Probabilistyczne modele propagacji w grafach.

Bartosz Łabuz

26 października 2025

Spis treści

Rozdział 1

Wstęp

1.1 Motywacja i zastosowania

Propagację wirusów podczas epidemii ludzkość obserwowała już od starożytności. W dzisiejszych czasach, wraz z rozwojem internetu i mediów społecznościowych, mamy możliwość doświadczyć również dynamicznej propagacji informacji. Aby efektywnie rozprzestrzenić informacje, nie można robić tego "na ślepo", lecz trzeba wykorzystać wiedzę teoretyczną. Najbardziej naturalną metodą matematycznej reprezentacji relacji międzyludzkich są grafy: wierzchołkami grafu są ludzie, a krawędzie określają, czy dane osoby mają ze sobą kontakt. Połączenie teori grafów z rachunkiem prawdopodobieństwa pozwala stworzyć dokładny i praktyczny model propagacji informacji.

1.2 Cel pracy

Celem niniejszej pracy jest

- teoretyczna analiza procesów losowej propagacji w grafach,
- wyznaczenie rozkładu prawdopodobieństwa propagacji na wybranych rodzinach grafów,
- symulacja propagacji w środowisku komputerowym w celu zweryfikowania wyników teoretycznych.

1.3 Zakres pracy

Praca obejmuje:

- wstęp teoretyczny z zakresu teorii grafów i rachunku prawdopodobieństwa,
- opis badanych modeli propagacji: SI, SIR, SIS,
- implementację symulacji w Pythonie,
- analizę wyników i wnioski dotyczące wpływu struktury grafu na propagację.

Rozdział 2

Podstawy matematyczne

2.1 Notacja

Przez $\mathbb{N}$ oznaczamy zbiór $\{0, 1, 2, \dots\}$, a przez $\mathbb{N}_+ = \{1, 2, 3, \dots\}$. Moc zbioru A oznaczamy $|A|$. Logarytm naturalny z x oznaczamy $\log(x)$. Dla $n \in \mathbb{N}_+$ przez $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ oznaczamy n'tą liczbę harmoniczną.

Niech $G = (V, E)$ będzie grafem prostym nieskierowanym. Stopień wierzchołka $v \in V$ oznaczamy $\deg(v)$. Zbiór sąsiadów $v \in V$ oznaczamy $\mathrm{N}(v)$. Odległość między u i v oznaczamy $\mathrm{d}(u, v)$ dla $u, v \in V$. Ekscentryczność $v \in V$ oznaczamy $\epsilon(v) = \max_{u \in V} \mathrm{d}(u, v)$. Przez $\delta(G)$ i $\Delta(G)$ oznaczamy odpowiednio minimalny i maksymalny stopień wierzchołka w grafie G.

Jeśli $\mathbb{P}$ jest miarą prawdopodobieństwa na przestrzeni Ω to prawdopodobieństwo zdarzenia A oznaczamy $\mathbb{P}[A]$. Dla zmiennej losowej $X : \Omega \to \mathbb{R}$ jej wartość oczekiwaną oznaczamy $\mathbb{E}[X]$ a jej wariancje $\mathrm{Var}[X]$. Funkcję masy prawdopodobieństwa (PMF) oznaczamy $\mathbb{P}[X = t]$ a dystrybuante (CDF) oznaczamy $F_X(t) = \mathbb{P}[X \leq t]$ dla $t \in \mathbb{R}$.

2.2 Rodziny grafów

Graf ścieżkowy

Dla $n \in \mathbb{N}_+$ graf ścieżkowy ma zbiór wierzchołków $V = \{1, 2, \dots, n\}$ oraz zbiór krawędzi $E = \{\{i, i+1\} : i \in \{1, 2, \dots, n-1\}\}$. Oznaczamy go przez P_n.

Graf gwiazda

Dla $n \in \mathbb{N}_+$ graf gwiazda ma zbiór wierzchołków $V = \{0, 1, \dots, n\}$ oraz zbiór krawędzi $E = \{\{0, i\} : i \in \{1, 2, \dots, n\}\}$. Oznaczamy go przez S_n.

Graf cykliczny

Dla $n \in \mathbb{N}_+$ graf cykliczny ma zbiór wierzchołków $V = \{1, 2, \ldots, n\}$ oraz zbiór krawędzi $E = \{\{i, i+1\} : i \in \{1, 2, \ldots, n-1\}\} \cup \{\{n, 1\}\}$. Oznaczamy go przez C_n.

Graf pełny

Dla $n \in \mathbb{N}_+$ graf pełny ma zbiór wierzchołków $V = \{1, 2, \ldots, n\}$ oraz zbiór krawędzi $E = \{\{i, j\} : i, j \in \{1, 2, \ldots, n\} \wedge i \neq j\}$. Oznaczamy go przez K_n.

2.3 Rozkłady prawdopodobieństwa

Rozkład Bernoulliego

Próba Bernoulliego to doświadczenie losowe, którego wynik może być jednym z dwóch:

- sukces z prawdopodobieństwem $p \in (0; 1)$
- porażka z prawdopodobieństwem $1 - p$

Zmienna losowa X przyjmująca wartość 1 w przypadku sukcesu i 0 w przypadku porażki ma rozkład Bernoulliego. Oznaczamy $X \sim \mathrm{Ber}(p)$.

Rozkład dwumianowy

Rozkład dwumianowy opisuje liczbę sukcesów w n próbach Bernoulliego. Niech X będzie zmienną losową przyjmującą wartości w $\{0, 1, \ldots, n\}$, a każda próba ma prawdopodobieństwo sukcesu $p \in (0; 1)$. Wtedy:

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

Wartość oczekiwana i wariancja mają postać:

$$\mathbb{E}[X] = np, \quad \mathrm{Var}[X] = np(1-p)$$

Oznaczamy $X \sim \mathrm{Bin}(n, p)$.

Rozkład geometryczny

Rozkład geometryczny opisuje liczbę prób Bernoulliego potrzebnych do uzyskania pierwszego sukcesu. Niech X będzie zmienną losową przyjmującą wartości w $\mathbb{N}_+$, a każda próba ma prawdopodobieństwo sukcesu $p \in (0; 1)$. Wtedy:

$$\mathbb{P}[X = k] = p(1-p)^{k-1}, \quad k \in \mathbb{N}_+.$$

Dystrybuanta jest równa:

$$\mathbb{P}[X \leq t] = 1 = (1-p)^t$$

Wartość oczekiwana i wariancja mają postać:

$$\mathbb{E}[X] = \frac{1}{p}, \quad \mathrm{Var}[X] = \frac{1-p}{p^2}$$

Oznaczamy $X \sim \mathrm{Geo}(p)$.

Rozkład ujemny dwumianowy

Rozkład ujemny dwumianowy opisuje liczbę prób Bernoulliego potrzebnych do uzyskania m sukcesów. Niech X oznacza liczbę prób, przy czym każda próba ma prawdopodobieństwo sukcesu $p \in (0;1)$, a liczba sukcesów $m \in \mathbb{N}_+$ jest ustalona. Wtedy:

$$\mathbb{P}[X = k] = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k \geq m.$$

Wartość oczekiwana i wariancja mają postać:

$$\mathbb{E}[X] = \frac{m}{p}, \quad \mathrm{Var}[X] = \frac{m(1-p)}{p^2}$$

Oznaczamy $X \sim \mathrm{NegBin}(m, p)$.

Rozkład normalny

Zdefiniujmy funkcje

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad \mathbf{\Phi}(t) = \int_{-\infty}^{t} \varphi(x)\,\mathrm{d}x$$

Niech $\mu \in \mathbb{R}$ oraz $\sigma > 0$. Zmienna losowa X ma rozkład normalny, jeśli jej funkcja gęstości wyraża się wzorem:

$$f_X(t) = \frac{1}{\sigma} \cdot \varphi\Big(\frac{t-\mu}{\sigma}\Big), \quad t \in \mathbb{R}.$$

Dystrybuanta jest równa:

$$\mathbb{P}[X \leq t] = \mathbf{\Phi}\Big(\frac{t-\mu}{\sigma}\Big), \quad t \in \mathbb{R}.$$

Wartość oczekiwana i wariancja:

$$\mathbb{E}[X] = \mu, \quad \mathrm{Var}[X] = \sigma^2.$$

Oznaczenie: $X \sim \mathcal{N}(\mu, \sigma^2)$.

Jeśli $\mu = 0$ oraz $\sigma = 1$ to mówimy, że X ma rozkład standardowy normalny. Zauważmy, że φ oraz $\mathbf{\Phi}$ są odpowiednio PDF jak i CDF takiego rozkładu.

2.4 Fakty, sumy i nierówności

Fakt 1. *Niech $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ będą IID o CDF równej F_X. Zdefiniujmy zmienną losową $Y = \max\{X_1, X_2, \ldots, X_n\}$. Wtedy*

$$F_Y(t) = F_X^n(t)$$

Fakt 2. *Niech $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ będą IID o CDF równej F_X. Zdefiniujmy zmienną losową $Y = \min\{X_1, X_2, \ldots, X_n\}$. Wtedy*

$$F_Y(t) = 1 - (1 - F_X(t))^n$$

Fakt 3. *Niech $X \sim \mathrm{Bin}(n, p)$ oraz $Y \sim \mathrm{Bin}(m, p)$ będą niezależnymi zmiennymi losowymi. Wtedy*

$$X + Y \sim \mathrm{Bin}(n + m, p)$$

Fakt 4. *Niech $X_1, X_2, \ldots, X_m \sim \mathrm{Geo}(p)$ będą IID oraz $Y = X_1 + \cdots + X_m$. Wtedy*

$$Y \sim \mathrm{NegBin}(m, p)$$

Fakt 5. *Niech $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ oraz $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ będą niezależnymi zmiennymi losowymi. Wtedy*

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Suma 1. *Niech $n \in \mathbb{N}$ oraz $x, y \in \mathbb{R}$. Wtedy*

$$\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x + y)^n$$

Suma 2. *Niech $n \in \mathbb{N}$ oraz $x, y \in \mathbb{R}$. Wtedy*

$$\sum_{k=0}^{n} k \cdot \binom{n}{k} x^k y^{n-k} = nx(x + y)^{n-1}$$

Suma 3. *Niech $n \in \mathbb{N}$ oraz $x \in \mathbb{R} \setminus \{1\}$. Wtedy*

$$\sum_{k=0}^{n} x^k = \frac{1 - x^{n+1}}{1 - x}$$

Suma 4. *Niech $n \in \mathbb{N}$ oraz $x \in \mathbb{R} \setminus \{1\}$. Wtedy*

$$\sum_{k=0}^{n} k \cdot x^k = \frac{x}{(1 - x)^2} \cdot (nx^{n+1} - (n + 1)x^n + 1)$$

Suma 5. *Niech $x \in (-1;1)$. Wtedy*

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

Suma 6. *Niech $x \in (-1;1)$. Wtedy*

$$\sum_{k=0}^{\infty} k \cdot x^k = \frac{x}{(1-x)^2}$$

Nierówność 1. *Niech $a, b \in \mathbb{N}$, $a < b$ oraz $f : [a;b] \to \mathbb{R}$ będzie funkcją ciągłą i malejąca. Wtedy*

$$\int_a^b f(x)\,\mathrm{d}x \leq \sum_{k=a}^{b} f(k) \leq f(a) + \int_a^b f(x)\,\mathrm{d}x$$

Nierówność 2. *Niech $n \in \mathbb{N}_+$. Wtedy*

$$H_n \leq 1 + \log(n)$$

Nierówność 3. *Niech $x \in (0;1)$. Wtedy*

$$\frac{1}{\log(\frac{1}{1-x})} \leq \frac{1}{x}$$

Nierówność 4 (Nierówność między średnimi)**.** *Niech $x_1, x_2, \ldots, x_n \geq 0$. Wtedy*

$$\log(x_1 \cdots x_n) \leq n \cdot \log\left(\frac{x_1 + \cdots + x_n}{n}\right)$$

Równość zachodzi wtedy i tylko wtedy gdy $x_1 = \cdots = x_n$.

