AGH

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

Metody analizy i wizualizacji dużych zbiorów danych

Laboratorium 6 - IVGA

Jan Gołda, Wojciech Basiura - WIET 2018/2019

1. Opis zagadnienia

Laboratorium 6 poświęcone było wizualizacji grafów w programie Ivga.

Zadaniem domowym była wizualizacja zbioru History2 oraz analiza otrzymanych wyników.

1.1. Zbiór danych

Badanym przez nas zbiorem danych jest History2, który został nam udostępniony przez prowadzącego.

Zbiór ten reprezentuje odnośniki pomiędzy artykułami na Wikipedii z kategorii historia.

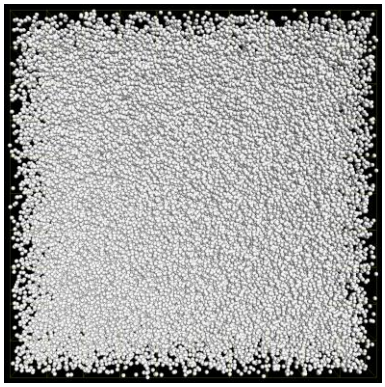
Graf ma następujące rozmiary:

98715 wierzchołków

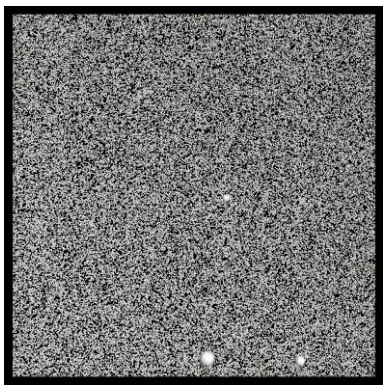
600486 krawędzi

2. Wizualizacja

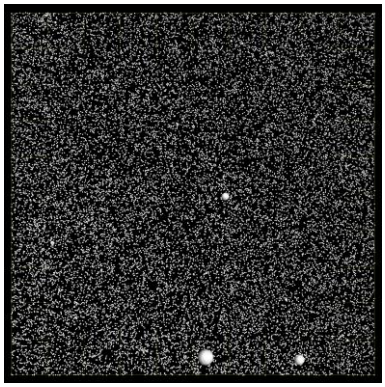
Po załadowaniu danych grap przedstawiał się następująco:



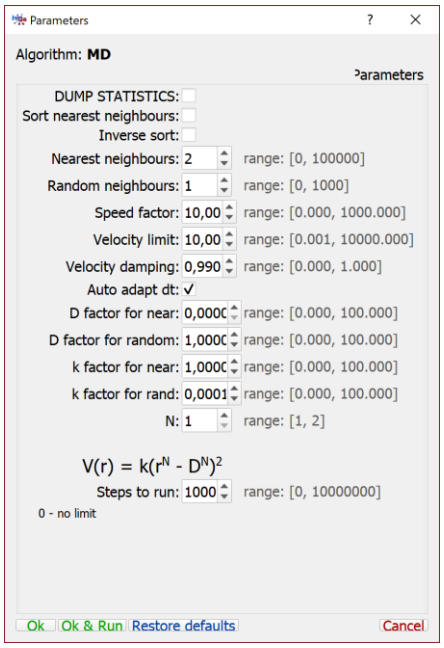
W celu poprawienia czytelności grafu zmieniliśmy część opcji, w tym wybraliśmy „2D” oraz ustawiliśmy „Node radius mode” na „Degree of node”:



Aby poprawić czytelność grafu usunęliśmy, korzystając z funkcji *„Select Advanced…” -> „Degree of node”*, **72641** wierzchołków o stopniu mniejszym niż **10**, w rezultacie otrzymując graf o **26074** wierzchołkach:

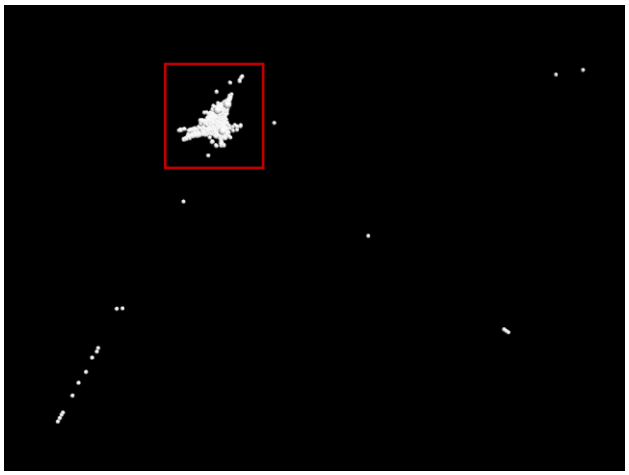


Następnie uruchomiliśmy algorytm MD z parametrami:

