Analiza statystyczna grafu przy użyciu standardowych narzędzi.

Gadawski Łukasz

17 listopada 2015

1 Wykorzystane dane

Numer mojego indeksu to 236655. W związku z tym reszta z dzielenia numeru przez 13 równa jest liczbie 3, a więc do mojej analizy przypisany został temat "Połączenia nerwowe nicienia *Caenorhabditis elegans*".

Jako dane wejściowe wykorzystano opis grafu w formacie *GraphML*. W przypadku pakietu *networkx* jest zdefiniowana gotowa metoda umożliwiająca wczytanie grafu w tym formacie. Do wykonania operacji w programie *Pajek* w załączonym skrypcie stworzono konwerter grafu w tym formacie do pliku w formacie *.net*, który może zostać wczytany przez program *Pajek*.

2 Networkx

Wykonanie funkcji *info()* na wczytanym grafie, powoduje wypisującej krótkiego podsumowania informacji o grafie:

```
Name: C. Elegans neural network
Type: MultiDiGraph
Number of nodes: 297
Number of edges: 2359
Average in degree:   7.9428
Average out degree:   7.9428
Is directed:  True
```

Typ *MultiDiGraph* oznacza, że wczytany graf jest grafem skierowanym umożliwiającym przechowywanie duplikujących się krawędzi. Następnie podana jest liczba wierzchołków oraz krawędzi, a także średnie stopnie wejściowe oraz wyjściowe wierzchołków.

Aby usunąć zduplikowane krawędzie oraz przekształcić na graf nieskierowany należy stworzyć graf typu *Graph* z wczytanego grafu. Po takiej operacji otrzymujemy następujące informacje na temat wczytanego grafu:

```
Name: C. Elegans neural network
Type: Graph
Number of nodes: 297
Number of edges: 2148
Average degree:  14.4646
Is directed:  False
```

Co jest zgodne z naszymi oczekiwaniami. Zmniejszyła się nieznacznie liczba krawędzi co świadczy o tym, że w grafie wejściowym występowały zduplikowane łuki, które zostały przekształcone na krawędzie.

Poprzez wykorzystanie odpowiedniej metody *networkx.connected_components(graph)* możliwe jest wyznaczenie składowych spójnych grafu, w tym przypadku występuje jedna składowa spójna. Rząd największej z nich wynosi 297 wierzchołków, natomiast jej rozmiar wynosi 2148 krawędzi. Stworzony w ramach projektu skrypt wypisuje te dane:

```
Number of connected components:  1
Range (nodes) of largest connected component:  297  nodes
Size (edges) of largest connected components:  2148  edges
```

Następnie zaimplementowano wyznaczenie 5 wierzchołków o największej wartości wymaganych współczynników, uzyskując następujące wartości (każda para zawiera etykietę wierzchołka oraz wartość danego współczynnika):

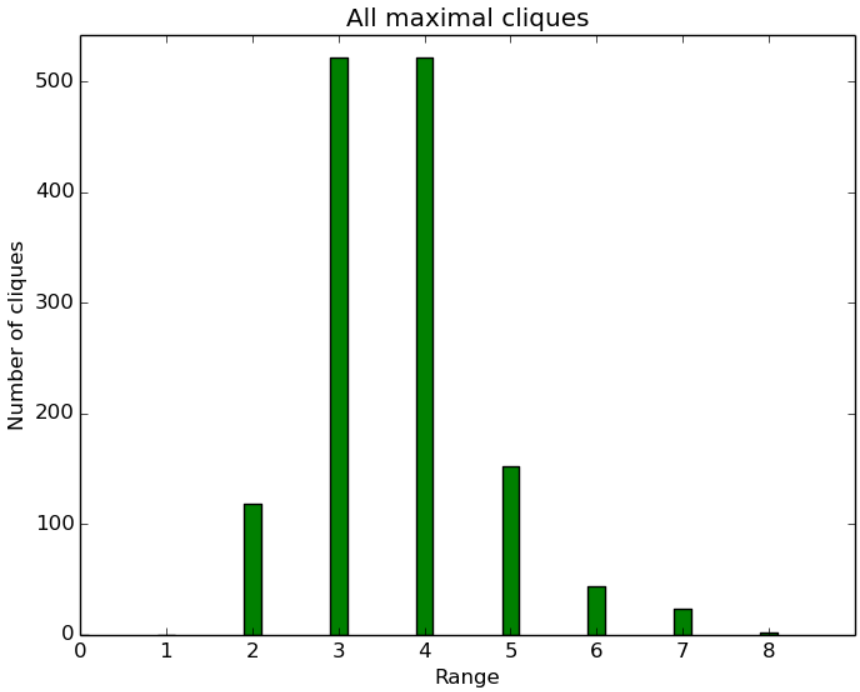
- bliskości:
  $[('n44', 0.30339720167657325),$
  $('n12', 0.05759006668189359),$
  $('n2', 0.04801549330577613),$
  $('n190', 0.044489436955661085),$
  $('n4', 0.031163449440266965)]$
- pośrednictwa:
  $[('n44', 0.597979797979798),$
  $('n12', 0.5220458553791887),$
  $('n2', 0.5165794066317626),$
  $('n86', 0.5085910652920962),$
  $('n3', 0.5008460236886633)]$
- rangi:
  $[('n44', 0.08673438527286116),$
  $('n2', 0.021427745128270354),$
  $('n12', 0.019912816778276513),$
  $('n172', 0.010832449312916758),$
  $('n86', 0.010682766687318981)]$

