Zaawansowane technologie w bazach danych

Analiza sieci społecznych w portalu last.fm

Justyna Plewa

Paweł Pierzchała

1 Cel projektu

Celem projektu jest analiza społeczności tworzących się w portalach internetowych. W analizowanym serwisie wyszukujemy społeczności oraz określamy ich strukturę. Sprawdzamy jak ta struktura zmienia się w czasie.

2 Portal last.fm

Last.fm jest internetową radiostacją, system muzycznych rekomendacji oraz portalem społecznościowym. Każdy z użytkowników ma listę odtwarzanych utworów, którą może aktualizować na „żywo” używając plugin-ów do popularnych odtwarzaczy plików mp3. Gromadzone są również dane o koncertach użytkownika. Ponad to serwis udostępnia funkcje typowe dla innych portali społecznościowych takie jak znajomi, galerie, komentarze.

Portal udostępnia publiczne dane użytkowników w formacie XML po przez web-service, którego ograniczeniem jest liczba 5 zapytań na sekundę.

W projekcie analizujemy następujące powiązania między użytkownikami:

- Lista znajomych
- Ulubione utwory – dwaj użytkownicy są powiązani, jeżeli mają taki sam ulubiony utwór
- Koncerty – dwaj użytkownicy są powiązani, jeżeli byli na tym samym koncercie

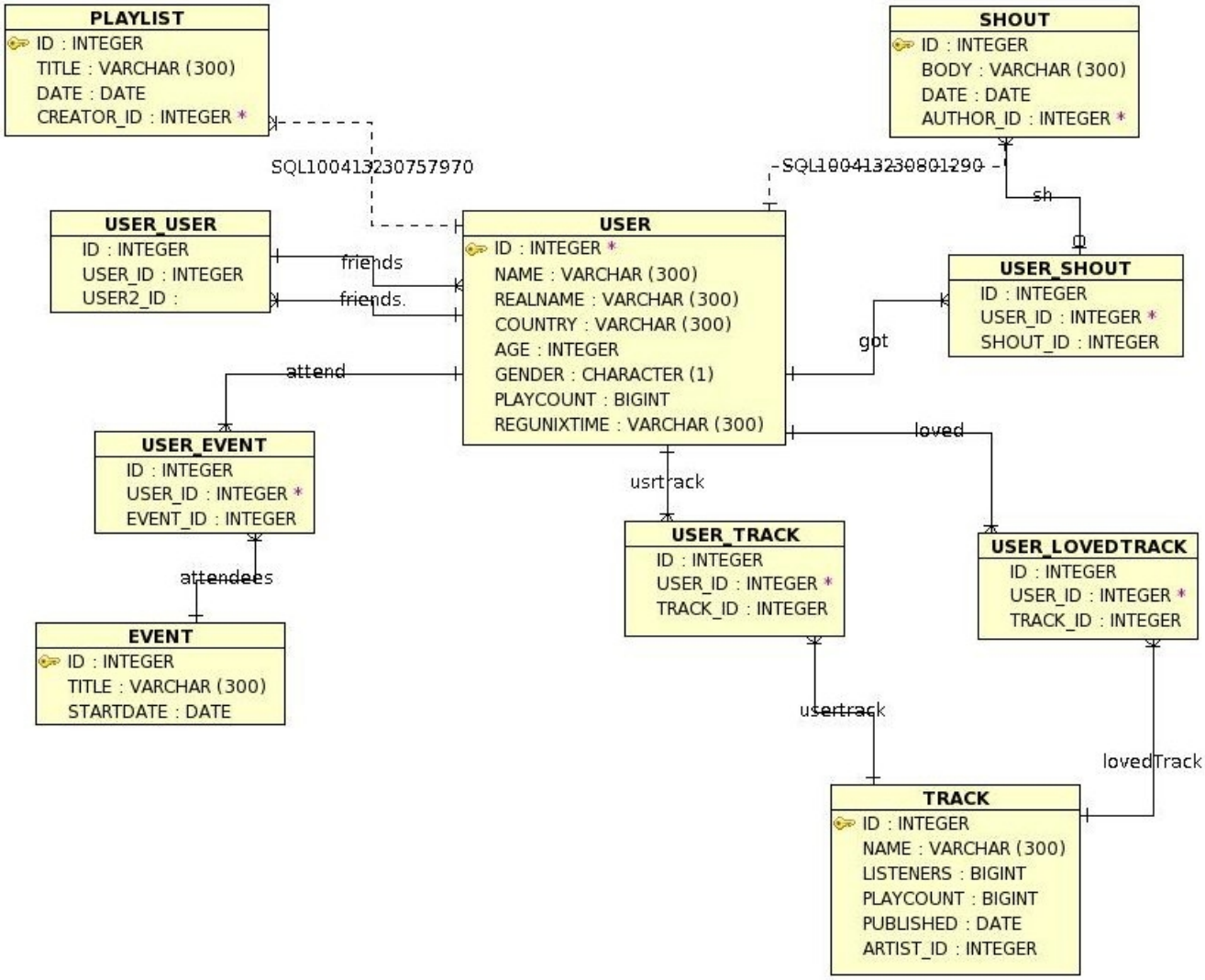
3 Technologie

Projekt został zaimplementowany w języku Java, z użyciem technologii:

- Hibernate
- Jung – wykorzystaliśmy struktury danych, moduł do wizualizacji, alalgorytm klastrujący oraz narzędzia do obliczania miar sieci społecznych
- Baza danych DB2
- last.fm API bindings for Java do pobierania danych z portalu Last.fm

3.1 Struktura bazy danych

Schemat zawierający najważniejsze tabele w bazie:



3.2 Problemy

Początkowo planowaliśmy pobrać dane przy użyciu skryptów napisanych w języku Ruby, okazało się, że plugin którego chcieliśmy użyć nie umożliwiał nam na pobranie interesujących nas danych.

W bibliotece „last.fm API bindings for Java” znajdował się błąd który uniemożliwiał pobieranie koncertów użytkownika z przeszłości. Po pobraniu źródeł biblioteki udało nam się ją naprawić.

4 Przeprowadzone analizy

Na podstawie informacji o powiązaniach miedzy użytkownikami tworzony jest nieskierowany graf sieci społecznej. Na tej sieci prowadzone jest klastrowanie przy pomocy algorytmu EdgeBetweennessClusterer.

> *EdgeBetweennessClusterer usuwa zadaną liczbę krawędzi o najwyższej wartości Betweenness (po każdorazowym usunięci miara obliczana jest na nowo). Klastry w tak zredukowanym grafie wyszukiwane są przy użyciu algorytmu WeakComponentClusterer.*

WeakComponentClusterer znajduje wszystkie składowe grafy, będące maksymalnym grafami w których między każdą parą wierzchołków istnieje ścieżka wewnątrz.

Dla każdego wierzchołka grafu obliczane są wartości jego:

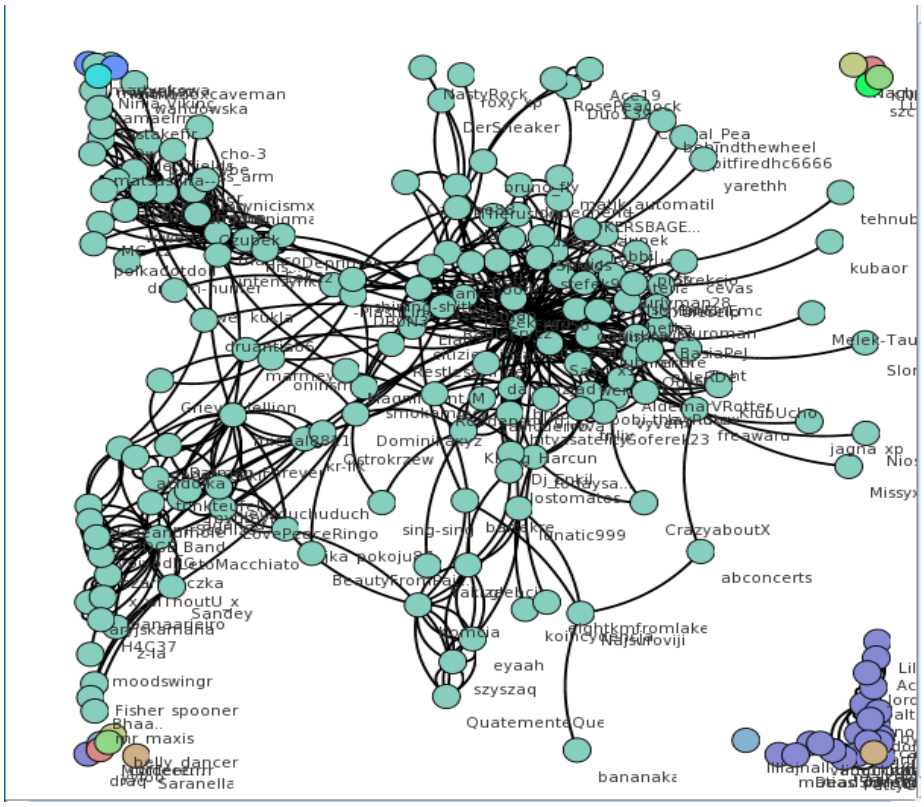
- PageRank – *określa ważność węzła na podstawie liczności i ważności węzłów doń prowadzących.*
- Betweenness Centrality – *istotność węzła jest wyznaczana na podstawie liczby najkrótszych ścieżek przechodzących przez dany węzeł*

