Probabilistyczne modele propagacji w grafach.

Bartosz Łabuz

21 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 liczb naturalnych $\{0, 1, 2, \dots\}$, 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 $\mathrm{Var}[X]$. Funckję masy prawdopodobieństwa oznaczamy $\mathbb{P}[X = t]$ a dystrybuante X oznaczamy $F_X(t)$ dla $t \in \mathbb{R}$. Jeśli zmienne losowe $X_1, X_2, \dots, X_n$ są niezależne i o jednakowych rozkładach to mówimy, że są IID.

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 $E = \{\{0, i\} : i \in \{1, 2, \ldots, n\}\}$. Oznaczamy go przez S_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.

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.

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

2.4 Tożsamości i nierówności

Fakt 1. *Niech $a, b \in \mathbb{N}$, $a < b$ oraz $f : [a;b] \to \mathbb{R}$ będzie funkcją ciągłą i monotoniczą. Jeśli f jest rosnąca to*

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

Jeśli f jest malejąca to

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

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

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

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

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

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

$$\sqrt[n]{x_1 \cdots x_n} \leq \frac{x_1 + \cdots + x_n}{n}$$

Równoważnie możemy zapisać

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

Fakt 5. *Niech $n \in \mathbb{N}$ oraz $x, y \in \mathbb{R}$. Wtedy*

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

Fakt 6. *Niech $n \in \mathbb{N}$ oraz $x, y \in \mathbb{R}$. Wtedy*

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

Fakt 7. *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 8. *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 9. *Niech $X_1, X_2, \ldots, X_m$ będą niezależnymi zmiennymi losowymi o rozkładzie geometrycznym* $\mathrm{Geo}(p)$ *oraz $Y = X_1 + X_2 + \cdots + X_m$. Wtedy*

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

Fakt 10. *Niech $X, Y : \Omega \to \mathbb{R}$ będą zmiennymi losowymi takim, że dla każdego $\omega \in \Omega$ zachodzi $X(\omega) \leq Y(\omega)$. Wtedy.*

$$\mathbb{E}[X] \leq \mathbb{E}[Y]$$

Fakt 11 (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.

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 dwoch wierzchołkach u, v połączonych krawędzią. Za wierzchołek startowy wybierzmy u. W tym przypadku 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, $X_v \sim \mathrm{Geo}(p)$. Jeśli chodzi o rozkład Y_t to mamy:

- $\mathbb{P}[Y_t = 1] = q^t$, bo próba zarażenia musiałaby nie udać się t razy
- $\mathbb{P}[Y_t = 2] = 1 - q^t$

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 $Z \sim \mathrm{Geo}(p)$ oraz $\mathbb{E}[Z] = \frac{1}{p}$.

4.2 Analiza dla grafów ścieżkowych

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

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

gdzie $U_1, U_2, \ldots, U_n \sim \mathrm{Geo}(p)$ oraz $U_1, U_2, \ldots, U_n$ są niezależne. Widzimy zatem, że

$$X_k \sim U_1 + U_2 + \cdots + U_{k-1}$$

a więc z Faktu 9 X_k ma rozkład ujemny dwumianowy o parametrach $(k-1, p)$,

$$X_k \sim \text{NegBin}(k-1, p).$$

Ponadto mamy:

$$\mathbb{E}[X_k] = \frac{k-1}{p}, \quad \text{Var}[X_k] = \frac{(k-1)(1-p)}{p^2}$$

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

$$Y_t = \min\{n, 1 + B_t\}, \quad \text{gdzie} \quad B_t \sim \text{Bin}(t, p)$$

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:

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

W pierwszej sumie podstawiamy $j = k-1$, co pozwala nam złączyć obie sumy i otrzymać:

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

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

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

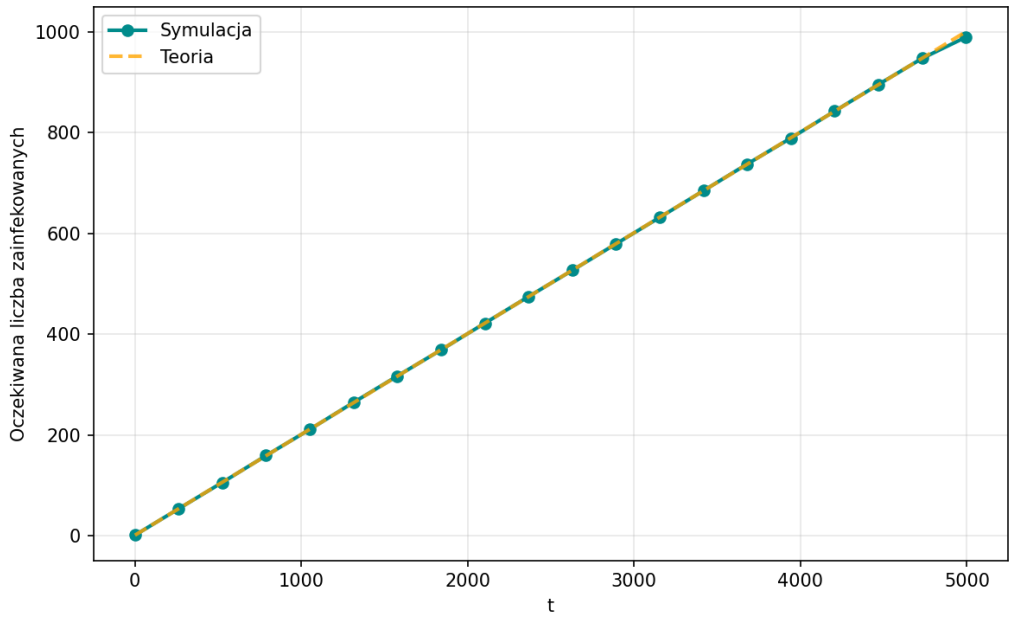
gdzie sumy sumujemy korzystając z Faktu 5 oraz Faktu 6. Stąd

