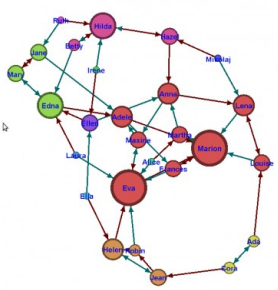
Analiza sieci społecznościowych w narzędziu Gephi

Celem pierwszej części ćwiczenia jest zapoznanie się z narzędziem Gephi i wykonanie prostych analiz oraz wizualizacji sieci. W ramach ćwiczenia studenci zapoznają się z algorytmami rozkładu grafów, poznają ogólny interfejs narzędzia i wykonują samodzielnie ćwiczenie dotyczące wizualizacji grafu powiązań między postaciami z powieści W.Hugo "Nędznicy"

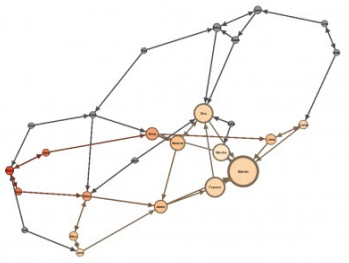
podstawowa obsługa Gephi

1. Załaduj zbiór danych dining.gephi i otwórz go w narzędziu Gephi. Upewnij się, że znajdujesz się na zakładce „*Overview*”. W oknie **`Layout`** rozwiń listę i wybierz rozkład o nazwie ForceAtlas 2. W opcjach rozkładu zaznacz opcję **`Dissuade Hubs`** i ustaw parametr Gravity na 5.0. Uruchom rozkład i zaobserwuj wyniki, po krótkiej chwili zatrzymując działanie algorytmu.
2. Włącz wyświetlanie etykiet i dostosuj ich rozmiar do wykresu, możesz także zmienić kolor czcionki do wyświetlania etykiet.
3. W oknie **`Partition`** kliknij w przycisk **`Edges`** i odświetl listę atrybutów, które można wykorzystać do partycjonowania zbioru krawędzi. Wybierz z listy wartość *choice* i kliknij przycisk **`Apply`**. Zauważ zmianę, jaka się dokonała w wizualizacji grafu.
4. Wykorzystaj dodatkowe narzędzia z belki bocznej do zmiany wyglądu grafu. Pokoloruj wierzchołek *Jane* na czerwono i zwiększ nieco jego rozmiar. Włącz narzędzie Dragging, zwiększ promień selekcji i przetestuj jego działanie. Wyznacz najkrótszą ścieżkę między *Hildą* i *Evą*.
5. Dodaj nowy wierzchołek do grafu, etykietuj go swoim imieniem i utwórz dwie krawędzie prowadzące z „Twojego” wierzchołka do dwóch dowolnych innych wierzchołków. Następnie, przejdź na zakładkę „*Data Laboratory*” i sprawdź identyfikator „Twojego” wierzchołka. Kliknij na przycisku **`Edges`** i zaktualizuj krawędzie prowadzące od „Twojego” wierzchołka, wskazując na jedną z krawędzi jako krawędź pierwszego wyboru, a drugą krawędź jako krawędź drugiego wyboru. Powróć do wizualizacji sieci i sprawdź, czy zmiany zostały uwzględnione. Odśwież wizualizację.
6. Przejdź do okna **`Ranking`** i zaznacz przycisk **`Nodes`**. Zauważ, że stopnie wyjściowe i wejściowe wierzchołków zostały już policzone. Zaznacz chęć zróżnicowania wielkości węzłów w zależności od stopnia wejściowego, ustawiając minimalny rozmiar wierzchołka na 2 a maksymalny rozmiar wierzchołka na 10. Kliknij na link Spline i sprawdź, w jaki sposób funkcja odwzorowująca stopień wierzchołka na jego wielkość zmienia wizualizację sieci.
7. Wyświetl rozkład stopni wejściowych i wyjściowych wierzchołków klikając we właściwym miejscu w oknie **`Statistics`**. Zwróć uwagę, że wyliczone wartości zostały automatycznie dodane do tabeli w zakładce „*Data Laboratory*”.
8. Wyznacz silnie spójne komponenty występujące w sieci (posłuż się przyciskiem **`Connected Components`** w oknie **`Statistics`**). Wykorzystaj znalezione komponenty do pokolorowania wierzchołków. Jeśli zaproponowane kolory są zbyt do siebie podobne, kliknij prawym klawiszem myszy i z menu kontekstowego wybierz opcję **`Randomize`**. Obejrzyj uzyskaną wizualizację. W chwili obecnej Twoja sieć powinna wyglądać mniej więcej w ten sposób:

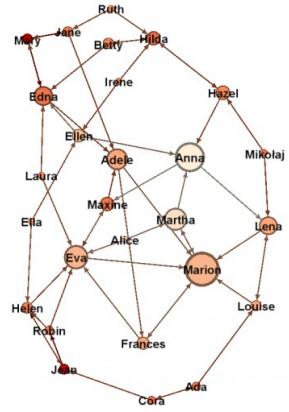


9. Zresetuj wielkości i kolory węzłów i etykiet. Wyznacz odległości między wszystkimi parami wierzchołków klikając na przycisk **`Avg. Path Length`** w oknie **`Statistics`**. Upewnij się, że traktujesz krawędzie jako łuki posiadające kierunek. Nie włączaj normalizacji wyznaczanych miar centralności. Obejrzyj uzyskane rozkłady.

10. Wykorzystaj narzędzie do rysowania odległości od zadanego wierzchołka. Wyświetl mapę odległości wierzchołków od wierzchołka *Hilda*.

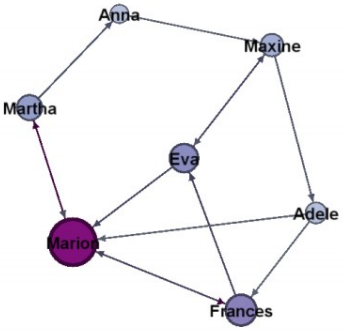


11. Przejdź do okna **`Ranking`** i zaznacz przycisk **`Nodes`**. Zaznacz chęć zmiany koloru wierzchołka w zależności od wskazanego kryterium, a z listy jako kryterium wybierz centralność wg. miary bliskości. Zmień używany schemat kolorów i zastosuj kryterium. Obejrzyj wynik w wizualizacji, wyświetl także ranking wierzchołków w postaci tabelarycznej.



12. Jako kolejne kryterium wybierz centralność wg. pośrednictwa i wskaż rozmiar wierzchołka jako modyfikowaną cechę. Zastosuj kryterium do sieci i zaobserwuj wynik. Twoja sieć powinna wyglądać mniej więcej tak:

13. Krawędzi w analizowanej sieci można interpretować jako jednoznacze głosowanie wierzchołka A na wierzchołek B. W takiej sytuacji dobre przybliżenie globalnego rankingu ważności wierzchołków produkuje algorytm PageRank. Uruchom ten algorytm z prawdopodobieństwem losowego przeskoczenia do innego węzła równym 15% i progiem zbieżności w wysokości 0.001. Użyj wyznaczonej miary PageRank aby narysować sieć w taki sposób, aby wielkość i kolor wierzchołka odpowiadały wyznaczonej centralności wg. PageRank. Zidentyfikuj 5 najważniejszych dziewczyn w sieci.
14. Wybierz tylko te wierzchołki w grafie, które posiadają miarę centralności wg. PageRank powyżej 0.05 i utwórz z nich nową sieć. W tym celu przejdź na zakładkę „*Filters*” i rozwiń gałąź **`Attributes→Range`**, wybierz węzeł **`PageRank`** i przenieś metodą Drag & Drop do okna **`Queries`**. Wykorzystaj suwak aby ograniczyć zbiór wierzchołków do pożądanego progu miary PageRank. Twoja sieć powinna wyglądać tak:



15. Otwórz ponownie pełną sieć i zresetuj kolory i wielkości wierzchołków. Następnie wybierz dowolne kryterium centralności i użyj go do zróżnicowania wyglądu wierzchołków (przykładowo: rozmiar wierzchołka reprezentuje pośrednictwo, kolor wierzchołka odpowiada komponentowi). Przejdź na zakładkę „*Preview*”. Z listy rozwijanej wybierz dowolny szablon, a następnie dokonaj niewielkich zmian. Wygeneruj i obejrzyj plik wynikowy.

zadanie samodzielne

Pobierz plik les.miserables.gephi potrzebny do wykonania ćwiczenia. Plik zawiera informacje o współwystępowaniu poszczególnych postaci w kolejnych scenach w powieści Wiktora Hugo pt. Nędznicy. Zapoznaj się z fabułą powieści: http://pl.wikipedia.org/wiki/N%C4%99dznicy

1. Wczytaj sieć do narzędzia Gephi. Eksperymentalnie dobierz taki rozkład, który zwiększa czytelność. Zauważ, że wierzchołki są już przydzielone do klas modularności.
2. Wyświetl sieć w taki sposób, aby uwypuklić centralność wierzchołków według ich pośrednictwa.
3. Wyświetl sieć z wyznaczonymi odległościami od wierzchołka Gavroche.
4. Zbuduj podsieć zawierającą 10% najważniejszych wierzchołków, gdzie miarą ważności wierzchołka jest jego stopień.
5. Przygotuj najbardziej atrakcyjną wizualizację sieci. Jeśli chcesz, możesz ograniczyć sieć do pewnego podzbioru wierzchołków. Eksperymentuj z doborem kolorów, kształtów, czcionek, tła.