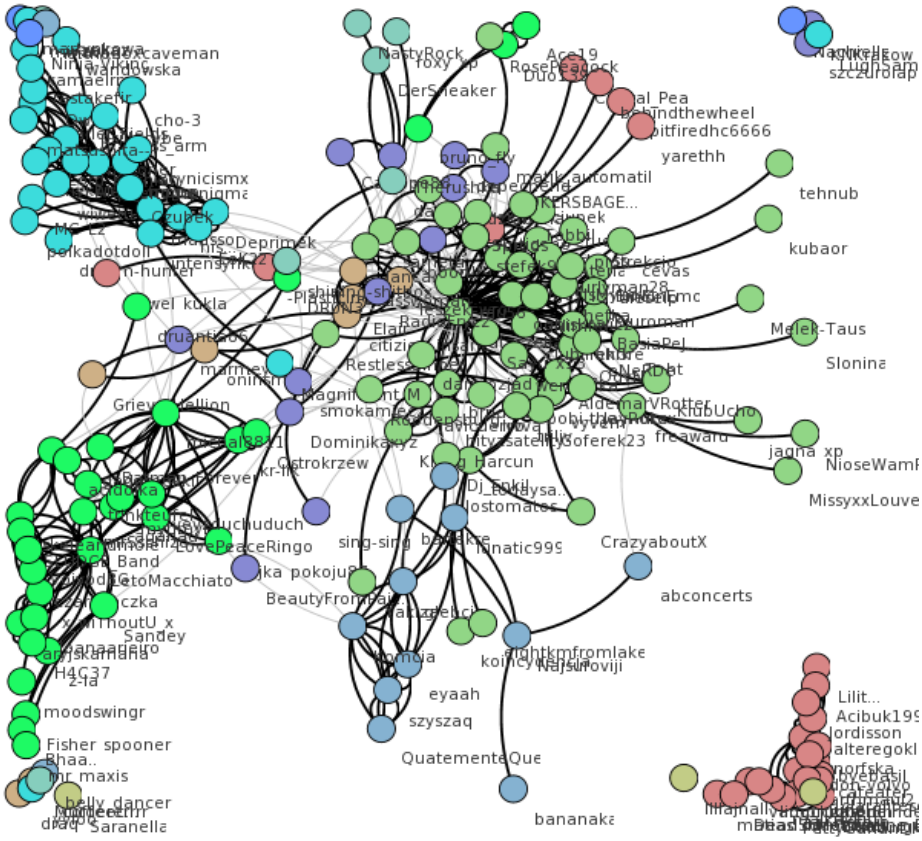
Porównujemy wyniki klastrowania różnych sieci powiązań, aby określić ich pokrycie.

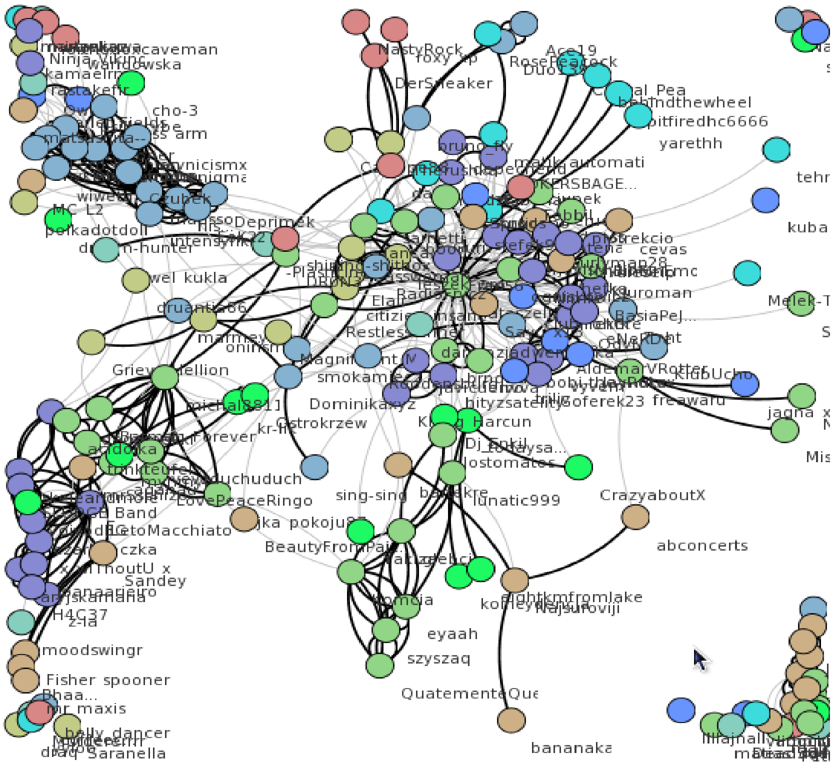
Dla dwóch wyników klastrowania A i B, dla każdego klastra z grupy A znajdujemy klaster z grupy B dla którego ich przecięcie jest największe. Określamy stosunek liczebności nowo powstałej grupy i klastra A, który jest pokryciem jednego klastra. Pokrycie jest średnią wszystkich pojedynczych pokryć.