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

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

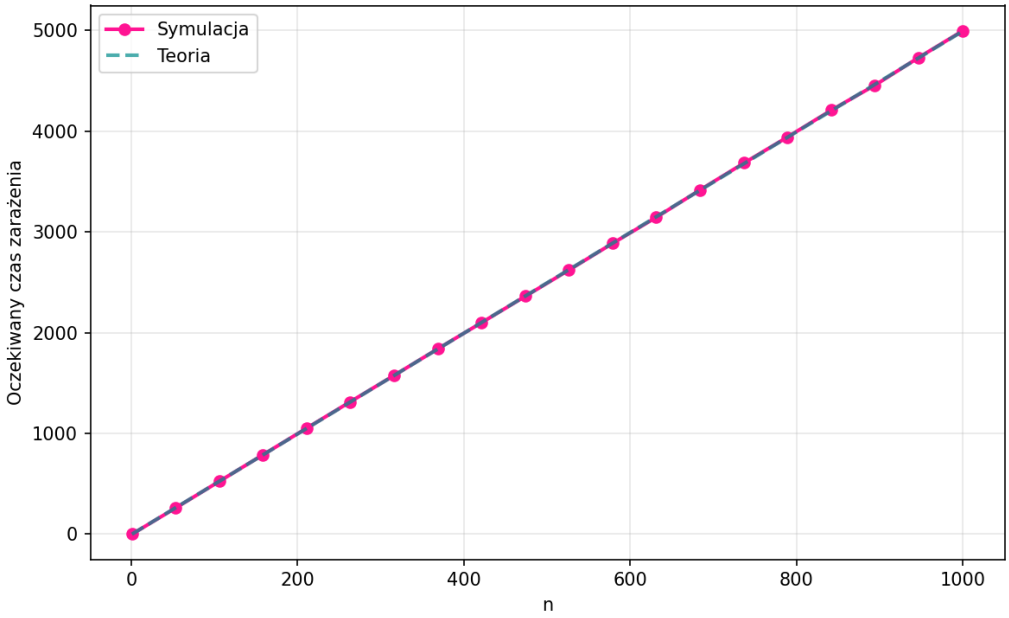


Rysunek 1: $\mathbb{E}[Y_t]$ dla $\mathrm{P_n}$ w funkcji t

się z przewidywanymi kształtami, to jest $1 + tp$ dla $\mathbb{E}[Y_t]$ oraz $\frac{n-1}{p}$ dla $\mathbb{E}[Z]$.

4.3 Analiza dla grafów gwiazd

Następnie rozpatrzmy rodzinę grafów gwiazd S_n. Niech źródłem będzie centralny wierzchołek grafu, to jest $s = 0$. Propagacja rozchodzi się tutaj po



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

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 $\mathbb{P}[X_v \leq t] = 1 - q^t$. Podsumowując mamy

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

Stąd oczywiście otrzymujemy

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

Przejdźmy teraz do zmiennej Z. Przypomnijmy, że $Z = \max\{X_1, X_2, \ldots, X_n\}$. Skoro zmienne te są IID, to z Faktu 7 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 $\mathbb{E}[Z] = \sum_{k=0}^{\infty} 1 - (1-q^k)^n$. Fakt 1 umożliwia oszacowanie tej sumy. Połóżmy $f(x) = 1 - (1-e^{-\alpha x})^n$ gdzie $\alpha = -\log(q)$. 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 $u = 1 - e^{-\alpha x}$. Wtedy $\mathrm{d}u = \alpha e^{-\alpha x}\, \mathrm{d}x$, a więc $\mathrm{d}x = \frac{1}{\alpha} \cdot \frac{1}{1-u}\, \mathrm{d}u$. Ponadto $u(0) = 0$, $u(\infty) = 1$ (bo $\alpha > 0$). Zatem całka ma postać:

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

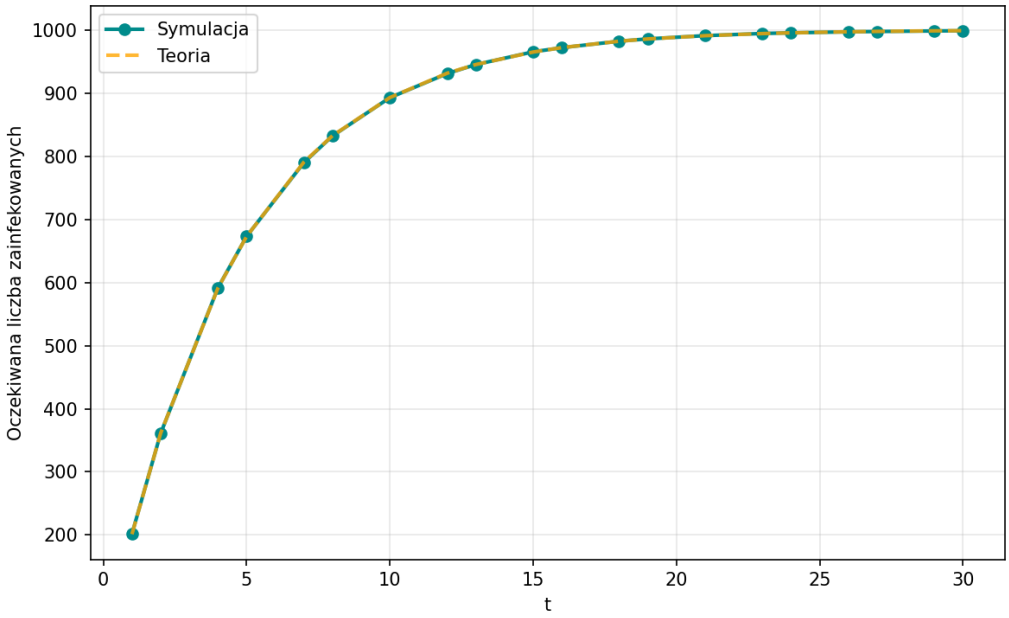
Zauważmy, że $-\log(q) = \log(\frac{1}{1-p})$ a więc ostatecznie dostajemy:

