Gephi

Metody analizy i wizualizacji
dużych zbiorów danych

Niezbędne materiały

- Gephi: https://gephi.org/users/download/
- zbiory: https://nofile.io/f/3rtznfoqZ75/grafy.rar

Complex Networks

Sieć złożona (complex network) to graf o nietrywialnych własnościach topologicznych - własnościach, które często występują w grafach modelujących rzeczywiste systemy.

Przykładami sieci złożonych są np. WWW, sieci społecznościowe, sieci współpracy, sieci powiązań pomiędzy artykułami.

Complex networks

Typy sieci:

- **prosta** - krawędź istnieje lub nie,
- **skierowana** - krawędź ma kierunek (reprezentowane np. poprzez strzałki)
- **znakowa** - krawędź ma znak (+-)
- **ważona** - krawędź jest powiązane z wartością oznaczającą jej wagę
- **z cechami węzłów** - węzły mogą mieć wagę lub kolor

Complex networks: small-world

"A ***small-world network*** *is a type of mathematical graph in which most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a small number of hops or steps.*"

Six degrees of Kevin Bacon

Bacon Number	# of people
0	1
1	2769
2	305215
3	1021901
4	253177
5	20060
6	2033
7	297
8	25
9	7

Średni “Bacon number”: **2.994**

Tylko **329** z 1 605 485 aktorów miało “Bacon number” większy niż **7**.

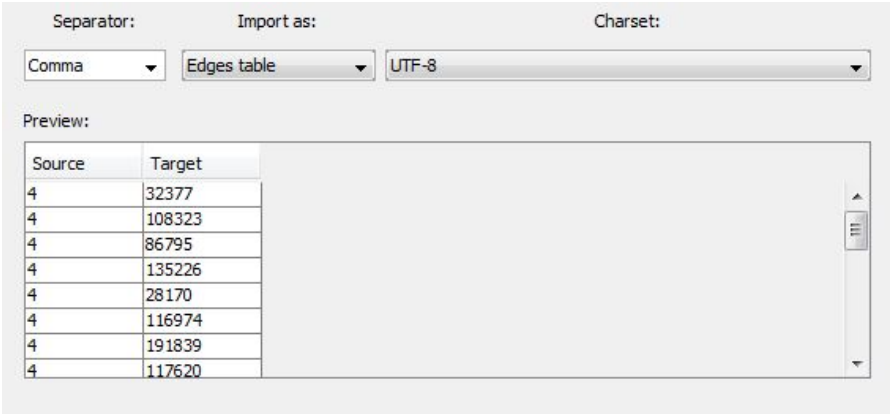
Complex networks: scale-free

"A **scale-free network** *is a connected graph or network with the property that the number of links* **k** *originating from a given node exhibits a power law distribution* $P(k) \sim k^{-\gamma}$"

"Gephi is a tool for data analysts and scientists keen to explore and understand graphs. Like Photoshop™ but for graph data, the user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden patterns. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research."

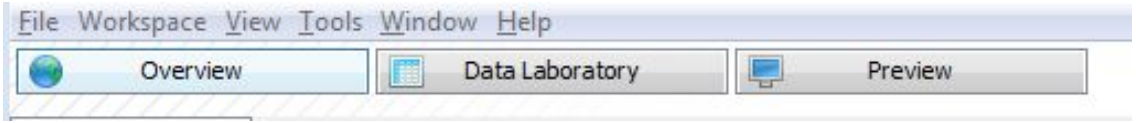
Budowanie grafu

1. Uruchom **Gephi**
2. ***File*** -> ***Import spreadsheet...*** -> ***art/edges_a.csv***

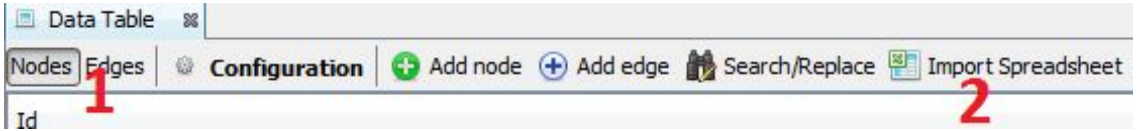


Budowanie grafu

3. *Data Laboratory*

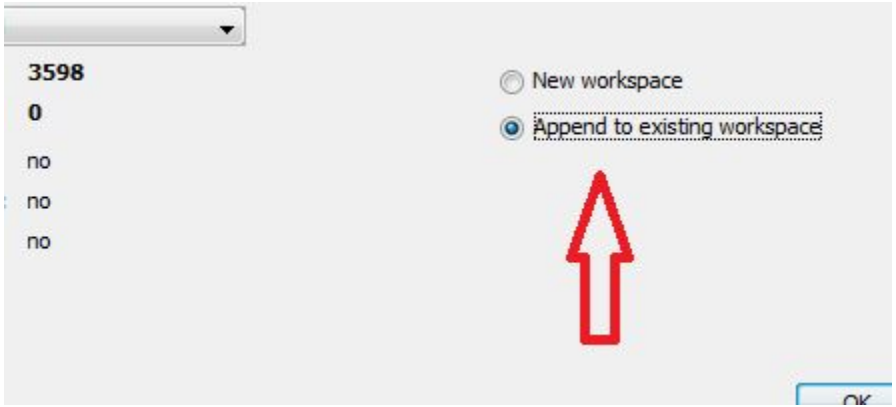


4. *Nodes -> Import Spreadsheet*



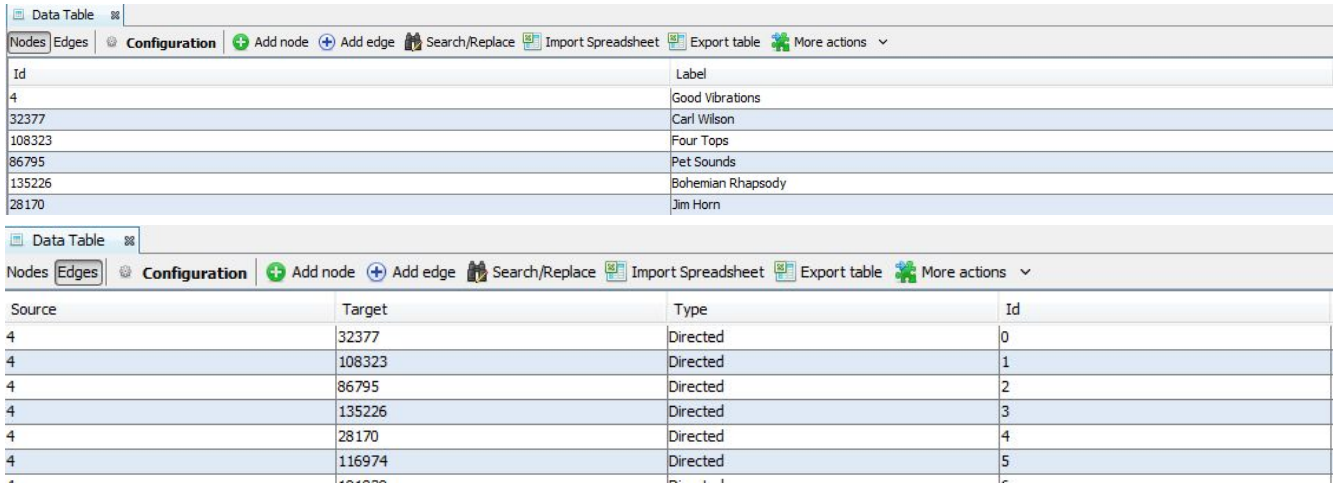
Budowanie grafu

5. Zaznacz opcję ***Append to existing workspace***



Budowanie grafu

6. Upewnij się, że otrzymałeś poniższą strukturę



Data Table

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions

Id	Label
4	Good Vibrations
32377	Carl Wilson
108323	Four Tops
86795	Pet Sounds
135226	Bohemian Rhapsody
28170	Jim Horn