Nierówność 5 (Nierówność Markova)**.** *Niech X będzie zmienną losową taką, że $X \geq 0$ oraz $\mathbb{E}[X] < \infty$. Wtedy dla dowolnego $t > 0$*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}}{t}$$

Nierówność 6 (Nierówność Cauchy'ego-Schwarza)**.** *Niech X, Y będą zmiennymi losowymi takimi, że $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. Wtedy*

$$\mathbb{E}[X \cdot Y] \leq \sqrt{\mathbb{E}[X^2]} \cdot \sqrt{\mathbb{E}[Y^2]}$$

Nierówność 7 (Nierówność Jensena dla wartości oczekiwanej)**.** *Niech* $n \in \mathbb{N}_+$ *oraz* $g : \mathbb{R}^n \to \mathbb{R}$ *będzie funkcją wypukłą zaś* $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{N}$ *będą zmiennymi losowymi (niekoniecznie niezależnymi). Wtedy*

$$g(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]) \leq \mathbb{E}[g(X_1, \ldots, X_n)]$$

Jeśli g *jest wklęsła to nierówność zachodzi w drugą stronę. W szczególności, ponieważ* max *jest funkcją wypukłą, a* min *wklęsłą, mamy:*

$$\max(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]) \leq \mathbb{E}[\max(X_1, \ldots, X_n)]$$

$$\min(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]) \geq \mathbb{E}[\min(X_1, \ldots, X_n)]$$

Rozdział 3

Modele propagacji losowej

Dany jest graf spójny nieskierowany $G = (V, E)$. Propagacja na takim grafie jest procesem stochastycznym. Zakładamy, że czas dla tego procesu jest dyskretny i mierzony w jednostkach naturalnych, zatem za zbiór chwil przyjmujemy $\mathbb{N}$. Niech $\mathcal{Q}$ będzie skończonym zbiorem stanów, jakie mogą przyjmować wierzchołki G. W każdej chwili $t \in \mathbb{N}$ każdy wierzchołek $v \in V$ znajduje się w pewnym stanie $Q \in \mathcal{Q}$. Definiujemy zmienną losową $\mathbf{X} : \mathbb{N} \times V \to \mathcal{Q}$, taką, że $\mathbf{X}_t(v) = Q$ wtedy i tylko wtedy, gdy wierzchołek v w chwili t znajduje się w stanie Q.

3.1 Model SI

Model **Susceptible–Infected (SI)** opisuje propagację w sieci, w której każdy wierzchołek znajduje się w jednym z dwóch stanów: podatny (S) lub zainfekowany (I). Początkowo ustalony wierzchołek $s \in V$ znajduje się w stanie I, natomiast pozostałe wierzchołki są w stanie S. Mamy więc $\mathcal{Q} = \{S, I\}$. W każdej jednostce czasu dowolny zainfekowany wierzchołek może zarazić każdego swojego sąsiada z prawdopodobieństwem p, dla ustalonego $p \in (0; 1)$. Wierzchołek raz zainfekowany pozostaje w tym stanie na zawsze. W modelu **SI** liczba zainfekowanych wierzchołków jest funkcją niemalejącą w czasie. Dla upraszczenia notacji kładziemy:

- $q = 1 - p$
- $\mathcal{S}_t = \{v \in V : \mathbf{X}_t(v) = S\}$
- $\mathcal{I}_t = \{v \in V : \mathbf{X}_t(v) = I\}$

Rozkład prawdopodobieństwa w tym modelu jest definiowany przez następujące zależności:

$$\mathbf{X}_0(v) = \begin{cases} I, & \text{jeśli } v = s \\ S, & \text{jeśli } v \neq s \end{cases}$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = S] = 1 - \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = S \mid \mathbf{X}_t(u) = S] = \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = I] = 1$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = S \mid \mathbf{X}_t(u) = I] = 0$$

Zdefiniujmy teraz zmienne losowe opisujące istotne własności. Dla każdego $v \in V$ definiujmy zmienną losową

$$X_v = \min\{t \in \mathbb{N} : \mathbf{X}_t(v) = I\}$$

która określa pierwszą chwilę czasu zarażenia wierzchołka v. Jeśli taka chwila nie istnieje (tzn. w danym przebiegu procesu wierzchołek v nigdy się nie zarazi), to przyjmujemy $X_v = \infty$. Zauważmy, że dla każdego $t \in \mathbb{N}$ zachodzi

$$\mathbb{P}[\mathbf{X}_t(v) = I] = \mathbb{P}[X_v \leq t]$$

Następnie dla każdego $t \in \mathbb{N}$ definiujemy zmienną losową

$$Y_t = |\mathcal{I}_t|$$

oznaczającą liczbę zainfekowanych wierzchołków w chwili t. Dodatkowo definiujemy zmienną losową opisującą czas całkowitego zarażenia grafu:

$$Z = \max_{v \in V} X_v$$

W modelu **SI** interesują nas następujące wielkości:

- rozkład prawdopodobieństwa zmiennych X_v, Y_t oraz Z
- wartości oczekiwane zmiennych, $\mathbb{E}[X_v]$, $\mathbb{E}[Y_t]$ oraz $\mathbb{E}[Z]$
- ograniczenia dolne, górne oraz asymptotyka powyższych wartości oczekiwanych kiedy wyznaczenie ich dokładnej wartości nie będzie możliwe

3.2 Model SIS

Model **Susceptible–Infected–Susceptible (SIS)** rozszerza model **SI** o powracanie wierzchołków zarażonych do stanu podatnego. Wierzchołek zainfekowany może powrócić do stanu podatnego z prawdopodobieństwem $\alpha \in (0;1)$. Tutaj mamy również $\mathcal{Q} = \{S, I\}$. W modelu **SIS** liczba zainfekowanych wierzchołków może oscylować w czasie i nie musi osiągnąć stanu pełnego zakażenia. Dla upraszczenia notacji kładziemy $\beta = 1 - \alpha$. Rozkład prawdopodobieństwa w tym modelu jest definiowany przez następujące zależności:

$$\mathbf{X}_0(v) = \begin{cases} I, & \text{jeśli } v = s \\ S, & \text{jeśli } v \neq s \end{cases}$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = S] = 1 - \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = S \mid \mathbf{X}_t(u) = S] = \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = I] = \beta$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = S \mid \mathbf{X}_t(u) = I] = \alpha$$

3.3 Model SIR

Model **Susceptible–Infected–Recovered (SIR)** rozszerza model **SI** o dodanie trzeciego stanu. Stanem tym jest R (Recovered). Mamy zatem $\mathcal{Q} = \{S, I, R\}$. Stan R jest trwały — wierzchołek, który wyzdrowiał, nie może już ani się zarazić, ani nikogo zakazić. Zarażony wierzchołek może przejść z I do stanu R z prawdopodobieństwem $\gamma \in (0;1)$. Dla upraszczenia notacji kładziemy

- $\delta = 1 - \gamma$
- $\mathcal{R}_t = \{v \in V : \mathbf{X}_t(v) = R\}$

Rozkład prawdopodobieństwa w tym modelu jest definiowany przez następujące zależności:

$$\mathbf{X}_0(v) = \begin{cases} I, & \text{jeśli } v = s \\ S, & \text{jeśli } v \neq s \end{cases}$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = S] = 1 - \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = S \mid \mathbf{X}_t(u) = S] = \prod_{v \in \mathrm{N}(u) \cap \mathcal{I}_t} q$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = R \mid \mathbf{X}_t(u) = I] = \gamma$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = I \mid \mathbf{X}_t(u) = I] = \delta$$

$$\mathbb{P}[\mathbf{X}_{t+1}(u) = Q \mid \mathbf{X}_t(u) = R] = \begin{cases} 1, & \text{dla } Q = R \\ 0, & \text{dla } Q \in \{S, I\} \end{cases}$$

Rozdział 4

Analiza modelu SI

4.1 Dwa wierzchołki, jedna krawędź

Na samym początku rozważmy najprostrzy graf, czyli o dwóch wierzchołkach u, v połączonych krawędzią. Za wierzchołek startowy wybierzmy u. Istnieją tylko dwa możliwe stany systemu: (I, S) oraz (I, I). Przejście ze stanu (I, S) do (I, I) następuje z prawdopodobieństwem p w każdej jednostce czasu. Zatem czas zarażenia drugiego wierzchołka X_v ma rozkład geometryczny, $X_v \sim \text{Geo}(p)$.

Rozważmy teraz rozkład Y_t. Mamy $\mathbb{P}[Y_t = 1] = q^t$, bo próba zarażenia musiałaby nie udać się t razy oraz $\mathbb{P}[Y_t = 2] = 1 - q^t$. Stąd $\mathbb{E}[Y_t] = 1 \cdot q^t + 2 \cdot (1 - q^t) = 2 - q^t$.

Jeśli chodzi o zmienną Z to zachodzi $Z = \max\{X_u, X_v\} = X_v$ a więc również $Z \sim \text{Geo}(p)$ oraz $\mathbb{E}[Z] = \frac{1}{p}$.