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

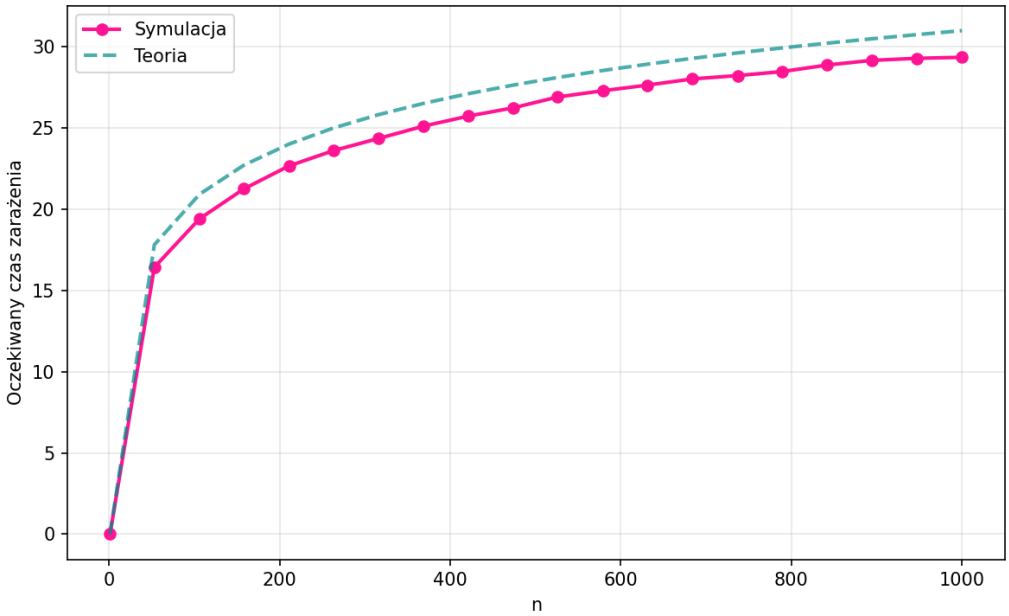
Stąd mamy asymptotyczny przewidywany czas zarażenia grafu S_n:

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

Przeprowadźmy teraz symulacje. Ustalmy $p = 0.2$, $n = 1000$. Dla każdego $t \in \{1, 2, \ldots, \log(n)\}$ 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.



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

4.4 Ograniczenia na czasu 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 Z oznaczymy czas całkowitego zarażenia G i odpowiednio przez Z' dla G' to wtedy zachodzi nierówność $\mathbb{E}[Z] \leq \mathbb{E}[Z']$.*

Intuicyjnie sprawa jest oczywista. Mając mniej krawędzi w grafie potrzebujemy więcej czasu na rozprzestrzenienie się w nim informacji.

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 $t \in \mathbb{N}$. Ponadto mamy

$$Z = \min\{t \in \mathbb{N} : \mathcal{I}_t = V\}, \quad Z' = \min\{t \in \mathbb{N} : \mathcal{I}'_t = V\}$$

Niech $Z' = \tau$. Wtedy $\mathcal{I}'_\tau = V$ a więc $V \subseteq \mathcal{I}_t$. Zatem $\mathcal{I}_\tau = V$ i $Z \leq \tau$. Korzystając z Faktu 10 dostajemy $\mathbb{E}[Z] \leq \mathbb{E}[Z']$. □

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. Ozanczmy przez $\ell = \epsilon(s)$. Wtedy zachodzi*