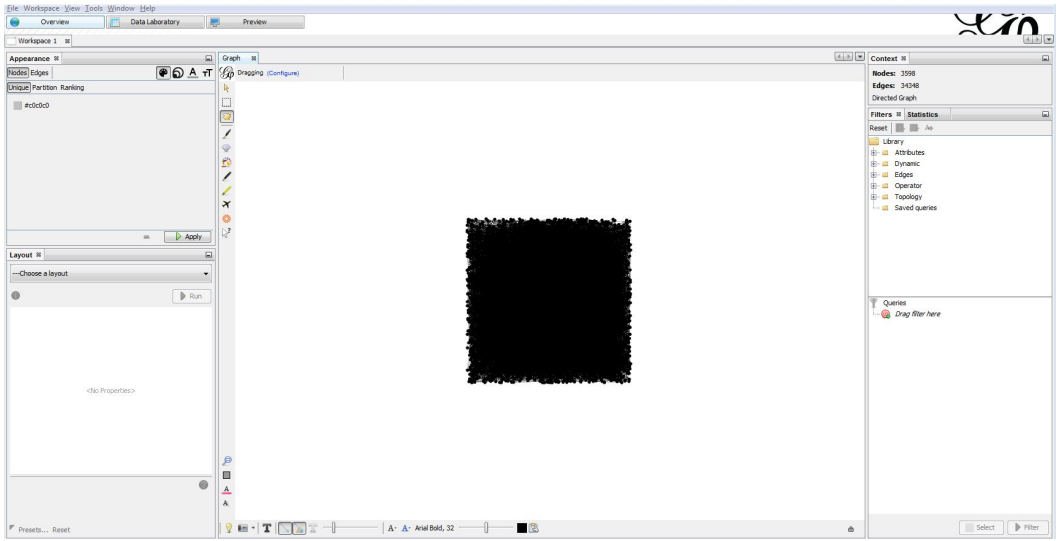
Data Table

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions

Source	Target	Type	Id
4	32377	Directed	0
4	108323	Directed	1
4	86795	Directed	2
4	135226	Directed	3
4	28170	Directed	4
4	116974	Directed	5

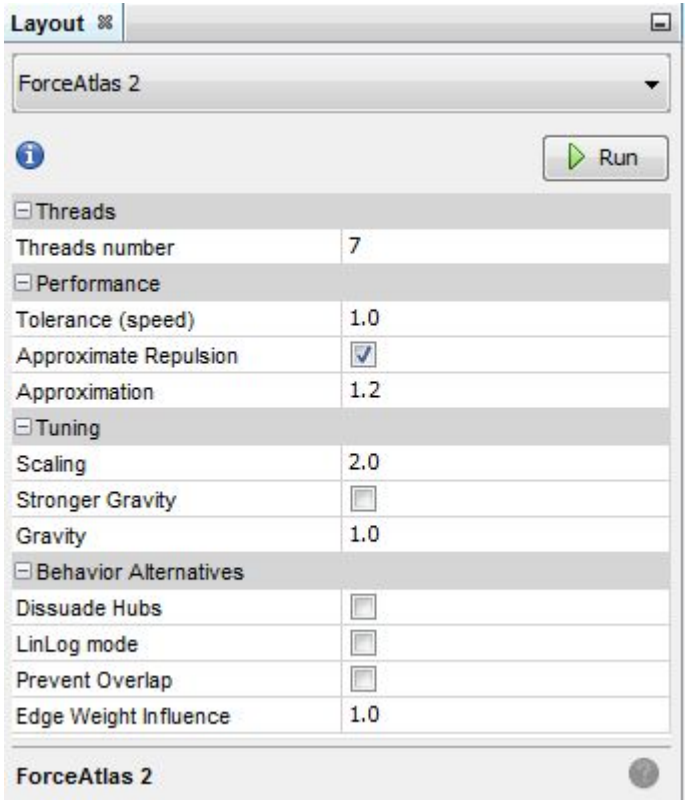
Budowanie grafu

7. Wróć do widoku **Overview**

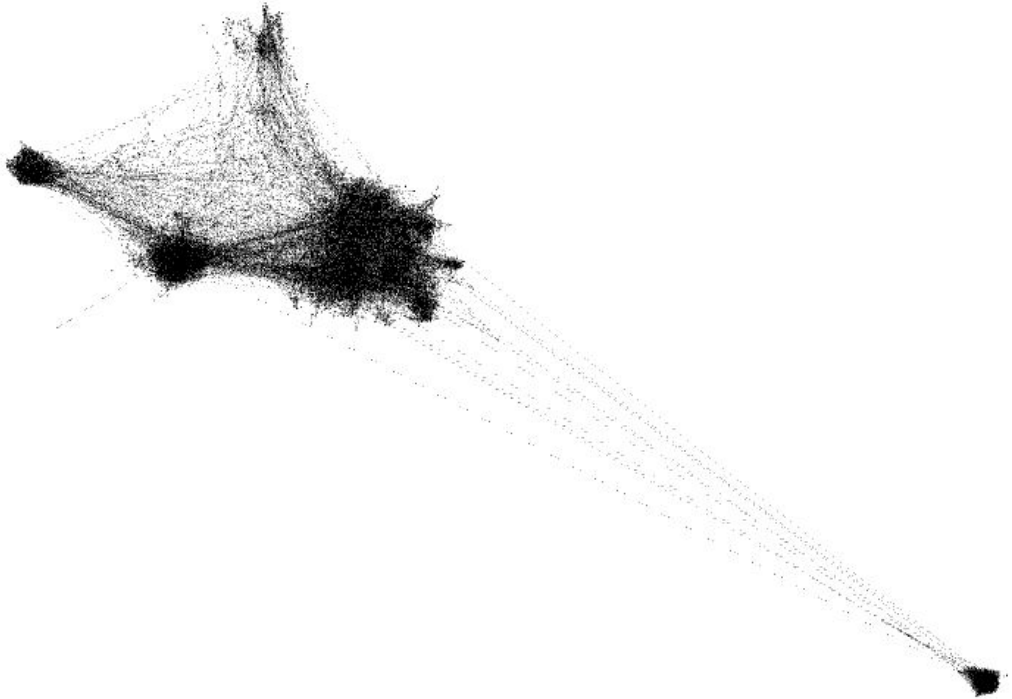


Layout grafu

1. Po lewej stronie znajdziesz pole ***Layout***
2. Wybierz opcję ***ForceAtlas 2***
3. Naciśnij ***Run***
4. Zaobserwuj zmianę wizualizacji grafu



ForceAtlas 2



ForceAtlas 2

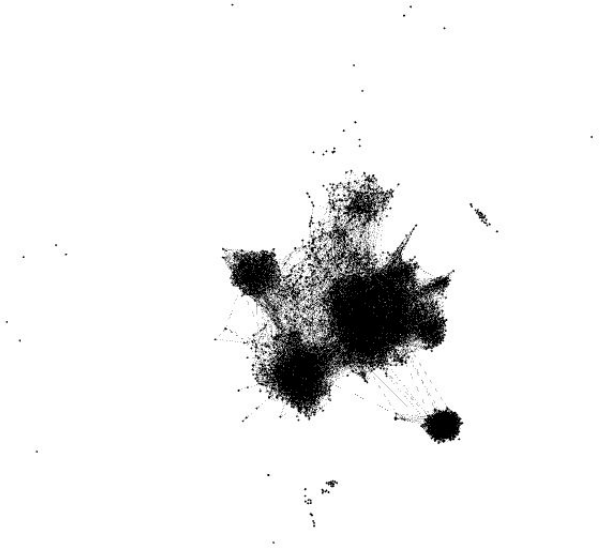
“ForceAtlas2 is a force directed layout: it simulates a physical system in order to spatialize a network. Nodes repulse each other like charged particles, while edges attract their nodes, like springs. These forces create a movement that converges to a balanced state. This final configuration is expected to help the interpretation of the data.”

ForceAtlas 2

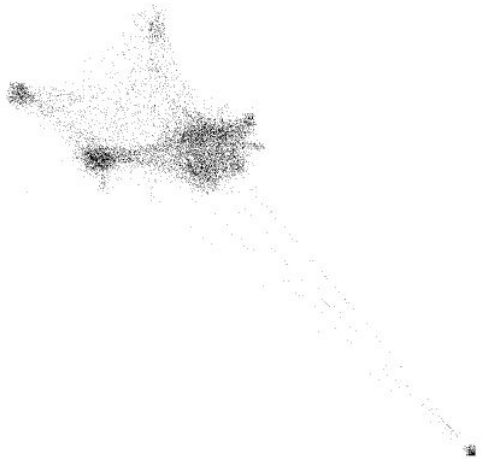
- “*Its very essence is to turn structural proximities into visual proximities*”,
- użyteczne w analizie *social networks* i *communities*,
- rezultat jest niedeterministyczny, a współrzędne węzła nie odzwierciedlają żadnej konkretnej wartości
- pozycja węzła jest możliwa do zinterpretowania jedynie w porównaniu do innych węzłów.