4.2 Trójkąt

Popatrzmy teraz na nieco większy graf — trójkąt. Niech jeden z wierzchołków będzie źródłem s, pozostałe zaś u, v. Aby poinformować u musimy uzyskać sukces bezpośrednio od s lub zarazić v a potem u. Możemy zapisać więc $X_u = \min\{A, B\}$ gdzie $A \sim \text{Geo}(p)$ oraz $B \sim \text{NegBin}(2, p)$. Wiemy, że $\mathbb{P}[A \leq t] = 1 - q^t$. Z kolei

$$\begin{aligned}\mathbb{P}[B \leq t] &= \sum_{k=2}^{t} (k-1) \cdot p^2 q^{k-2} = \frac{p^2}{q} \cdot \frac{q}{(1-q)^2} \cdot ((t-1)q^t - tq^{t-1} + 1) \\ &= 1 - q^t - tpq^{t-1}\end{aligned}$$

gdzie skorzystaliśmy z Sumy 4. Dalej, z Faktu 2 mamy

$$\mathbb{P}[X_u \leq t] = 1 - (1 - (1 - q^t)) \cdot (1 - (1 - q^t - tpq^{t-1})) = 1 - q^{2t} - tpq^{2t-1}$$

Jeśli chodzi o liczbę zainfekowanych po t krokach, to skoro mamy trzy wierzchołki to i trzy wartości do policzenia. Oczywiście $\mathbb{P}[Y_t = 1] = q^{2t}$. Aby po t chwilach tylko dwa węzły był zainfekowane musimy zarazić któryś z wierzchołków po $1 \leq k \leq t$ rundach z prawdopodobieństwem $2pq \cdot q^{2\cdot(k-1)}$ a następnie uzyskać $t - k$ porażek. Na każdą z nich mamy szanse równą q^2. Podsumowując

$$\mathbb{P}[Y_t = 2] = \sum_{k=1}^{t} 2pq \cdot q^{2\cdot(k-1)} \cdot q^{2\cdot(t-k)} = 2tpq^{2t-1}$$

Na koniec mamy $\mathbb{P}[Y_t = 3] = 1 - q^{2t} - 2tpq^{2t-1}$. Ponadto

$$\mathbb{E}[Y_t] = 1 \cdot q^{2t} + 2 \cdot 2tpq^{2t-1} + 3 \cdot (1 - q^{2t} - 2tpq^{2t-1}) = 3 - 2q^{2t} - 2tpq^{2t-1}$$

Widzimy, że $\mathbb{E}[Y_t] \to 3$ przy $t \to \infty$ co jest zgodne z intuicją.

Propagacja może sie zakończyć na dwa sposoby. Pierwszym z nich jest sytuacja, w której przez $t - 1$ żadne zakażenie nie zaszło a w chwili t zarażą sie oba wierzchołkie. Prawdopodobieństwo tego przypadku wynosi $p^2q^{2\cdot(t-1)}$. Druga możliwość to taka gdzie w k'tym kroku (dla $1 \leq k \leq t-1$) zaraził się jeden wierzchołków, szansa $2\cdot pq\cdot q^{2\cdot(k-1)}$, a potem przez kolejne $t-1-k$ kroków trzeci nie został zarażony, na co mamy prawdopodobieństwo $(q^2)^{t-1-k}$, do chwili t. To ostatnie przejście ma $1 - q^2$ szans. Łącznie dostajemy

$$\begin{aligned}\mathbb{P}[Z = t] &= p^2q^{2t-2} + \sum_{k=1}^{t-1}(2pqq^{2k-2}) \cdot (q^{2t-2k-2}) \cdot (1 - q^2) \\ &= p^2q^{2t-2} + 2pq^{2t-3} \cdot (t-1)(1 - q^2)\end{aligned}$$

Mamy też $\mathbb{P}[Z > t] = \mathbb{P}[Y_t \neq 3] = q^{2t} + 2tpq^{2t-1}$. Wartość oczekiwana więc wynosi:

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{t=0}^{\infty} \mathbb{P}[Z > t] = \sum_{t=0}^{\infty} q^{2t} + 2tpq^{2t-1} \\ &= \frac{1}{1-q^2} + \frac{2p}{q} \cdot \frac{q^2}{(1-q^2)^2} = \frac{-3q^2 + 2q + 1}{(1-q^2)^2} = \frac{4 - 3p}{p(2-p)^2}\end{aligned}$$

Wykonaliśmy dość sporo obliczeń jak na tak mały graf. Możemy zauważyć więc, że istnienie cykli w grafie znacząco komplikuje sytuacje jeśli chodzi o model SI .

4.3 Grafy ścieżkowe

Jako pierwszą rodzinę grafów rozważmy grafy ścieżkowe P_n. Załóżmy, że proces zaczyna się w wierzchołku $s = 1$. Zatem infekcja rozchodzi się po grafie "od lewej do prawej". Dla tej rodziny grafów uda nam się wyznaczyć dokładny rozkład prawdopodobieństwa. Zauważmy, że czasy zarażenia kolejnych wierzchołków tworzą ciąg zmiennych losowych

$$X_1 = 0, \quad X_v = X_{v-1} + U_k, \quad v \in \{2, 3, \ldots, n\},$$

gdzie $U_2, U_3, \ldots, U_n \sim \mathrm{Geo}(p)$ oraz $U_2, U_3, \ldots, U_n$ są niezależne. Widzimy zatem, że $X_v \sim U_1 + U_2 + \cdots + U_{v-1}$ a więc z Faktu 4 X_v ma rozkład ujemny dwumianowy,

$$X_v \sim \mathrm{NegBin}(v-1, p).$$

Ponadto mamy:

$$\mathbb{E}[X_v] = \frac{v-1}{p}, \quad \mathrm{Var}[X_v] = \frac{(v-1)(1-p)}{p^2}$$

Ustalmy $t \in \mathbb{N}$ i przejdźmy do obliczania rozkładu Y_t. Zauważmy, że liczba dodatkowych zakażeń poza startowym wierzchołkiem do czasu t to po prostu liczba sukcesów w t niezależnych prób Bernoulliego. Musimy jednak pamiętać, że Y_t nie może przekroczyć n. Zatem mamy dokładnie

$$Y_t = \min\{n, 1 + B_t\}, \qquad B_t \sim \mathrm{Bin}(t, p)$$

Pozwala to na wyznaczenie PMF dla Y_t:
Dla $1 \leq k \leq n-1$ mamy:

$$\mathbb{P}[Y_t = k] = \mathbb{P}[B_t = k-1] = \binom{t}{k-1} p^{k-1} q^{t-k+1}$$

oraz dla $k = n$ mamy:

$$\mathbb{P}[Y_t = n] = \mathbb{P}[B_t \geq n-1] = \sum_{j=n-1}^{t} \binom{t}{j} p^j q^{t-j}$$

Przejdźmy teraz do obliczania wartości oczekiwanej Y_t:

$$\begin{aligned}
\mathbb{E}[Y_t] &= \sum_{k=1}^{n-1} k \cdot \mathbb{P}[Y_t = k] + n \cdot \mathbb{P}[Y_t = n] \\
&= \sum_{k=1}^{n-1} k \cdot \binom{t}{k-1} p^{k-1} q^{t-k+1} + n \cdot \sum_{j=n-1}^{t} \binom{t}{j} p^j q^{t-j} \\
&= \sum_{j=0}^{t} \min\{n, 1+j\} \cdot \binom{t}{j} p^j q^{t-j}
\end{aligned}$$

Policzmy teraz asymptotykę dla $n \to \infty$. Wtedy $n > 1 + j$ dla wszystkich $0 \leq j \leq t$, a więc:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[Y_t] &= \sum_{j=0}^{t}(1+j)\binom{t}{j}p^j q^{t-j} = \sum_{j=0}^{t}\binom{t}{j}p^j q^{t-j} + \sum_{j=0}^{t} j\binom{t}{j}p^j q^{t-j} \\ &= (p+q)^t + tp(p+q)^{t-1} = 1 + tp\end{aligned}$$

gdzie sumy sumujemy korzystając z Sumy 1 oraz Sumy 2. Stąd

$$\lim_{n\to\infty} \mathbb{E}[Y_t] = 1 + tp$$

Czas całkowitego zainfekowania grafu P_n to $Z = \max\{X_1, X_2, \ldots, X_n\} = X_n$. Zatem rozkład zmiennej Z jest już nam znany, $Z \sim \mathrm{NegBin}(n-1, p)$, a wartość oczekiwana wynosi

$$\mathbb{E}[Z] = \frac{n-1}{p}$$

Sprawdźmy, czy nasze obliczenia teoretyczne zgadadzają się z empirycznie wyznaczonymi wartościami. Ustalmy $p = 0.2$, $n = 1000$. Dla każdego $t \in \{1, 2, \ldots, \frac{n-1}{p}\}$ przeprowadźmy 2000 symulacji propagacji na grafie P_n w celu estymacji $\mathbb{E}[Y_t]$. Następnie dla $n \in \{1, 2, \ldots, 1000\}$ tą samą liczbą symulacji oszacujmy $\mathbb{E}[Z]$. Wyniki eksperymentu niemal idealnie pokrywają się z przewidywanymi kształtami, to jest $1 + tp$ dla $\mathbb{E}[Y_t]$ oraz $\frac{n-1}{p}$ dla $\mathbb{E}[Z]$.

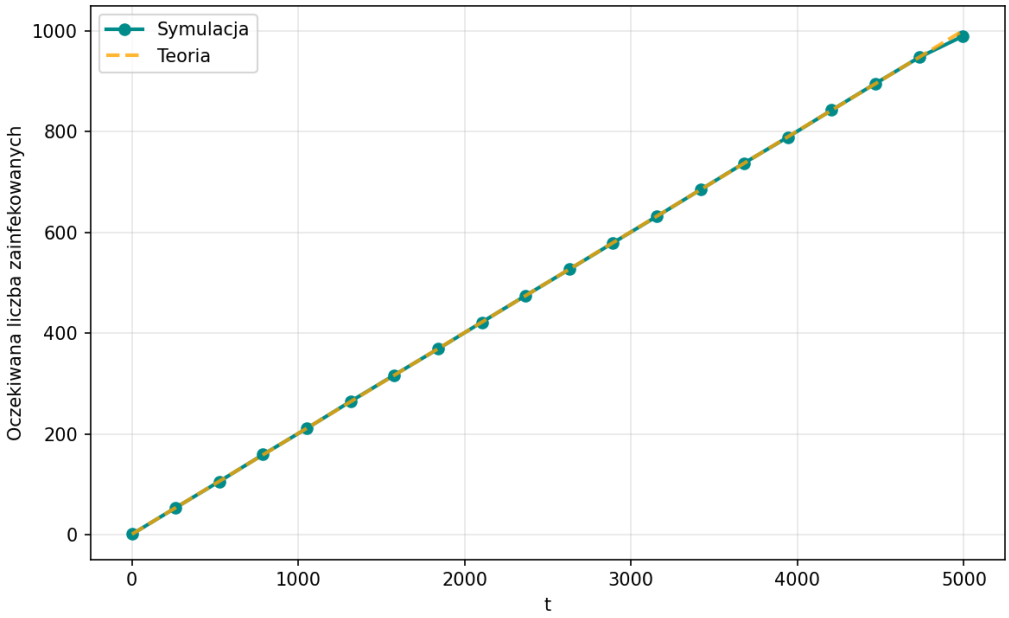
4.4 Grafy gwiezdne

Następnie rozpatrzmy rodzinę grafów gwiazd S_n. Niech źródłem będzie centralny wierzchołek grafu, to jest $s = 0$. Propagacja rozchodzi się tutaj po każdym ramieniu gwiazdy niezależnie. Stąd mamy $X_v \sim \mathrm{Geo}(p)$ dla każdego $v \in \{1, 2, \ldots, n\}$. Ponadto zmienne $X_1, X_2, \ldots, X_n$ są od siebie niezależne. Mamy więc