Dokonano wyznaczenie klik w rozpatrywanym grafie. Uzyskano następujące wartości, gdzie pierwsza wartość odpowiada rozmiarowi kliki, natomiast druga ilości klik danego rzędu:

- 0: 0
- 1: 0
- 2: 119
- 3: 522
- 4: 522
- 5: 153

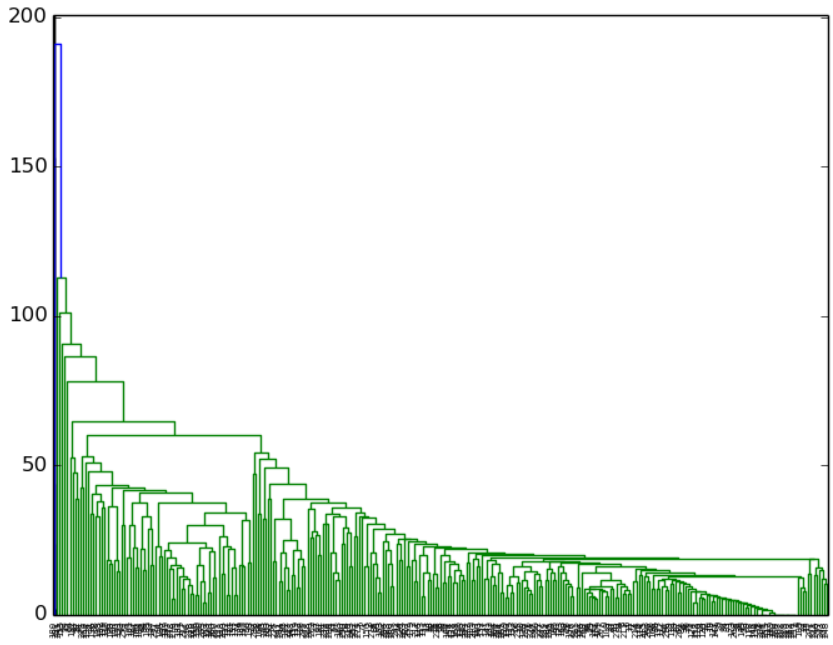
- 6: 44
- 7: 24
- 8: 2

Zobrazowano ten rozkład również na poniższym wykresie:



Rysunek 1: Rozkład liczby klik w zależności od rzędu.

Na koniec dokonano grupowania aglomeracyjnego metodą *comlete linkage* (średnica nowej grupy). Danymi wejściowymi do wykorzystanej metody jest macierz reprezentująca przekształcenie dla każdej pary wierzchołków wartości odległości pomiędzy każdą z par. Uzyskany dendrogram przedstawiono poniżej:



Rysunek 2: Dendrogram.

Arbitralnie podzieliłbym graf na linii odległości między wierzchołkami o wartości 60. Pozwoliłoby to na uzyskanie 7 grup.

