Metody analizy sieci złożonych

Zadanie 2 – podstawowa analiza sieci w Python (NetworkX) i R (igraph).

Zadanie 2

Zadanie numer 2 jest realizowane indywidualnie z terminem wysyłki rozwiązania do kolejnych zajęć. Średni czas realizacji zadania: 4 godziny. Celem zadania jest przejście od analizy wizualnej sieci (Zadanie 1) do analizy numerycznej (oba podejścia się uzupełniają).

Dane

Z repozytorium Konect http://konect.cc/networks/ pobierz następujący zbiór danych:

S1. Komunikacja mailowa przedsiębiorstwa produkcyjnego http://konect.cc/networks/radoslaw_email/ - jest to lista zawierająca metadane dot. korespondencji e-mailowej, gdzie każdy wiersz jest postaci:

<id_nadawcy, id_odbiorcy, waga, unix_timestamp>

Waga wynosi zawsze jeden, można ją zignorować. Z Jeśli jakaś wiadomość od jednego nadawcy ma ten sam znacznik czasowy (timestamp) co kilka kolejnych rekordów, oznacza to, że wiadomość była wysłana do kilku odbiorców, przy czym zbiór danych nie zawiera informacji kto w wiadomości był TO/CC/BCC.

Python i NetworkX (w trakcie zajęć)

NetworkX (https://networkx.org/) to pakiet do języka programowania Python, służący do tworzenia, manipulowania i badania struktury, dynamiki i funkcji sieci złożonych [NetworkX].

Oficjalny przewodnik użytkownika znajduje się pod adresem
https://networkx.org/documentation/stable/tutorial.html

Inne przykłady

- https://github.com/CambridgeUniversityPress/FirstCourseNetworkScience/tree/master/tutorials
- https://github.com/jdfoote/Intro-to-Programming-and-Data-Science/blob/summer2020/extra_topics/network_analysis.ipynb

1. **Grafy losowe (model Erdős-Rényi)**
    a. Wygeneruj sieć Erdős-Rényi o N=100 i p=0.05.
    b. Wylistuj wszystkie wierzchołki i krawędzie.
    c. Oblicz stopień (degree) każdego węzła a następnie stwórz histogram stopni węzłów.
    d. Ile jest komponentów (connected components) w grafie?

e. Zwizualizuj graf w taki sposób, aby rozmiar węzłów odpowiadał mierze PageRank.

2. **Sieci bezskalowe (model Barabási-Albert)**

   a. Wygeneruj graf wedle modelu Barabási-Albert z N=1000 i m_0=m=3

   b. Zwizualizuj graf layoutem Kamada-Kawai

   c. Znajdź najbardziej centralny węzeł według miary pośrednictwa (betweenness), jaki ma numer?

   d. Jaka jest średnica grafu?

   e. Jakie różnice widzisz pomiędzy grafem Barabási-Albert i Erdős-Rényi.

3. **Praca z rzeczywistymi danymi**

   a. Zaimportuj zbiór out.radoslaw_email_email i zachowaj tylko pierwsze dwie kolumny (dodatkowo przeskocz dwa pierwsze wiersze), następnie stwórz z zaimportowanych danych graf nieskierowany.

   b. Sprawdź ile wierzchołków i krawędzi ma Twój graf a następnie pozbądź się wielokrotnych krawędzi i pętli. O ile spadła liczba krawędzi?

   c. Jaki jest stopień każdego węzła? Stwórz histogram stopni węzłów. Podobnie stwórz histogram dla miary pośrednictwa (betweenness) oraz bliskości (closeness). Zweryfikuj czy wierzchołki, które mają największe wartości tych miar, są równie wysoko we wszystkich rankingach (możesz użyć np. miary Kendalla).

   d. Ile jest komponentów (connected components) znajduje się w grafie? Czy przy takiej reprezentacji danych mogą być wierzchołki bez krawędzi?

   e. Jaka jest średnica grafu i czy jest dużo inna od średniej długości ścieżki w grafie?

   f. Stwórz graf na nowo, ale tym razem jako skierowany. Ustaw wagi krawędzi w grafie w taki sposób, aby waga pomiędzy wierzchołkiem v_1 a v_2 była wyrażona jako liczba wiadomości wysłanych przez v_1 do v_2 w stosunku do wszystkich wiadomości wysłanych przez węzeł v_1.

   g. Sprawdź czy w takim grafie skierowanym można dostać się z każdego wierzchołka do każdego innego.

