDEMIAS (1)

El enfoque clásico de modelado de epidemias **no tiene en cuenta explícitamente que la propagación se produce en una red compleja**

Asume que se da un **mezclado homogéneo**, lo que implica que un individuo puede ser infectado por cualquier otro individuo de la población con una probabilidad β en cada unidad de tiempo

Realmente, las epidemias se propagan a través de los contactos de las personas, es decir, a través de los enlaces de su red social. Por tanto, **hay que tener en cuenta el papel de la red en el proceso epidémico**

Las redes son el sustrato sobre el que se desarrolla el comportamiento dinámico del sistema. Al mismo tiempo, los distintos procesos dinámicos afectan a la evolución de la estructura de la red

REDES COMPLEJAS Y MODELADO DE EPIDEMIAS (2): Tipos

Fenómeno	Red compleja	Agente
Enfermedad venérea	Red sexual	Patógenos
Otras enfermedades infecciosas	Red de contactos, red de transporte	Patógenos
Chinches	Red de hoteles – viajeros	Chinches
Malaria	Red de mosquitos – humanos	<i>Plasmodium</i>
Propagación de rumores	Red de comunicaciones	Información, memes
Difusión de innovaciones	Red de comunicaciones	Ideas
Gusanos de Internet	Internet	Malware (códigos binarios)
Virus de teléfonos móviles	Red social / Red Bluetooth de proximidad	Malware (códigos binarios)

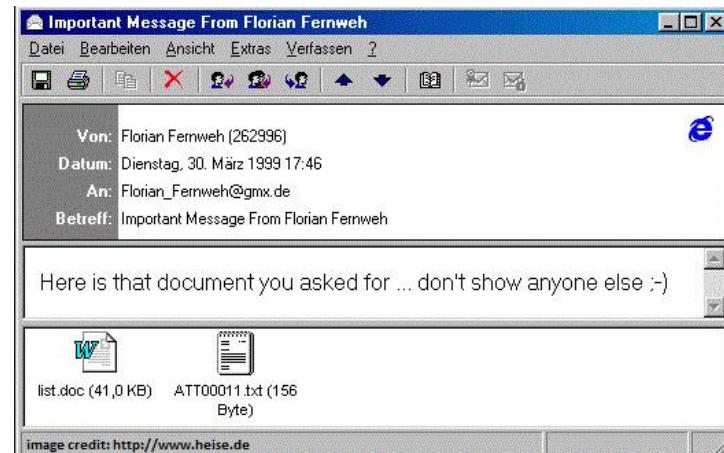
REDES COMPLEJAS Y MODELADO DE EPIDEMIAS (3): Ejemplo

Marzo 1999: El virus Melissa infecta ordenadores a través del gestor de e-mail MS Outlook

El usuario recibía un correo con un documento Word con virus en el attachment. Al abrirlo, el virus se reenviaba a los primeros 50 contactos de la libreta de direcciones

Se detectó por primera vez el viernes 26 de Marzo. El lunes 29 de Marzo el virus había infectado **más de 100.000 ordenadores**

Compañías como Microsoft, Intel o Lucent Technologies tuvieron que bloquear sus conexiones a Internet a causa de Melissa



REDES COMPLEJAS Y MODELADO DE EPIDEMIAS (3)

La estructura de la red, su evolución a lo largo del tiempo y su uso están mutuamente correlacionados y se deben estudiar conjuntamente:

La **topología de la red** influye los procesos que ocurren en el sistema complejo:

- ¿A qué estado convergen los nodos?
- ¿Cuánto se tarda en llegar a dicho estado?
- ¿Cómo se puede inmunizar un sistema complejo con una topología de red concreta?

El **mecanismo del proceso de difusión** también influye en el proceso global:

- **Contagio simple vs. contagio complejo:** En cada unidad de tiempo,
 - contagio simple: Cada “amigo” (nodo conectado a ti) infectado te infecta con una cierta probabilidad
 - contagio complejo: sólo realizas una acción si una cantidad/porcentaje de tus “amigos” lo hacen (**umbrales**)

MODELO SIR EN REDES COMPLEJAS: FUNDAMENTOS (1)

Los modelos epidémicos basados en redes se comportan como los clásicos, la diferencia es que consideran los contactos definidos por la red para la propagación

Todas variantes del modelo SIR pueden generalizarse usando redes complejas pero no es fácil resolver analíticamente modelos de este tipo para redes complejas generales

La alternativa es simular el proceso en el ordenador mediante técnicas de modelado social (modelado basado en agentes, ABM)

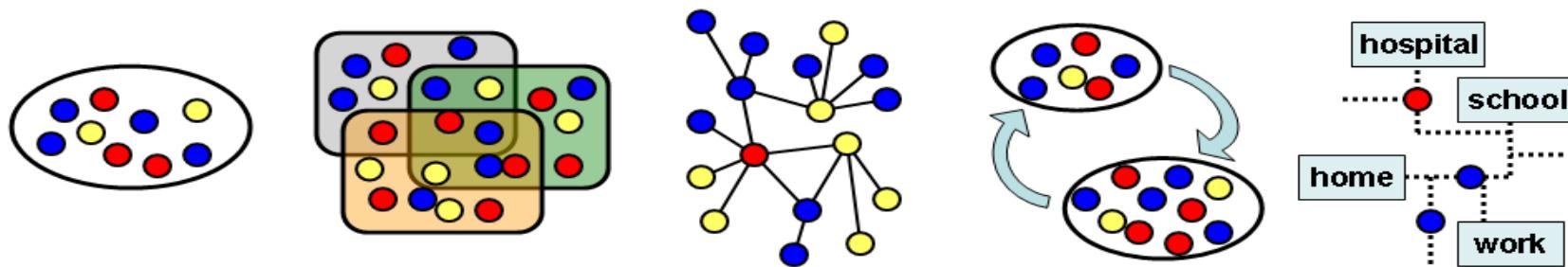
Esta metodología es mucho más potente. No se consideran globalmente los ratios de cada estado sino que se trabaja a nivel de cada individuo (agente)

Se conocen sus propiedades individuales, que pueden variar de unos a otros, y se consideran las interacciones locales (comportamiento emergente) y la aleatoriedad

P.ej. **se puede modelar el progreso del brote** en distintos escenarios con individuos de distintos tipos, distintas conexiones entre ellos, etc.

MODELO SIR EN REDES COMPLEJAS: FUNDAMENTOS (2)

En realidad, los distintos modelos tienen una complejidad creciente según se vayan considerando nuevas características más complejas de la realidad:



Homogeneous mixing

Social structure

Contact network models

Multi-scale models

Agent Based models

Simple



Realista

Habilidad para advertir/explicar
tendencias a nivel de población

El realismo del modelo hace que
se pierda transparencia
La validación es más compleja

MODELO SIR EN REDES COMPLEJAS: APLICABILIDAD

- Modelos epidémicos y modelos de propagación de virus informáticos:
Susceptibles, Infectados y Recobrados

Pastor-Satorras y Vespignani. *Epidemic spreading in scale-free networks*. Physical Review Letters 86 (2001) 3200–3203