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

Dowód. Dla $0 \leq j \leq \ell$ kładziemy $A_j = \{v \in V : \mathrm{d}(s, v) = j\}$ oraz $n_j = |A_j|$. Oczywiście $n_0 = 1$ a więc $n_1 + \cdots + n_\ell = n - 1$. 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 $1 \leq j \leq \ell$. Wtedy:*

$$\mathbb{E}[U_j] \leq \frac{H_{n_j}}{\log(\frac{1}{1-p})} + 1$$

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_{n_j} bo $n_j = |A_j|$. Pamiętajamy, że $\frac{H_{n_j}}{\log(\frac{1}{1-p})} + 1$ jest ograniczeniem górnym na całkowity czas zarażenia grafu gwiazdy oraz wykorzystując Twierdzenie 1 dostajemy porządany wynik. □

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

$$\begin{aligned}
\mathbb{E}[Z] \leq \mathbb{E}[T_\ell] &= \mathbb{E}\left[\sum_{j=1}^{\ell} U_j\right] = \sum_{j=1}^{\ell} \mathbb{E}[U_j] \leq \sum_{j=1}^{\ell} \frac{H_{n_j}}{\log(\frac{1}{1-p})} + 1 \\
&= \ell + \frac{1}{\log(\frac{1}{1-p})} \sum_{j=1}^{\ell} H_{n_j} \leq \ell + \frac{1}{\log(\frac{1}{1-p})} \sum_{j=1}^{\ell} 1 + \log(n_j) \\
&= \ell + \frac{\ell}{\log(\frac{1}{1-p})} \cdot \left(1 + \sum_{j=1}^{\ell} \log(n_j)\right) \\
&= \ell + \frac{\ell}{\log(\frac{1}{1-p})} \cdot \left(1 + \log\left(\prod_{j=1}^{\ell} n_j\right)\right) \\
&\leq \ell + \frac{\ell}{\log(\frac{1}{1-p})} \cdot \left(1 + \log\left(\frac{1}{\ell}\sum_{j=1}^{\ell} n_j\right)\right) \\
&= \ell + \frac{\ell}{\log(\frac{1}{1-p})} \cdot \left(1 + \log\left(\frac{n-1}{\ell}\right)\right) = \ell + \ell \cdot \frac{\log(\frac{n-1}{\ell}) + 1}{\log(\frac{1}{1-p})}.
\end{aligned}$$

gdzie w linijce pierwszej wykorzystujemy monotoniczności propagacji, w drugiej Fakt 2 a w piątej z nierówności między średnimi (4). □

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

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

Przypomnijmy, że faktyczna wartość oczekiwana jest równa $\frac{n-1}{p}$ więc oszacowanie jest dość ostre. Z kolei dla rodziny S_n mamy $\ell = 1$ 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.

4.5 Analiza dla drzew

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 2. *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:

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

□

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 2 otrzymujemy

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

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

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

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

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

Oszacujmy teraz średni czas całkowity czas propagacji drzewa. Niech $\{u_1, \ldots, u_m\}$ będą liściami w G. Wtedy mamy $Z = \max_{1 \leq i \leq m} X_{u_i}$. Zauważmy, że $\epsilon(s) = \max_{1 \leq i \leq m} \mathrm{d}_{u_i}$ i jest to wysokość drzewa. Oznaczmy ją przez h. Z nierówności Jensena (Fakt 11) dla funkcji $\max\{x_1, \ldots, x_m\}$ otrzymujemy

$$\mathbb{E}[Z] = \mathbb{E}[\max_{1 \leq i \leq m} X_{u_i}] \geq \max_{1 \leq i \leq m} \mathbb{E}[X_{u_i}] = \max_{1 \leq i \leq m} \frac{\mathrm{d}_{u_i}}{p} = \frac{h}{p}$$