ForceAtlas 2

Gravity = 100.0

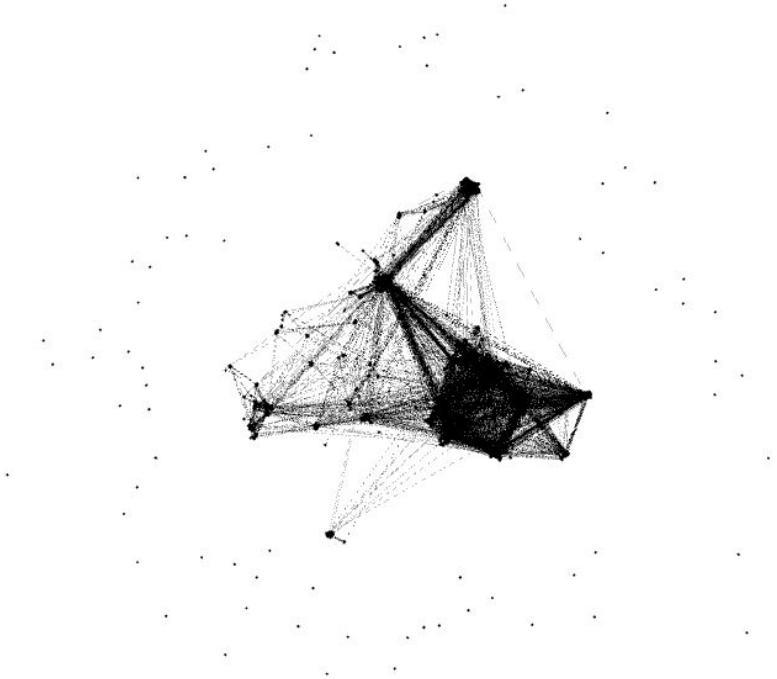


Scaling = 30.0

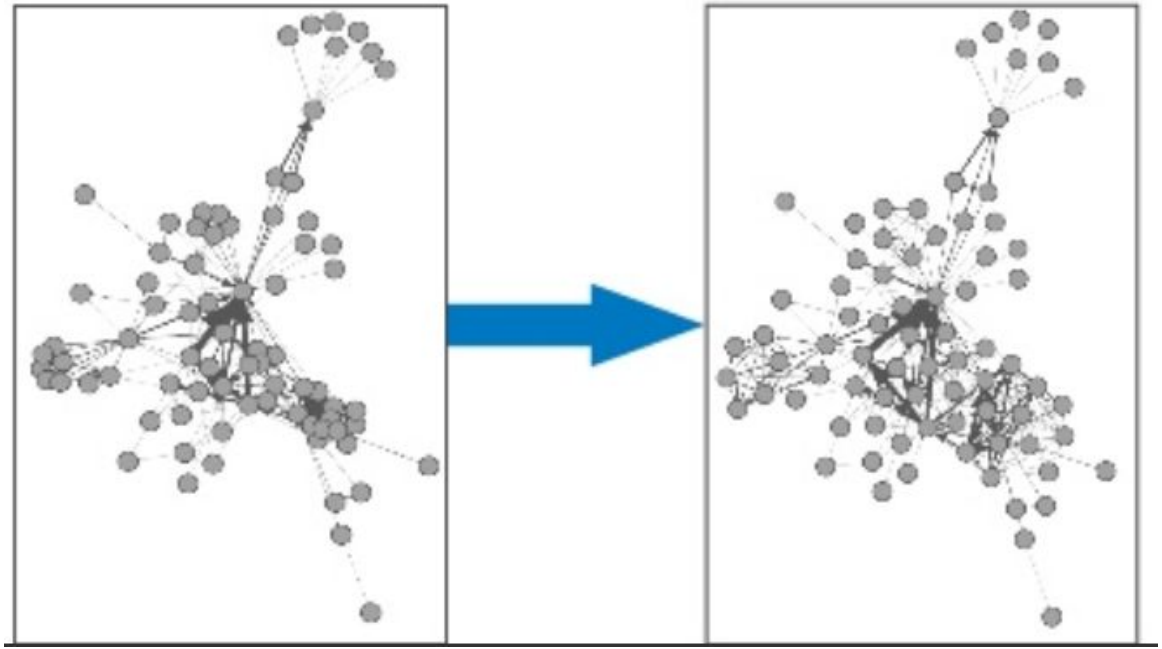


OpenOrd

Celuje w lepsze rozpoznanie klastrów. Może być uruchamiany równoległe w celu przyspieszenia obliczeń. Bazuje na algorytmie Frutchermana-Reingolda, pracuje z ustaloną liczbą iteracji i korzysta z symulowanego wyżarzania.



Nonoverlap



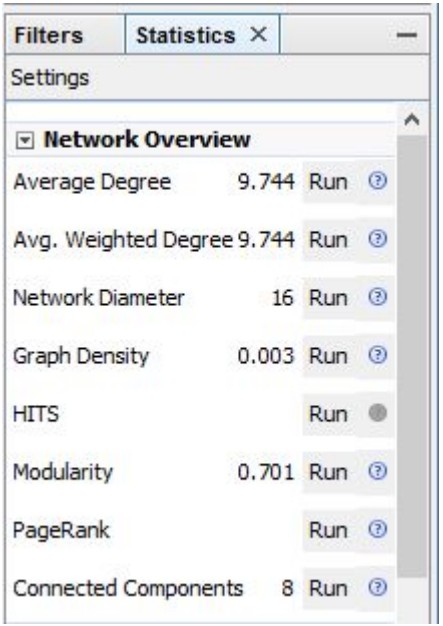
Deskryptory grafu

Te prostsze:

- Average Degree
- Average Weighted Degree
- Network Diameter
- Graph Density

I te bardziej zaawansowane:

- HITS
- Modularity
- PageRank
- Connected Components



Deskryptory: Average Degree

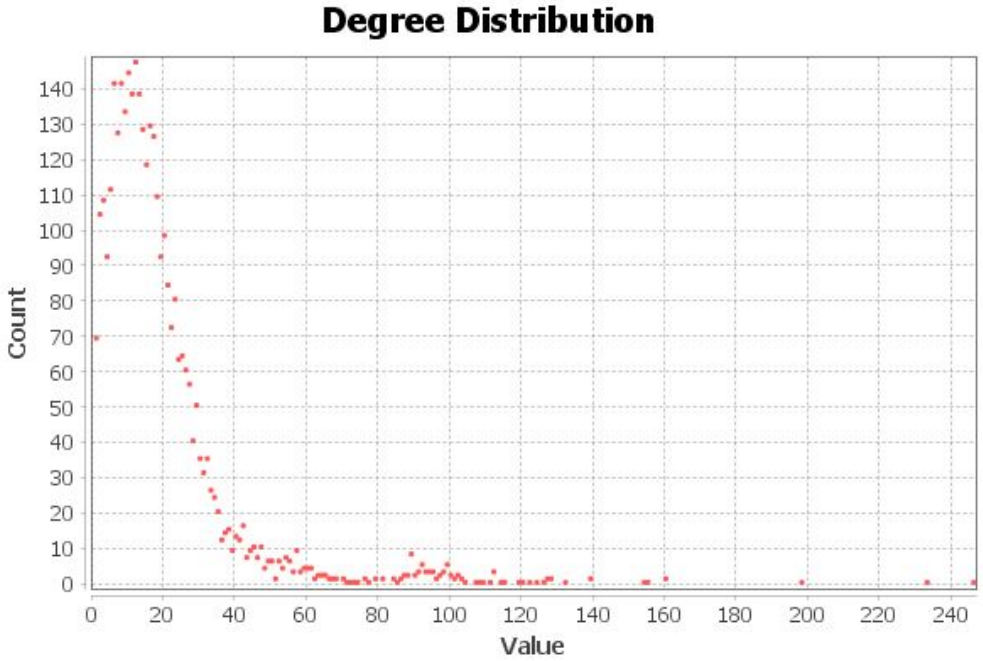
Co dostajemy:

- średnią wartość stopnia wierzchołka grafu
- trzy wykresy: Degree Distribution, In-Degree Distribution, Out-Degree Distribution

Degree Report

Results:

Average Degree: 9.744



Deskryptory: Network Diameter

Dwie możliwości rozpatrywania grafu:

- directed
- undirected

Graph Distance Report

Parameters:

Network Interpretation: directed

Results:

Diameter: 16
Radius: 0
Average Path length: 5.420045517483211

Graph Distance Report

Parameters:

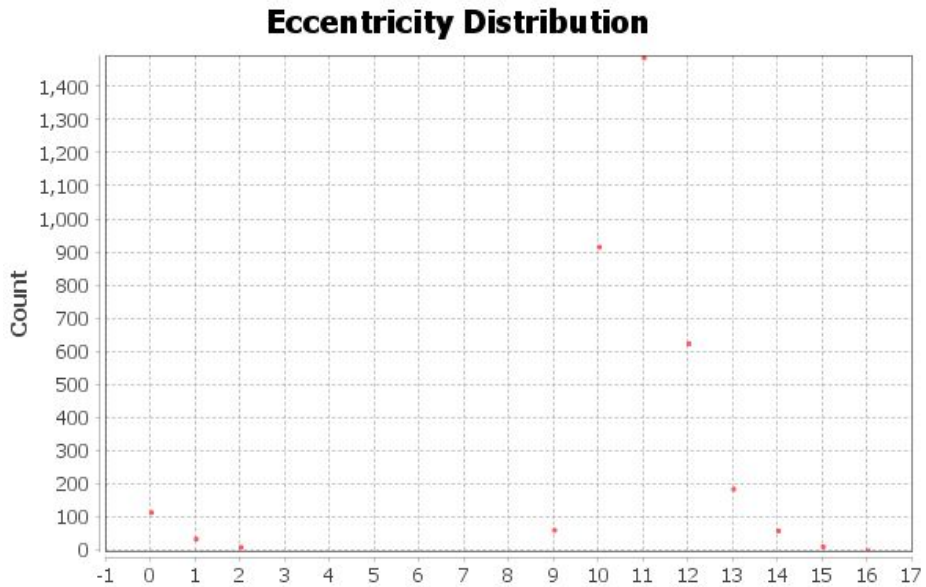
Network Interpretation: undirected

Results:

Diameter: 11
Radius: 1
Average Path length: 3.9862296999648805

Deskryptory: Network Diameter c.d.