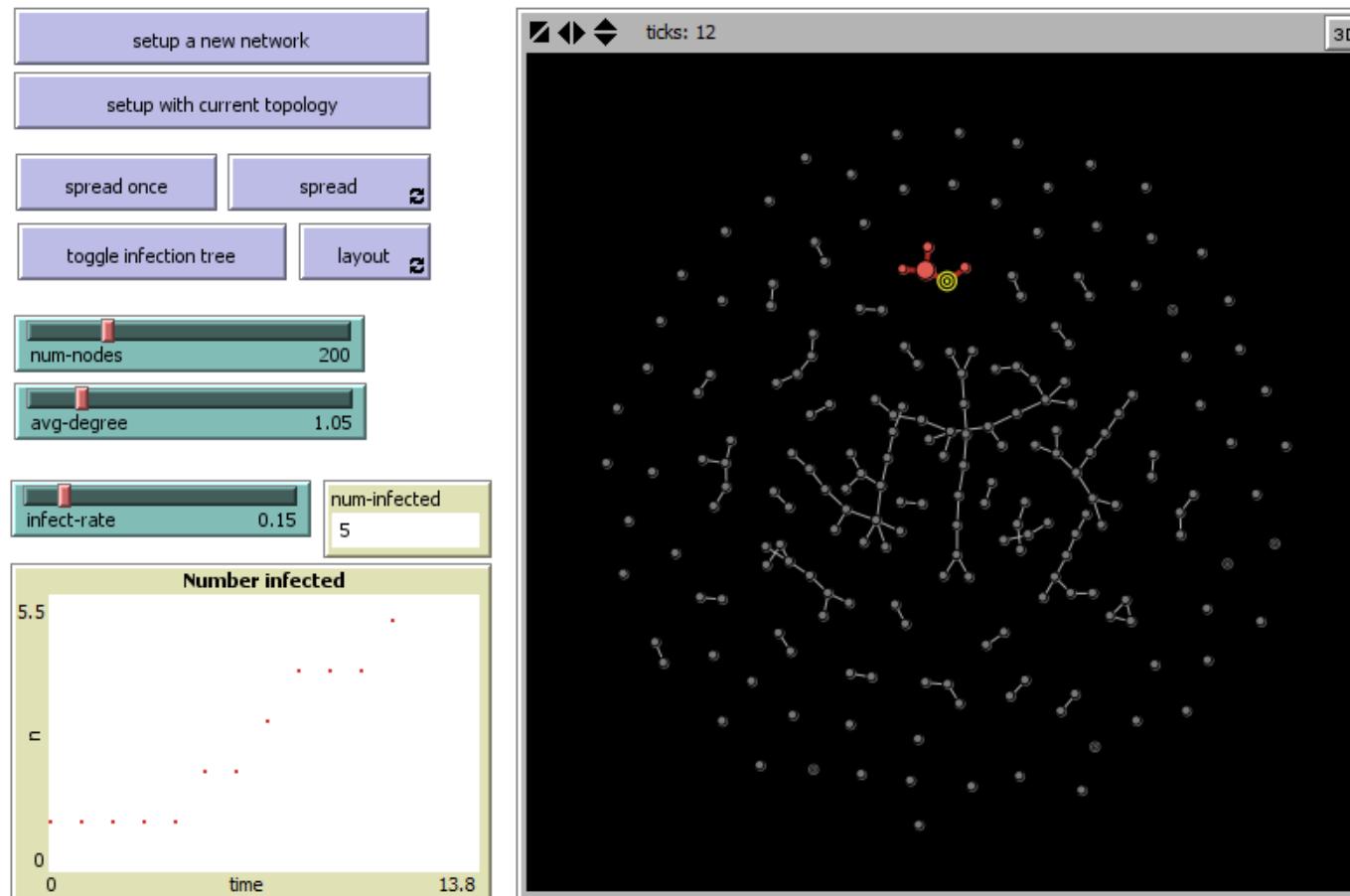
El modelo SIR es muy apropiado para las fases iniciales de un ataque de un virus de ordenador:

Pastor-Satorras y Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press. 2004

Tabah. *Literature dynamics: Studies on growth, diffusion, and epidemics*. Annual Review of Information Science & Technology 34 (1999) 249-286

- Modelos de propagación de rumores: Ignorantes, Difusores y Represores
- Daley, Gani y Cannings. *Epidemic modeling: An introduction*. Cambridge University Press. 1999
- Modelos de difusión de conocimiento: Innovadores, Incubadores y Adoptadores

<http://www.ladamic.com/netlearn/NetLogo501/ERDiffusion.html>



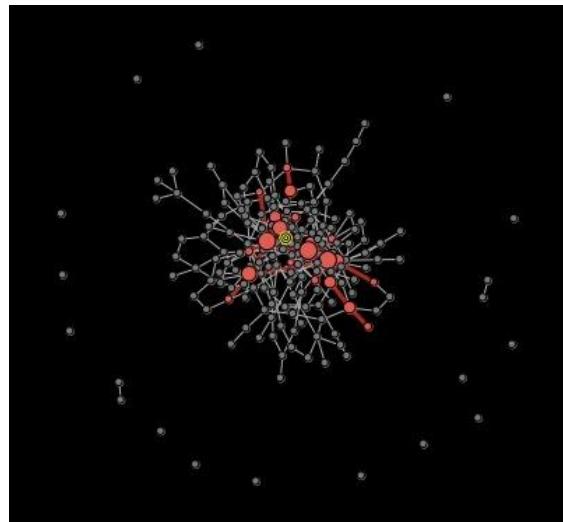
- Cuando $t \rightarrow \infty$, cada nodo susceptible con posibilidad de infectarse queda infectado
- La única condición es que exista un camino en la red entre dicho nodo y cualquier otro nodo infectado, de forma que la enfermedad pueda alcanzarlo
- **A diferencia del modelo clásico**, si partimos de un único nodo infectado, no todos los nodos se infectan, **sólo los que pertenecen a la misma componente conexa**
- Como la mayoría de las redes reales tienen una componente gigante y muchas componentes conexas pequeñas, **la epidemia se extiende en mayor o menor medida dependiendo de la localización del nodo inicial infectado**
- Si se escoge aleatoriamente, **la probabilidad de que pertenezca a la componente gigante y se produzca una pandemia en la población es $S=N_G/N$**
- Analizando las componentes pequeñas, se puede conocer la probabilidad de las pequeñas epidemias

INFLUENCIA DE LA CONECTIVIDAD Y LA DENSIDAD

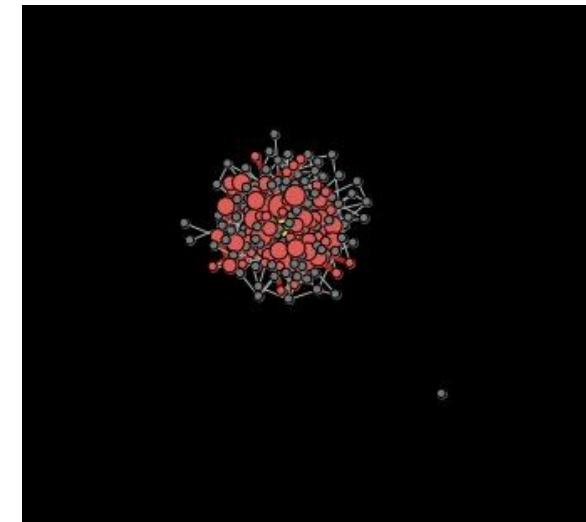
Con los mismos parámetros del modelo, ¡la difusión depende de la topología de la red!

Nodos infectados después de 10 pasos, ratio de infección = 0.15

grado medio $\langle k \rangle = 2.5$



grado medio $\langle k \rangle = 10$



PREGUNTA: Cuando aumenta la densidad de la red,
¿la propagación es igual, más rápida o más lenta?

<http://www.ladamic.com/netlearn/NetLogo501/BADiffusion.html>

Ratio de transmisión $\beta=1$. Nodos infectados tras cuatro unidades de tiempo:



PREGUNTA: Cuando los nodos tienen un acoplamiento preferencial, ¿la velocidad de propagación es igual, más rápida o más lenta?

<http://www.openabm.org/files/books/3443/ch13-SIRonnetwork.html>



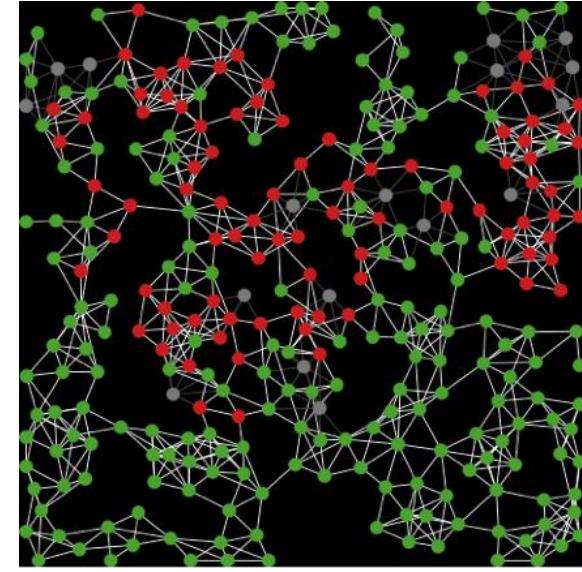
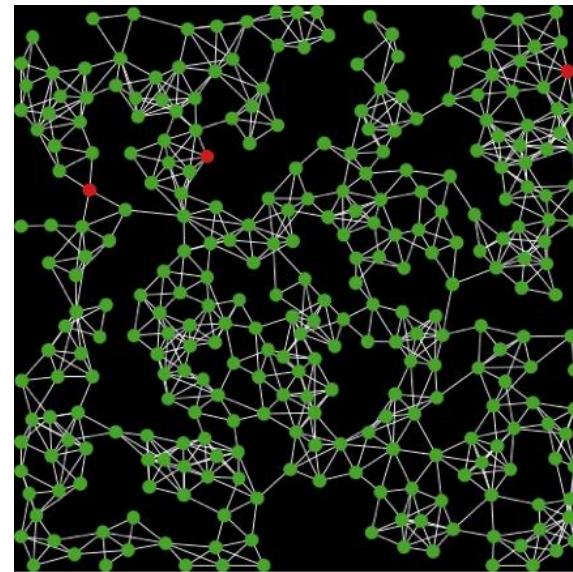
- **S→I:** Cada agente infectado puede infectar un agente susceptible conectado a él de acuerdo a la probabilidad (ratio) de infección β
- **I→R|S:** Cada agente infectado puede recuperarse con una probabilidad de recuperación μ . En ese caso, puede pasar a dos estados distintos:
 - **I→R:** El agente se vuelve inmune con probabilidad de inmunidad ι
 - **I→S:** El agente no se vuelve inmune y pasa de nuevo a susceptible

Ejemplo:

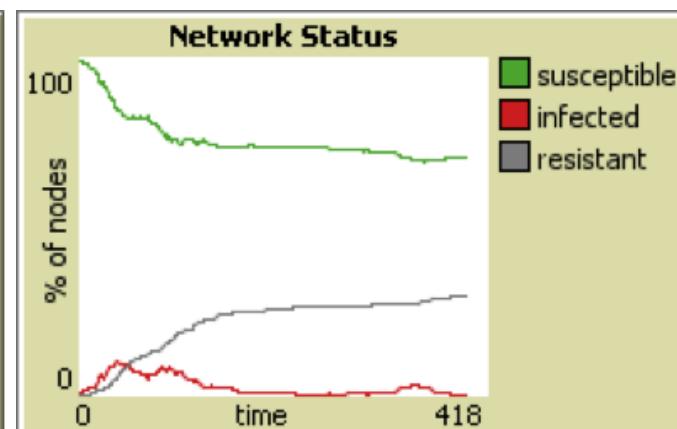
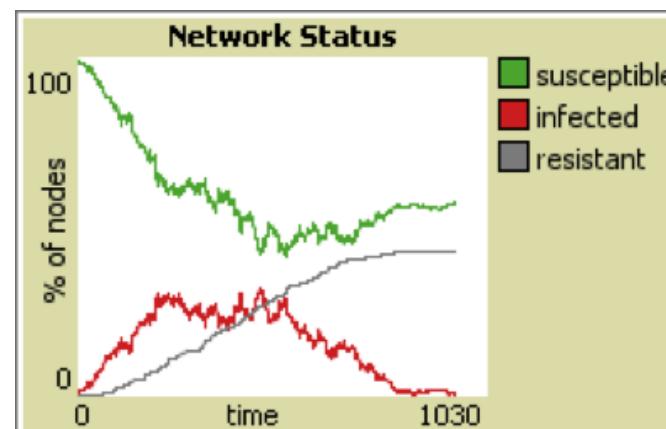
$\beta=0.025$, $\mu=0.05$, $\tau=0.025$

3 individuos infectados en la población inicial

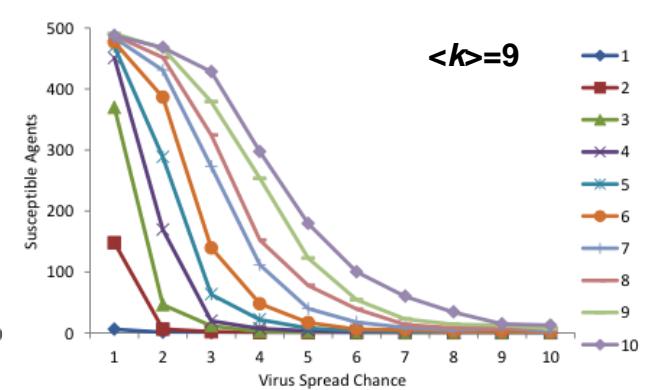
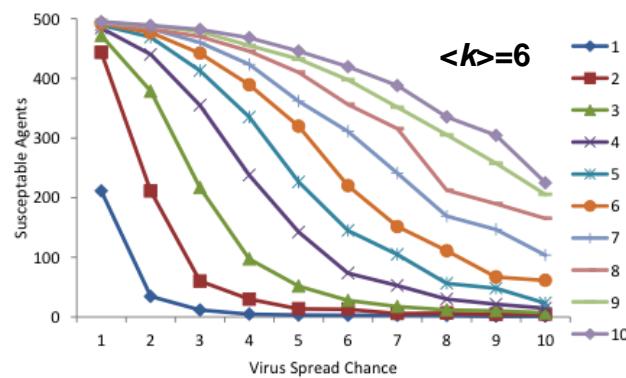
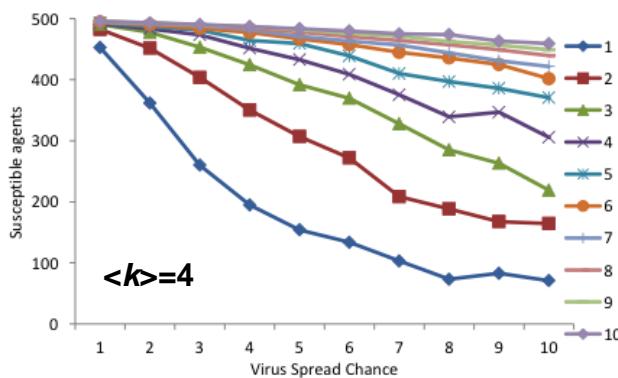
214 pasos de simulación

**Estudio del umbral epidemiológico:**

$\beta=0.025$, $\mu=\{0.05, 0.5\}$

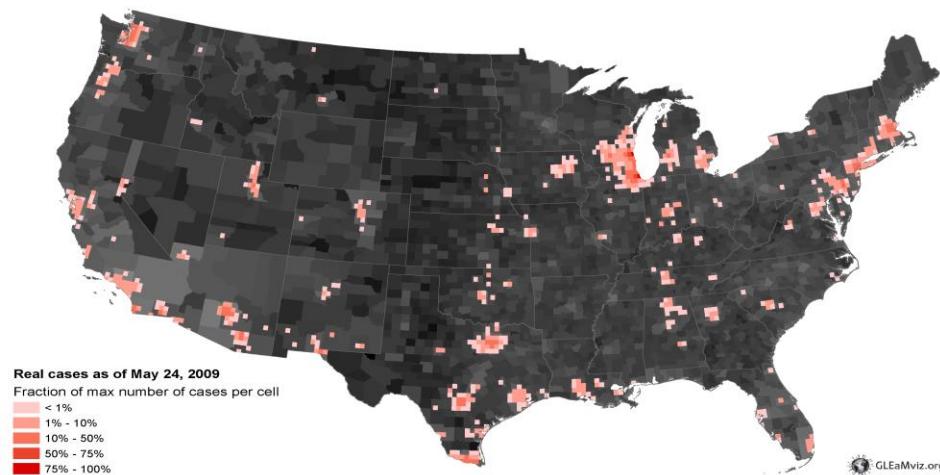


- Estudio de la influencia de la topología de la red (grado medio $\langle k \rangle$) y de los valores de las probabilidades de infección β y recuperación μ
- Modelo SIR: Los agentes recuperados, no pueden volver a ser susceptibles (modelo SIRS anterior con $\tau=0$)
- Tres redes ER con $\langle k \rangle=\{4, 6, 9\}$. $\beta = \mu = \{0.01, 0.02, \dots, 0.1\}$. 500 agentes. 100 simulaciones con 1000 pasos (**Montecarlo**)