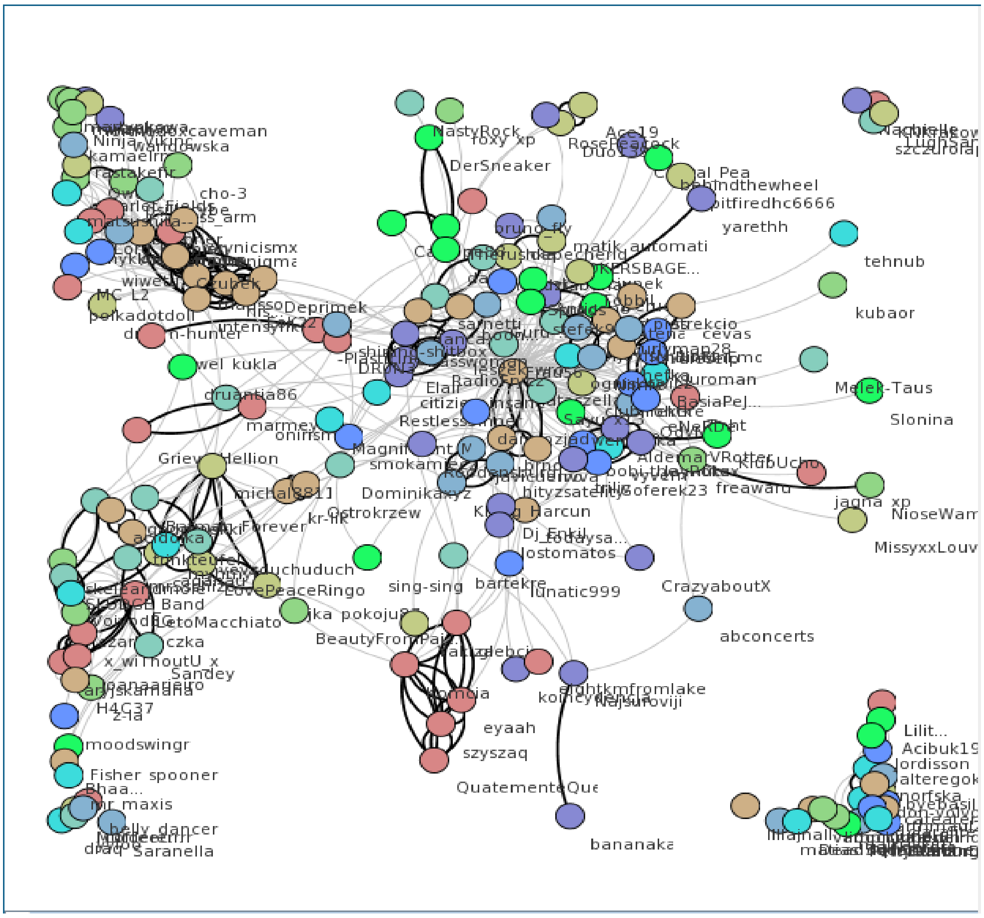
5 Wyniki eksperymentów

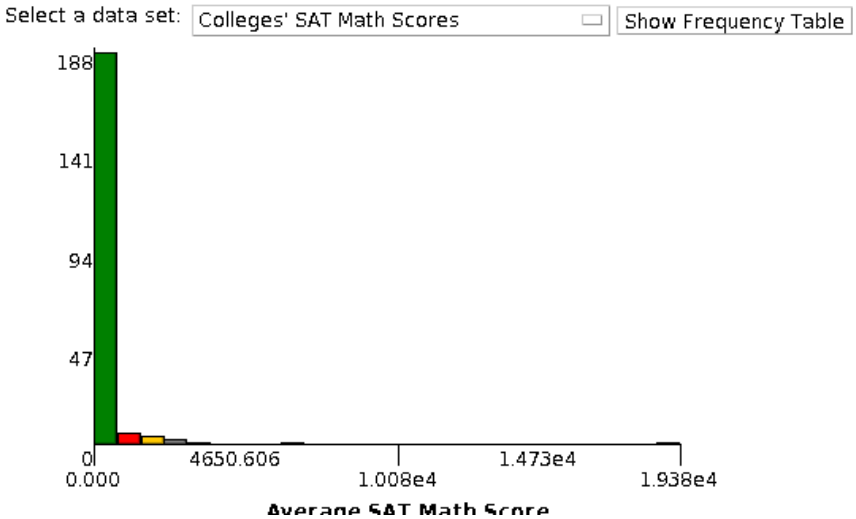
5.1 Klastrowanie znajomych



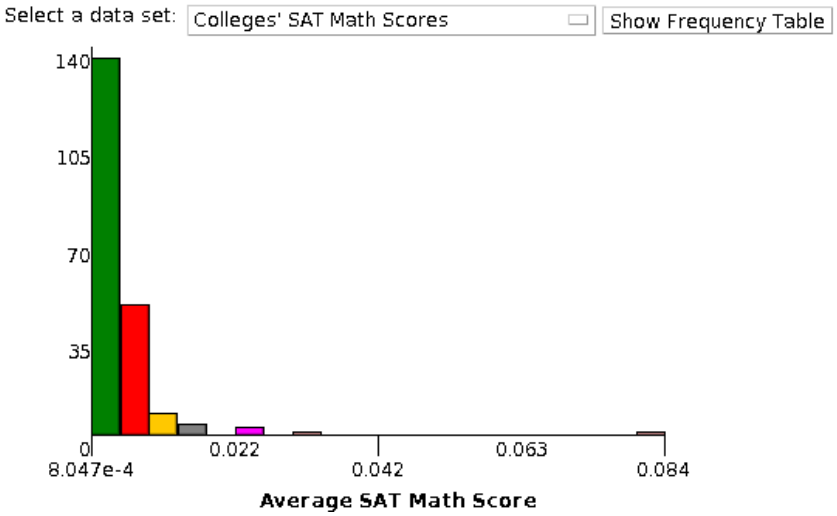



NastyRock
foxy_xp
Ace19
DerSneaker
pitfiredhc6666
yarethh
matik_automati
Ninja Viking
cho-3
polkadotdoll
wel kukla
druantia86
michal8811
Dominikaxyz
Ostrokrzew
kr-lik
LovePeaceRingo
sing-sing
lunatic999
CrazyaboutX
abconcerts
freaward
eyaah
szyszaq
QuatementeQue
bananaka
moodswingr
Fisher spooner
mr maxis
belly dancer
Saranella
H4C37
Sandey
eighthkmfromlake
kuba

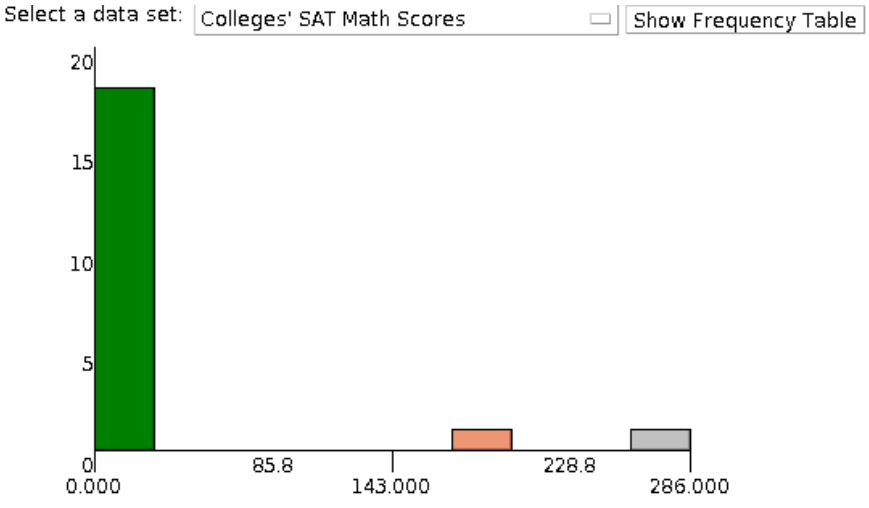

NastyRock
foxy_xp
Ace19
DerSneaker
pitfiredhc6666
yarethh
matik_automati
tehnub
kubaor
Melek-Taus
Slonina
Ninja Viking
cho-3
polkadotdoll
wel kukla
druantia86
michal8811
Dominikaxyz
Ostrokrzew
kr-lik
LovePeaceRingo
sing-sing
lunatic999
CrazyaboutX
abconcerts
freaward
jagna_xp
eyaah
szyszaq
QuatementeQue
bananaka
moodswingr
Fisher spooner
mr maxis
belly dancer
Saranella
H4C37
Sandey
eighthkmfromlake
Lilit...