W drugiej części ćwiczenia studenci uczą się korzystać z algorytmów badania modularności sieci oraz analizują sieci dynamiczne oraz ich własności.

zaawansowana wizualizacja

1. Otwórz menu **`Tools`** i wybierz opcję **`Plugins`**. W zakładce **`Settings`** upewnij się, że zaznaczone są opcje **`Gephi Thirdparties plugins`** oraz **`Gephi Update Center`**. Przejdź do zakładki **`Available plugins`** i zainstaluj następujące wtyczki:
   - Circular Layout
   - GeoLayout

   W razie konieczności zrestartuj program.
2. Pobierz zbiór danych les.miserables.gephi i otwórz go w Gephi. Wybierz layout **`Force Atlas`** i uruchom go korzystając z domyślnych parametrów.
3. Zmień siłę odpychania na 10 000 i zaobserwuj wynik. Ustaw siłę stabilizacji na 100 000, a następnie wykorzystaj narzędzie do przenoszenia aby zmienić lokalizację zbioru punktów. Zaobserwuj wynik.
4. 4. Algorytm Force Atlas ma złożoność rzędu $O(n^2)$ i bazuje na balansie sił, biorąc pod uwagę wagę krawędzi. Nadaje się do sieci o rozmiarze do 10 000 wierzchołków. Uruchom jeszcze raz ten rozkład, zmieniając (w trakcie działania algorytmu) następujące parametry:
   - **`Autostab strength`** = 2000 (większe wartości powodują wolniejszy ruch wierzchołków)
   - **`Repulsion strength`** = 1000 (siła wzajemnego odpychania się przez wierzchołki)
   - **`Attraction strength`** = 1 (siła, z jaką połączone wierzchołki się przyciągają)
   - **`Gravity`** = 100 (ogólna siła przyciągania wszystkich wierzchołków w kierunku centrum sieci aby uniknąć nadmiernego rozrzucenia wierzchołków)
   - **`Attraction distrib`**. = true (powoduje przesunięcie hubów w kierunku peryferiów a autorytetów w kierunku centrum sieci).
5. W trakcie działania algorytmu kliknij na dowolnym wierzchołku prawym klawiszem myszy i z menu kontekstowego wybierz opcję **`Settle`**. Zaobserwuj wynik.

6. Algorytm Fruchterman-Reingold symuluje sieć jako zbiór cząstek obdarzonych masą (wierzchołki) i połączonych za pomocą sprężyn (krawędzie) o określonych siłach (wagi krawędzi). Ostateczny rozkład wierzchołków próbuje minimalizować całkowitą energię układu. Złożoność algorytmu to $O(n^2)$, a główną wadą jest długi czas obliczeń. Algorytm nadaje się do sieci o liczbie wierzchołków nieprzekraczającej 1000. Uruchom ten rozkład, zmieniając (w trakcie działania algorytmu) parametry **`Area`** i **`Gravity`**

7. Algorytm Yifan-Hu Multilevel to bardzo szybki algorytm wykorzystujący połączenie idei ukierunkowania wierzchołków przez siły reprezentujące krawędzie oraz odpychania wierzchołków przez odległe klastry. Klastry wierzchołków są aproksymowane do jednego meta-wierzchołka przez algorytm Barnesa-Huta. Złożoność algorytmu to O(n log(n)). Algorytm dobrze się sprawdza dla sieci o rozmiarze od 100 do 100 000 wierzchołków. Nie uwzględnia wag krawędzi. Uruchom rozkład Yifan-Hu Multilevel zmieniając (w trakcie działania algorytmu) parametry **`Step ratio`** = 0.99 (większe wartości powodują poprawę jakości kosztem szybkości działania), **`Optimal distance`** = 200 (naturalna długość „sprężyn”, większe wartości powodują większe rozrzucenie wierzchołków), **`Theta`** = 1.0 (parametr algorytmu Barnesa-Huta, mniejsze wartości powodują większą precyzję wyliczeń)
8. Algorytm OpenOrd Layout służy przede wszystkim do takiego rozmieszczenia wierzchołków, które wizualnie najbardziej separuje klastry występujące w sieci. Podstawą algorytmu jest algorytm Fruchtmana-Reingolda realizowany iteracyjnie z wykorzystaniem techniki symulowanego wyżarzania. Złożoność algorytmu jest rzędu O(n log(n)), algorytm radzi sobie z

sieciami do 1 000 000 wierzchołków. Uruchom rozkład OpenOrd Layout zmieniając (w trakcie działania algorytmu) następujące parametry: **`Edge cut`** = 0.95 (procentowa odległość między dwoma najbardziej odległymi wierzchołkami w sieci, większe wartości prowadzą do większej separacji klastrów) oraz **`Num iterations`** = 100, 800 (ściąganie i rozciąganie klastrów)

