EID Parte 1

Vicente Rivera

I. Marco Teórico

El modelo SIR (*Susceptible–Infectado–Recuperado*) es un modelo dinámico clásico, propuesto por Kermack y McKendrick (1927), que describe la evolución temporal de una población dividida en tres compartimentos: $S(t)$ (susceptibles), $I(t)$ (infectados) y $R(t)$ (recuperados). Su relevancia trasciende la epidemiología: al establecer una analogía entre poblaciones biológicas y redes de computadores, el SIR entrega un marco matemático útil para analizar brotes de *malware* y evaluar medidas de contención.

El modelo plantea un sistema de ecuaciones diferenciales que describe cómo cambia el tamaño de cada grupo con el tiempo:

$$\frac{dS}{dt} = -\beta \frac{SI}{N}, \tag{1}$$

$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I, \tag{2}$$

$$\frac{dR}{dt} = \gamma I, \tag{3}$$

donde $N = S(t) + I(t) + R(t)$ es el tamaño poblacional (constante), $\beta > 0$ es la tasa efectiva de contagio (contacto capaz de producir infección) y $\gamma > 0$ es la tasa de recuperación (remoción de infectados hacia recuperados).

Un parámetro clave es el **número básico de reproducción** R_0, que representa el número esperado de casos secundarios generados por un individuo infeccioso en una población completamente susceptible. Para el sistema anterior,

$$R_0 = \frac{\beta}{\gamma} \frac{S(0)}{N},$$

de modo que, si inicialmente $S(0) \approx N$, entonces $R_0 \approx \beta/\gamma$. Si $R_0 > 1$, la infección tiende a crecer (brote); si $R_0 < 1$, decae.

a) Analogía con ciberseguridad.: En el contexto de *malware*:

- **Susceptibles** S: dispositivos vulnerables pero no comprometidos.
- **Infectados** I: dispositivos comprometidos que pueden propagar el malware.
- **Recuperados** R: dispositivos limpiados, parcheados o aislados que ya no propagan.

Los parámetros se reinterpretan como:

- β: eficacia del vector de infección (frecuencia de contactos "exitosos" vía phishing, exploits, medios extraíbles, etc.).
- γ: eficacia de la respuesta (detección, limpieza, aislamiento, aplicación de parches).

Bajo esta analogía, R_0 cuantifica la capacidad de expansión inicial de un brote digital y permite valorar el impacto de medidas como endurecimiento de políticas, campañas de concientización, segmentación de red o despliegue de firmas antivirus.

b) Supuestos y alcance.: El SIR asume mezcla homogénea (todos "contactan" con todos al mismo ritmo) y parámetros constantes en el tiempo. Estas hipótesis simplifican el fenómeno pero son razonables para análisis iniciales y *what-if*. Extensiones consideradas más adelante (tasas dependientes del tiempo o del estado, y modelos logísticos/Bernoulli) relajan estos supuestos para capturar saturación, ventanas horarias de operación y cambios de política.

Nota sobre R_0. En la literatura del SIR con término $\beta\, SI/N$, la forma general es $R_0 = (\beta/\gamma)\, S(0)/N$. Cuando la población es casi totalmente susceptible al inicio ($S(0) \approx N$), se usa la aproximación $R_0 \approx \beta/\gamma$.

II. Modelos con una Variable

A. Tipos de malware

Resumen breve de las principales amenazas y su modo de operación, para fijar vocabulario y contexto antes del modelado.

- **Virus**: Los virus informáticos son programas maliciosos diseñados para dañar, infiltrarse o obtener acceso no autorizado a sistemas y redes.
- **Worm**: Se propagan automáticamente a través de redes informáticas; su objetivo es consumir ancho de banda, sobrecargar servidores, crear puertas traseras o distribuir otros malware.
- **Troyano**: Se disfrazan de software legítimo para engañar a los usuarios y ejecutarse en sus sistemas; su objetivo es robar información personal, instalar backdoors o descargar malware adicional.
- **Ransomware**: Cifran archivos críticos del sistema y exigen un rescate para restaurar el acceso; su objetivo es la extorsión.
- **Spyware**: Se instalan en el sistema sin consentimiento, generalmente empaquetados con software legítimo. Monitorean la actividad del usuario y su objetivo es robar información sensible para uso malintencionado.
- **Botnets**: Convierten dispositivos infectados en "zombis" controlados remotamente por un servidor. Su objetivo es realizar ataques DDoS, enviar spam o minar criptomonedas.