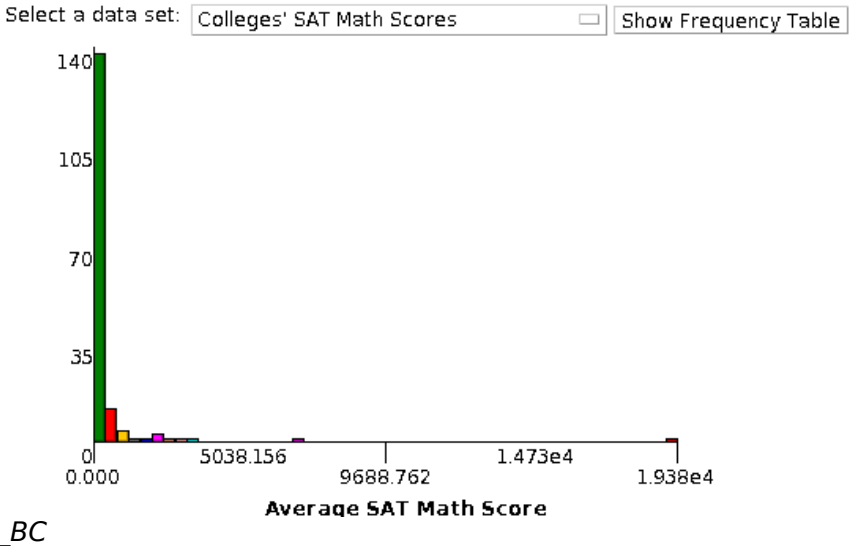
Rys 1: BC



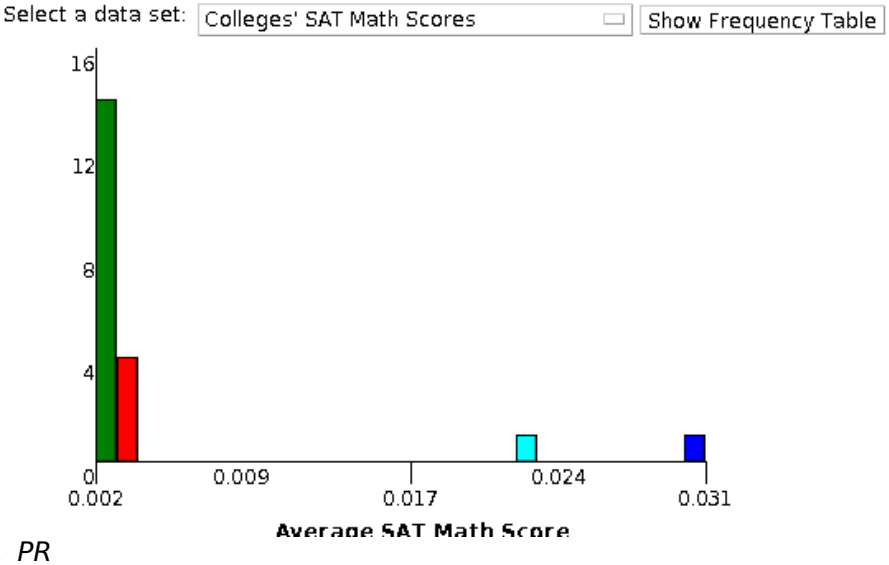
Rys 2: PR



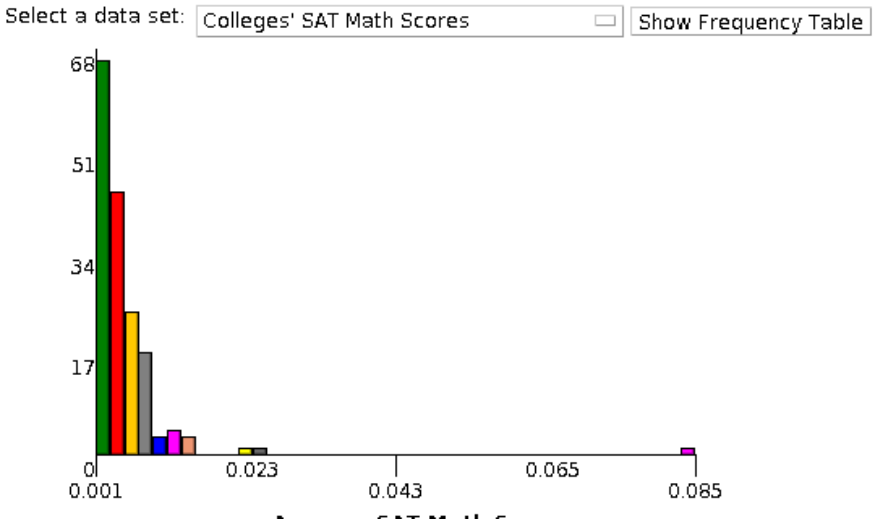
Rys 3: Klaster1_BC



Rys 4: Klaster12_BC

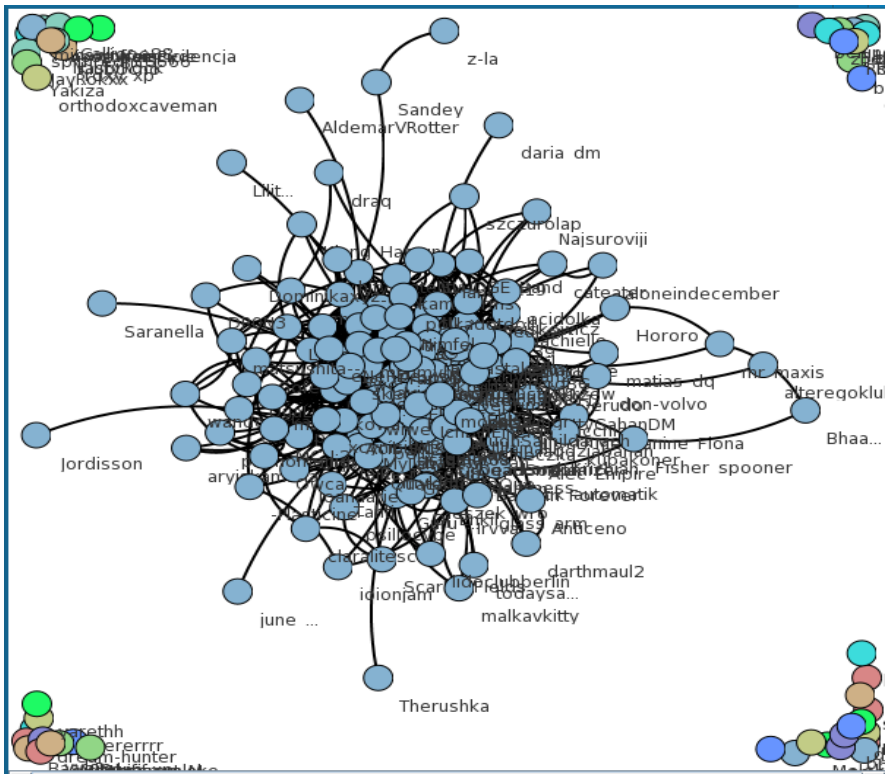


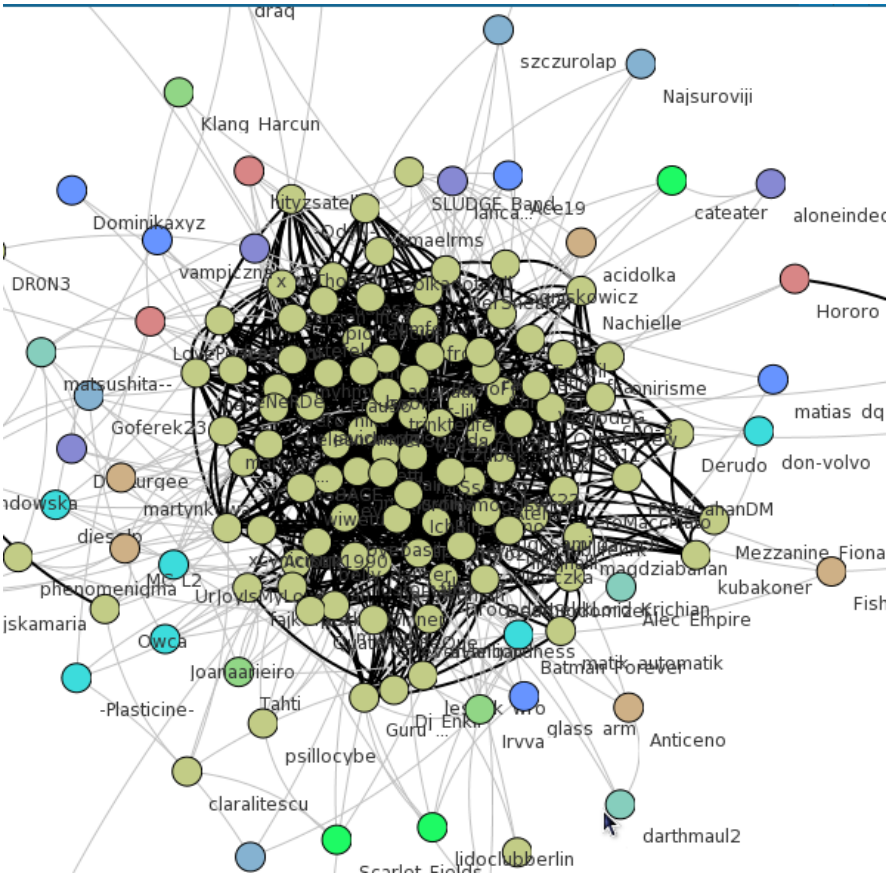
Rys 5: Klaster1_PR

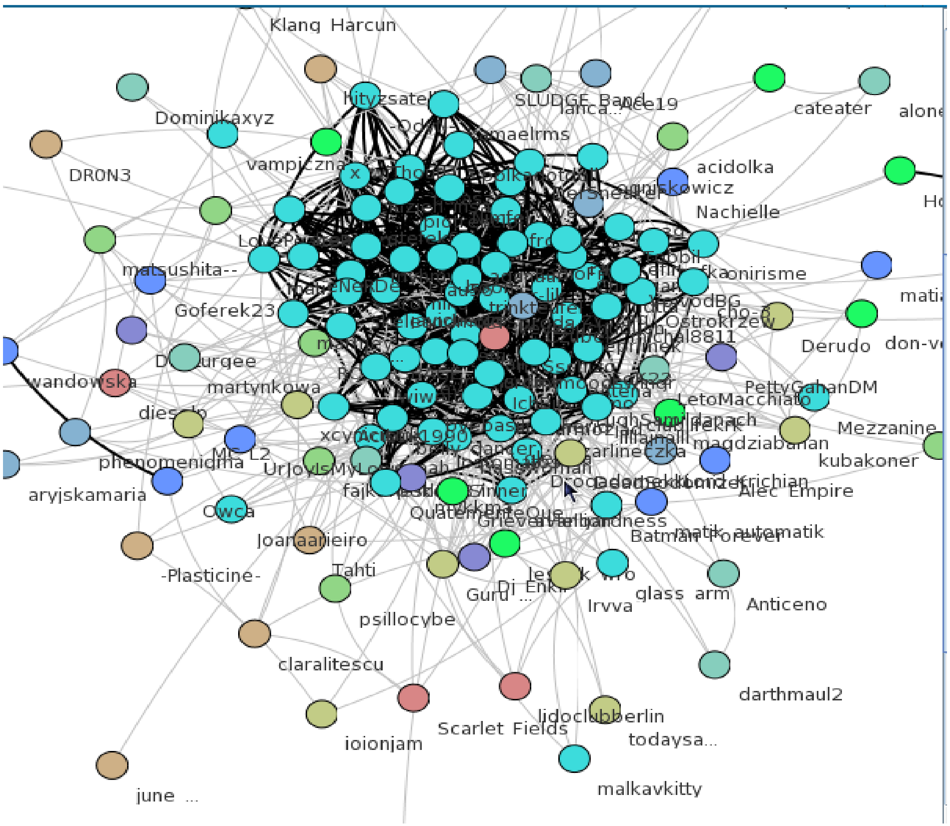


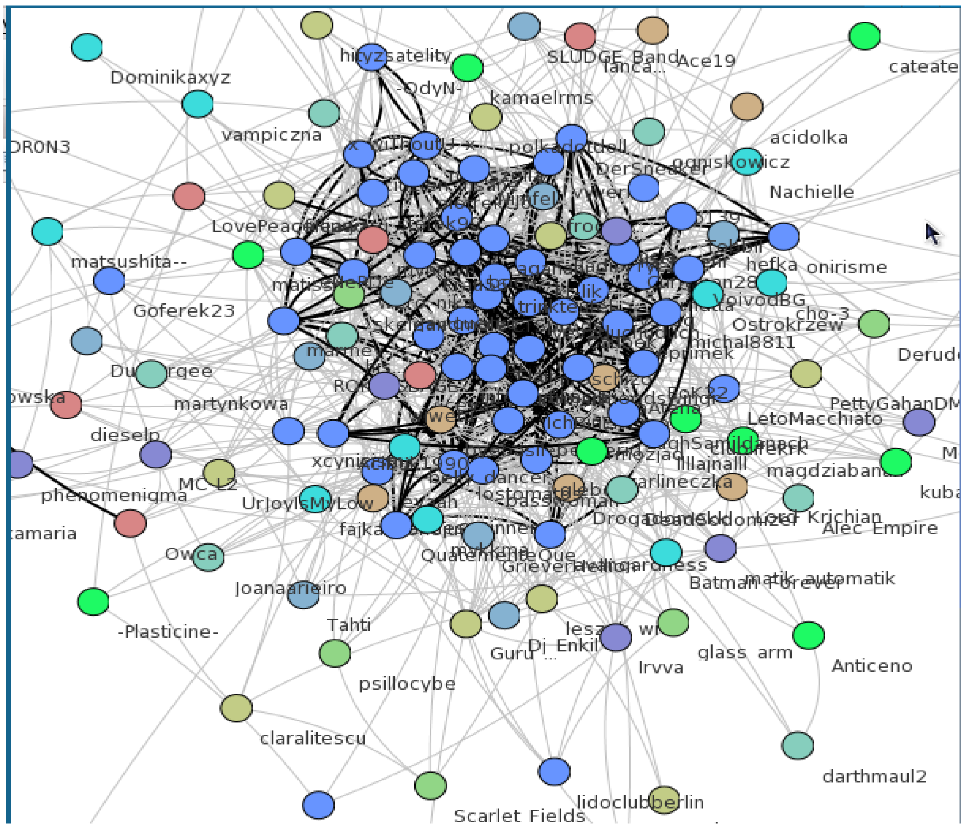
Rys 6: Klaster12_PR

5.2 Klastrowanie ulubionych utworów

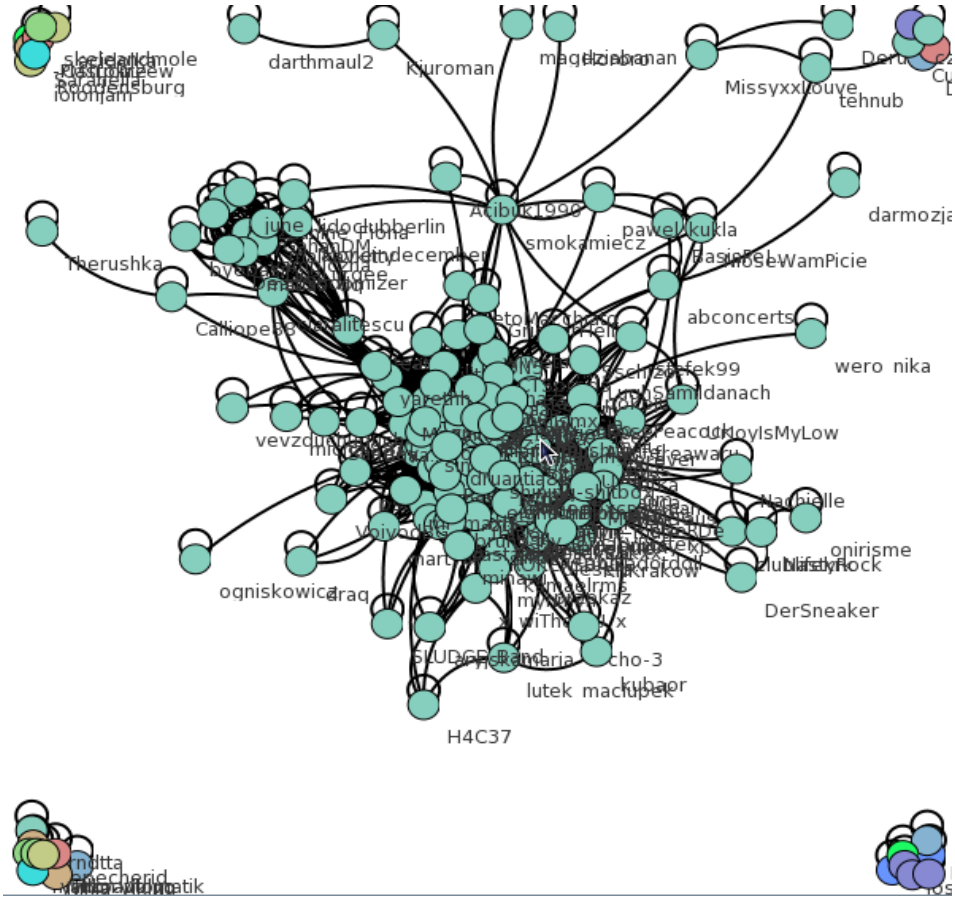


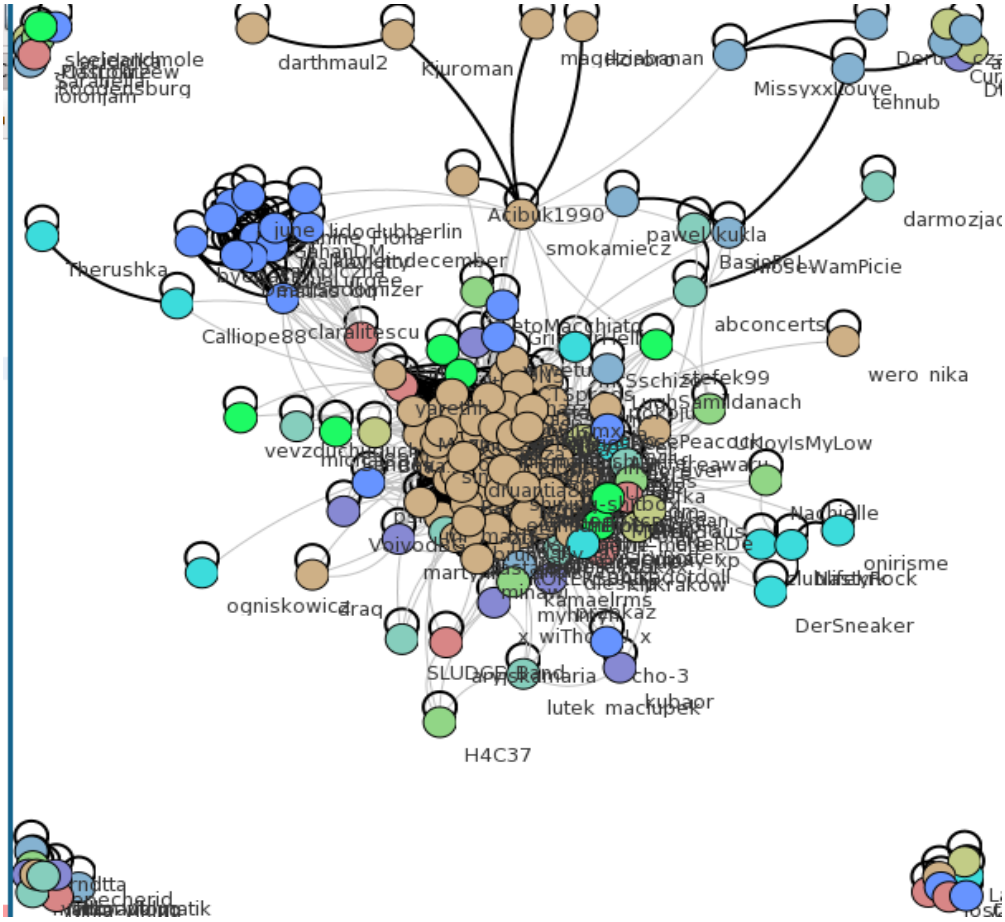