Oprócz wcześniej pokazanych wartości otrzymujemy wykresy, które są opisane jako:



Betweenness Centrality: Measures how often a node appears on shortest paths between nodes in the network.

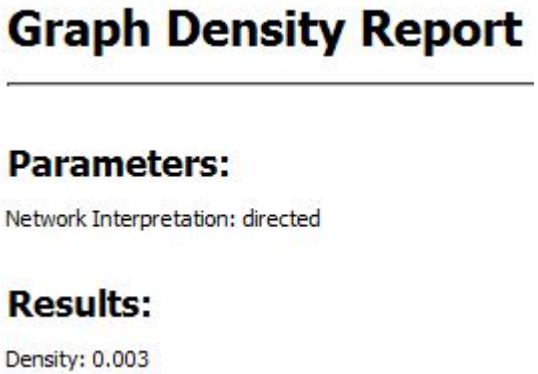
Closeness Centrality: The average distance from a given starting node to all other nodes in the network.

Eccentricity: The distance from a given starting node to the farthest node from it in the network.

Deskryptory: Graph Density

Gęstość grafu - stosunek liczby krawędzi do największej możliwej liczby krawędzi

Również dostępne są pomiary dla dwóch wersji grafu: skierowanego i nieskierowanego.



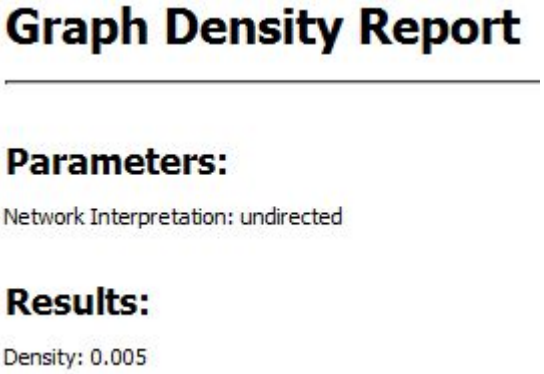
Graph Density Report

Parameters:

Network Interpretation: directed

Results:

Density: 0.003



Graph Density Report

Parameters:

Network Interpretation: undirected

Results:

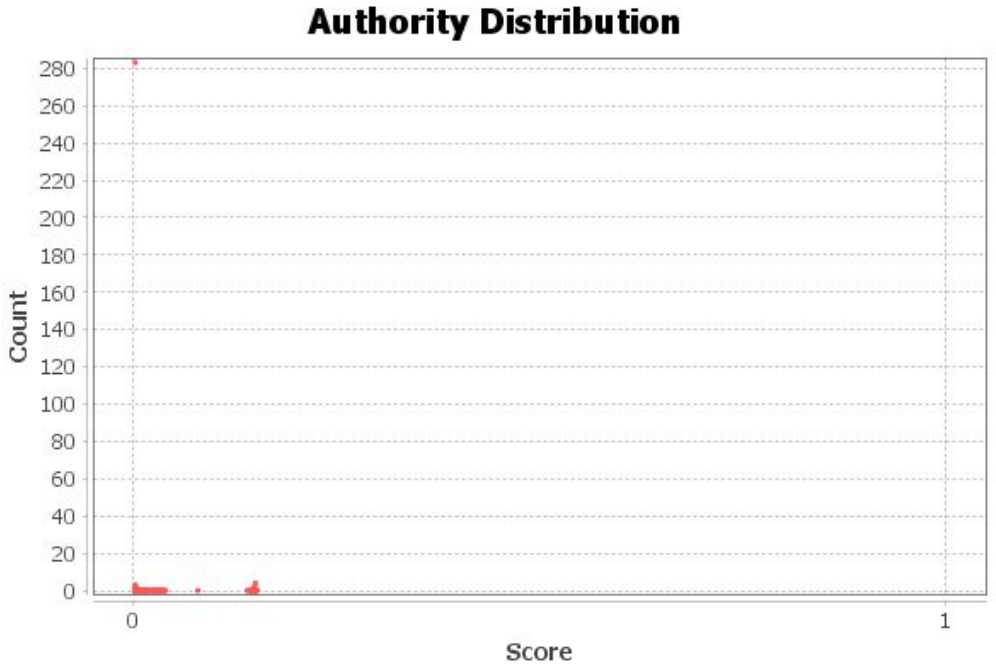
Density: 0.005

Deskryptory: HITS

HITS - Hypertext Induced Topic Selection

Authority - dokument cytowany na który wskazują inne dokumenty

Hub - dokument cytujący wskazujący na inne dokumenty



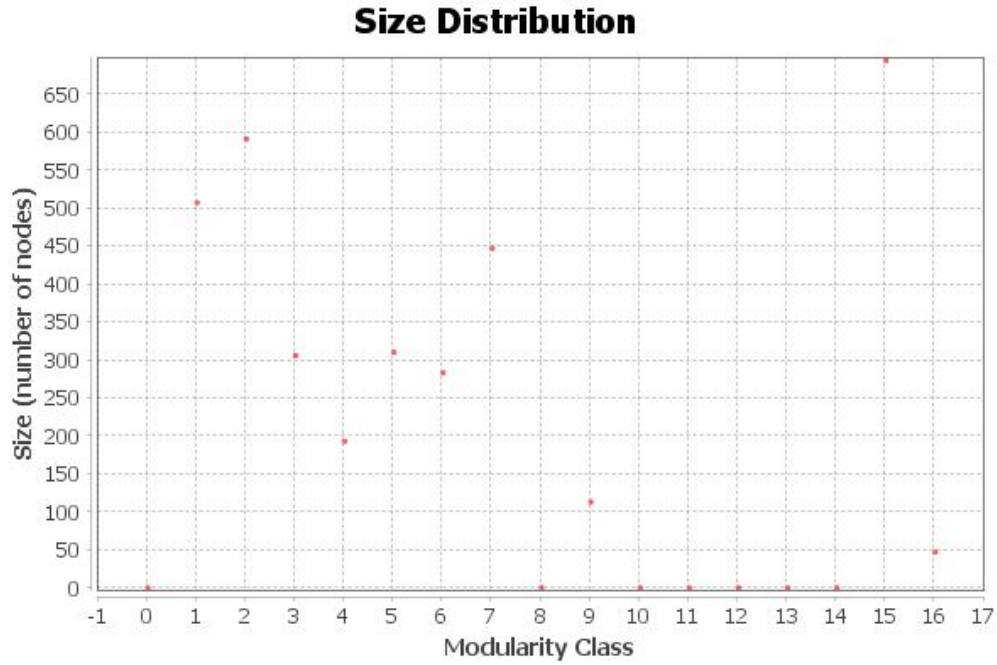
Deskryptory: Modularity

Możliwość wyznaczenia pod-sieci (communities), które mogą odpowiadać konkretnej grupie węzłów w świecie rzeczywistym, który jest modelowany grafem.

Ilość wyznaczonych grup regulowana parametrem *resolution*.