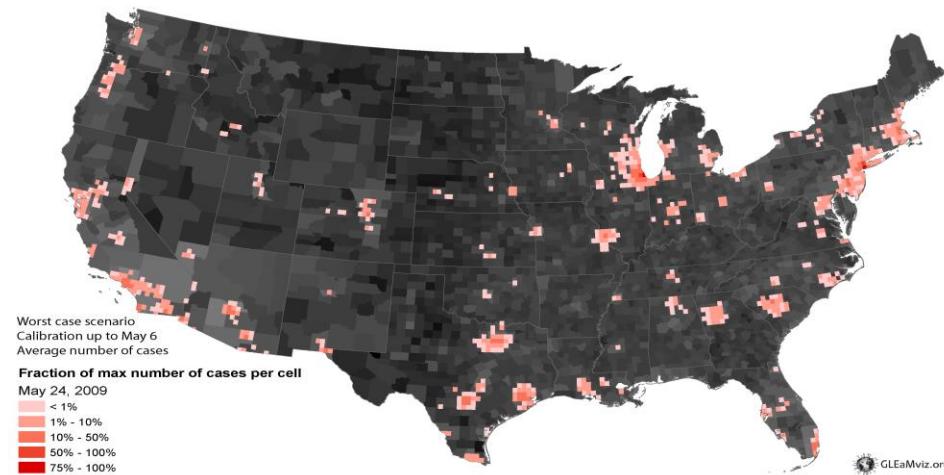


- Claramente, a mayor grado medio, mayor facilidad para la difusión del virus

Real



Pronosticada

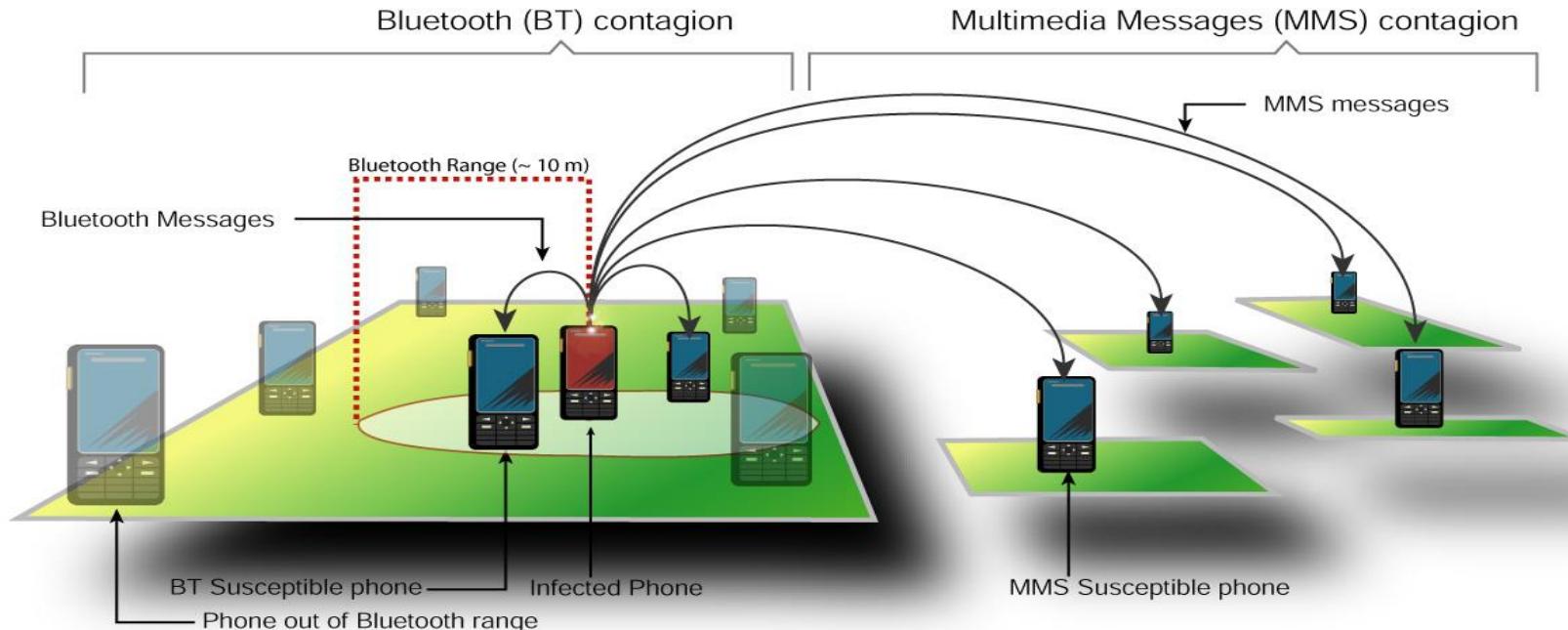


Primera pandemia analizada con redes complejas y ABM

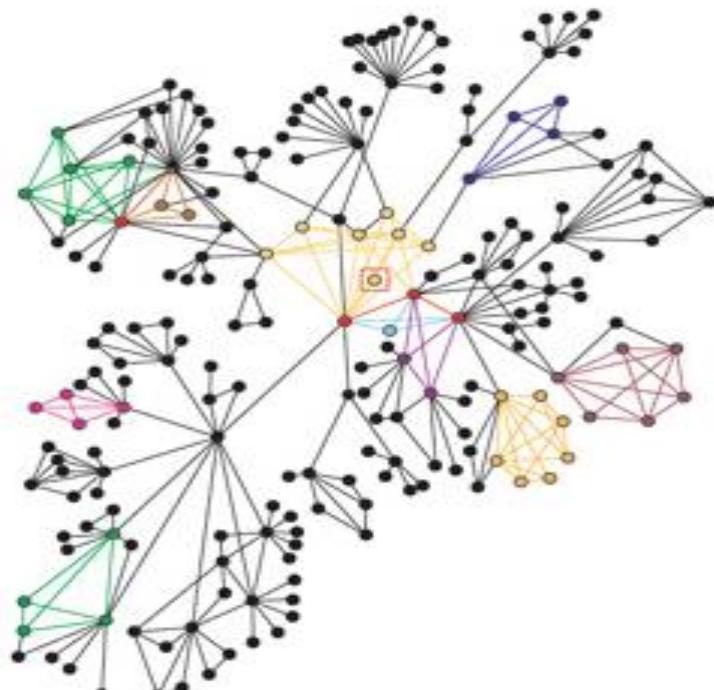


<http://www.gleamviz.org/simulator/>

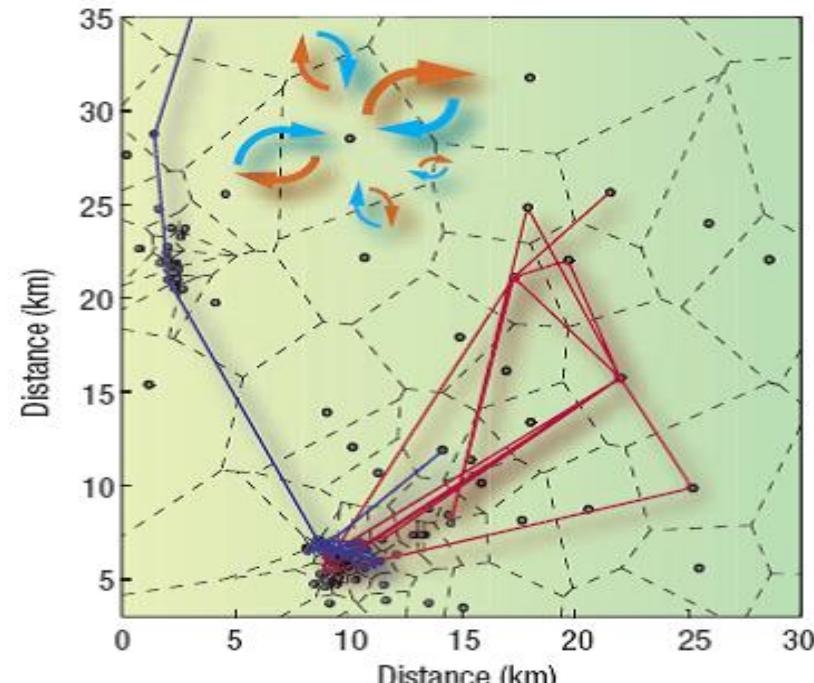
¿Cómo se transmiten los virus de móviles? Virus Bluetooth y MMS



Virus Bluetooth y MMS:



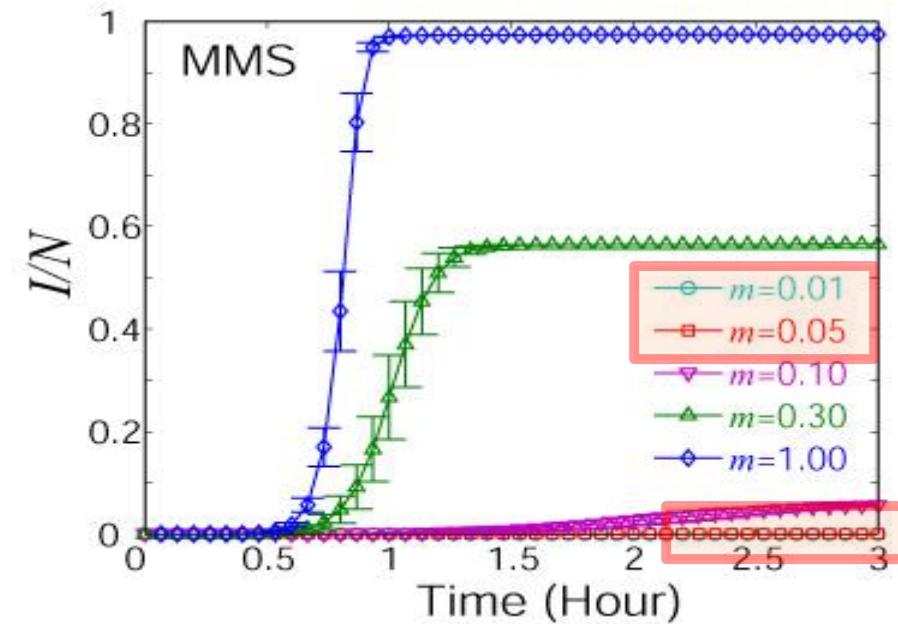
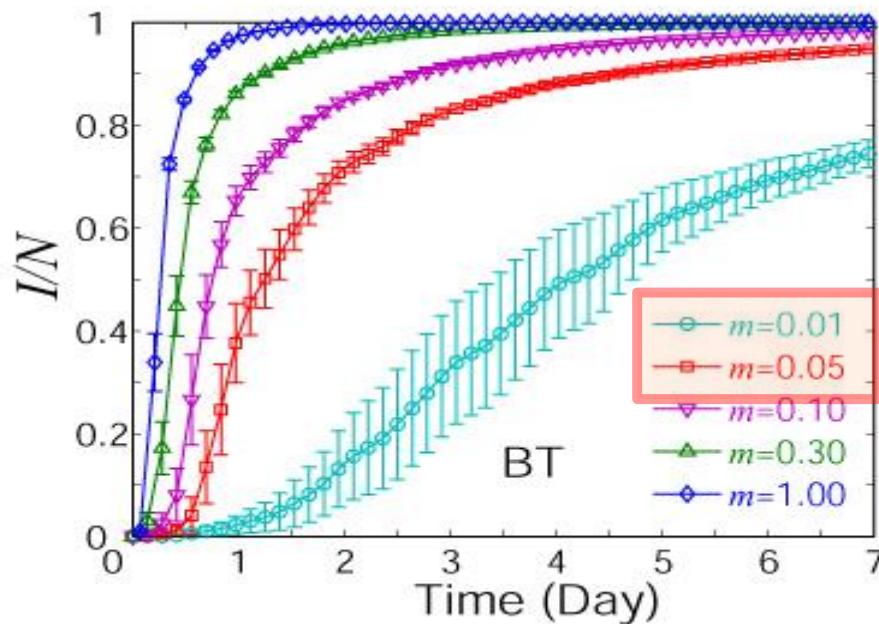
Red social (virus MMS)