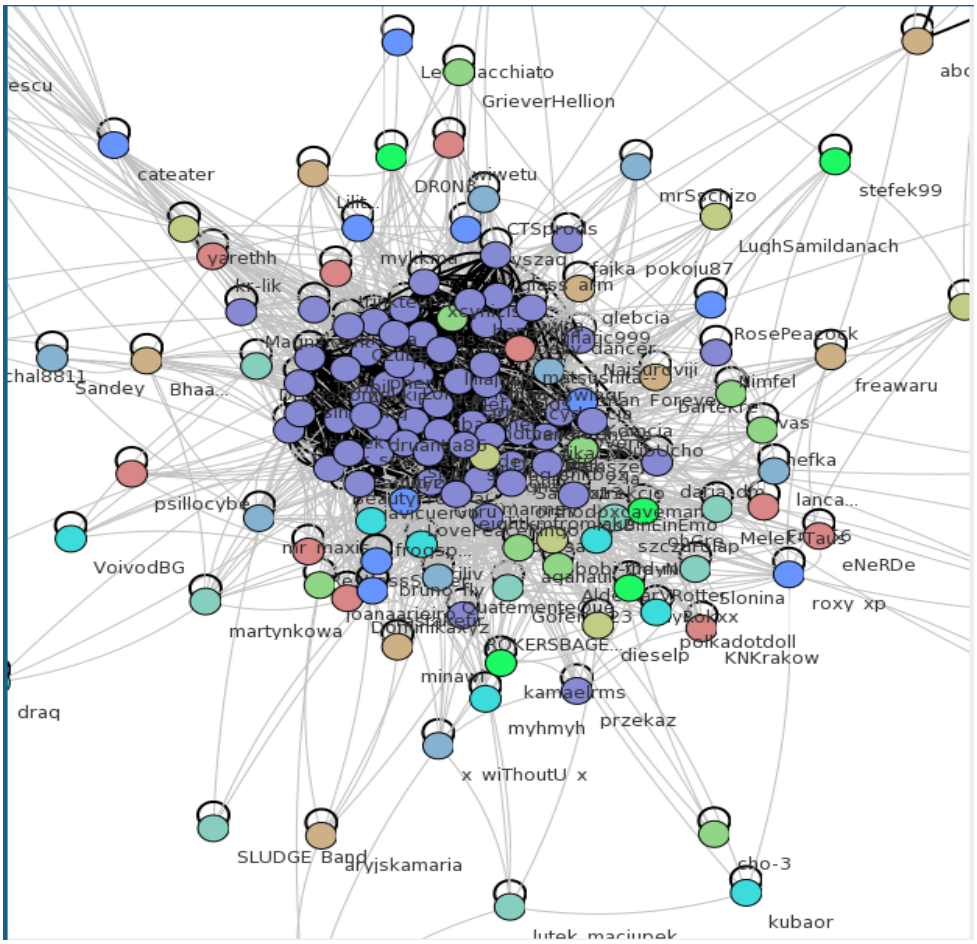

Klang Harcun
SLUDGE Band
Ace19
cateater
Dominikaxyz
DRON3
vampiczna
kamaelrms
acidolka
Nachielle
onirisme
matsushita--
Goferek23
Derudo
wandowska
martynkowa
PettyGahanDM
LetoMacchiato
Mezzanine
magdziabanan
kubakoner
phenomenigma
MC L2
aryjskamaria
Alec Empire
Owca
matik automatik
Batman Forever
-Plasticine-
Tahti
Guru ...
Dj Enkil
Irvva
glass arm
Anticeno
psillocybe
claralitescu
darthmaul2
lidoclubberlin
Scarlet Fields
ioionjam
todaysa...
june ...
malkavkitty