R i igraph (w trakcie zajęć lub po zajęciach)

Igraph (https://igraph.org/) to zestaw bibliotek do tworzenia i manipulowania grafami oraz analizy sieci złożonych. Jest napisany w C i istnieje również jako pakiety do języka Python i **R**. Istnieje ponadto interfejs dla oprogramowania Mathematica.

Pełna dokumentacja znajduje się na stronie https://igraph.org/r/#docs można się także do niej dostać bezpośrednio z biblioteki używając komend *library(help="igraph")* lub *help("igraph")*

1. **Grafy losowe (model Erdős-Rényi)**

   a. Wygeneruj sieć Erdős-Rényi o N=100 i p=0.03.

   b. Wylistuj wszystkie wierzchołki i krawędzie.

c. Ustaw wagi wszystkich krawędzi na losowe z zakresu 0.01 do 1

d. Oblicz współczynnik grupowania (Clustering coefficient) każdego węzła a następnie stwórz histogram współczynników stopni węzła.

e. Ile jest komponentów (connected components) w grafie?

f. Zwizualizuj graf w taki sposób, aby rozmiar węzłów odpowiadał stopniowi węzła (degree).

2. **Sieci bezskalowe (model Barabási-Albert)**

   a. Wygeneruj graf wedle modelu Barabási-Albert z N=1000 i m_0=m=3

   b. Zwizualizuj graf layoutem Fruchterman & Reingold

   c. Znajdź najbardziej centralny węzeł według miary bliskości (closeness), jaki ma numer?

   d. Jaka jest średnica grafu?

   e. Jakie różnice widzisz pomiędzy grafem Barabási-Albert i Erdős-Rényi.

3. **Praca z rzeczywistymi danymi**

   a. Pobierz sieć z repozytorium http://konect.cc/networks/ o zbliżonej do S1 liczbie wierzchołków i krawędzi (ale nie S1).

   b. Zaimportuj zbiór i stwórz graf nieskierowany.

   c. Sprawdź ile wierzchołków i krawędzi ma Twój graf a następnie pozbądź się wielokrotnych krawędzi i pętli. O ile spadła liczba krawędzi?

   d. Jaki jest stopień każdego węzła? Stwórz histogram stopni węzłów. Podobnie stwórz histogram dla miary pośrednictwa (betweenness) oraz bliskości (closeness). Zweryfikuj czy wierzchołki, które mają największe wartości tych miar, są równie wysoko we wszystkich rankingach (możesz użyć np. miary Kendalla).

   e. Ile jest komponentów (connected components) znajduje się w grafie?

   f. Jaka jest średnica grafu i czy jest dużo inna od średniej długości ścieżki w grafie?

   g. Stwórz histogram długości najkrótszych ścieżek (shorthest paths) pomiędzy wszystkimi wierzchołkami oraz histogram współczynników grupowania (Clustering coefficient).

Dyskusja po zadaniu 2

Czy widzisz jakieś zasadnicze różnice w pracy z oboma pakietami? Czy jakiś ma cechy, które uważasz za szczególnie interesujące w porównaniu z drugim?

Raport

Do prowadzącego przesyłamy raport zawierający kod, odpowiedzi na pytania, komentarze, wizualizacje i wykresy najlepiej w formie notatnika (np. Jupyter Notebook dla Pythona i R) ale może też być w formie skryptu Pythona/R z odpowiednimi komentarzami lub w formie PDFa.

Literatura

[Jankowski2018] Jankowski J., Szymanski B., Kazienko P., Michalski R., Bródka P. (2018) Probing Limits of Information Spread with Sequential Seeding. Scientific Reports, 8(1), 13996,

[Kempe2003] Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In ACM SIGKDD pp. 137-146.

[NetworkX] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008