Movilidad humana (virus Bluetooth)

Onella et al. PNAS (2007). Palla et al. Nature 446: 664 (2007)
González, Hidalgo y Barabasi. Nature 453: 779 (2008)

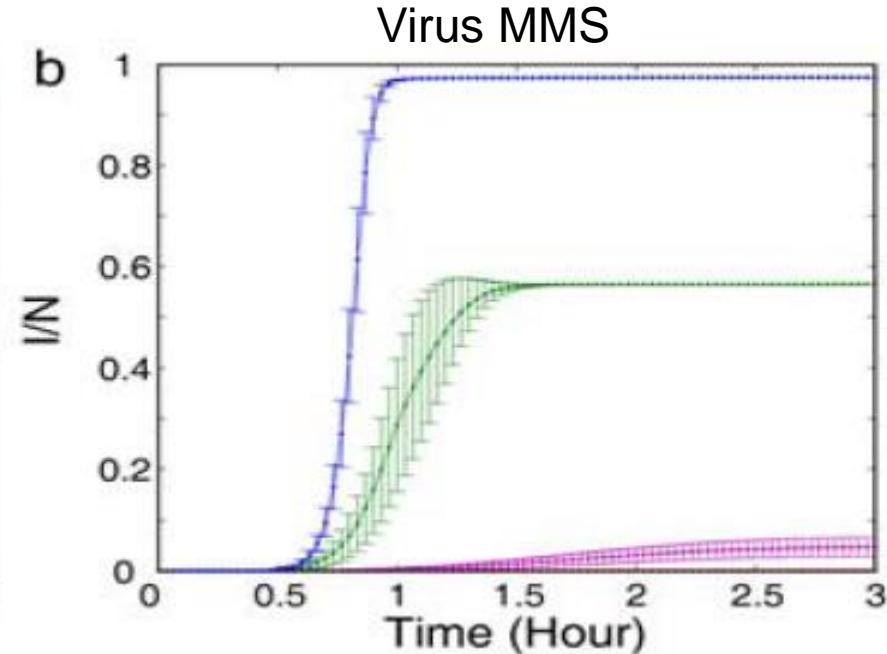
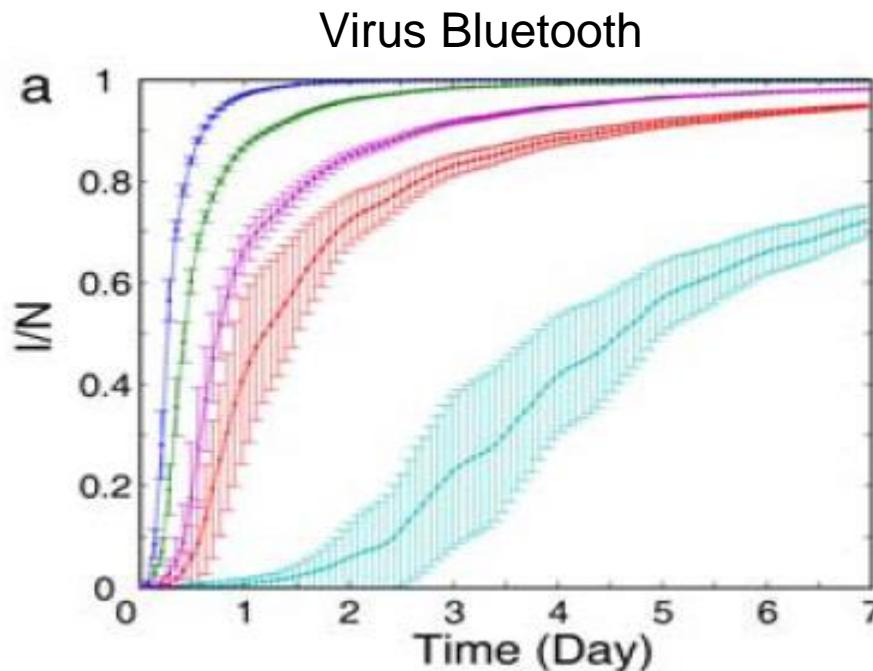
Patrones de difusión de los virus Bluetooth y MMS:



m : cuota de mercado del sistema operativo y/o el auricular que el virus puede infectar

Los *smart phones* tenían un $m=0.05$ (5%) del mercado completo de los móviles
El SO más extendido: Symbian, ~70% de todos los *smart phones*: $m_{max} \approx 0.03$

Patrones de difusión espacial de los virus Bluetooth y MMS:



Guiado por la movilidad humana:

Lento, pero puede alcanzar a todos los usuarios con el tiempo

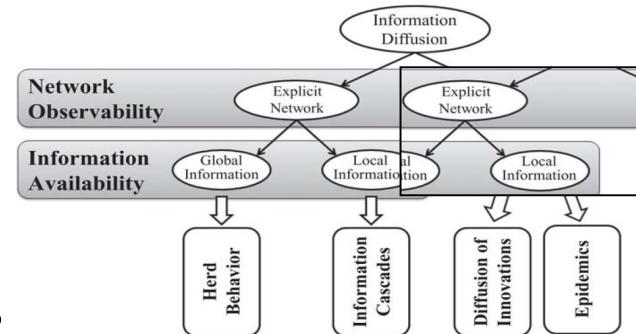
Guiado por la red social:

Rápido, pero sólo puede alcanzar una fracción finita de usuarios (la componente gigante)

OTROS MODELOS DE DIFUSIÓN DE INFORMACIÓN EN REDES

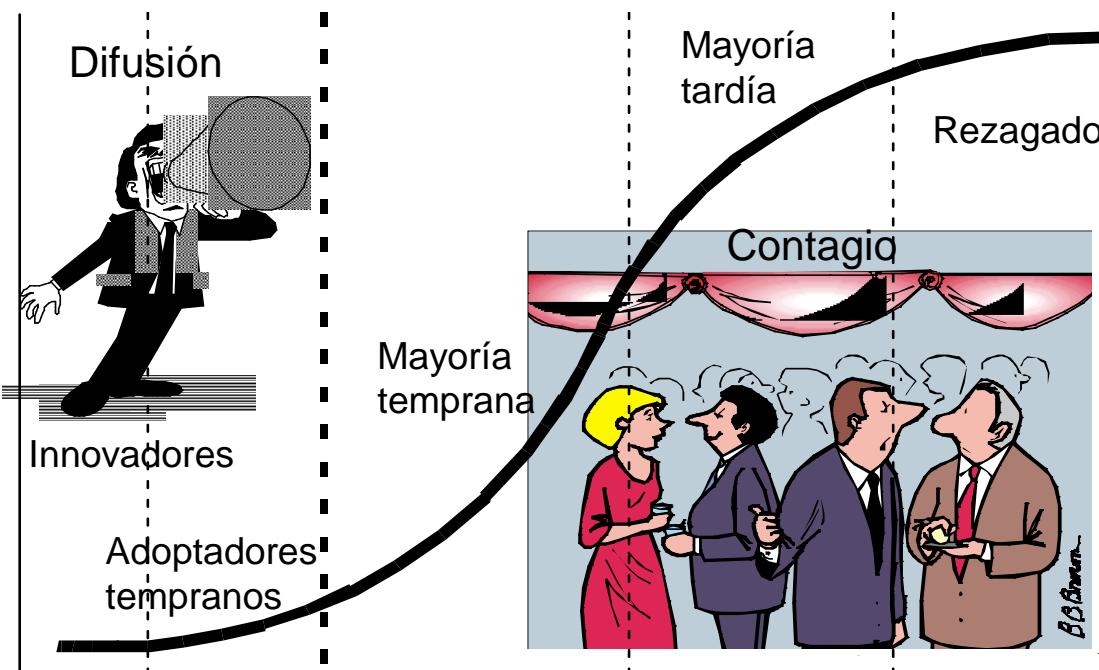
**Easley y Kleinberg. Cascading Behavior in Networks. Networks, Crowds, and Markets:
Reasoning about a Highly Connected World (Cap. 19). Cambridge University Press, 2010**

- Los modelos de difusión de información existentes permiten modelar distintas situaciones:
 - **Cascadas de Información y Modelos Epidémicos en redes:** la difusión se produce sólo vía los amigos (**contagio/decisión con información local**)
 - **Difusión de innovaciones en redes:** Tres variantes con **información global y local**: sólo innovación (global), sólo imitación (local) y mixto (global y local)
- En los modelos epidémicos y de cascada *centrados en el emisor*, los individuos no toman la decisión por si mismos. En el resto sí
- Los **contagios** pueden ser **simples** (individuales por probabilidad) y **complejos** (umbrales)