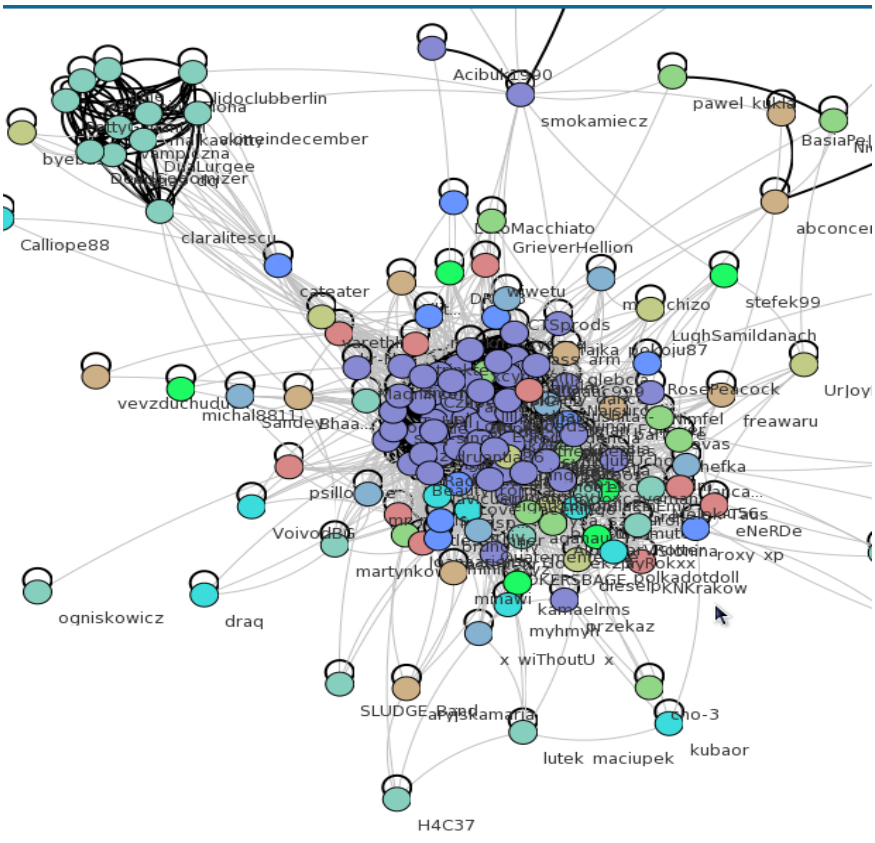

SLUDGE Band
Ace19
cateate
Dominikaxyz
-OdyN-
kamaelrms
DRON3
vampiczna
acidolka
Nachielle
LovePeaceHate
matsushita--
onirisme
VoivodBG
Goferek23
Ostrokrzew
michal8811
martynkowa
PettyGahanDM
LetoMacchiato
dieselp
magdziabanan
MC L2
phenomenigma
UrJoyIsMyLow
Alec Empire
Owca
QuatermenteQue
matik automatik
Batman Forever
Joanaarieiro
Tahti
Guru ...
Dj Enkil
Irvva
glass arm
Anticeno
psillocybe
claralitescu
darthmaul2
lidoclubberlin
Scarlet Fields

5.3 Klastrowanie koncertów





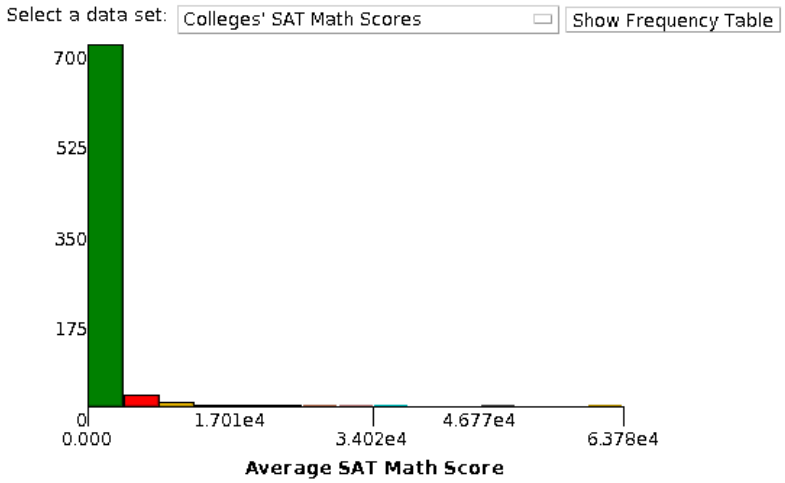

GrieverHellion
cateater
stefek99
LughSamildanach
RosePeacock
freawaru
psillocybe
VoivodBG
eNeRDe
martynkowa
dieselp
KNKrakow
draq
myhmyh
przekaz
x wiThoutU x
SLUDGE Band
cho-3
kubaor


smokamiecz
GrieverHellion
cateater
stefek99
LughSamildanach
RosePeacock
freawaru
eNeRDe
roxy xp
KNKrakow
kamaelrms
myhmyh
przekaz
x wiThoutU x
SLUDGE Band
cho-3
kubaor
lutek maciupek
H4C37

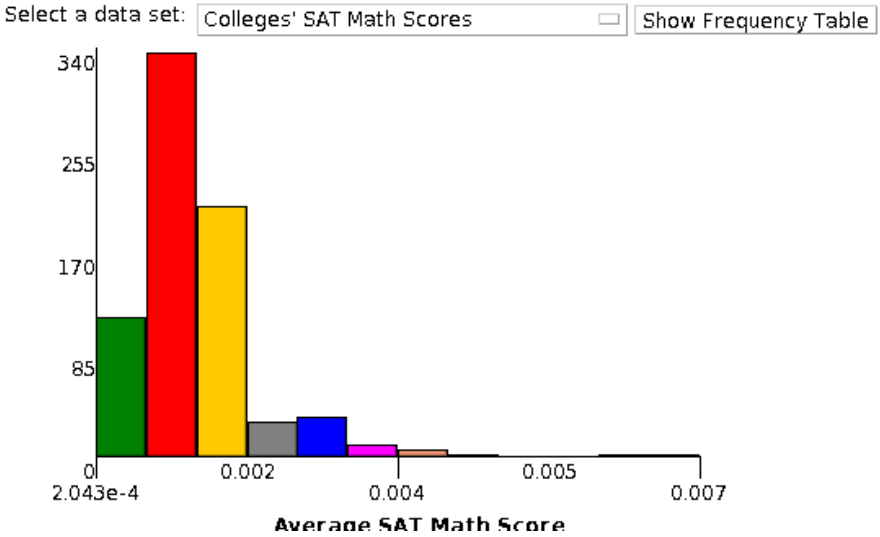
5.4 Klastrowanie w odcinkach czasu(porównanie)