Results:

Modularity: 0.715
Modularity with resolution: 0.715
Number of Communities: 17

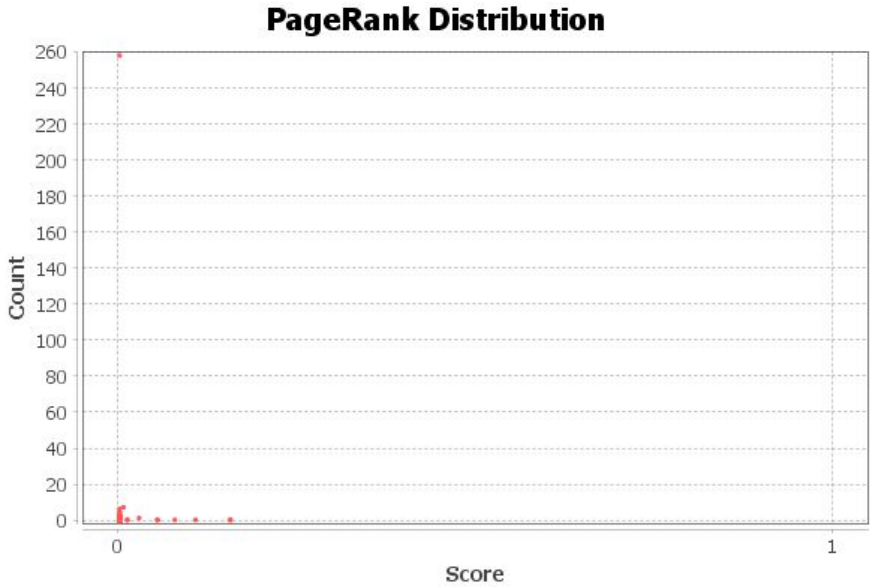


Deskryptory: PageRank

Parameters:

Epsilon = 0.001
Probability = 1.0

Results:



Metoda dawniej stosowana w wyszukiwarce Google.

Celem jest ocena jakości stron internetowych - jak często strona jest linkowana przez inne.

W Gephi:

PageRank

Ranks nodes "pages" according to how often a user following links will non-randomly reach the node "page".

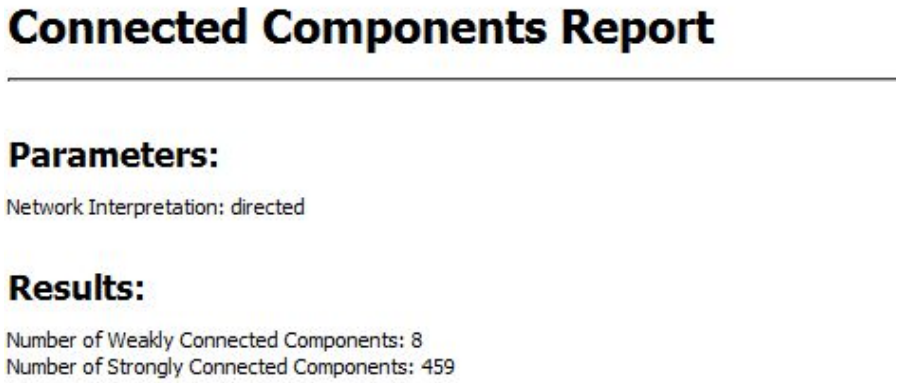
Deskryptory: Connected Components

Dostępne opcje:

- directed - detekcja strongly & weakly connected components
- undirected - detekcja weakly connected components

Czym są connected components? - http://historicaldataninjas.com/social-network-analysis-for-dummies/

Strongly connected components are groups of nodes in which the nodes can all be reached through directed edges. There are also weakly connected components where the direction of the edges is not taken into consideration, so each node can be reached through any kind of edge.



Connected Components Report

Parameters:

Network Interpretation: directed

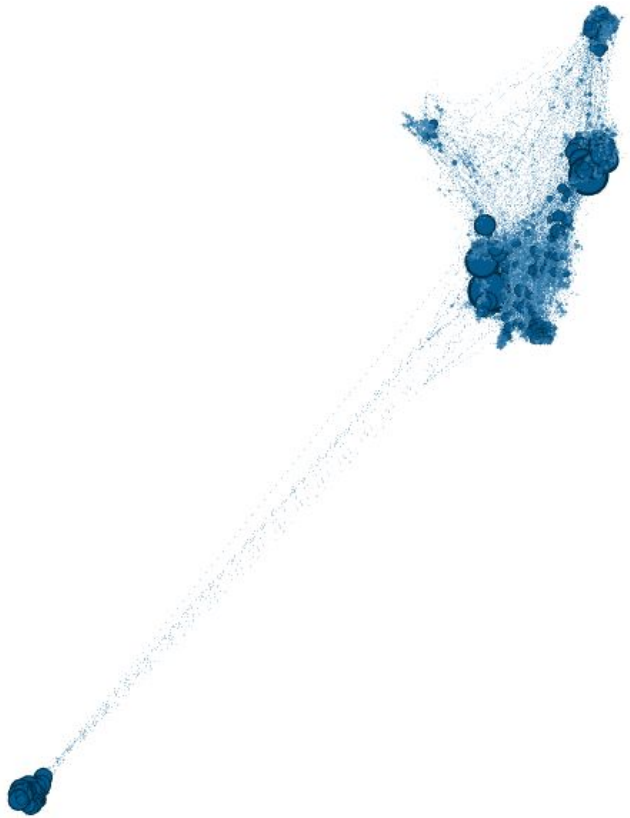
Results:

Number of Weakly Connected Components: 8
Number of Strongly Connected Components: 459

Wygląd grafu

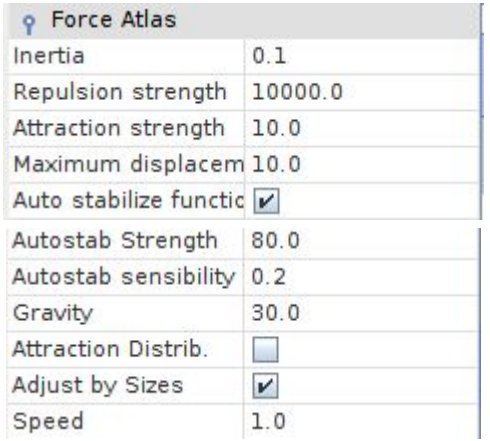
1. Po lewej stronie znajdziesz pole ***Appearance***
2. Z ikon po prawej stronie pola wybierz ***Color***
3. Wybierz ***Nodes*** -> ***Ranking*** -> ***Degree***
4. Ustal kolor, np. ***niebieski***
5. Z ikon po prawej stronie pola wybierz ***Size***
6. Wybierz ***Nodes*** -> ***Ranking*** -> ***Degree***
7. Ustal wartości, np. ***Min size: 1, Max size: 200***

Wygląd grafu



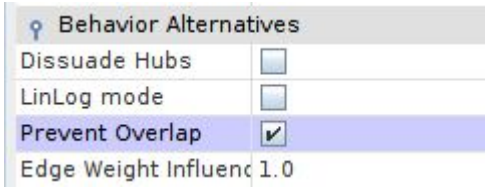
Separacja wierzchołków

Używając layoutu Force Atlas można uzyskać separację wierzchołków ustawiając odpowiednio opcje: Repulsion strength (np. na 10 000) oraz zaznaczając pole Adjust by Sizes w panelu Layout.



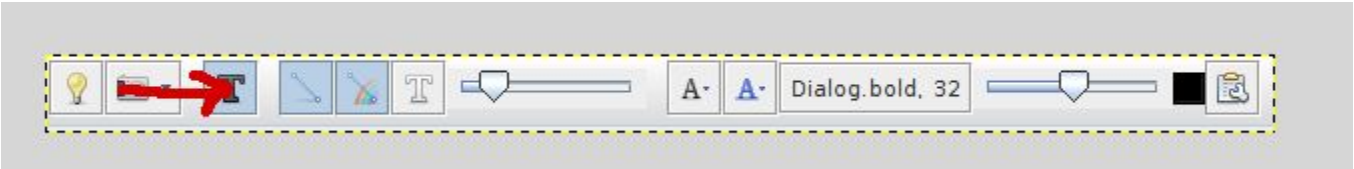
Separacja wierzchołków

Z kolei aby wykonać separację wierzchołków w layoucie Force Atlas 2
należy w panelu Layout w części Behaviour Alternatives zaznaczyć pole Prevent Overlap.