DIFUSIÓN DE LA INNOVACIÓN

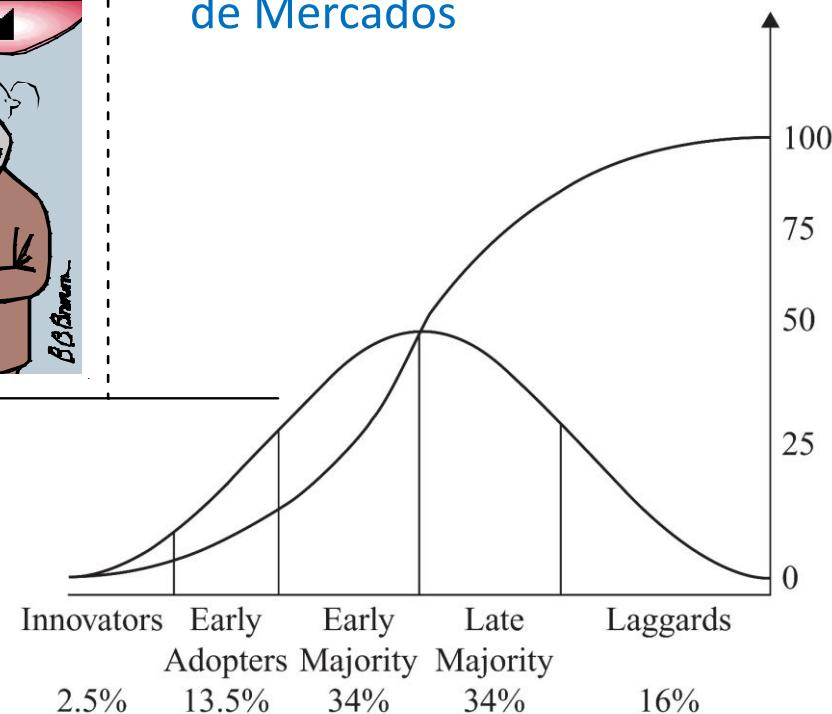
Curva de Adopción



La **Teoría de la Difusión de Innovaciones** estudia cómo y por qué se difunden esas innovaciones

Una **innovación** es una idea, práctica u objeto percibido como nuevo por un individuo

Muy estudiado en **Investigación de Mercados**



- La Teoría de la Difusión de Innovaciones estudia cómo y por qué se difunden y adoptan esas innovaciones
- Modelo clásico de Bass de 1969 (sin redes):

$$\frac{dA(t)}{dt} = i(t)[P - A(t)]. \quad i(t) = \alpha + \beta A(t)$$

α : factor de innovación (influencia externa, global)

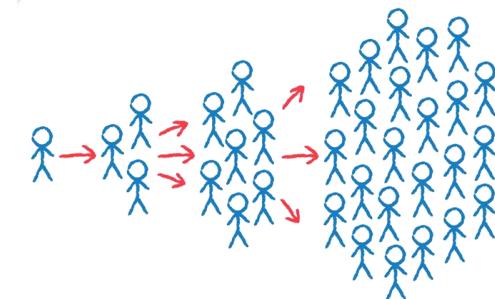
β : factor de imitación (influencia de los adoptadores, boca-a-oreja, local)

P: tamaño de la población

A(t): Ratio de adoptadores

- El modelo clásico es analítico: se asume mezclado homogéneo y sólo se consideran los ratios (obtención de las curvas de adopción)
- En los modelos basados en redes (ABMs) se consideran los adoptadores de entre los contactos del nodo (red social personal). Están centrados en el emisor

- En los medios sociales, los individuos suelen reenviar el contenido publicado por otros usuarios, recibido directamente de los vecinos directos (*amigos*)
- Las **cascadas de información** se producen cuando **la información se propaga únicamente a través de los amigos**
- No hay ninguna información global externa que influya en la propagación (p.ej. la viralidad de un tweet (*trending topics*), la publicidad de un producto, ...)
- Las decisiones son binarias, los nodos pueden ser activos o inactivos
- Los nodos vecinos **influyen** al individuo para que adopte el comportamiento, la innovación o la decisión
- **Hay modelos centrados en el emisor y en el receptor**



1996-97: Hotmail fue uno de las primeras compañías en usar marketing viral con gran éxito

Sólo tuvo que añadir la frase “*Get your free Email at Hotmail*” en el pie de cada correo enviado por sus usuarios

Consiguió **12 millones de usuarios en 18 meses**

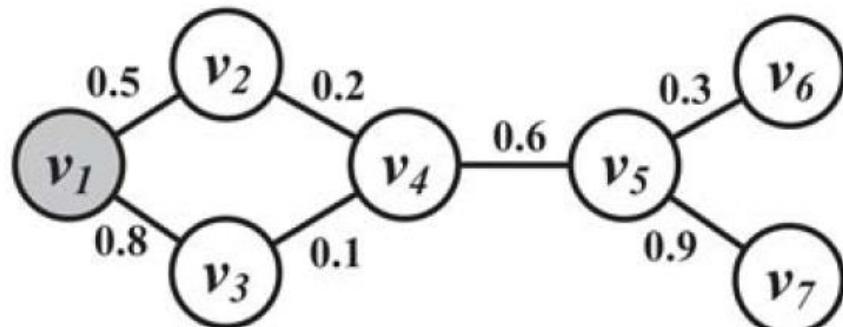
Fue el crecimiento mayor y más rápido de una empresa basada en usuarios en la época. Cuando alcanzó los 66 millones de usuarios, **la compañía creaba 270.000 nuevas cuentas cada día**



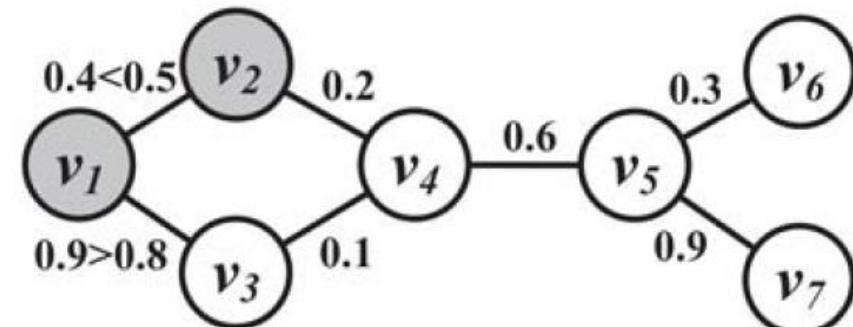
Hotmail®

--
Get your free Email at [Hotmail](#)

- El **Independent Cascade Model** es uno de los modelos más conocidos
- Es un **modelo centrado en el emisor** en el que cada individuo tiene una probabilidad de influir a cada uno de sus vecinos
- Una vez activado, el nodo puede influir a sus vecinos **una sola vez**, en un proceso progresivo
- **En principio**, no existe la posibilidad de desactivarse



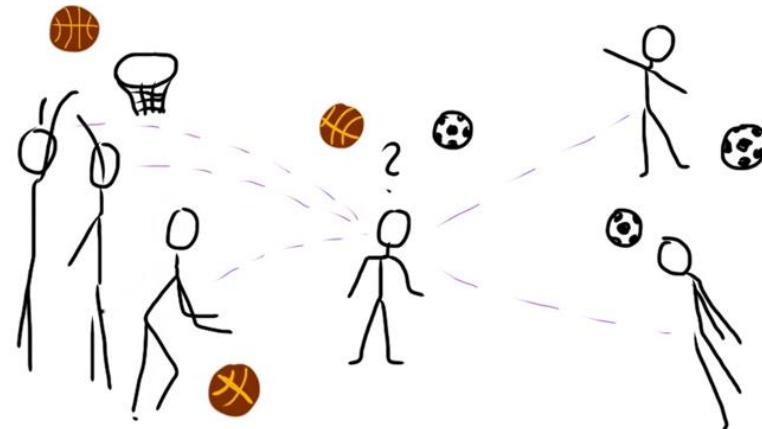
Step 1



Step 2

- En los **modelos de contagio complejo**, el contagio no se produce por un único contacto infectado sino requiere dos o más (Ej. credibilidad de una leyenda urbana)
- En los **modelos de umbral** la adopción requiere superar un umbral asociado a un porcentaje de contactos infectados (20%, 30%, ...). Los umbrales pueden ser comunes a toda la población o particulares a cada agente
- Pueden incluir información sobre las ventajas de la adopción (**recompensas**). Ej. dos opciones, futbol o baloncesto, d contactos:

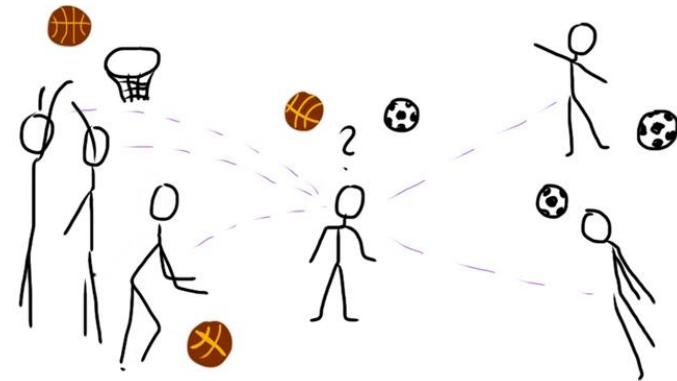
Un ratio $p = 3/5$
juegan a
baloncesto



Un ratio $p = 2/5$
juegan a fútbol

¿Qué elección proporciona una mayor recompensa?