Największą wysokość ma drzewo, które jest ścieżka i wtedy $h = n-1$. Zgodnie z poprzednimi wyliczeniami jest to dobre oszacowanie. Aby ogarniczyć $\mathbb{E}[Z]$ z góry skorzystamy z Twierdzenia 2:

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

4.6 Analiza dla grafów pełnych

Graf pełny K_n intuicyjnie powinien mieć najszybszą propagację ze względu na maksymalną liczbę krawędzi. Początkowo rozkład X_v pokrywa się z rozkładem gwiazdy natomiast w każdej kolejnej rundzie mocno się komplikuje. Jeśli bowiem $Y_t = a$ to $\mathbb{P}[X_v = t+1] = 1 - q^a$. Ta olbrzymia liczba powiązań uniemożliwia 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 $B \sim \mathrm{Bin}(n - a, 1 - q^n)$. Zatem

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

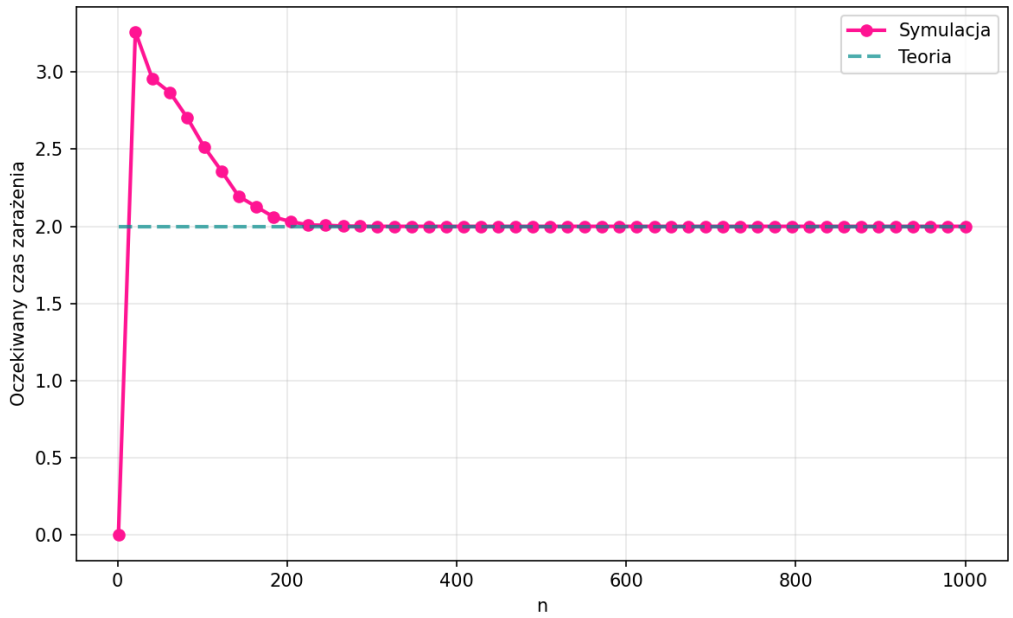
Skoro rozkład Y_1 pokrywa się z tym dla S_{n-1} to

$$\mathbb{E}[Y_1] \approx \frac{H_{n-1}}{\log(\frac{1}{1-p})} \approx \frac{\log(n)}{\log(\frac{1}{q})} = \log_q\left(\frac{1}{n}\right)$$

Powinniśmy się spodziewać, że również $a \approx \mathbb{E}[Y_1]$ a co za tym idzie $q^a \approx \frac{1}{n}$ jak również

$$\mathbb{E}[Y_2 \mid Y_1 = a] \approx n \cdot (1 - \frac{1}{n}) + \frac{1}{n} \cdot \log_q \left(\frac{1}{n}\right) \approx n - 1$$

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 $n > 200$ mamy $\mathbb{E}[Z] \approx 2$. Możemy więc wysunąć hipotezę:



Rysunek 5: $\mathbb{E}[Z]$ dla $\mathrm{K_n}$ w funkcji n

Dla grafu K_n mamy:

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