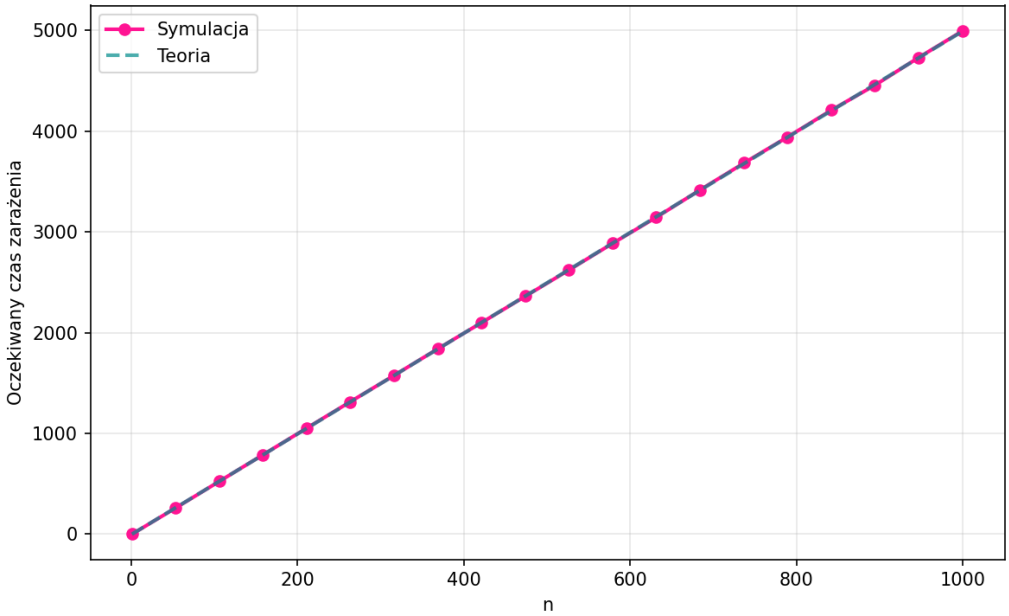
$$\mathbb{E}[X_v] = \frac{1}{p}, \quad \mathrm{Var}[X_v] = \frac{1-p}{p^2}$$

Kwestia Y_t jest również dość prosta. Skoro propagacja działa na każdym wierzchołku niezależnie to zmienna Y_t jest wynikiem n prób Bernoulliego. Sukces pojedynczej próby to prawdopodobieństwo, że zmienna X_v o rozkładzie geometrycznym po conajwyżej t jednostkach czasu osiągnie swój sukces. A więc jest to $\mathbb{P}[X_v \leq t] = 1 - q^t$. W takim razie mamy

$$Y_t = 1 + B_t, \quad B_t \sim \mathrm{Bin}(n, 1 - q^t)$$



Rysunek 1: $\mathbb{E}[Y_t]$ dla $\mathrm{P_n}$ w funkcji t



Rysunek 2: $\mathbb{E}[Z]$ dla $\mathrm{P_n}$ w funkcji n

Stąd oczywiście otrzymujemy

$$\mathbb{E}[Y_t] = 1 + n \cdot (1 - q^t)$$

Przejdźmy teraz do zmiennej Z. Mamy $Z = \max\{X_1, X_2, \ldots, X_n\}$. Skoro zmienne te są IID, to z Faktu 1 mamy

$$\mathbb{P}[Z \leq t] = (1 - q^t)^n$$

Policzmy teraz wartość oczekiwaną całkowitego zainfekowania grafu.

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{k=1}^{\infty} \mathbb{P}[Z \geq k] = \sum_{k=1}^{\infty} 1 - \mathbb{P}[Z \leq k-1] = \sum_{k=1}^{\infty} 1 - (1 - q^{k-1})^n \\
&= \sum_{k=0}^{\infty} 1 - (1 - q^k)^n = \sum_{k=0}^{\infty} \left(1 - \sum_{j=0}^{n} \binom{n}{j} (-1)^j q^{kj}\right) \\
&= \sum_{k=0}^{\infty} \sum_{j=1}^{n} \binom{n}{j} (-1)^{j+1} q^{kj} = \sum_{j=1}^{n} \sum_{k=0}^{\infty} \binom{n}{j} (-1)^{j+1} (q^j)^k \\
&= \sum_{j=1}^{n} \binom{n}{j} \frac{(-1)^{j+1}}{1 - q^j}
\end{aligned}$$

Nie jest to jednak przyzwoity wynik i nie ma postaci zwartej. Spróbujmy zatem wyznaczyć asymptotykę $\mathbb{E}[Z]$. Mamy $\mathbb{E}[Z] = \sum_{k=0}^{\infty} 1 - (1 - q^k)^n$. ‘ ’Nierówność 1 umożliwia oszacowanie tej sumy. Połóżmy $f(x) = 1 - (1 - e^{-\lambda x})^n$ gdzie $\lambda = -\log(q)$. Oczywiście $f(0) = 1$, $f(\infty) = 0$ oraz f jest malejąca a więc

$$\int_0^{\infty} f(x)\, \mathrm{d}x \leq \mathbb{E}[Z] \leq 1 + \int_0^{\infty} f(x)\, \mathrm{d}x$$

Podstawiamy $u = 1 - e^{-\lambda x}$. Wtedy $\mathrm{d}u = \lambda e^{-\lambda x}\, \mathrm{d}x$, a więc $\mathrm{d}x = \frac{1}{\lambda} \cdot \frac{1}{1-u}\, \mathrm{d}u$. Ponadto $u(0) = 0$, $u(\infty) = 1$ (bo $\lambda > 0$). Zatem całka ma postać:

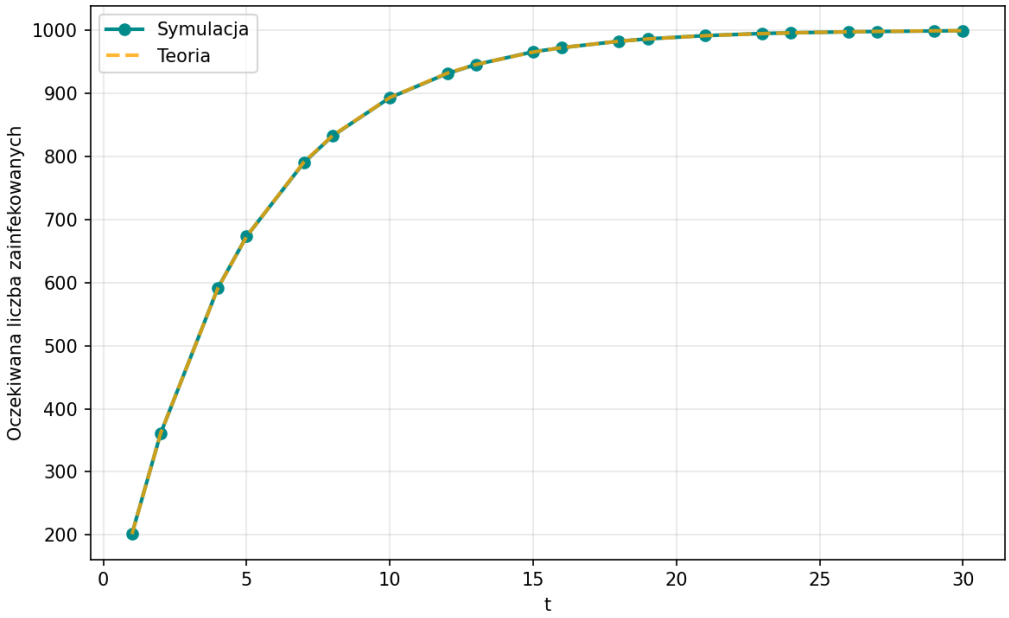
$$\frac{1}{\lambda} \int_0^1 \frac{1 - u^n}{1 - u}\, \mathrm{d}u = \frac{1}{\lambda} \int_0^1 \sum_{j=0}^{n-1} u^j\, \mathrm{d}u = \frac{1}{\lambda} \sum_{j=0}^{n-1} \frac{1}{j+1} = \frac{H_n}{\lambda}$$

Zauważmy, że $-\log(q) = \log(\frac{1}{1-p})$ a więc ostatecznie dostajemy:

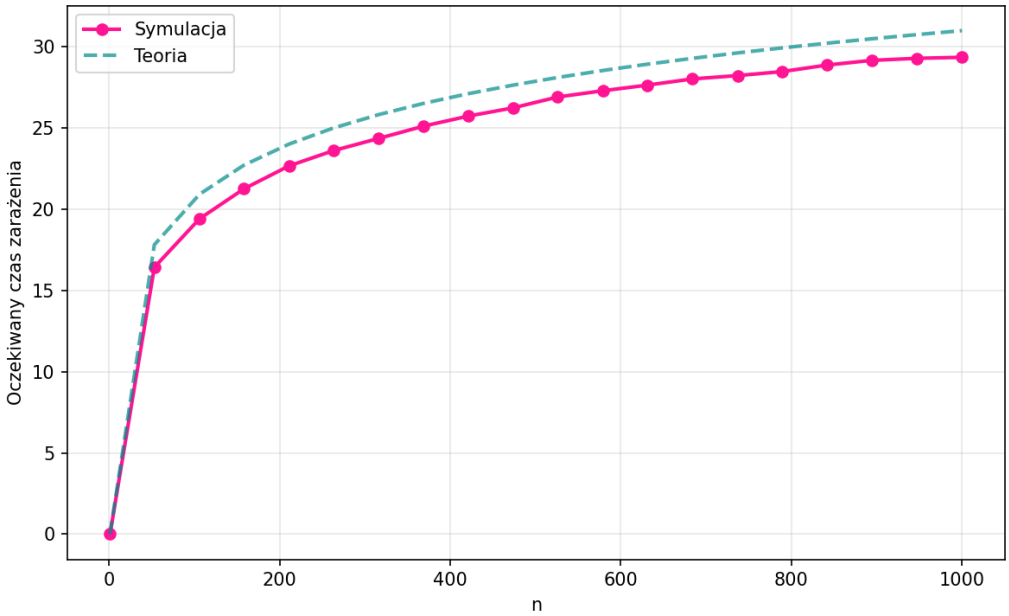
$$\frac{H_n}{\log(\frac{1}{1-p})} \leq \mathbb{E}[Z] \leq \frac{H_n}{\log(\frac{1}{1-p})} + 1$$

Stąd mamy asymptotyczny przewidywany czas zarażenia grafu S_n:

$$\mathbb{E}[Z] \sim \frac{H_n}{\log(\frac{1}{1-p})}$$



Rysunek 3: $\mathbb{E}[Y_t]$ dla $\mathrm{S_n}$ w funkcji t



Rysunek 4: $\mathbb{E}[Z]$ dla $\mathrm{S_n}$ w funkcji n

Przeprowadźmy teraz symulacje. Ustalmy $p = 0.2$, $n = 1000$. Dla każ-

dego $t \in \{1, 2, \ldots, \log(n)\}$ wykonajmy 2000 powtórzeń propagacji na S_n. Potem dla $n \in \{1, 2, \ldots, 1000\}$ tak samo dla $\mathbb{E}[Z]$. Dla $\mathbb{E}[Y_t]$ po raz kolejny mamy idealne dopasowanie. Dla $\mathbb{E}[Z]$ zaś numeryczna wartość jest o około 1 mniejsza niż przewidywana, co jest dobrym wynikiem.