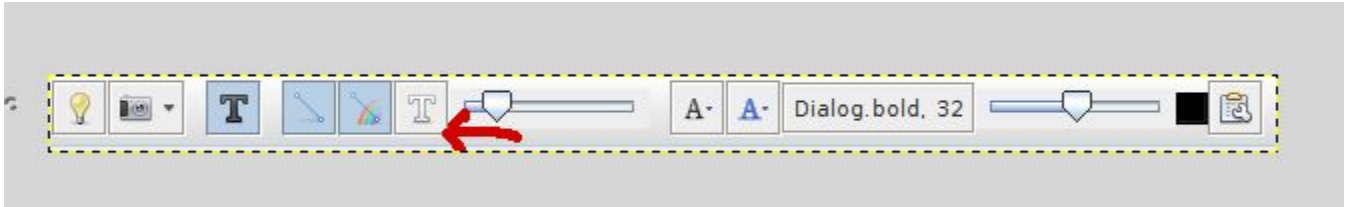


Etykiety

Aby pokazać na grafie etykiety wierzchołków należy włączyć opcję Show Node Labels znajdującą się na samym dole okna programu:



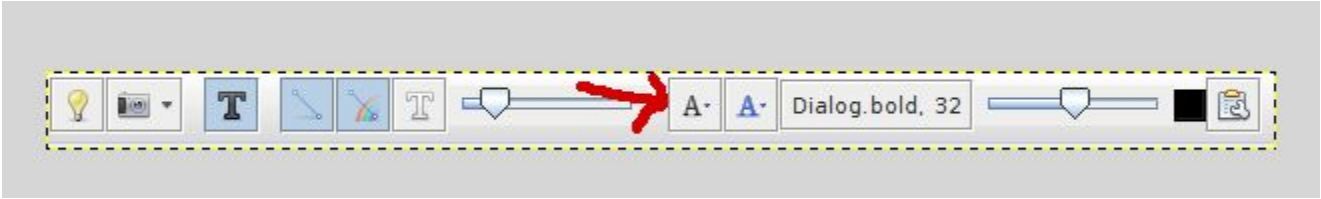
Z kolei aby włączyć etykiety dla krawędzi trzeba zaznaczyć pole Show Edge Labels:



Etykiety

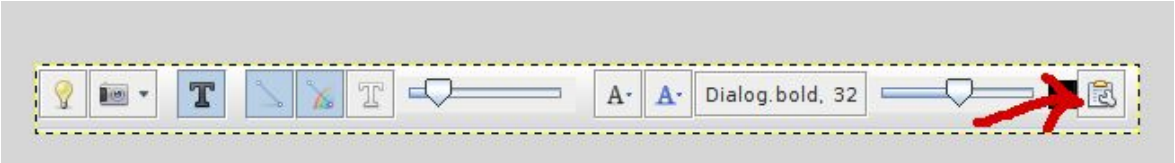
Suwak przy literze T pozwala na dostosowanie wielkości etykiety do wagi krawędzi.

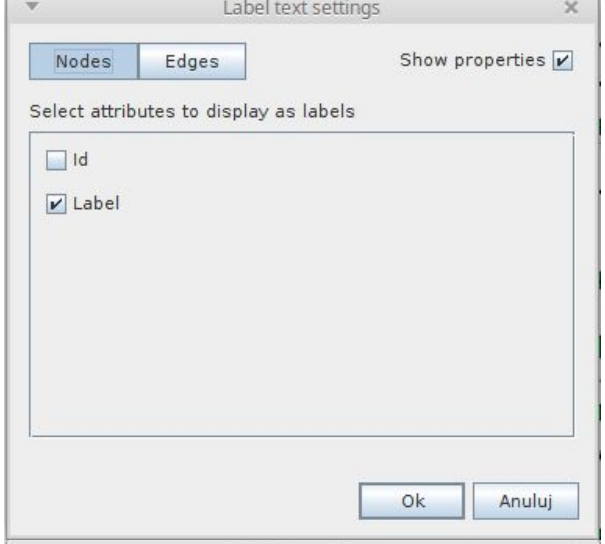
Z kolei zaznaczony element umożliwia dostosowanie rozmiaru etykiet dla wierzchołków, dostarczając między innymi ustalonego rozmiaru czy rozmiaru zależnego od wielkości wierzchołka.

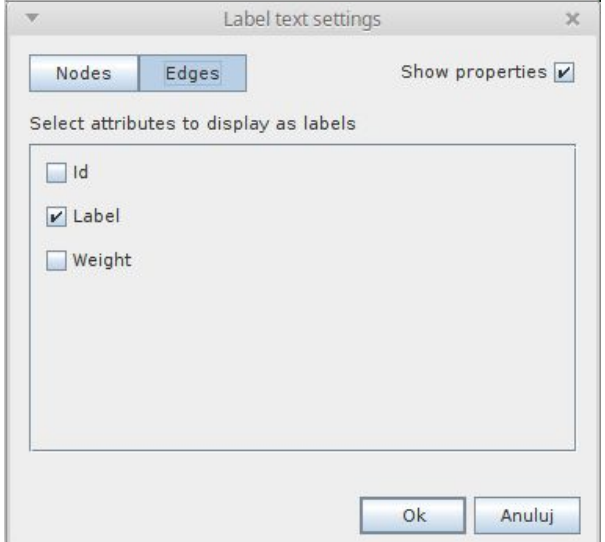


Etykiety

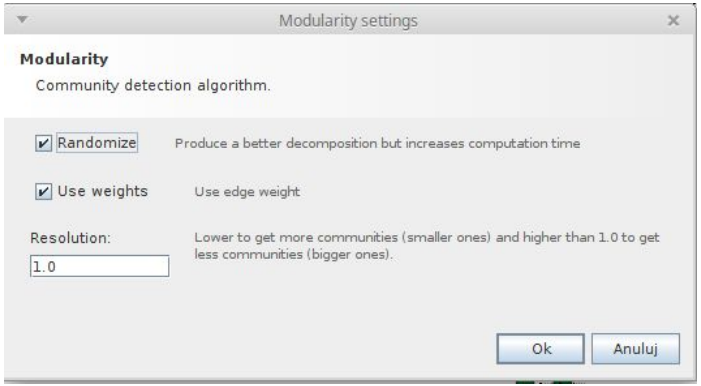
Zaznaczony element umożliwia wyspecyfikowanie
jakie atrybuty dla wierzchołków (Id, Label)
i krawędzi (Id, Label, Weight)
powinny zostać uwzględnione w wizualizacji.







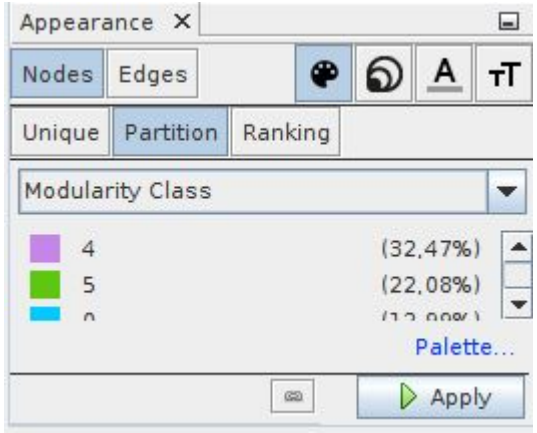
Klastry

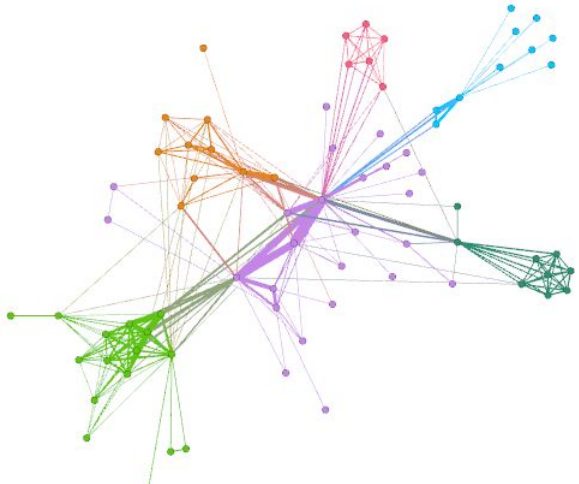


Aby zobaczyć jak wierzchołki tworzą klastry, czyli grupy o podobnych cechach należy w zakładce Statistics wybrać opcję Modularity, którą należy uruchomić wciskając Run. Następnie wybierając Randomize uruchamiamy tworzenie klastrów.

Klastry

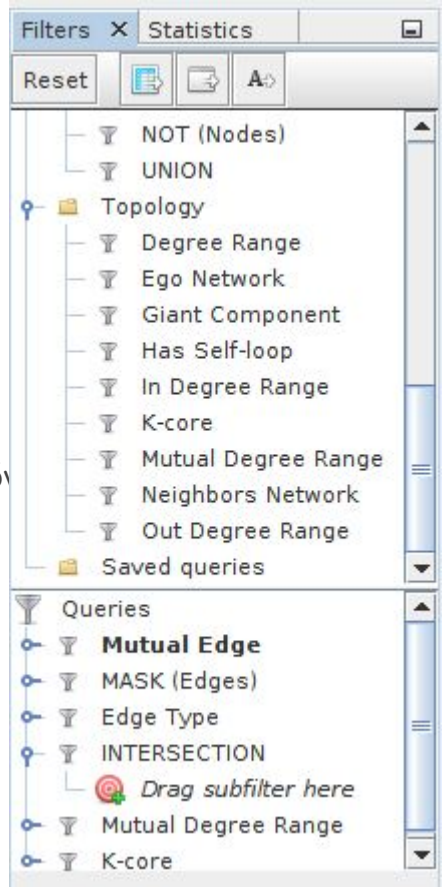
W zakładce Partition w lewej strony pojawiła się opcja Modularity Class wraz z kolorami dla kolejnych klastrów (communities): Wciskając Apply włączamy kolorowanie dla grafu. Na mniejszym grafie (Nędznicy) uzyskujemy następujący efekt:





Filtry

W zakładce Filters wybieramy konkretny filtr, który chcemy zastosować na grafie aby uzyskać podgraf o określonych cechach wspólnych. Z katalogu Topology wskazujemy filtr Degree Range, który pozwala na wyfiltrowanie tylko tych wierzchołków których stopień zawiera się w podanym zakresie. Filtr Ego Network pozwala wyszukać sąsiadów podanego wierzchołka do określonej głębokości sąsiedztwa (1, 2, 3, MAX). Filtr K-Core pozwala na znalezienie podgrafu, w którym wszystkie wierzchołki mają stopień równy przynajmniej k. Istnieją też filtry In Degree Range i Out Degree Range, którą wyszukują wierzchołki o stopniu z zadanego zakresu odpowiednio dla krawędzi wchodzących i wychodzących. Ponadto w innych katalogach znajdują się filtry dla krawędzi, czy dla operatorów możliwych do zastosowania na grafie (np. INTERSECTION).