Koncerty od 03.01.2009 vs Koncert 03.01.2010

Pokrycie: 1.0



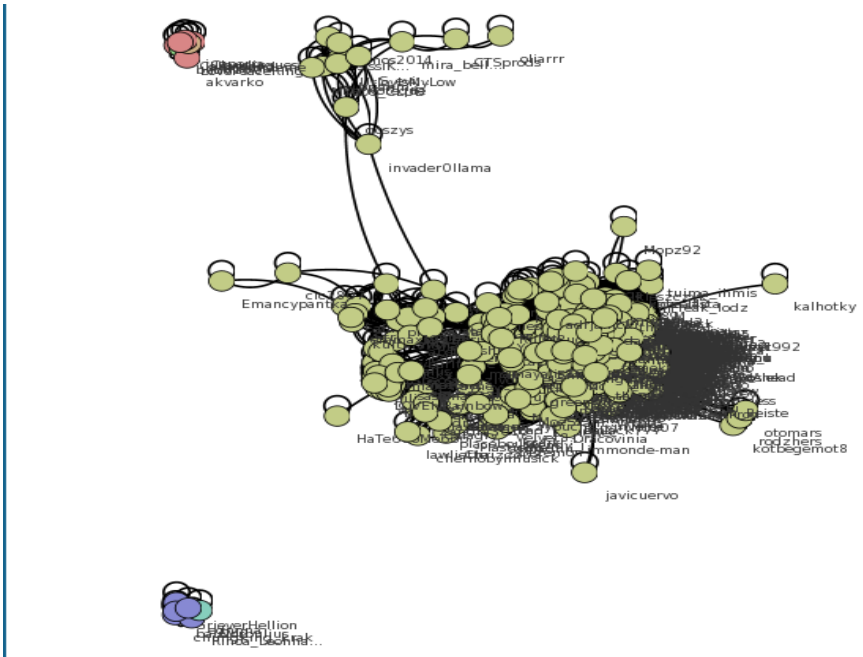
Rys 7: BC



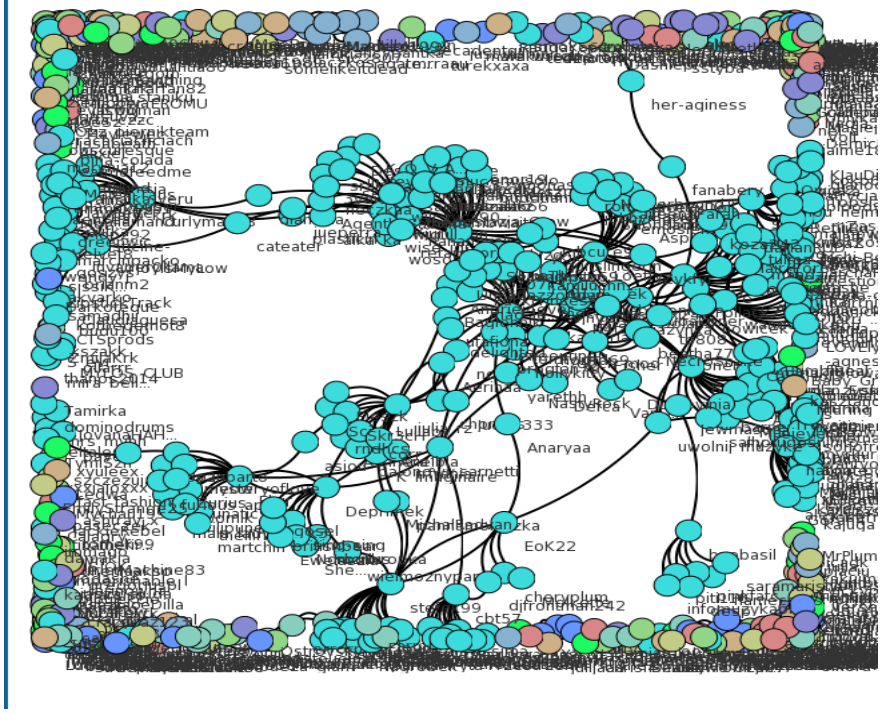
Rys 8: PR

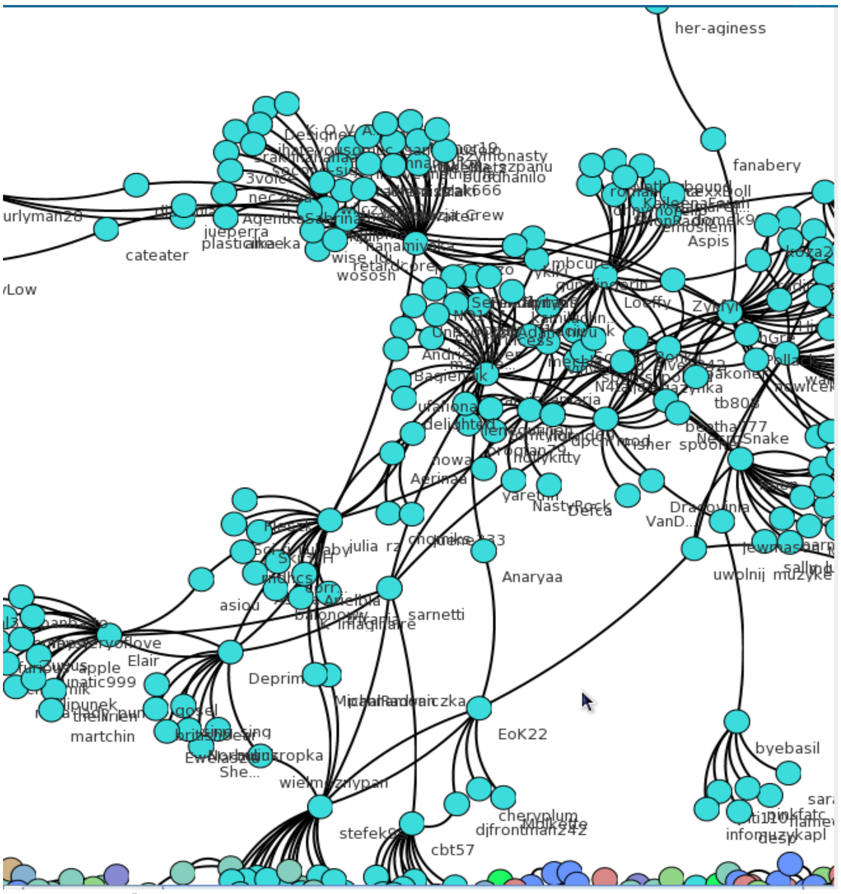
5.5 Porównanie różnych rodzajów powiązań