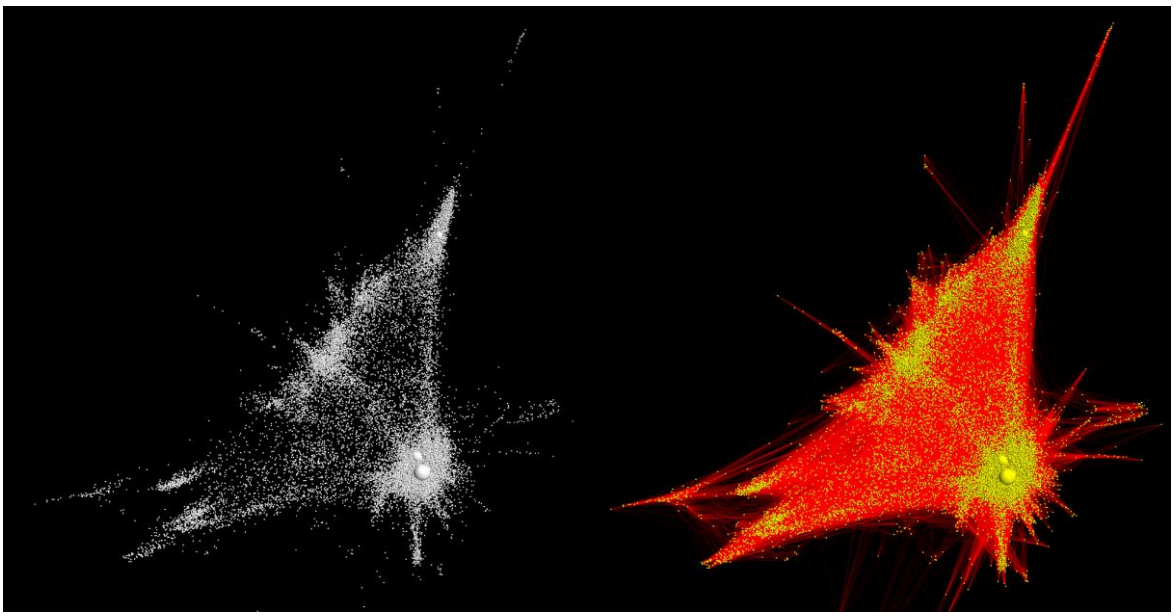


Czego rezultat zaowocował powstaniem następującej struktury:

Po ręcznym sprawdzeniu co reprezentują poszczególne grupy zdecydowaliśmy się skupić na głównym obszarze:



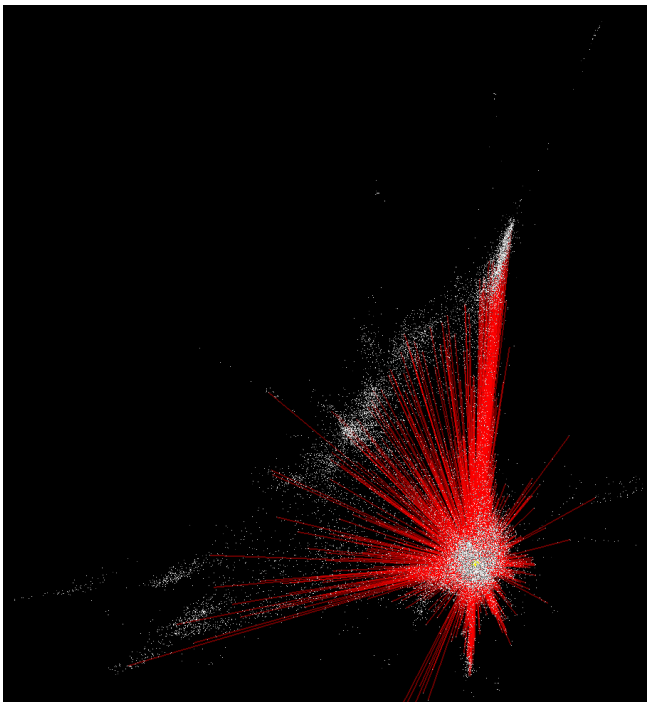
Który po kolejnych korektach co do sposobu wyświetlania wyglądał tak (bez/z krawędziami):