B. Selección del tipo de virus y parametrización (Troyano)

1) Justificación: Se selecciona un **troyano de acceso remoto (RAT)** porque: (i) su propagación depende de la interacción humana (phishing, descargas, ejecución manual), lo que lo vuelve ideal para contrastar modelos con tasas constantes vs. variables; (ii) es una amenaza persistente en entornos corporativos y domésticos; (iii) permite estudiar el efecto de medidas de contención (parches, segmentación, capacitación).

2) Características operativas del troyano modelo:

- **Vector principal:** phishing con adjuntos maliciosos.
- **Vectores secundarios:** descargas de software ilegal y explotación de vulnerabilidades de navegador.
- **Objetivo:** apertura de backdoor para robo de información y control remoto.

3) Variables y unidades:

- $I(t)$: computadores infectados en el tiempo t [equipos].
- $r(t)$: tasa de *nuevas* infecciones [infecciones/hora].
- N: tamaño de la red al inicio [equipos].

4) Supuestos del escenario:

- $N = 1000$ equipos susceptibles inicialmente; condición inicial $I(0) = 1$.
- Actividad principal en la jornada laboral (08:00–18:00); fuera de ese rango la actividad disminuye pero no es nula.
- Etapa temprana: el modelo lineal se usa mientras $I \ll N$ (el agotamiento de susceptibles es despreciable).

5) Tasa de propagación propuesta: Se usa una tasa por tramos:

$$r(t) = \begin{cases} 3 \text{ infecciones/hora}, & \text{si } t \text{ está en } [08{:}00, 18{:}00], \\ \rho \text{ infecciones/hora}, & \text{fuera de ese horario, con } 0 < \rho \ll 3. \end{cases}$$

Con $\rho = 0.2$ como valor de referencia, el orden de magnitud durante la jornada es ≈ 30 nuevas infecciones/día laboral (3×10), coherente con una *oleada* de phishing.

a) Descomposición de r (interpretabilidad).: Puede estimarse como

$$r \approx (\text{correos/hora}) \cdot p_{\text{apertura}} \cdot p_{\text{click}} \cdot p_{\text{bypass AV}} \cdot f_S,$$

donde $f_S = S/N$ es la fracción susceptible. Esta forma facilita analizar el impacto de medidas: campañas de concientización ($\downarrow p_{\text{click}}$), filtrado ($\downarrow$ correos/hora), parches/EDR ($\downarrow p_{\text{bypass AV}}$).

C. Modelo lineal con tasa (aprox.) constante de nuevas infecciones

Planteamos un modelo lineal para el número de computadores infectados $I(t)$ donde la tasa de *nuevas* infecciones no depende de I (etapa temprana, $I \ll N$). Sea t el tiempo medido en horas desde el inicio de la jornada (08:00 $\Rightarrow t = 0$). La ecuación diferencial es

$$\frac{dI}{dt} = r(t), \qquad I(0) = I_0, \tag{4}$$

con $I_0 = 1$. Usamos la tasa por tramos del escenario:

$$r(t) = \begin{cases} 3 \text{ infecciones/hora}, & \text{si } t \in [0, 10] \text{ (jornada 08:00–18:00)}, \\ \rho \text{ infecciones/hora}, & \text{si } t \in (10, 24] \text{ (fuera de jornada)}, \end{cases}$$

con $0 < \rho \ll 3$.