- **Events**

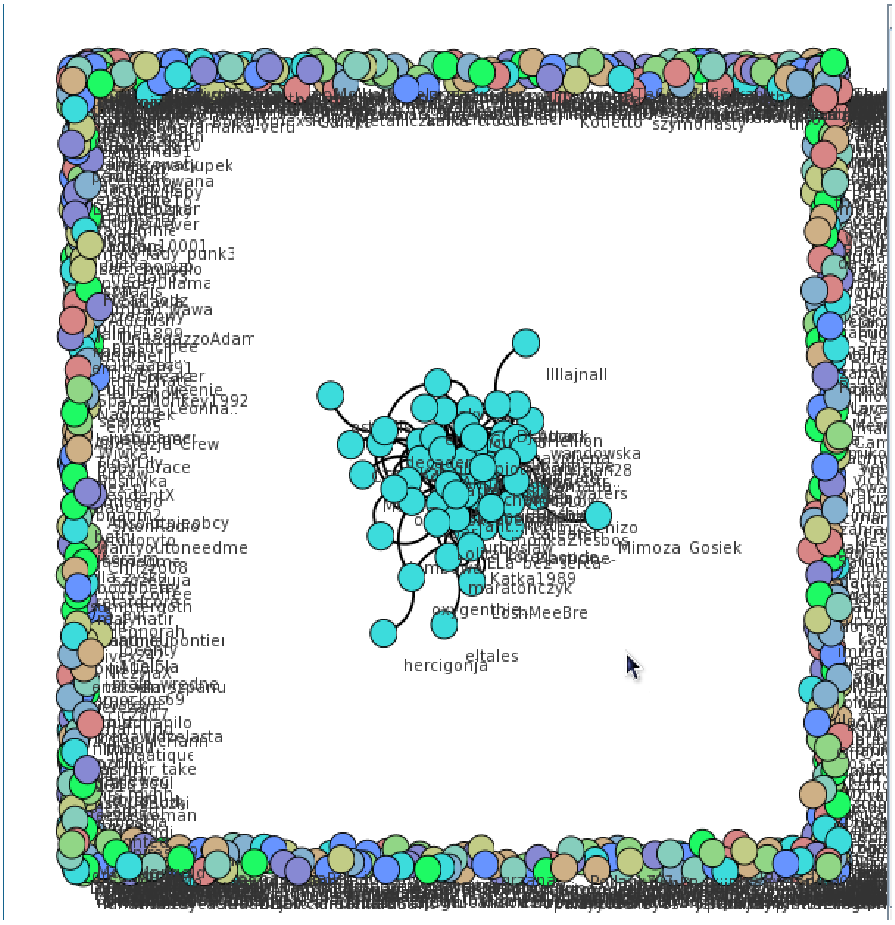

akvarko
invader0llama
Mopz92
Emancypantka
kalhotky
javicuervo

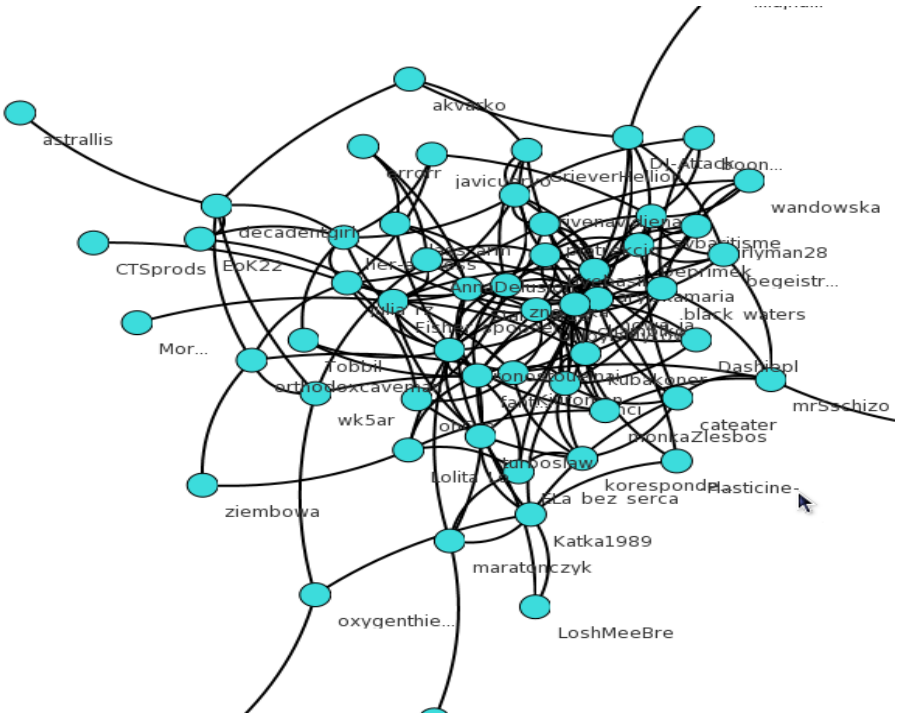
- **Friends**


her-aginess
Anaryaa
EoK22
cateater

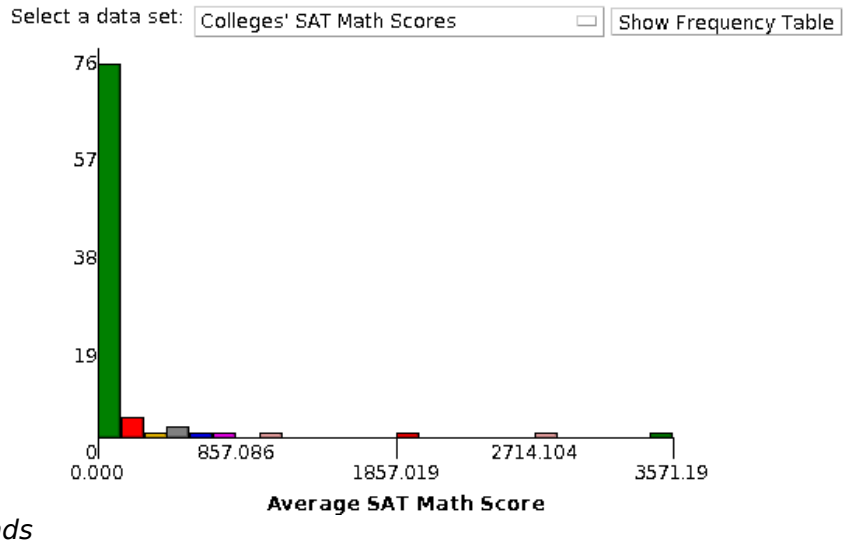

her-aginess
fanabery
cateater
Aspis
tb808
Anaryaa
sarnetti
martchin
EoK22
byebasil
cbt57

- **Loved**

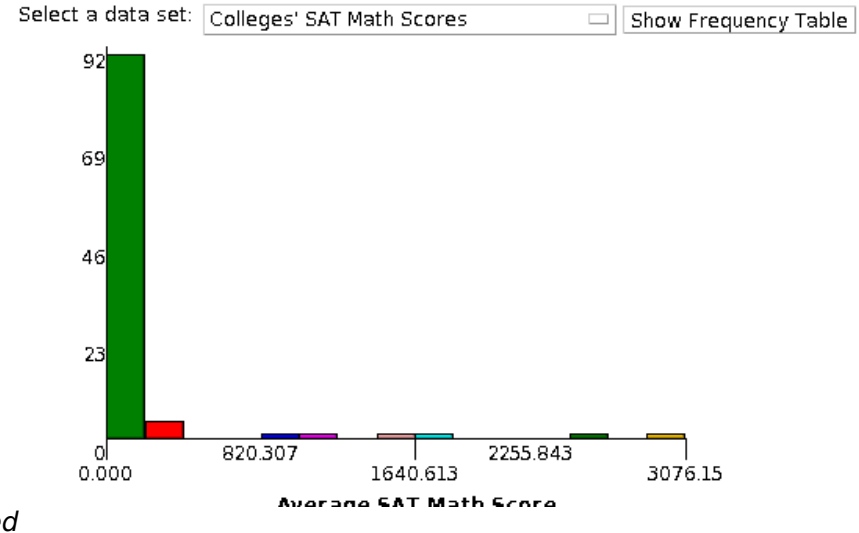

Illlajnall
Mimoza Gosiek
Katka1989
maratonczyk
hercigonja
eltales



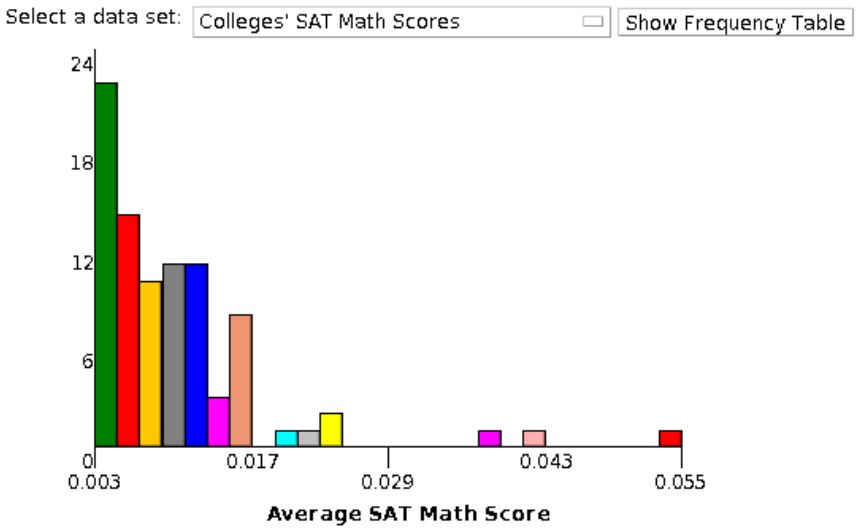
- **Friends vs Loved**



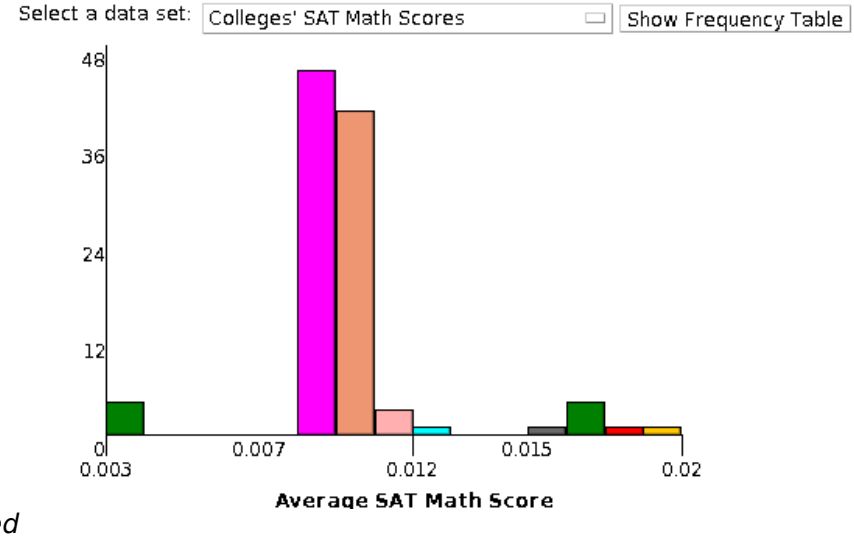
Rys 1: BC_Friends



Rys 2: BC_Loved



Rys 3: PR_Friends



Rys 4: PR_Loved

*Pokrycie:**0.7058823529411765***

6 Wnioski

7 Dokumentacja kodu

7.1 Pakiet analysis

W tej części projektu znajdują się kod umożliwiający tworzenie oraz analizowanie sieci społecznych.

- AnalysisHelper – *klasa wspomagająca wyszukiwanie części wspólnych społeczności*
  - ExtractSolidCommunities – *metoda zwracająca części wspólne dwóch wyników klastrowania*
  - ExtractSolidCommunitiesFactor – *służy do określania współczynnika pokrycia*

- EvParams – *klasa opisująca parametry okresy klastrowania koncertów*

- GraphFactory – *klasa zwierająca metody umożliwiające tworzenie grafów różnego rodzaju oraz raportowanie wyników*
  - CreateEventsGraph – *tworzenie grafu powiązań koncertowych, można określić liczbę użytkowników oraz parametr okresu poprzez klasę EvParams*
  - CreateFriendsGraph – *tworzenia grafu sieci na podstawie informacji o znajomych*
  - CreateLovedGraph – *tworzenie grafu na podstawie ulubionych utworów użytkowników*
  - Report – *generowanie raportu opisującego każdy klaster i każdego użytkownika*

- MathHelper – *Klasa zawiera metody pomocnicze do obliczania średniej oraz odchylenia standardowego.*

- TextModeAnalyzer – *Klasa służąca do uruchamiana analiz w trybie tekstowym.*

- UserLabeller – *Klasa która etykietuje użytkowników w wizualizacji*

- VisualAnalyzer – *Analizator w trybie graficznym*

7.2 Pakiet crawler

Pakiet ten zawiera klasy służące do pobierania danych z serwisu Last.fm.

- Crawler – *zawiera informacje o kluczu API last.fm*
- EventCrawler – *pobiera koncerty oraz ich uczestników*
- LovedCrawler – *pobiera ulubione utwory użytkowników*
- UserCrawler – pobiera znajomych

7.3 Pakiet hiberex

Klasy pakietu hiberex odpowiedzialne są za mapowanie danych z bazy danych.

- AdditionalFunc – *klasa zawierająca metody uproszczające zapytania do bazy danych*

- SessionFactoryUtil **-** *fabryka sesji, niezbędna do działania hibernate*

- Klasy odpowiadające rekordom w bazie danych
  - Shout
  - User
  - Artist
  - Pair
  - Tag
  - Venue
  - Event
  - Playlist
  - Group
  - Track

Bibliografia

[1] „Community structure in social and biological networks” http://www.pnas.org/content/99/12/7821.full.pdf
[2] „JUNG” http://jung.sourceforge.net/