3 Pajek

Wczytanie grafu w programu *Pajek* było możliwe dzięki wykorzystaniu skryptu napisanym w języku python, który umożliwia zapis grafu do formatu **.net**. Uzyskano następujące podsumowanie:

```
Number of vertices (n): 297
-----------------------------------------------------------
                                        Arcs          Edges
-----------------------------------------------------------
Number of lines with value=1            1023              0
Number of lines with value#1            1336              0
-----------------------------------------------------------
Total number of lines                   2359              0
-----------------------------------------------------------
Number of loops                            0              0
Number of multiple lines                  14              0
-----------------------------------------------------------

Density1 [loops allowed]     = 0.02674330
```

```
Density2 [no loops allowed] = 0.02683365

Average Degree = 15.88552189
```

Wyniki są zbieżne z wynikami uzyskanymi przez wczytanie grafu przy użyciu pakietu *networkx*, zarówno liczba łuków jak i średni stopień wierzchołka są z dokładnością do zaokrąglenie identyczne. W podsumowaniu zawarta została również informacja o zduplikowanych łukach, których występuje 14.

Usunięcie zduplikowanych łuków zostało zrealizowane poprzez transformację wczytanej sieci i usunięcie powielonych łuków. Operacje została przeprowadzona z ustawieniem dla pozostawionej krawędzi maksymalnej wartości, spośród łuków który były wielokrotne dla danej pary wierzchołków. Podstawowe informacje o grafie po tej operacji prezentują się następująco:

```
Number of vertices (n): 297
------------------------------------------------------------
                                        Arcs           Edges
------------------------------------------------------------
Number of lines with value=1            1013               0
Number of lines with value#1            1332               0
------------------------------------------------------------
Total number of lines                   2345               0
------------------------------------------------------------
Number of loops                            0               0
Number of multiple lines                   0               0
------------------------------------------------------------

Density1 [loops allowed]    = 0.02658459
Density2 [no loops allowed] = 0.02667440

Average Degree = 15.79124579
```

Następnie dokanano transformację sieci i przekształcenie łuków na krawędzie, tym samym uzyskując graf nieskierowany. Po tej operacji uległa zmianie liczba krawędzi, a także średni stopień wierzchołka:

```
Number of vertices (n): 297
------------------------------------------------------------
                                        Arcs           Edges
------------------------------------------------------------
Number of lines with value=1               0             878
Number of lines with value#1               0            1270
------------------------------------------------------------
Total number of lines                      0            2148
------------------------------------------------------------
Number of loops                            0               0
Number of multiple lines                   0               0
------------------------------------------------------------

Density1 [loops allowed]    = 0.04870251
Density2 [no loops allowed] = 0.04886705
```

```
Average Degree = 14.46464646
```