Solución en una jornada (0–24 h): Integrando por tramos:

$$I(t) = \begin{cases} I_0 + 3\,t, & 0 \le t \le 10, \\ I_0 + 30 + \rho\,(t - 10), & 10 < t \le 24. \end{cases}$$

En particular, al cierre del día ($t = 24$):

$$I(24) = I_0 + \underbrace{30}_{\text{jornada}} + \underbrace{14\,\rho}_{\text{fuera de jornada}}.$$

Si se adopta $\rho = 0.2$, el incremento diario esperado es $30 + 14 \cdot 0.2 = 32.8$ infecciones/día.

Extensión a varios días: Sea $m = \left\lfloor \frac{t}{24} \right\rfloor$ el número de días transcurridos completos y $\tau = t - 24m \in [0, 24)$ la hora dentro del día actual. Por recurrencia,

$$I(24m) = I_0 + m\,(30 + 14\,\rho).$$

Luego, para el día m-ésimo:

$$I(t) = \begin{cases} I(24m) + 3\,\tau, & 0 \le \tau \le 10, \\ I(24m) + 30 + \rho\,(\tau - 10), & 10 < \tau < 24. \end{cases}$$

Caso de referencia: tasa constante pura: Si se asume $r(t) \equiv r$ constante todo el tiempo (p. ej. $r = 3$ infecciones/hora),

$$\frac{dI}{dt} = r, \quad I(0) = I_0 \;\Rightarrow\; I(t) = I_0 + r\,t.$$

Con $r = 3$ e $I_0 = 1$, queda $I(t) = 1 + 3t$.

Interpretación:

- I_0 fija el nivel inicial (intercepto).
- $r(t)$ controla la pendiente por tramos y captura la ventana operativa (actividad humana y de campaña).
- La validez del modelo lineal requiere $I \ll N$; cuando S se agota o hay retroalimentación por concientización/contención, se necesita un modelo con saturación (ver Sección 2.5).

D. Discusión: tasa dependiente del tiempo o del nivel de infección

Hasta ahora asumimos una tasa por tramos $r(t)$ aproximadamente constante. En escenarios realistas de troyanos (campañas de *phishing*, parches, concientización), es razonable que r dependa de t (medidas en el tiempo) o de I (retroalimentación por saturación o alerta).

a) Dependencia temporal $r(t)$.: Casos plausibles:

- **Horario y semana laboral:** función periódica diaria/semanal. Ejemplo diario:

$$r(t) = \begin{cases} r_{\text{lab}}, & t \in [08{:}00, 18{:}00], \\ \rho, & \text{fuera de jornada}, \; 0 < \rho \ll r_{\text{lab}}. \end{cases}$$

- **Oleada de campaña con decaimiento:** tras un envío masivo de *phishing* en $t = t_0$,

$$r(t) = r_{\text{pico}}\, \mathrm{e}^{-\alpha\,(t - t_0)} \quad \text{para } t \ge t_0,$$

con $\alpha > 0$ reflejando que los usuarios y defensas se adaptan.

- **Intervención en t_c:** endurecimiento de filtros/EDR que reduce la tasa en un factor $\eta \in (0,1)$:

$$r(t) = \begin{cases} r_0, & t < t_c, \\ (1-\eta)\, r_0, & t \geq t_c. \end{cases}$$

Efecto cualitativo: un $r(t)$ decreciente suaviza la pendiente de $I(t)$ y puede llevar de crecimiento casi lineal a mesetas; los escalones en r generan quiebres de pendiente visibles.

b) Dependencia en el estado $r(I)$.: Casos plausibles:

- **Saturación por agotamiento de susceptibles:** $r(I) = r_0 \left(1 - \dfrac{I}{K}\right)$, donde $K \leq N$ es una *capacidad efectiva* (por segmentación, políticas, etc.). Conduce al modelo logístico de 2.5.
- **Concientización progresiva:** $r(I) = \dfrac{r_0}{1+cI}$ con $c > 0$, simulando que a mayor número de incidentes, más usuarios y equipos quedan alerta/aislados.

Efecto cualitativo: si $r(I)$ decrece con I, el crecimiento inicial es rápido pero se frena ($\ddot{I} < 0$) y tiende a una meseta; esto representa mejor campañas que "se agotan" por aprendizaje del sistema.