4.5 Ograniczenia na czas zarażenia

Po rozważeniu dwóch rodzin grafów dostrzegamy znaczną różnicę w wartościach oczekiwnaych zmiennych Y_t oraz Z. Dla grafów ścieżkowych minimalna liczb rund potrzebnych do zainfekowania całego grafu wynosi $t = n - 1$ natomiast dla gwiazd jest to zaledwie $t = 1$. Widzimy, że w pewnym sensie najlepszy przypadek sprzyjający szybkiej propagacji jest taki, w którym źródło s jest połączone z wszyskimi pozostałymi wierzchołkami grafu. Z drugiej strony najgorsza sytuacja ma miejsce, jeśli istnieje daleko oddalony węzeł, szczególnie z mało liczbą ścieżek do niego prowadzących, tak jak dla grafów ścieżkowych. Teraz postaramy się uogólnić tą obserwację.

Twierdzenie 1. *Niech $G = (V, E)$ będzie grafem spójnym a $G' = (V, E')$ spójnym podgrafem G. Załóżmy, że $\mathbf{X}$ jest procesem stochastycznym w modelu SI na G oraz G' jednocześnie z tym samym źródłem $s \in V$. Jeśli przez X'_v, Y'_t oraz Z' oznaczymy zmienne losowe w G' odpowiadające tym w G to zachodzą następujące nierówności:*

$$X_v \leq X'_v, \quad Y_t \geq Y'_t, \quad Z \leq Z'$$

Dowód. Oznaczmy przez $\mathcal{I}_t$ zainfekowane wierzchołki w G a przez $\mathcal{I}'_t$ w G'. Wtedy $\mathcal{I}'_t \subseteq \mathcal{I}_t$ dla dowolnego $t \in \mathbb{N}$. Ustalmy $v \in V$. Niech $X'_v = a$. Wtedy $v \in \mathcal{I}'_a$ jak i $v \in \mathcal{I}_a$. Stąd $X_v \leq a = X'_v$. Ustalmy teraz $t \in \mathbb{N}$. Oczywiście skoro $\mathcal{I}'_t \subseteq \mathcal{I}_t$ to i $|\mathcal{I}'_t| \leq |\mathcal{I}_t|$ a co za tym idzie $Y'_t \leq Y_t$. Dalej niech $Z' = b$. Wtedy $\mathcal{I}'_b = V$ a więc $V \subseteq \mathcal{I}_t$. Zatem $\mathcal{I}_b = V$ i $Z \leq b = Z'$. □

Intuicyjnie sprawa jest oczywista. Mając mniej krawędzi w grafie potrzebujemy więcej czasu na rozprzestrzenienie się w nim informacji. W praktyce oznacza to, że jeżeli znamy średni czas zainfekowania dowolnego podgrafu G to znamy ogarniczenie górne na czas dla całego grafu. Postarajmy się teraz oszacować sensownie z góry wartość $\mathbb{E}[Z]$ dla dowolnego grafu.

Twierdzenie 2. *Niech $G = (V, E)$ będzie grafem o n wierzchołkach zaś $s \in V$ będzie ustalonym źródłem. Oznaczmy $\lambda = \log(\frac{1}{1-p})$ oraz $h = \epsilon(s)$. Wtedy zachodzi*

$$\mathbb{E}[Z] \leq h + \frac{h}{\lambda} \cdot \left(\log\left(\frac{n-1}{h}\right) + 1 \right)$$

Dowód. Dla $0 \leq j \leq h$ kładziemy $A_j = \{v \in V : \mathrm{d}(s, v) = j\}$ oraz $a_j = |A_j|$. Oczywiście $a_0 = 1$ a więc $a_1 + \cdots + a_h = n - 1$. Dalej zdefiniujmy zmienne losowe $T_j = \min\{t \in \mathbb{N} : A_j \subseteq \mathcal{I}_t\}$. Zmienna T_j określa czas potrzebny do zainfekowania wierzchołków oddalonych o j od źródła. Udowodnijmy teraz przydatny lemmat.

Lemat 1. *Niech $U_j = T_j - T_{j-1}$ dla $1 \leq j \leq h$. Wtedy:*

$$\mathbb{E}[U_j] \leq \frac{H_{a_j}}{\lambda} + 1$$

Dowód. U_j to czas potrzebny na zarażenie wierzchołków A_j podczas gdy A_{j-1} są już zarażone. Wybierzmy podgraf G' w taki sposób, żeby każdy wierzchołek z A_j był połączony dokładnie jedną krawędzią z którymś z wierzchołków ze zbioru A_{j-1}. Wtedy rozkład propagacji na G' jest izomorficzny z tym dla S_{a_j} bo $a_j = |A_j|$. Z Twierdzenia 1 wnioskujemy, że zmienna U_j jest ograniczona przez całkowity czas zarażenia grafu gwiazdy. Wartośc oczekiwana w tym drugim przypadku jest mniejsza niż $\frac{1}{\lambda} H_{a_j} + 1$. Z monotoniczności wartości oczekiwanej dostajemy porządany wynik. □

Wróćmy do udowadniania ograniczenia na $\mathbb{E}[Z]$. Mamy $T_h = \sum_{j=1}^{h} U_j$ a więc:

$$\begin{aligned}
\mathbb{E}[Z] = \mathbb{E}[T_h] &= \mathbb{E}\Big[\sum_{j=1}^{h} U_j\Big] = \sum_{j=1}^{h} \mathbb{E}[U_j] \leq \sum_{j=1}^{h} \frac{H_{a_j}}{\lambda} + 1 \\
&= h + \frac{1}{\lambda} \sum_{j=1}^{h} H_{a_j} \leq h + \frac{1}{\lambda} \sum_{j=1}^{h} 1 + \log(a_j) \\
&= h + \frac{h}{\lambda} \cdot \Big(1 + \sum_{j=1}^{h} \log(a_j)\Big) = h + \frac{h}{\lambda} \cdot \Big(1 + \log\Big(\prod_{j=1}^{h} a_j\Big)\Big) \\
&\leq h + \frac{h}{\lambda} \cdot \Big(1 + \log\Big(\frac{1}{h} \sum_{j=1}^{h} a_j\Big)\Big) = h + \frac{h}{\lambda} \cdot \Big(1 + \log\Big(\frac{n-1}{h}\Big)\Big)
\end{aligned}$$

gdzie w linijce pierwszej wykorzystujemy Lemat 1, w drugiej Nierówność 2 a w piątej z nierówności między średnimi (4). □

Porównajmy przed chwilą udowodnione twierdzenie z poprzednimi wynikami. Dla rodziny P_n mamy $h = n - 1$. Dodatkowo korzystając z Nierówności 3 mamy

$$\mathbb{E}[Z] \leq (n-1) \cdot \Big(1 + \frac{1}{p}\Big)$$

Faktyczna wartość oczekiwana jest równa $\frac{n-1}{p}$ więc oszacowanie jest dość ostre. Z kolei dla rodziny S_n mamy $h = 1$ oraz $n + 1$ wierzchołków a więc

$$\mathbb{E}[Z] \leq 1 + \frac{\log(n) + 1}{\log(\frac{1}{1-p})}$$

Ponownie oszacowanie jest dość dokładne. Wynik ten zdaje się być dobry dla grafów rzadkich.

4.6 Grafy cykliczne

Przejdźmy teraz do grafów cyklicznych. W rozważaniach dla trójkąta, to jest C_3 mogliśmy zauważyć, że cykl w tym grafie sprawiał trudności. W ogólnym przypadku nie jest lepiej. Rozważmy graf C_n.Niech źródłem będzie wierzchołek n. Ustalmy wierzchołek $v \in \{1, 2, \ldots, n-1\}$. Niech $a = \min\{v, n-v\}$ oraz $b = \max\{v, n-v\}$. Oczywiście $a \leq b$. Od źródła do tego wierzchołka są dwie ścieżki: jedna o długości a, druga o długości b. Propagacja rozchodzi się po nich równolegle i niezależnie. Dla $j \in \{1, 2, \ldots, n-1\}$ połóżmy $N_j \sim \mathrm{NegBin}(j, p)$. Zmienne te są niezależne. Mamy wtedy

$$X_v \sim \min\{N_a, N_b\}$$

Niech $F_j(t)$ będzie dystrybuantą zmiennej N_j. Z Faktu 2 mamy

$$\mathbb{P}[X_v \leq t] = 1 - (1 - F_a(t)) \cdot (1 - F_b(t))$$

Nie ma co liczyć na wyznaczenie eleganckiej postaci na PMF czy CDF dla X_v. Postaramy się więc przybliżyć wartość oczekiwaną dla dużych n. Z centralnego twierdzenia granicznego możemy przybliżyć $N_a \approx A,\ N_b \approx B$ dla $A \sim \mathcal{N}(\mu_a, \sigma_a^2),\ B \sim \mathcal{N}(\mu_b, \sigma_b^2)$ gdzie

$$\mu_a = \frac{a}{p}, \quad \sigma_a^2 = \frac{aq}{p^2}, \quad \mu_b = \frac{b}{p}, \quad \sigma_b^2 = \frac{bq}{p^2}$$