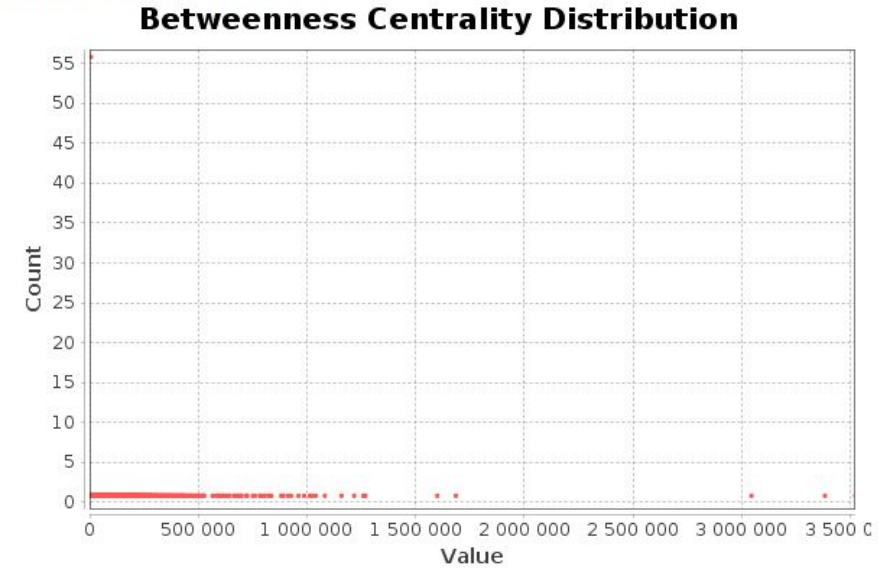


Wizualizacja zbioru MNIST w Gephi

Zastosowaliśmy layout Force Atlas 2 do wizualizacji zbioru MNIST dla k=20 najbliższych sąsiadów z opcją nie zachodzenia na siebie wierzchołków. Zbiór posiada 10 000 wierzchołków i 200 000 krawędzi. Deskryptor Average Degree dla otrzymanego grafu wynosi 20. Analiza grafu:

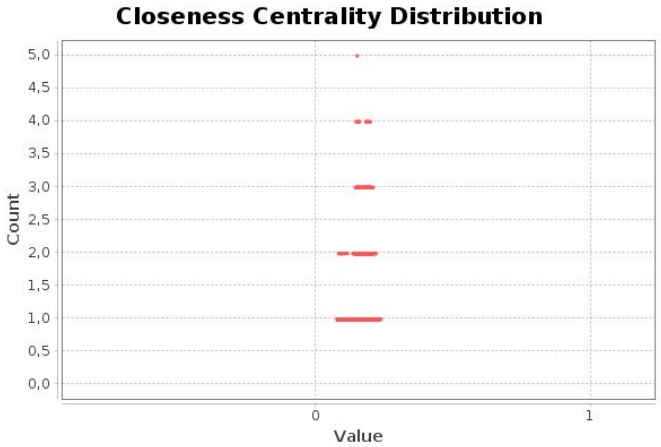
Results:

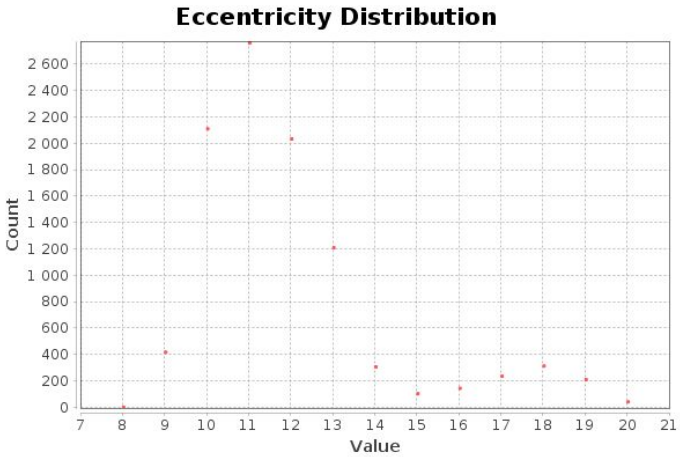
Diameter: 20
Radius: 8
Average Path length: 6.542103577518257



Zbiór mnist20knn:
https://drive.google.com/file/d/115wUPt6DOcg-lHmhfB0lpgpAxV6oUlcs/view?usp=sharing