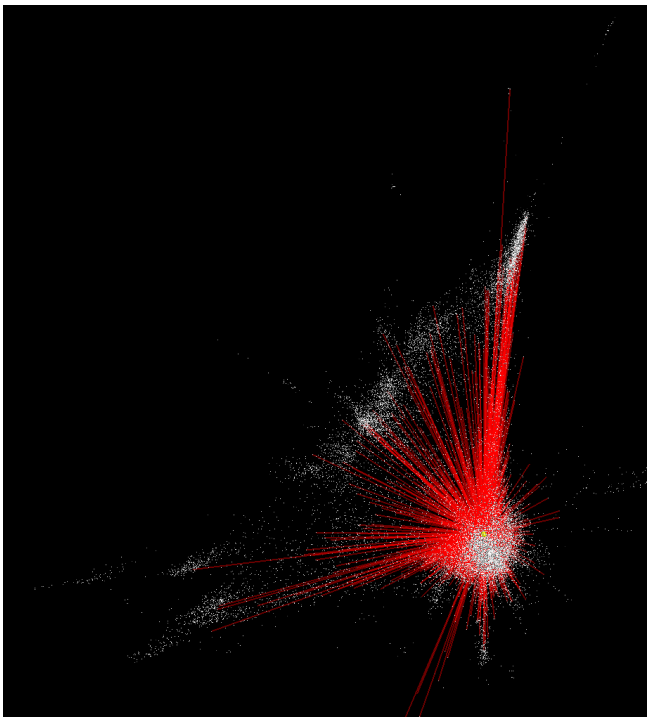
3. Analiza

Mając gotowy graf mogliśmy zacząć analizować dane. Najpierw, poprzez zaznaczenie krawędzi wychodzących, wyselekcjonowaliśmy kilka grup wokół wierzchołków o najwyższych stopniach.

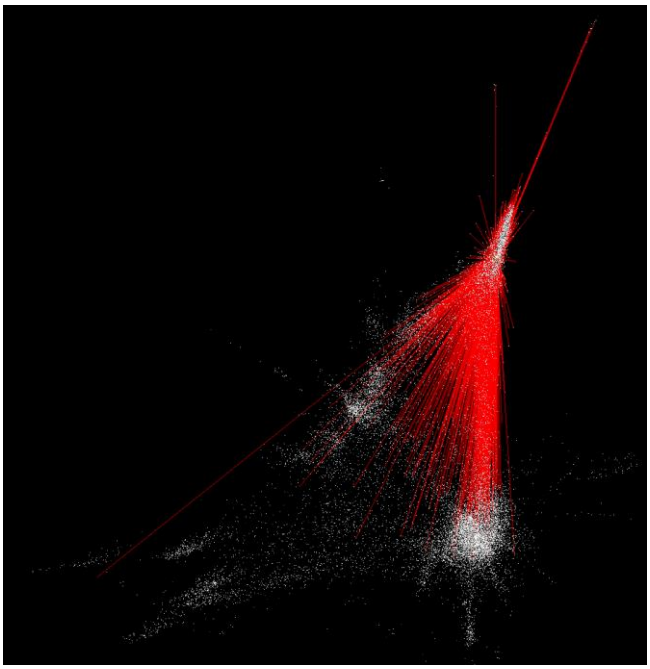
Druga Wojna Światowa (18211 odniesień):



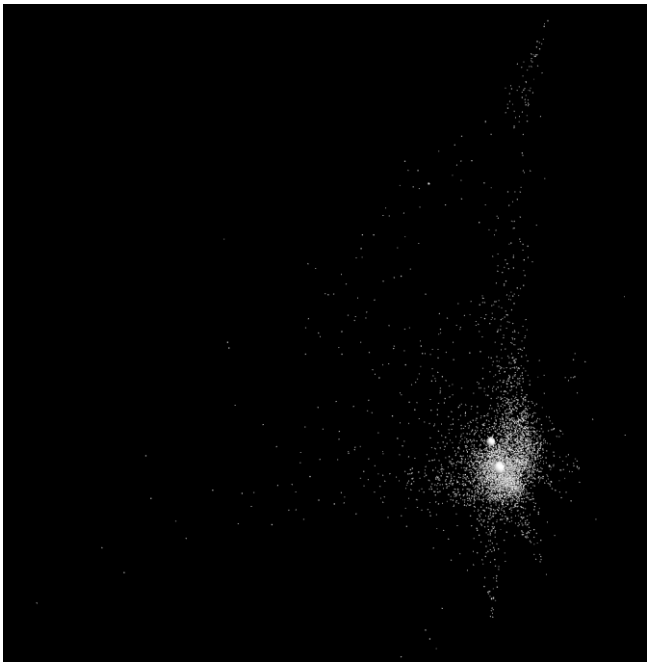
Pierwsza Wojna Światowa (11192 odniesień)