9. Algorytm *OpenOrd Layout* jest jednym z najlepszych algorytmów do wizualizacji dużych sieci. Pobierz plik internet_routers.gml.zip i otwórz go w nowym obszarze roboczym. Sprawdź, ile wierzchołków i krawędzi zawiera ta sieć, reprezentująca strukturę sieci Internet na poziomie systemów autonomicznych (jest to obraz z lipca 2006 r.) Wykorzystaj rozkład OpenOrd Layout w celu wizualizacji ogólnej struktury tej sieci.
10. Algorytm Force Atlas 2 to ulepszona wersja algorytmu Force Atlas, mogąca obsługiwać duże sieci do 1 000 000 wierzchołków. Złożoność algorytmu to O(n log(n)) dzięki zastosowaniu algorytmu Barnesa-Huta do agregacji wierzchołków. Algorytm wykorzystuje do rozkładu wagę krawędzi. Uruchom rozkład Force Atlas 2 zmieniając (w trakcie działania algorytmu) następujące parametry: **`LinLog mode`** = true (siła przyciągania wierzchołków rośnie liniowo, a siła odpychania wierzchołków rośnie logarytmicznie, w efekcie klastry stają się widoczniejsze), **`Scaling`** = 100 (większe wartości generują bardziej rozrzedzone sieci), **`Edge weight influence`** = 0, 0.5, 1 (wpływ wagi krawędzi na algorytm)
11. Algorytm Circular Layout jest jednym z najprostszych algorytmów. Wierzchołki są umieszczane na okręgu zgodnie z podanym kryterium. Złożoność algorytmu to O(n), algorytm działa dla dowolnie dużych sieci, ale rzadko udaje się wyprodukować przydatną wizualizację na podstawie surowych danych źródłowych. Wykorzystaj rozkład Circular Layout dokonując w nim następujących zmian: **`Order nodes by`** = Degree, **`Diameter size`** = 50 (wielkość okręgu)
12. Przydatnym sposobem wizualizacji jest algorytm Radial Axis Layout, umożliwiający dodatkowo grupowanie wierzchołków według zadanego kryterium. Upewnij się, że analizowana sieć ma wyznaczoną przynależność wierzchołków do modułów. Następnie wykorzystaj rozkład Radial Axis Layout ustalając następujące wartości parametrów: **`Group nodes by`** = Degree, **`Group nodes by`** = Modularity class, **`Order nodes by`** = Degree, **`Draw spar/axis spiral`** = true (w ten sposób bardziej uwypuklone zostają związki wewnątrz grupy), **`Ascending order`** = true (w ten sposób bardziej uwypuklone zostają związki pomiędzy grupami)
13. Załaduj zbiór danych airlines.gexf. Algorytm *Geo Layout* narysuje sieć w taki sposób, aby położenie każdego wierzchołka odpowiadało jego szerokości i długości geograficznej, przy zadanej metodzie odwzorowania na 2D. Wyświetl sieć, włącz wyświetlanie etykiet z nazwami lotnisk, a następnie sprawdź wpływ metod odwzorowania 2D na ostateczny wygląd sieci.

Zadanie samodzielne

Pobierz zbiór danych hero-social-network.gephi reprezentujący związki między superbohaterkami i superbohaterami ze wszechświata Marvel. Krawędź między dwoma postaciami reprezentuje fakt wystąpienia tych postaci w tym samym komiksie. Zbiór danych jest bardzo duży i obejmuje 10 000 postaci i 180 000 związków. Wykonaj następujące ćwiczenia:

- znajdź algorytm rozkładu najbardziej odpowiedni dla tego zbioru danych
- wyznacz miary centralności i ogranicz zbiór danych do takich wierzchołków, które reprezentują interesujące postacie
- przygotuj trzy wizualizacje podkreślające:
  - najważniejsze postaci pod względem liczby opublikowanych komiksów,
  - najważniejsze postaci pod względem łączenia ze sobą fragmentów wszechświata Marvel,
  - najważniejsze klastry postaci występujących często razem