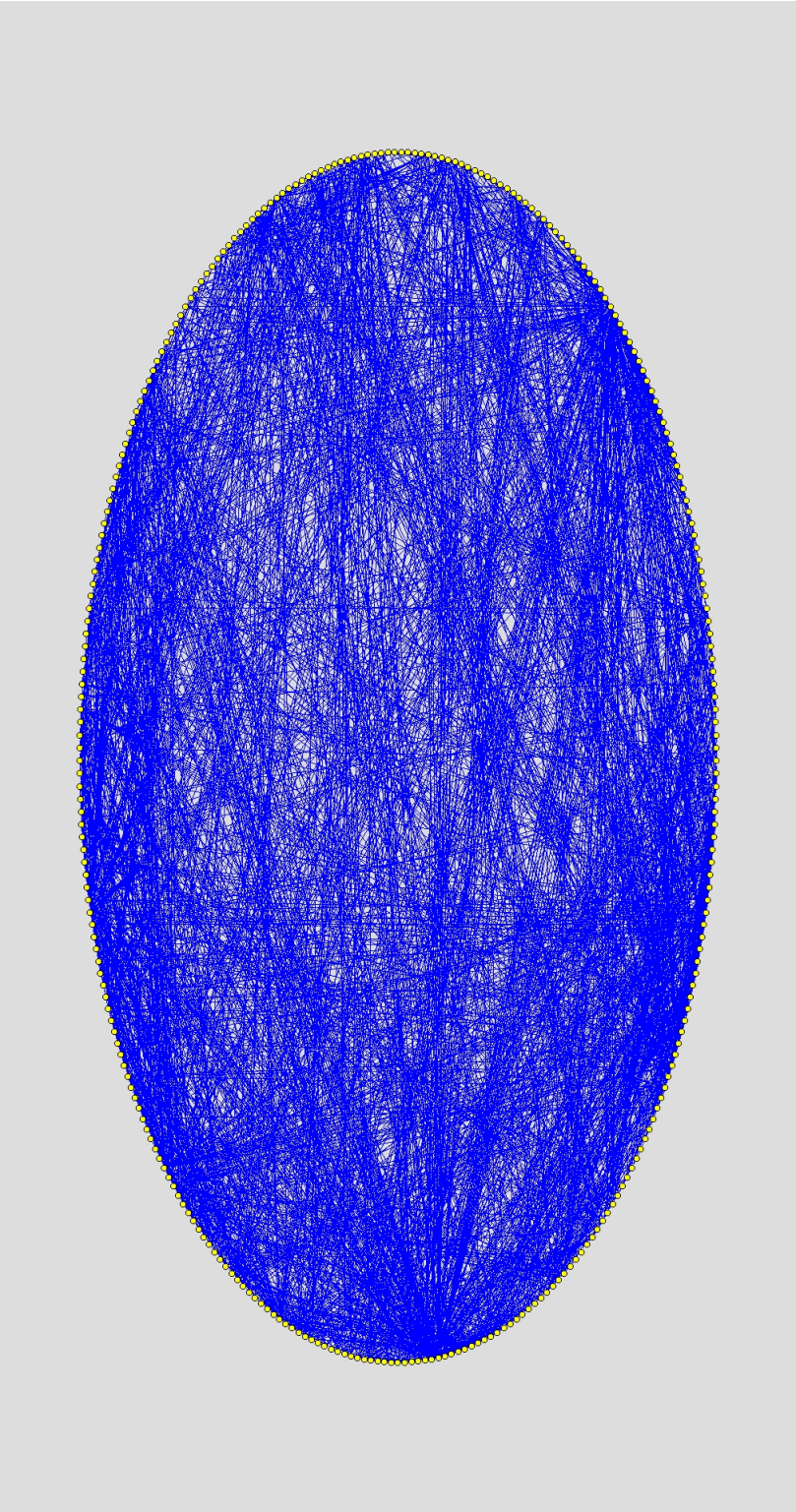
W dalszej kolejności została stworzona partycja, czyli grupa węzłów poprzez wyznaczenie składowych spójnych. W przypadku grafu nieskierowanego można wybrać zarówno opcję słabych jak i silnych składowych spójnych i tak uzyskany zostanie ten sam rezultat. W przypadku rozpatrywanego grafu istnieje jedna składowa spójna. Podsumowanie operacji przedstawiono poniżej:

```
================================================================================
3. Strong Components of N3 [>=1] (297, comp.=1)
================================================================================
Dimension: 297
The lowest value:  1
The highest value: 1

Frequency distribution of cluster values:

   Cluster      Freq     Freq%   CumFreq  CumFreq% Representative
 ------------------------------------------------------------------
         1       297  100.0000       297  100.0000 n38
 ------------------------------------------------------------------
       Sum       297  100.0000
```

Na koniec dokonano wykreślenie grafu, który przedstawia się następująco:



Rysunek 3: Wykreślony graf.

4 Porówanie czasów

Niestety dokładność pomiarów czasu wykonania operacji w przypadku programu *Pajek* są na tyle niedokładne, że w przypadku każdej operacji uzyskano wartość 0:00:00. W przypadku zadań wykonanych przy użyciu pakietu networkx czasy przedstawiają się następująco:

- usunięcie zduplikowanych krawędzi oraz przekształcenie na graf nieskierowany: 0.007080078125 sek,
- wyznaczenie składowych spójnych: 0.0146760940552 sek.

5 Załączniki

Do sprawzdania załączam skrypt w języku python umożliwiający uzyskanie przedstawionych rezultatów oraz dane na podstawie, których był wykonywany.