Zatem mamy

$$\mathbb{E}[X_v] \approx \mathbb{E}[\min\{A, B\}] = \mathbb{E}\Big[\frac{A + B - |A - B|}{2}\Big] = \frac{\mathbb{E}[A] + \mathbb{E}[B] - \mathbb{E}[|A - B|]}{2}$$

Połóżmy $C = A - B$. Korzystając z Faktu 5 mamy $C \sim \mathcal{N}(\mu_a - \mu_b, \sigma_a^2 + \sigma_b^2)$. Oznaczmy $\eta = \mu_a - \mu_b$ oraz $\xi = \sqrt{\sigma_a^2 + \sigma_b^2}$. Potrzebujemy teraz następującego lematu:

Lemat 2. *Niech $X \sim \mathcal{N}(\mu, \sigma^2)$. Wtedy*

$$\mathbb{E}[|X|] = 2\sigma \cdot \varphi\Big(\frac{\mu}{\sigma}\Big) + \mu \cdot (2\mathbf{\Phi}\Big(\frac{\mu}{\sigma}\Big) - 1)$$

Dowód.

$$\mathbb{E}[|X|] = \int_{-\infty}^{\infty} \frac{|x|}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big)\,\mathrm{d}x = \int_{0}^{\infty} \frac{x}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big)\,\mathrm{d}x - \int_{-\infty}^{0} \frac{x}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big)\,\mathrm{d}x$$

Oznaczmy $c = \frac{\mu}{\sigma}$ oraz podstawmy $z = \frac{x-\mu}{\sigma}$. Zatem $x = \mu + \sigma z$, $\mathrm{d}x = \sigma \mathrm{d}z$. Dla $x > 0$ mamy $z > -c$ zaś dla $x < 0$ mamy $z < -c$. Otrzymujemy więc

$$\begin{aligned}
&\int_{-c}^{\infty} (\mu + \sigma z)\varphi(z)\,\mathrm{d}z - \int_{-\infty}^{-c} (\mu + \sigma z)\varphi(z)\,\mathrm{d}z \\
&= \mu\int_{-c}^{\infty} \varphi(z)\,\mathrm{d}z + \sigma\int_{-c}^{\infty} z\varphi(z)\,\mathrm{d}z - \mu\int_{-\infty}^{-c} \varphi(z)\,\mathrm{d}z - \sigma\int_{-\infty}^{-c} z\varphi(z)\,\mathrm{d}z \\
&= \mu\Big(\int_{-c}^{\infty} \varphi(z)\,\mathrm{d}z - \int_{-\infty}^{-c} \varphi(z)\,\mathrm{d}z\Big) + \sigma\Big(\int_{-c}^{\infty} z\varphi(z)\,\mathrm{d}z - \int_{-\infty}^{-c} z\varphi(z)\,\mathrm{d}z\Big) \\
&= \mu\Big(\int_{-\infty}^{\infty} \varphi(z)\,\mathrm{d}z - 2\int_{-\infty}^{-c} \varphi(z)\,\mathrm{d}z\Big) + \sigma\Big(- \varphi(z)\big|_{-c}^{\infty} + \varphi(z)\big|_{-\infty}^{-c}\Big) \\
&= \mu \cdot \Big(1 - 2\mathbf{\Phi}(-c)\Big) + \sigma \cdot \Big(- \varphi(\infty) + \varphi(-c) + \varphi(-c) - \varphi(-\infty)\Big) \\
&= \mu \cdot \Big(2\mathbf{\Phi}(c) - 1\Big) + 2\sigma \cdot\ \varphi(c)
\end{aligned}$$

gdzie skorzystaliśmy z tożsamości $\mathbf{\Phi}(-x) = 1 - \mathbf{\Phi}(x)$, $\varphi(-x) = \varphi(x)$, $\varphi(\pm\infty) = 0$, $\int_{-\infty}^{\infty} \varphi(x)\,\mathrm{d}x = 1$ oraz $\int x\varphi(x)\,\mathrm{d}x = -\varphi(x)$. □

Z Lematu 2 dostajemy $\mathbb{E}[C] = 2\xi \cdot \varphi\Big(\frac{\eta}{\xi}\Big) + \eta \cdot (2\mathbf{\Phi}\Big(\frac{\eta}{\xi}\Big) - 1)$. Ostatecznie

$$\begin{aligned}
\mathbb{E}[X_v] &\approx \tfrac{1}{2}\Big(\mu_a + \mu_b - 2\xi\,\varphi\Big(\frac{\eta}{\xi}\Big) - \eta\Big(2\mathbf{\Phi}\Big(\frac{\eta}{\xi}\Big) - 1\Big)\Big) \\
&= \tfrac{\mu_a+\mu_b}{2} - \xi\,\varphi\Big(\frac{\eta}{\xi}\Big) - (\mu_a - \mu_b)\Big(\mathbf{\Phi}\Big(\frac{\eta}{\xi}\Big) - \tfrac{1}{2}\Big) \\
&= \mu_a\Big(1 - \mathbf{\Phi}\Big(\frac{\eta}{\xi}\Big)\Big) + \mu_b\,\mathbf{\Phi}\Big(\frac{\eta}{\xi}\Big) - \eta\,\varphi\Big(\frac{\eta}{\xi}\Big)
\end{aligned}$$

Przenalizujmy teraz zachowanie asymptotyczne otrzymanego wyrażenia. Skoro $a + b = n$ to niech $a = rn$, $b = (1-r)n$ dla pewnego $r \in (0;1)$. Dalej

$$\frac{\eta}{\xi} = \frac{\mu_a - \mu_b}{\sqrt{\sigma_a^2 + \sigma_b^2}} = \frac{\frac{a}{p} - \frac{b}{p}}{\sqrt{\frac{aq}{p^2} + \frac{bq}{p^2}}} = \frac{(2r-1)\sqrt{n}}{\sqrt{q}}$$

Musimy rozważyć dwa przypadki.

Jeśli $a < b$, co za tym idzie $r < \frac{1}{2}$ to $\frac{\eta}{\xi} \to -\infty$ wraz z $n \to \infty$. Wtedy też $\varphi\left(\frac{\eta}{\xi}\right) \to 0$ oraz $\mathbf{\Phi}\left(\frac{\eta}{\xi}\right) \to 0$ a więc $\mathbb{E}[X_v] \to \frac{a}{p}$.

Zaś gdy $a = b$ to $r = \frac{1}{2}$ jak i $\frac{\eta}{\xi} = 0$. Wiemy, że $\varphi(0) = \frac{1}{\sqrt{2\pi}}$ oraz $\mathbf{\Phi}(0) = \frac{1}{2}$. Wstawiając otrzymamy $\mathbb{E}[X_v] \to \frac{n}{2p} - \frac{\sqrt{np}}{p\sqrt{2\pi}}$. Podsumowując mamy następujący wynik:

$$\mathbb{E}[X_v] \sim \frac{\min\{v, n-v\}}{p}$$

Jest to całkowicie zgodne z intuicją. Wierzchołki w grafie C_n zachowują się podobnie jak w grafach P_n.

W celu wyznaczenie rozkładu Y_t dokonajmy obserwacji, że gdy dwie drogi zarażania spotkają się to propagacja dobiega końca. Każda z tych drug jak w przypadku grafu ścieżkowego ma rozkład dwumianowy. Możemy zapisać zatem

$$Y_t \sim \min\{n, 1 + L_t + R_t\}, \quad L_t, R_t \sim \mathrm{Bin}(t, p)$$

Z Faktu 3 mamy $L_t + R_t \sim \mathrm{Bin}(2t, p)$. Widzimy zatem, że rozkład Y_t dla grafu C_n pokrywa się ze zmienną Y_{2t} dla grafów typu P_n. Z wcześniejszego wyniku dla grafów ścieżek dostajemy

$$\lim_{n\to\infty} \mathbb{E}[Y_t] = 1 + 2tp$$

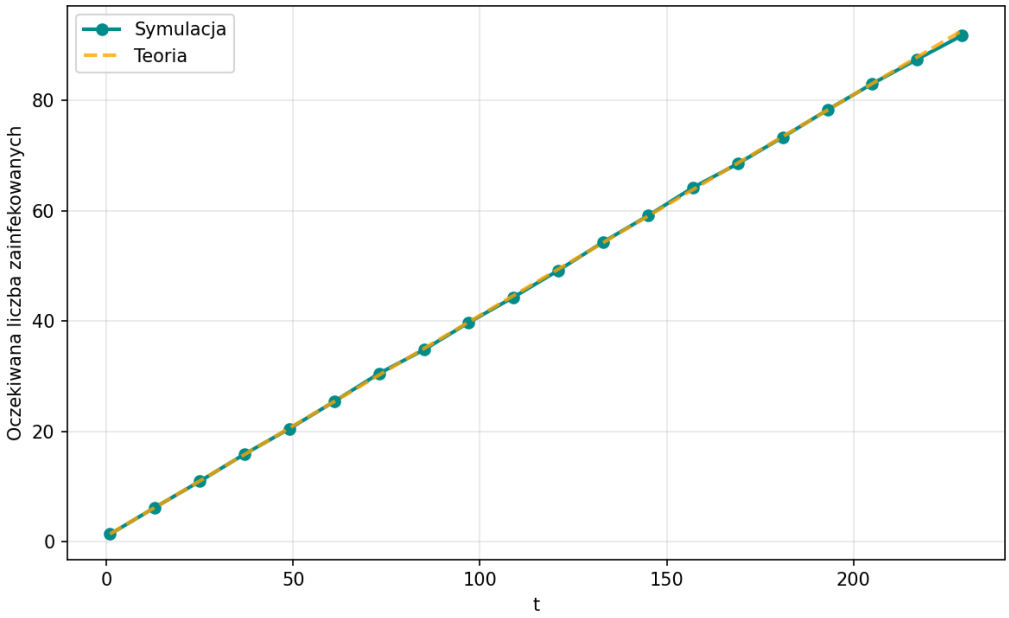
Teraz możemy wyznaczyć $\mathbb{E}[Z]$. Dla n patrzystego najdalej oddalony wierzchołek od źródła to $\frac{n}{2}$ a dla n nieparzystego to $\lfloor \frac{n\pm1}{2} \rfloor$. Asymptotycznie nie ma to znaczenia, możemy przyjąć $v = \frac{n}{2}$. Stąd

$$\mathbb{E}[Z] \approx \mathbb{E}[X_{\frac{n}{2}}] \approx \frac{n}{2p} - \frac{\sqrt{np}}{p\sqrt{2\pi}}$$