- Un actor tiene d vecinos
- Un ratio p juegan a baloncesto (B)
- Un ratio $1-p$ juegan a futbol (F)
- Si escoge B , recibe $p \cdot d \cdot r_b$
- Si escoge F , recibe $(1-p) \cdot d \cdot r_f$
- Así, escogerá B si: $p \cdot d \cdot r_b \geq (1-p) \cdot d \cdot r_f$ o $p \geq r_f / (r_b + r_f)$ (**umbral adopción q**)



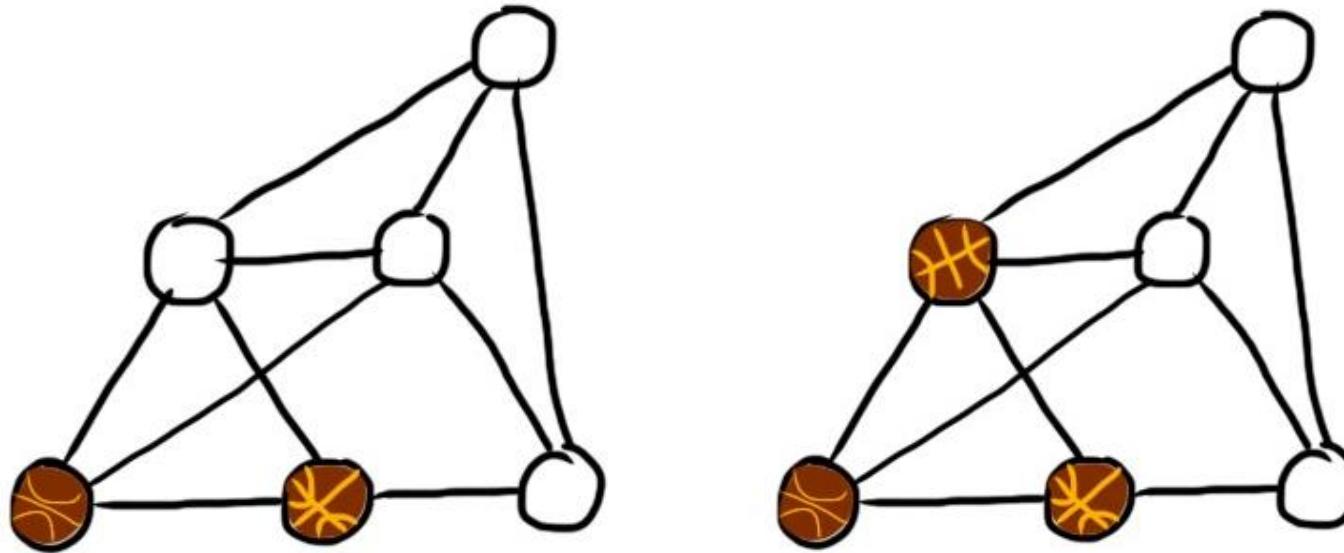
El sistema tiene dos estados de equilibrio posibles: todos adoptan F o B

¿Qué ocurre en el proceso intermedio?

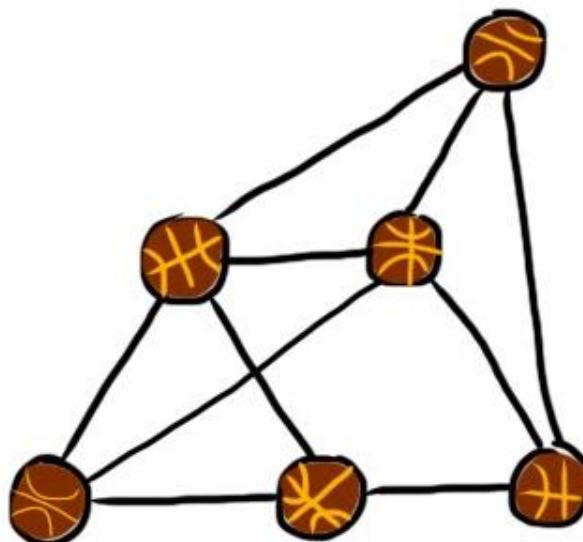
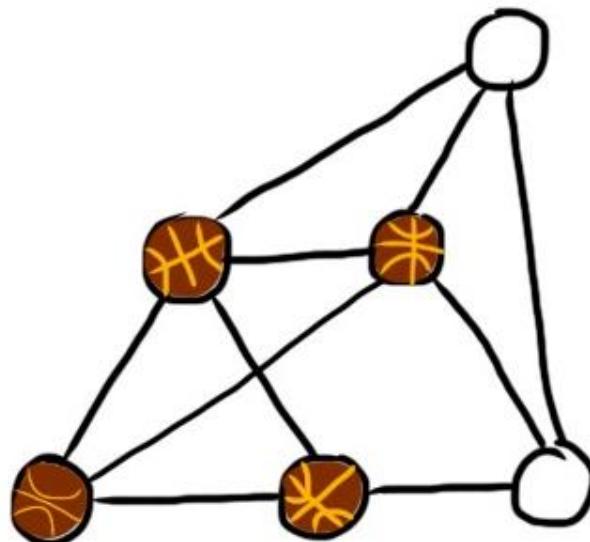
- ¿Qué pasa si dos nodos cambian su estado de forma aleatoria? ¿Se producirá una **propagación en cascada**?
- Ejemplo:
 - Valores de las recompensas: $r_b=3$, $r_f=2$
 - La recompensa para la interacción de dos nodos con comportamiento B es 3/2 veces mayor que la que obtendrían si ambos escogieran F
 - Los nodos cambiarán de F a B si al menos $q = r_f/(r_b+r_f) = 2/(3+2) = 2/5$ de sus vecinos están jugando a B (**umbral de adopción/cambio**)

¿Cómo se produce una propagación en cascada?

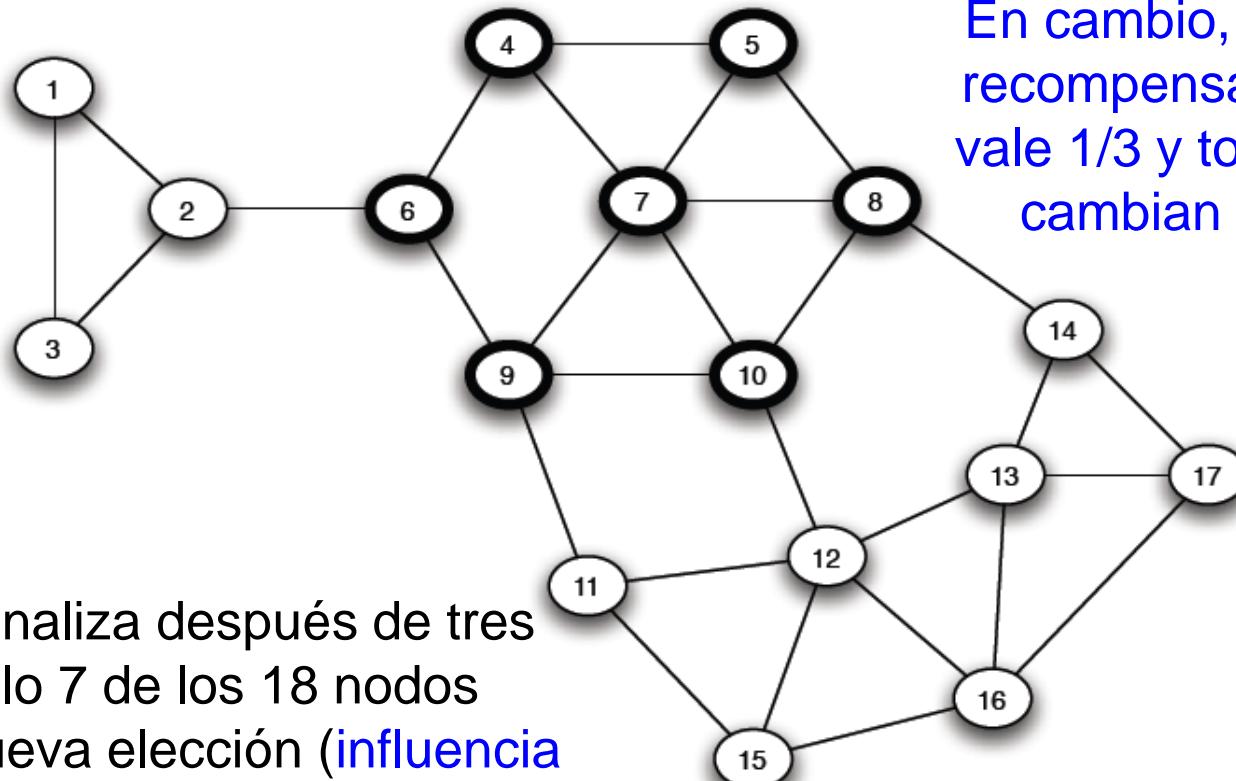
- Supongamos que dos nodos empiezan a jugar baloncesto:



¿Cuál será el siguiente nodo que se cambie al baloncesto?