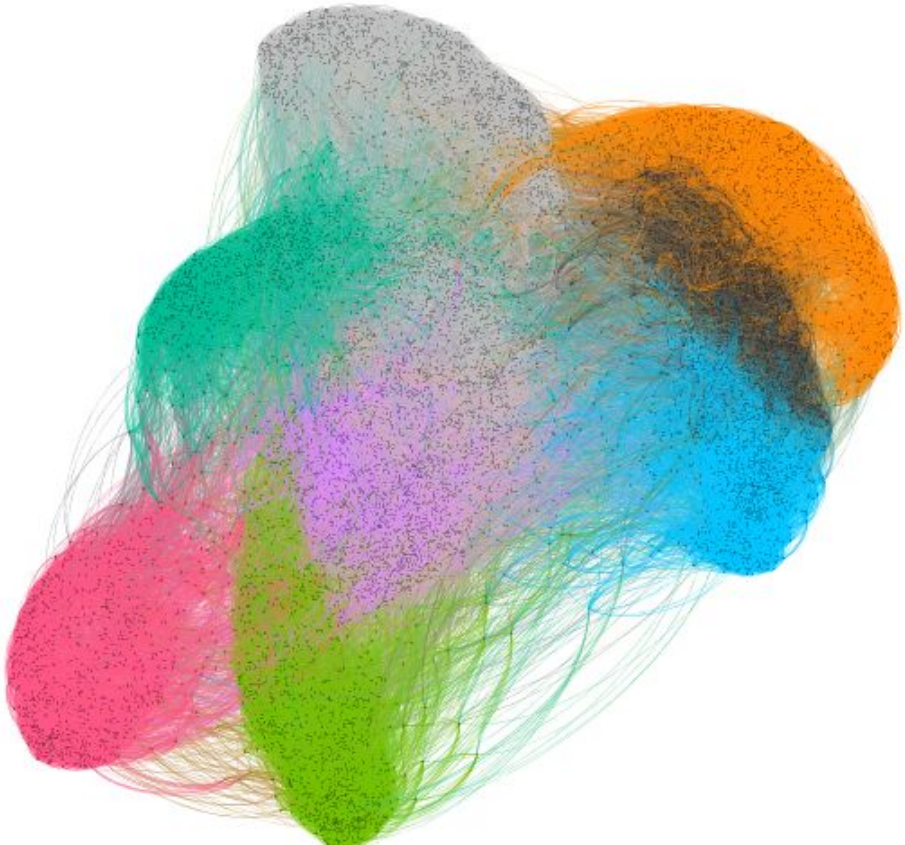
Wizualizacja zbioru MNIST w Gephi



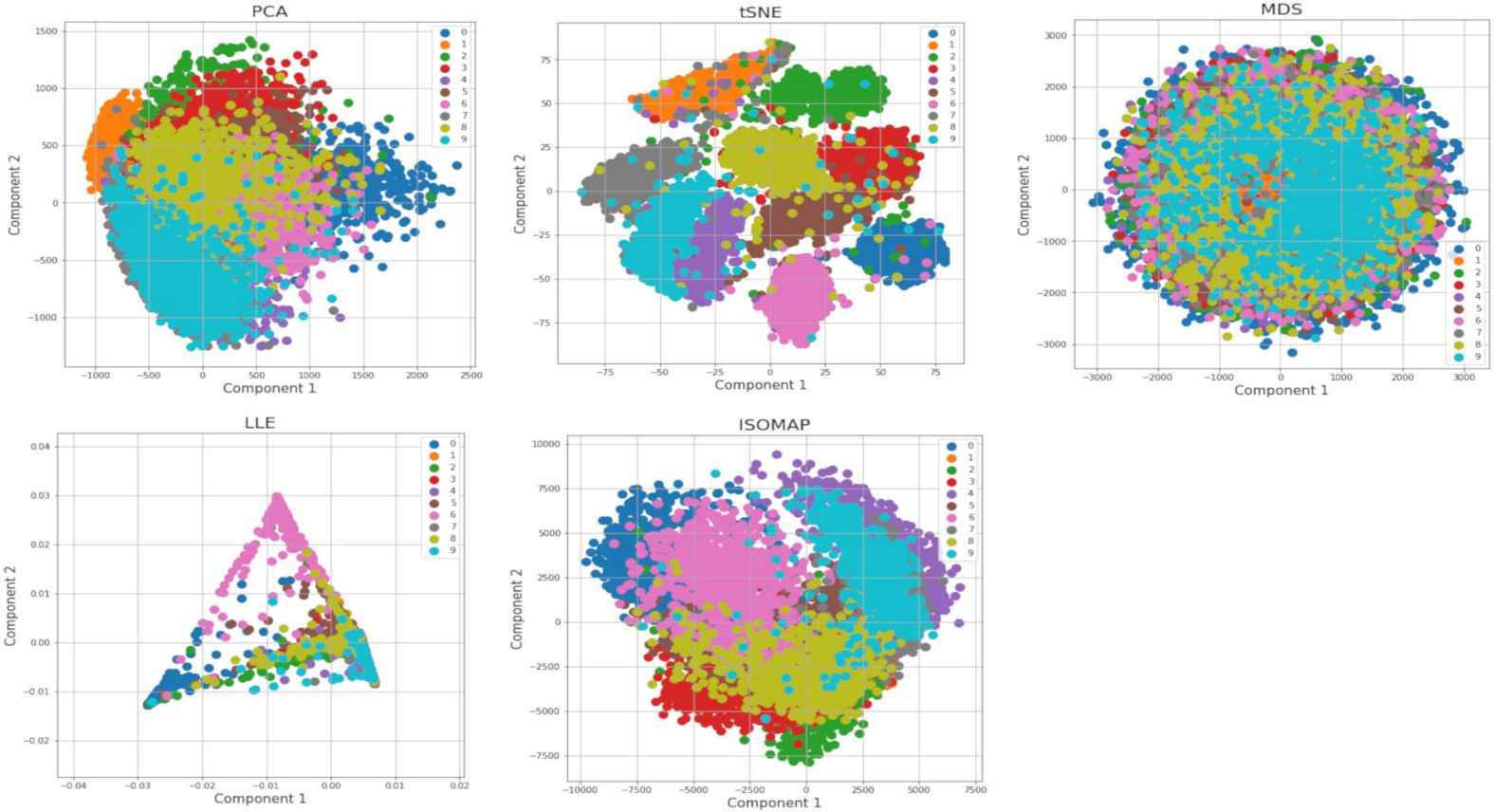


Dla uzyskania szybszej wizualizacji ustawiliśmy rozmiar wierzchołków bazując na atrybucie Degree na zakresie [4, 8], przez zaznaczenie tego w oknie Nodes->Ikona Size->Ranking->Degree.

Wizualizacja zbioru MNIST w Gephi



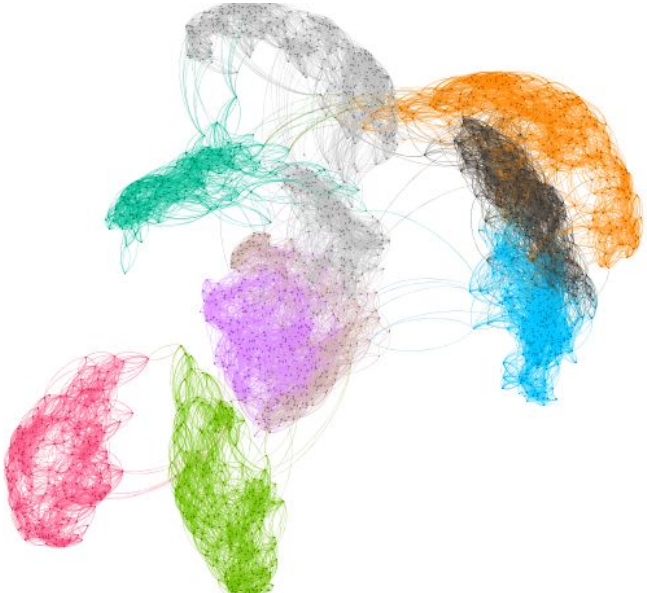
Wizualizacja zbioru MNIST

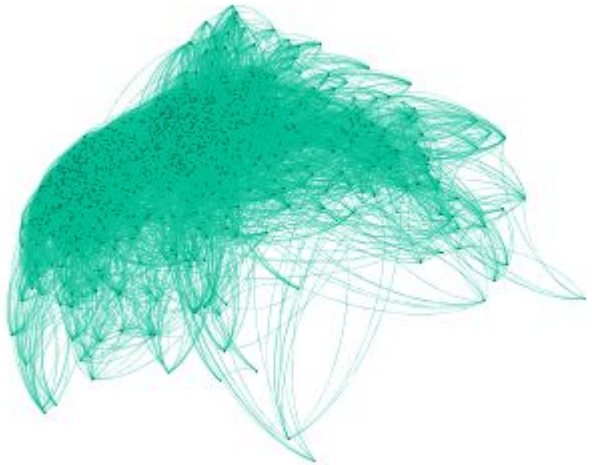


Porównanie wizualizacji zbioru MNIST różnymi metodami

Wizualizacja zbioru MNIST w Gephi

Zastosowanie filtru Degree Range ([47-96]), Filtr Attributes->Equal->Modularity Class (==5)



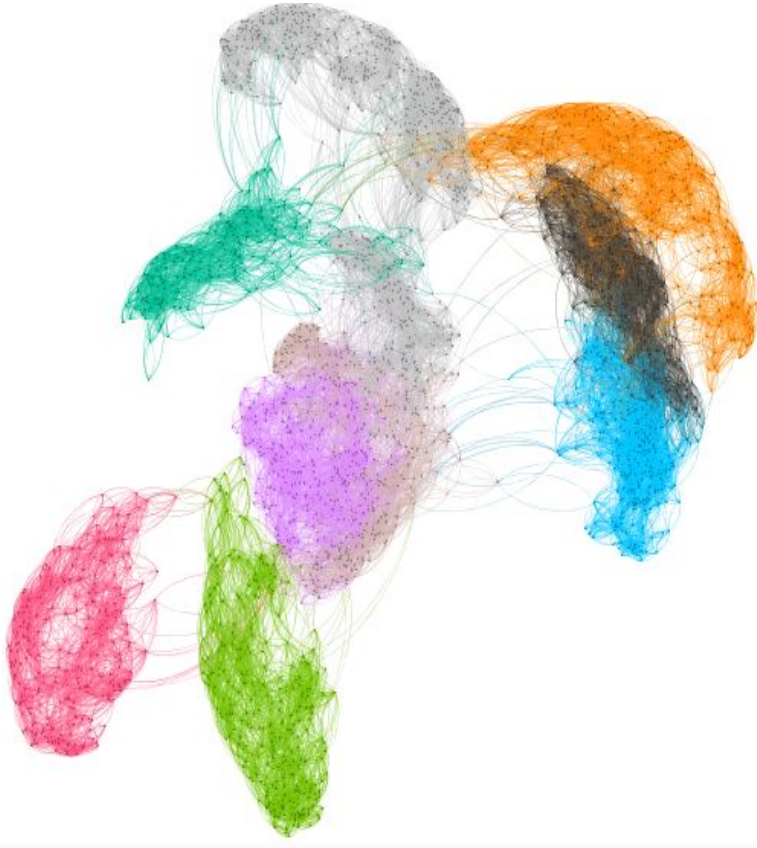


Wizualizacja zbioru MNIST w Gephi

Filtr Topology->Ego Network (pokazuje podgraf
wychodzący z wierzchołka o określonym ID
i do zadanej głębokości ścieżki).
Dla NodeID = 1920 i Depth = 3:

Wizualizacja zbioru MNIST w Gephi

Użycie filtru In-Degree Range (zakładka Topology)
z zakresu [25-76]:



Zadania

1. Wykonaj te same operacje na grafie ***history*** - import grafu, layout, deskryptory, wygląd grafu, klastry, filtry.
2. Zbuduj graf knn ze zbioru 20newsgroup, przeprowadź podobną analizę i porównaj wizualizację za pomocą t-SNE.