E. Modelo no lineal (logístico) y resolución analítica

Consideremos el modelo logístico para capturar saturación/concientización:

$$I'(t) = r\, I(t)\left(1 - \frac{I(t)}{K}\right), \qquad r, K > 0, \qquad I(0) = I_0 \in (0, K). \tag{5}$$

Separando variables y usando fracciones parciales,

$$\int \frac{dI}{I\left(1 - \frac{I}{K}\right)} = \int r\, dt \implies \int \left(\frac{1}{I} + \frac{1}{K-I}\right) dI = rt + C,$$

de donde

$$\ln\left(\frac{I}{K-I}\right) = rt + C.$$

Exponentiando y despejando $I(t)$ se obtiene la solución explícita:

$$I(t) = \frac{K}{1 + A\, \mathrm{e}^{-rt}}, \qquad A = \frac{K - I_0}{I_0}. \tag{6}$$

a) Interpretación de parámetros.:

- r controla la rapidez de la propagación inicial (pendiente en torno a $t = 0$).
- K es la *capacidad efectiva* o nivel máximo esperable dadas las barreras (segmentación de red, parches, EDR, políticas).
- I_0 fija el punto de partida (casos ya comprometidos al inicio).

Rasgos clave: crecimiento inicial casi exponencial ($I(t) \approx I_0\, \mathrm{e}^{rt}$ si $I_0 \ll K$) y *desaceleración* a medida que $I \to K$ ($\dot{I} \to 0$).

F. Comparación: lineal vs. logístico

a) Modelo lineal $I'(t) = r(t)$.:

- **Ventajas:** muy simple; estimación directa de r a partir de incrementos observados; útil en *etapas tempranas* ($I \ll N$) y para describir ventanas operativas (horario laboral) vía $r(t)$ por tramos.
- **Limitaciones:** no considera agotamiento de susceptibles ni retroalimentación por concientización/medidas; predice crecimiento indefinido si $r(t)$ no cae.

b) Modelo logístico $I'(t) = rI(1 - I/K)$.:

- **Ventajas:** incorpora saturación de forma parsimoniosa; reproduce brotes que se frenan al acercarse a K; parámetros interpretables (r rapidez inicial, K techo efectivo).
- **Limitaciones:** asume mezcla homogénea y parámetros constantes; no distingue explícitamente horarios ni shocks de campaña (aunque puede combinarse con $r = r(t)$).

c) ¿Cuál usar y cuándo?:

- **Fase inicial / día de campaña:** lineal por tramos con $r(t)$ alto en jornada y bajo fuera de jornada describe bien el incremento diario y permite *contabilidad operativa* (ej.: ≈ 30 infecciones por día laboral con $r_{\text{lab}} = 3$).
- **Horizonte de varios días con defensas activas:** el logístico es preferible: captura el frenado por agotamiento de susceptibles/concientización y evita extrapolaciones irreales.
- **Combinado realista:** usar $r = r(t)$ en el logístico para modelar campañas e intervenciones:

$$I'(t) = r(t)\, I(t)\left(1 - \frac{I(t)}{K}\right),$$

donde $r(t)$ puede decrecer tras la campaña o caer por un cambio de política en t_c.

G. Conclusión de la Sección 2

En el escenario del troyano considerado, el modelo lineal por tramos $I'(t) = r(t)$ es apropiado para el corto plazo (por jornada) y para estimar incrementos cuando $I \ll N$. Para horizontes de varios días, el modelo logístico $I'(t) = r\, I(1 - I/K)$ representa mejor la desaceleración por agotamiento de susceptibles y concientización, evitando extrapolaciones irreales. Operativamente, intervenciones que reducen la exposición o la efectividad del ataque disminuyen r; segmentación, parches y aislamiento reducen la capacidad efectiva K. En conjunto, (r, K) permiten evaluar ex ante el impacto de políticas de contención y decidir entre acciones de corto (ajustar $r(t)$ por ventana horaria y campañas) y mediano plazo (modular K mediante endurecimiento estructural).