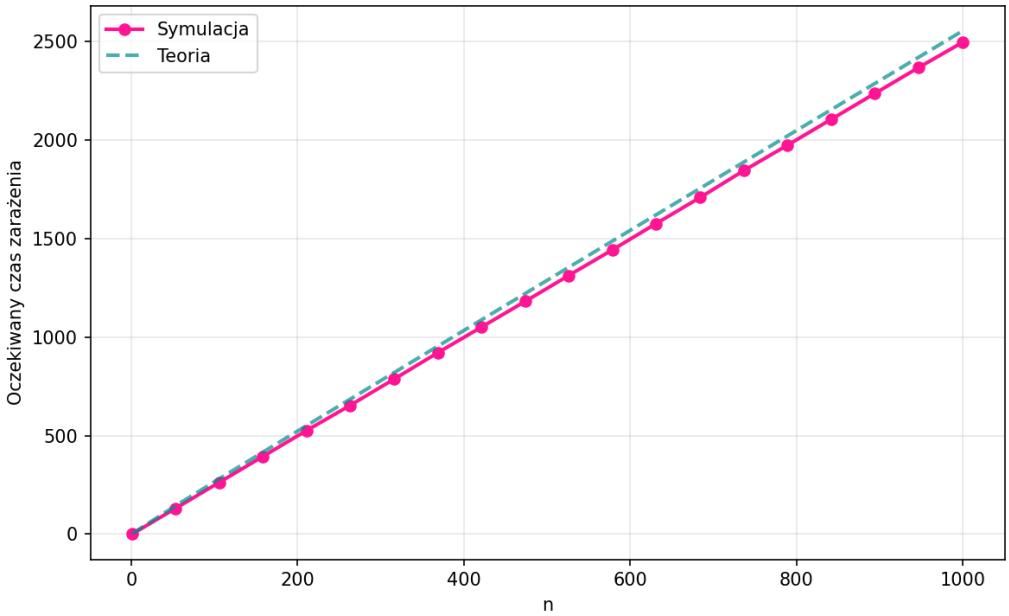
Również intuicyjny wynik. Aby się upewnić czy nie przesadziliśmy z szacowanie zwrócmy się ku symulacji. Dla $p = 0.2$, $n \in \{3, 4, \ldots, 1000\}$ policzmy wartość oczekiwaną całkowitego zarażenia po razy 2000 razy. Z wykresu widzimy, że empiryczny wynik pokrywa się asymptotycznie z teoretycznym.

4.7 Grafy pełne

Graf pełny K_n intuicyjnie powinien mieć najszybszą propagację ze względu na maksymalną liczbę krawędzi. Za źródło możemy przyjąć dowolny wierzchołek $s \in V$ ze względu na symetrie. Początkowo rozkład X_v pokrywa się



Rysunek 5: $\mathbb{E}[Y_t]$ dla $\mathrm{C_n}$ w funkcji t



Rysunek 6: $\mathbb{E}[Z]$ dla $\mathrm{C_n}$ w funkcji n

z rozkładem gwiazdy natomiast w każdej kolejnej rundzie mocno się kompli-

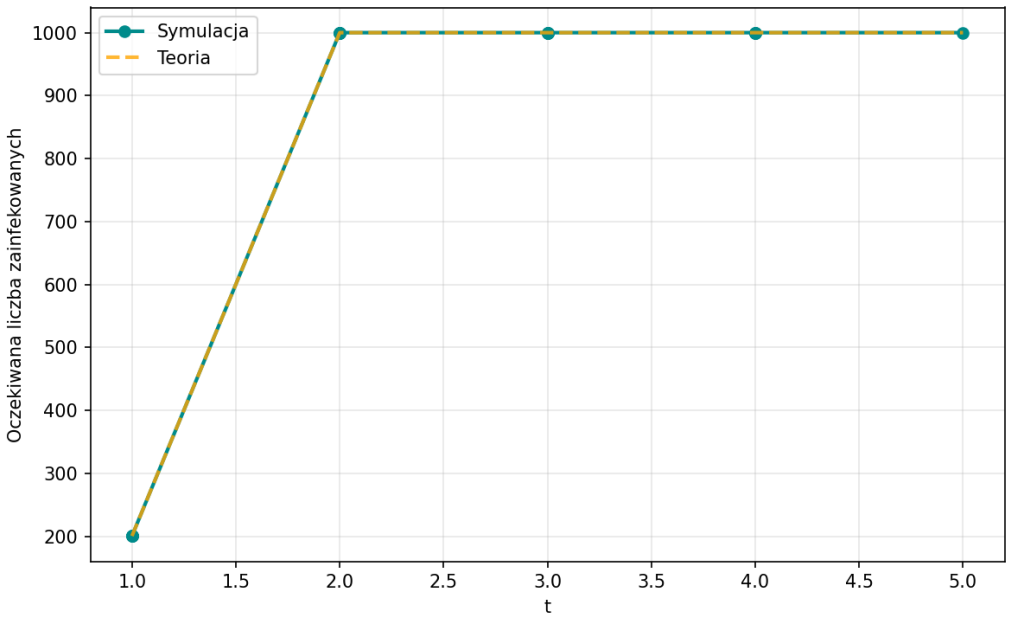
kuje. Zachodzi bowiem $Y_t = a$ to $\mathbb{P}[X_v = t+1 | Y_t = a] = 1 - q^a$. Nie mamy co liczyć na jakiekolwiek sensowne wyznaczenie rozkładu X_v. Podejdźmy do problemu narazie heurystycznie. Zauważmy, że jeśli $Y_1 = a$ to rozkład zmiennej Y_2 wynosi $Y_2 = a + B$ dla $B \sim \text{Bin}(n-a, 1-q^n)$. Zatem

$$\mathbb{E}[Y_2 \mid Y_1 = a] = n \cdot (1 - q^a) + aq^a$$

Mamy $Y_1 \sim \text{Bin}(n-1, p)$ oraz $\mathbb{E}[Y_1] = 1 + (n-1)p$. Możemy również założyć, że również $a \approx \mathbb{E}[Y_1]$ a co za tym idzie

$$\mathbb{E}[Y_2 \mid Y_1 = a] \approx n(1 - q^{1+(n-1)p}) + (1 + (n-1)p)q^{1+(n-1)p}$$

Jeśli $n \to \infty$ to wyrażenie to jest bliskie n. Spodziewamy się zatem, że zaledwie po dwóch rundach cały graf K_n będzie zainfekowany. Zweryfikujmy teraz ten heurystyczny argument symulacją w Pythonie. Ustalmy $p = 0.2$ i dla $n \in \{2, 3, \ldots, 1000\}$ odpalmy propagację. Widzimy, że dla $n > 200$

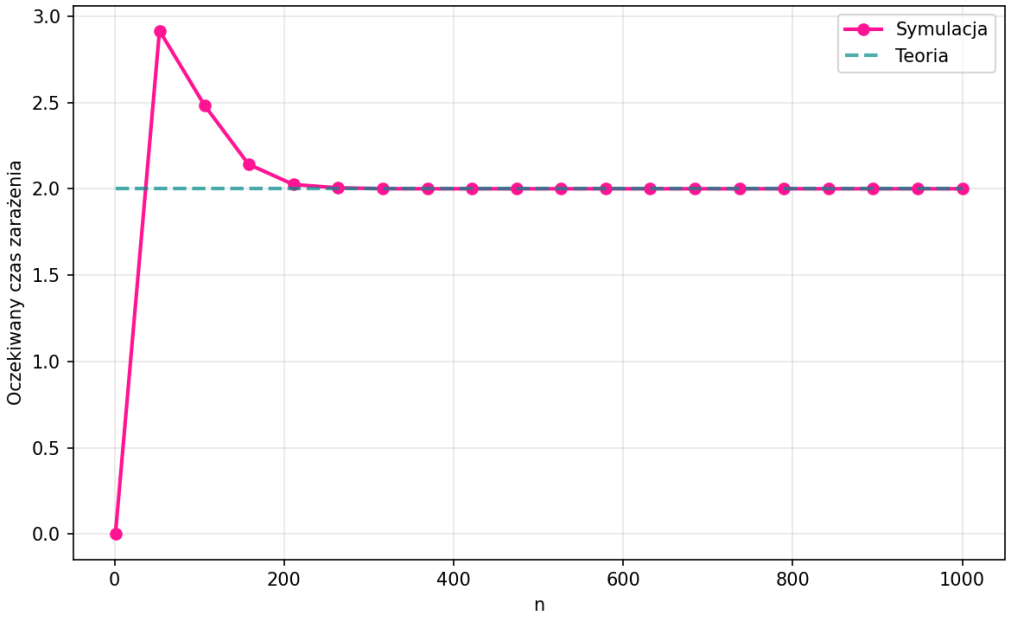


Rysunek 7: $\mathbb{E}[Y_t]$ dla K_n w funkcji t

mamy $\mathbb{E}[Z] \approx 2$. Możemy więc wysunąć hipotezę: Dla grafu K_n mamy:

$$\lim_{n \to \infty} \mathbb{E}[Z] = 2$$

Postarajmy się ją teraz udowodnić. Żeby to zrobić najpierw wyznaczmy asymptotyke $\mathbb{E}[Y_2]$. Oznaczamy $U = Y_1 = 1 + W$ gdzie $W \sim \text{Bin}(n-1, p)$. Z



Rysunek 8: $\mathbb{E}[Z]$ dla $\mathrm{K_n}$ w funkcji n

prawa całkowitej wartości oczekiwanej mamy

$$\mathbb{E}[Y_2] = n \cdot (1 - \mathbb{E}[q^U]) + \mathbb{E}[Uq^U]$$

Musimy wyznaczyć $\mathbb{E}[q^U]$ jak i $\mathbb{E}[Uq^U]$.