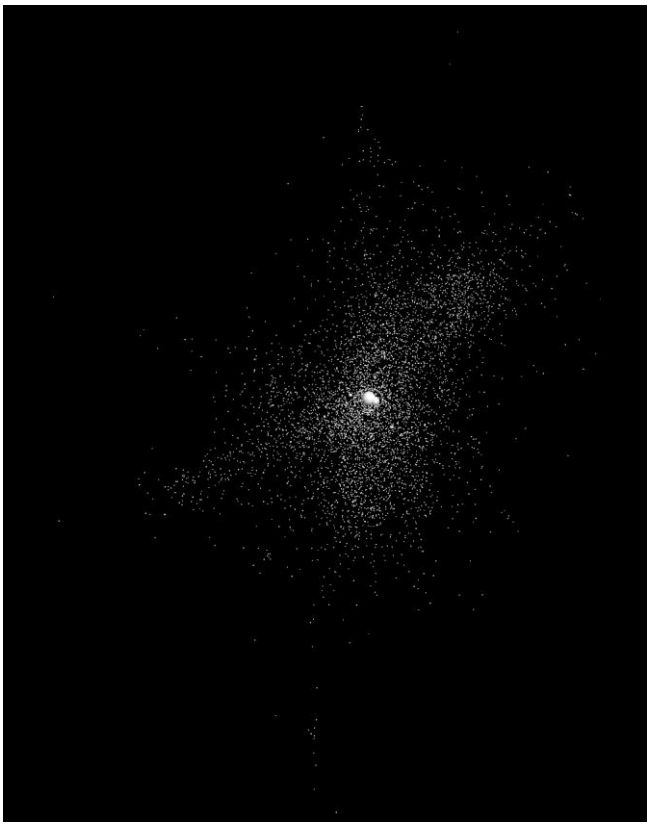
Amerykańska Wojna Domowa (8703 odniesienia):



W dalszej części postanowiliśmy się skupić na **Drugiej Wojnie Światowej**, poprzez usunięcie wierzchołków znajdujących się z odległości (wg ilości krawędzi) większej niż **1** od Drugiej Wojny Światowej.

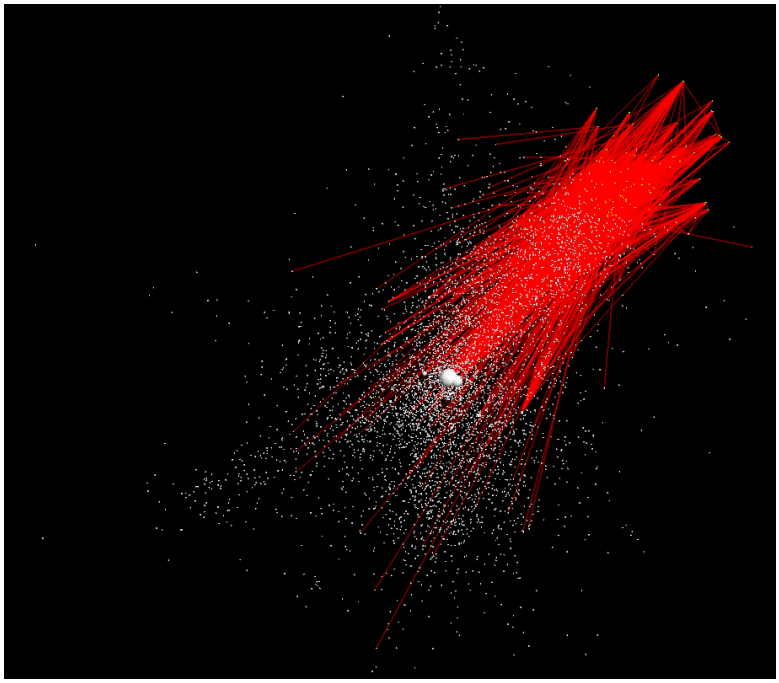


Jako że graf znacznie zmienił swój kształt z powodu usunięcia wierzchołków, ponownie uruchomiliśmy MD, tym razem otrzymując graf historii drugiej wojny światowej:



Z tak zwizualizowanego grafu można już wyciągać wnioski.

Dla przykładu: fragment grafu przestawiający wojnę Ameryki z Japonią (wojna na Pacyfiku).



4. Podsumowanie

Program w porównaniu do wcześniej testowanych, w szczególności Gephi, ma dosyć ubogi i nieczytelny interface, przez co na pierwszy rzut oka sprawia wrażenie gorszego.

Jednak jak zobaczyliśmy powyżej, jego funkcje pozwalają na wszystko co konieczne jeżeli idzie o prostą analizę graficzną/strukturalną grafu. Jeżeli dodatkowo weźmiemy pod uwagę jego moc obliczeniową (bez problemu poradził sobie z pełnym grafem History2 który ma prawie 100’000 wierzchołków) to dostrzeżemy jak potężne jest to narzędzie, szczególnie w porównaniu do Gephi, który przestawał sobie radzić już przy 4’000 wierzchołków.

Program ten na pewno jest warty uwagi, w szczególności gdy mamy do czynienia z bardzo dużymi grafami.