Otro ejemplo con los mismos parámetros en el que no se converge a una única elección (no se produce una cascada completa)



En cambio, si aumenta la
recompensa a de 3 a 4, q
vale 1/3 y todos los nodos
cambian de elección

El proceso finaliza después de tres pasos. Sólo 7 de los 18 nodos adoptan la nueva elección (influencia de la estructura de comunidades)

En procesos epidémicos, las intervenciones van orientadas a interferir con la propagación del patógeno en la red de contactos para retrasarla/detenerla

En otros procesos de difusión/adopción podemos estar interesados en **acelerar la propagación**. Un ejemplo claro es el **marketing viral (WoM)**

Se basa en que las conversaciones de consumidores sobre un producto son una herramienta más poderosa que la publicidad tradicional

Cuando no se puede aumentar la recompensa (calidad/precio del producto), una estrategia habitual es “**recompensar**” a unos pocos individuos (con un pequeño presupuesto de campaña) para promocionar el producto entre sus amigos **buscando una adopción grande y rápida** (una cascada que redunde en un aumento de ventas)



El problema consiste en decidir **qué individuos escoger** para maximizar cantidad y ratio de adopción del producto. Es similar a la decisión de qué individuos vacunar pero a la inversa

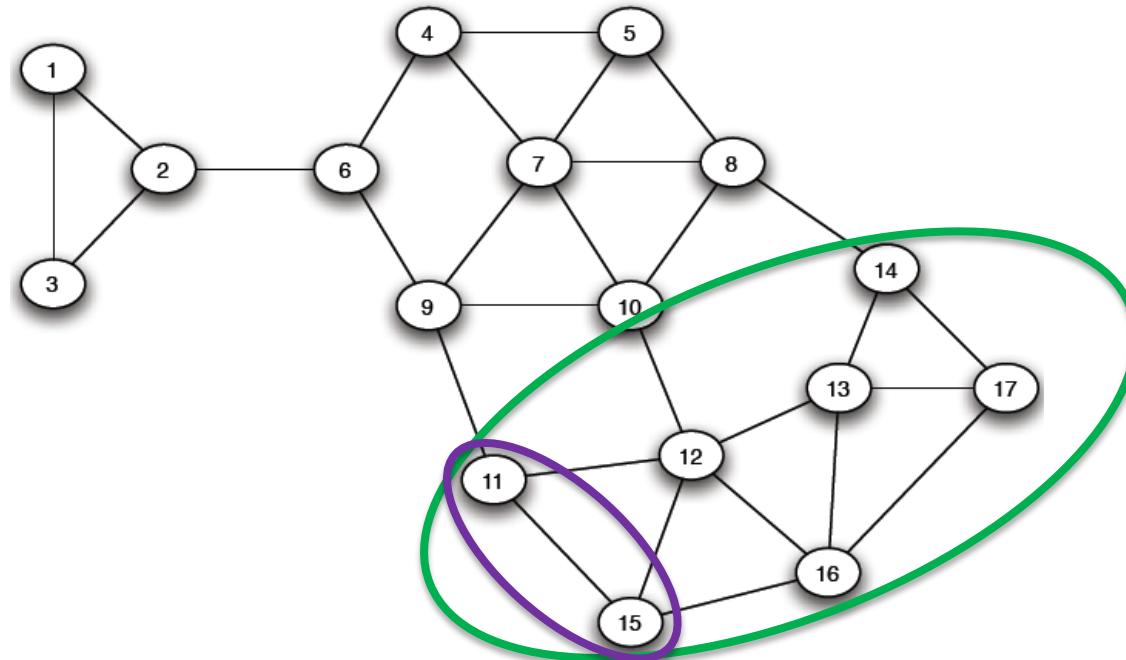
Cuantos más se recompensen en la campaña promocional, **mayor será la velocidad de adopción pero también mayor el coste** de la campaña

Los hubs difunden virus o ideas débilmente infecciosas.
Son una buena opción como semillas pero no la única
(estructura de comunidades)

Los **actores influyentes de la red (influentials)** según distintas medidas de análisis de redes sociales son una buena opción



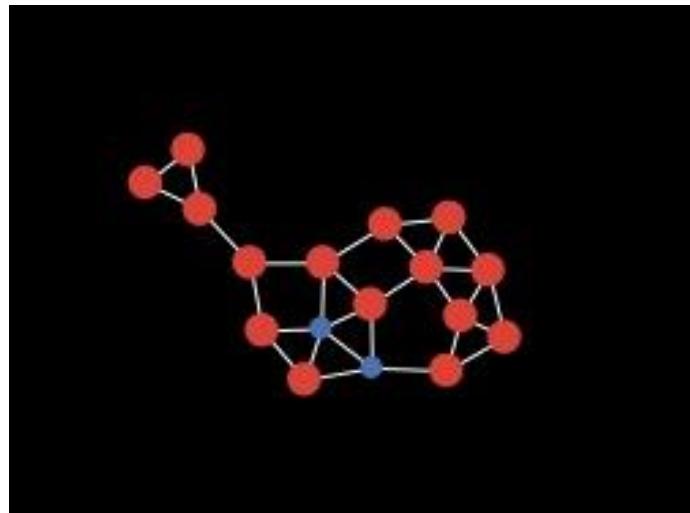
- La elección de las semillas es clave para producir el efecto cascada
- En el caso anterior, partiendo de los **nodos 12 y 13** se consigue convencer del 11 al 17 pero partir de **11 y 15** no produce ninguna “conversión” adicional



- La elección de los individuos es clave para producir el efecto cascada (actores influyentes de la red – *influentials*)

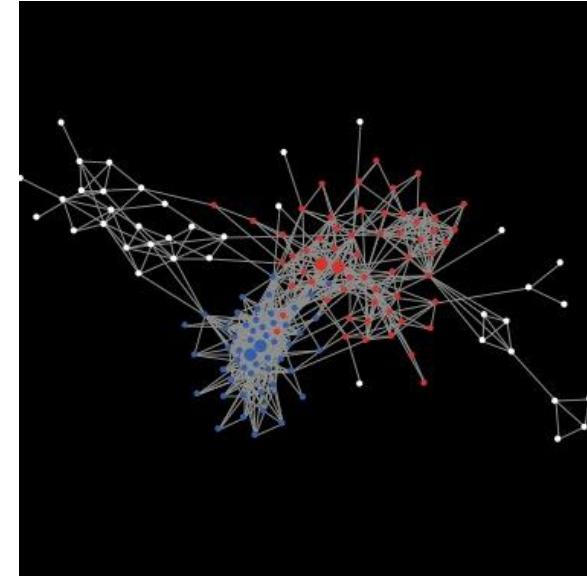
<http://www-personal.umich.edu/~ladamic/netlearn/NetLogo412/CascadeModel.html>

¿Se producirá una propagación en cascada?



<http://www-personal.umich.edu/~ladamic/netlearn/NetLogo412/CascadeModel.html>

- **El proceso de difusión depende de la estructura de la red**
- El modelo permite usar cuatro distintas. Una de ellas en la red de Facebook de *Lada Adamic*
- Se puede usar como un juego de dos jugadores. Cada persona tiene que escoger dos nodos:
 - El primero escoge un nodo y lo marca como azul
 - El segundo escoge otro nodo y lo marca como rojo
 - El primero escoge un nodo azul adicional
 - El segundo escoge un nodo rojo adicional
 - Se ejecuta el modelo con los parámetros prefijados



<http://www.ladamic.com/netlearn/NetLogo412/CascadeModel.html>

PREGUNTA: ¿Cuál es el papel de las comunidades en los procesos de contagio complejo?

- a) Posibilitar que las ideas se difundan en presencia de umbrales
- b) Crear “bolsas aisladas” impermeables a las ideas externas
- c) Permitir que distintas opiniones coexistan en distintas partes de la red

Referencias y Agradecimientos

Para elaborar las transparencias de este curso, he hecho uso de algunos materiales desarrollados por expertos en el área disponible en Internet:

- “Network Science Interactive Book Project” del Laszlo Barabasi Lab.
Northeastern University: <http://barabasilab.com/networksciencebook>
- Curso on-line “Social Network Analysis” de Lada Adamic, Coursera,
Universidad de Michigan: <https://www.coursera.org/course/sna>
- Transparencias del Capítulo 7 del libro “Social Media Mining” de
R. Zafarini, M.A. Abassi y H. Liu. <http://dmml.asu.edu/smm>