Lemat 3. *Niech* $X \sim \mathrm{Bin}(m, p)$. *Wtedy*

$$\mathbb{E}[z^X] = (q + pz)^m, \quad \mathbb{E}[Xz^X] = mpz(q + pz)^{m-1}$$

Dowód. Do obliczenia tych wartości posłuży nam Suma 1 jak i Suma 2.

$$\mathbb{E}[z^X] = \sum_{k=0}^{m} z^k \cdot \binom{m}{k} p^k q^{m-k} = \sum_{k=0}^{m} \binom{m}{k} (pz)^k q^{m-k} = (q + pz)^m$$

$$\mathbb{E}[Xz^X] = \sum_{k=0}^{m} kz^k \binom{m}{k} p^k q^{m-k} = \sum_{k=0}^{m} k \binom{m}{k} (pz)^k q^{m-k} = mpz(q + pz)^{m-1}$$

□

W naszym przypadku dostajemy

$$\mathbb{E}[q^U] = \mathbb{E}[q^{1+W}] = q \cdot \mathbb{E}[q^W] = q(q + pq)^{n-1} = q^n(1 + p)^{n-1}$$

Dla drugiej wartości mamy zaś

$$\begin{aligned}\mathbb{E}[Uq^U] &= \mathbb{E}[(1+W)q^{1+W}] = q(\mathbb{E}[q^W] + \mathbb{E}[Wq^W]) \\ &= q(q^{n-1}(1+p)^{n-1} + (n-1)pq^{n-1}(1+p)^{n-2}) \\ &= q^n(1+p)^{n-2}(1+p+(n-1)p) = q^n(1+p)^{n-2}(1+np)\end{aligned}$$

Podstawiając przed chwilą wyrażenia wzory do wzoru na $\mathbb{E}[Y_2]$ dostaniemy

$$\begin{aligned}\mathbb{E}[Y_2] &= n - nq^n(1+p)^{n-1} + q^n(1+p)^{n-2}(1+np) = \\ &= n - (n-1)q^n(1+p)^{n-2} = n - (n-1)(1+p)^{-2}(1-p^2)^n\end{aligned}$$

Połóżmy $\varepsilon_n = (n-1)(1+p)^{-2}(1-p^2)^n$. Wtedy $\mathbb{E}[Y_2] = n - \varepsilon_n$. Z nierówności Markova (5) otrzymujemy $\mathbb{P}[Z \geq 3] = \mathbb{P}[n - Y_2 \geq 1] \leq \mathbb{E}[n - Y_2] = \varepsilon_n$. Dalej zauważmy, że $\mathbb{P}[Z = 1] = p^{n-1}$ bo wszystkie próby zarażenia w rundzie pierwszej musiałby by się powieść. Ograniczmy teraz z dwóch stron $\mathbb{E}[Z]$. Z dołu mamy

$$\mathbb{E}[Z] = \sum_{k=1}^{\infty} \mathbb{P}[Z \geq k] \geq \mathbb{P}[Z \geq 1] + \mathbb{P}[Z \geq 2] = 1 + 1 - p^{n-1} = 2 - p^{n-1}$$

Zajmijmy się teraz oszacowaniem górnym. Zauważmy, że graf K_n zawiera P_n jako podgraf. Ustalmy jeden z tych podgrafów. Niech Z' będzie zmienną losową czasu całkowitego zarażenia dla tego podgrafu. Z Twierdzenia 1 mamy $Z \leq Z'$ a co za tym idzie $\mathbb{E}[Z^2] \leq \mathbb{E}[(Z')^2]$. Przypomnijmy, że $Z' \sim \mathrm{NegBin}(n-1, p)$ a więc $\mathbb{E}[(Z')^2] = \frac{(n-1)^2+(n-1)q}{p^2}$.

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{k=1}^{\infty} \mathbb{P}[Z \geq k] = \mathbb{P}[Z \geq 1] + \mathbb{P}[Z \geq 2] + \sum_{k=3}^{\infty} \mathbb{P}[Z \geq k] \\ &= 1 + 1 - p^{n-1} + \mathbb{E}[Z \cdot \mathbf{1}_{Z \geq 3}] \leq 2 - p^{n-1} + \sqrt{\mathbb{E}[Z^2]}\sqrt{\mathbb{E}[\mathbf{1}_{Z \geq 3}]} \\ &\leq 2 - p^{n-1} + \sqrt{\mathbb{E}[(Z')^2]}\sqrt{\mathbb{P}[Z \geq 3]} \\ &\leq 2 - p^{n-1} + \sqrt{\frac{(n-1)^2 + (n-1)q}{p^2}}\sqrt{\varepsilon_n}\end{aligned}$$

gdzie wykorzystaliśmy nierówność Cauchy'ego-Schwarza (6). Ostatecznie dostajemy

$$2 - p^{n-1} \leq \mathbb{E}[Z] \leq 2 - p^{n-1} + \sqrt{\frac{(n-1)^2 + (n-1)q}{p^2}}\sqrt{\varepsilon_n}$$

a zatem

$$\lim_{n\to\infty} \mathbb{E}[Z] = 2$$

Jeżeli zaledwie po dwóch rundach cały graf jest poinformowany to rozkłady X_v czy Y_t nie są dla nas istotne. Spójrzmy jeszcze na oszacowanie, które otrzymamy stosując Twierdzenie 2 dla grafu pełnego. Wynosi ono $1 + \frac{\log(n-1)+1}{\log(\frac{1}{1-p})}$. Ograniczenie to zdaje się nie być za dobre czego przyczną jest fakt, że K_n jest grafem gęstym.

4.8 Drzewa

Rozważmy drzewo $G = (V, E)$ oraz ustalony wierzchołek początkowy $s \in V$, który traktujemy jako korzeń drzewa. Dla $v \in V$ oznaczmy $d_v = \mathrm{d}(s, v)$. Ustalmy $v \in V$. Skoro G jest drzewem to istnieje dokładnie jedna ścieżka od s do v, powiedzmy $s, v_1, \ldots, v_k, v$. Ponieważ infekcja rozprzestrzenia się od korzenia s wzdłuż krawędzi drzewa, każde zakażenie wymaga sukcesu w niezależnym doświadczeniu Bernoulliego o prawdopodobieństwie p. W konsekwencji, aby infekcja dotarła z s do v, musi wystąpić d_v kolejnych sukcesów. Zatem rozkład X_v pokrywa się z rozkładem tej zmiennej dla grafu P_{d_v+1} na wierzchołkach $\{s, v_1, \ldots, v_k, v\}$. Stąd

$$X_v \sim \mathrm{NegBin}(\mathrm{d}_v, p)$$

oraz

$$\mathbb{E}[X_v] = \frac{\mathrm{d}_v}{p}, \quad \mathrm{Var}[X_v] = \frac{\mathrm{d}_v \cdot (1-p)}{p^2}$$

Lemat 4. *Dla dowolnego $t \in \mathbb{N}$ wartość oczekiwana zmiennej Y_t wyraża się wzorem*

$$\mathbb{E}[Y_t] = \sum_{v\in V} \mathbb{P}[X_v \leq t]$$

Dowód. Mamy $Y_t = |\{v \in V : X_v \leq t\}|$ zatem $Y_t = \sum_{v\in V} \mathbf{1}_{\{X_v \leq t\}}$. Nakładając na tą równość operator $\mathbb{E}$ otrzymujemy:

$$\mathbb{E}[Y_t] = \mathbb{E}\Big[\sum_{v\in V} \mathbf{1}_{\{X_v\leq t\}}\Big] = \sum_{v\in V} \mathbb{E}[\mathbf{1}_{\{X_v\leq t\}}] = \sum_{v\in V} \mathbb{P}[X_v \leq t]$$

□

Przejdźmy teraz to obliczania średniej liczby zainfekowanych wierzchołków w czasie t. Oznaczmy przez $F(t; m, p)$ dystrybuante zmiennej o rozkładzie $\mathrm{NegBin}(m, p)$. Z Lematu 4 otrzymujemy

$$\mathbb{E}[Y_t] = \sum_{v\in V} F(t; d_v, p)$$

Połóżmy $a_j = |\{v \in V : d_v = j\}|$ dla $0 \leq j \leq h$. Wtedy

$$\mathbb{E}[Y_t] = \sum_{j=0}^{h} a_j \cdot F(t; j, p)$$

Ponadto gdy $t < j \leq h$ to $F(t; j, p)$, bo żaden wierzchołek w odległości od korzenia większej niż liczba rund nie może zostać zarażony. Możemy więc zmniejszyć granice sumowania

$$\mathbb{E}[Y_t] = \sum_{j=0}^{\min\{h,t\}} a_j \cdot F(t; j, p)$$

Oszacujmy teraz średni czas całkowity czas propagacji drzewa. Niech $L = \{u_1, \ldots, u_m\}$ będzie zbiorem liści w G. Wtedy mamy $Z = \max_{u \in L} X_u$. Zauważmy, że $\epsilon(s) = \max_{u \in L} \mathrm{d}_u$ i jest to wysokość drzewa. Oznaczmy ją przez h. Z nierówności Jensena (7) otrzymujemy

$$\mathbb{E}[Z] = \mathbb{E}[\max_{u \in L} X_u] \geq \max_{u \in L} \mathbb{E}[X_u] = \max_{u \in L} \frac{\mathrm{d}_u}{p} = \frac{h}{p}$$

Aby ogarniczyć $\mathbb{E}[Z]$ z góry skorzystamy z Twierdzenia 2:

$$\mathbb{E}[Z] \leq h + h \cdot \frac{\log(\frac{n-1}{h}) + 1}{\log(\frac{1}{1-p})}$$

Ograniczenia te są różnych rzędów wielkości. Jednakże nie da się ich poprawić dla ogólnego drzewa znając tylko liczbę jego wierzchołków i wysokość. Ustalmy $n \in \mathbb{N}_+$ oraz $h \in \{1, 2, \ldots, n-1\}$ i poszukajmy drzew o n wierzchołkach i wysokości h osiągających zarówno dolne jak i górne ograniczenie na $\mathbb{E}[Z]$. Dla dolnej nierówności możemy wziąść drzewo składające się ze ścieżki długości h oraz $n-1-h$ liści bezpośrednio przy korzeniu. Wtedy $\mathbb{E}[Z] \approx \frac{h}{p}$. Aby znaleść drzewo osiągające górne ograniczenie musimy wrócic do dowodu Twierdzenia 2. Udowadniając granicę na wartość oczekiwaną korzystamy z trzech nierówności. Pierwsza z nich to Nierówność 2. Jest ona bardzo ciasna a ponadto niezależy od grafu. Druga z nich to nierównośc między średnią arytmetyczną a geometryczną (4). Aby uzyskać równość potrzebujemy mieć $a_1 = \cdots = a_h$. Czyli innymi słowy, nasze drzewo ma tyle samo węzłów na każdej głębokości. Połóżmy $a_1 = b$. Wtedy $hb = n - 1$ a więc $b = \lfloor \frac{n-1}{h} \rfloor$. Na koniec zostaje nierówność wynikająca z Lematu 1. Sam lemat daje nierówność, której nie da się poprawić, co wiemy poprzez analize dla grafów gwiazd. Lecz dla drzewa nędzie ona najmniej luźna, jeżeli każdy wierzchołek w warstwie A_j będzie mieć dokładnie jedną krawędź łaczącą go z wierzchołkiem w

warstiwe A_{j+1}, gdzie $0 \leq j \leq h-1$. Zatem drzewo składa się z korzenia oraz b rozłącznych ścieżek każda o długości h. I taki graf osiąga ograniczenie górne na $\mathbb{E}[Z]$. Widzimy zatem, że nasze ograniczenia nie są do poprawnienia bez dodatkowych parametrów grafu. Podsumowując możemy następująco szacować przewidywany czas całkowitego poinformowania drzewa o wysokości h:

$$\frac{h}{p} \leq \mathbb{E}[Z] \leq h + h \cdot \frac{\log(\frac{n-1}{h}) + 1}{\log(\frac{1}{1-p})}$$