Eksploracja Danych 2014/2015

Porównanie metod walidacji znajdowania grup w sieciach społecznych

SPRAWOZDANIE

Autorzy:
Radosław Trzcionkowski,
Łukasz Szczygłowski

Zawartość

1. Opis problemu 3
2. Opis zastosowanych metodyk i algorytmów 4
a. Zaimplementowane sposoby obliczania odległości 4
b. Zaimplementowane kryteria 4
c. Zaimplementowane sposoby wyliczania ewaluacji zewnętrznej 4
d. Zastosowane sposoby obliczania korelacji wektorów 4
3. Opis wykorzystanych danych 5
4. Przeprowadzone eksperymenty oraz wyniki 7
a. Eksperyment porównawczy kryterium 7
b. Eksperyment porównawczy algorytmów klastrujących 11
5. Wnioski oraz podobieństwa z artykułem źródłowym 18
6. Podsumowanie 19
7. Źródła 19

1. Opis problemu

Obserwowalny jest wyraźny trend rozwoju algorytmów znajdujących grupy w sieciach społecznościowych. Ich początkowe formy prostych heurystyk z dnia na dzień wyewoluowały na bardziej wyrafinowane, zorientowane na optymalizację zadanego kryterium, których wyniki są coraz dokładniejsze. Pomimo tak dobrze rozwijanych metodyk, problemem stała się bardzo uboga gama sposobów weryfikacji klastrowań uzyskanych przez wszelakie algorytmy. Prawdziwy problem pojawił się w momencie, kiedy dla zadanej, sklastrowanej różnym sposobem sieci, nie było znane jej optymalne, rzeczywiste podzielenie. Nie było możliwości bezpośredniego porównania dwóch sposobów podziału, co w efekcie wymusiło poszukiwanie sposobów, swoistych algorytmów, służących do określenia jakości pogrupowania zadanej sieci społecznej.

Celem heurystyk stojących za algorytmem klastrujących jest wyszukiwanie najbardziej satysfakcjonującego rozwiązania względem zadanego kryterium stopu. Większość obecnych algorytmów klastrujących bazuje swoją implementację na jednym ze znanych sposobów weryfikacji grupowania, modularności. Metoda ta ma podstawowy problem związany z limitem rozdzielczości, który w konsekwencji powoduje, że wyniki uzyskane podczas grupowania sprowadzają się do dużej ilości małych społeczności. Równolegle rozwijająca się gałąź uczenia maszynowego pokazała inne, również skuteczne metody walidacji znalezionych społeczności w sieci, takich jak: indeks Davies-Bouldin oraz Silhouette. Problem na tym etapie sprowadza się do znalezienia takiej metody walidacji grupowań, która - obok modularności - mogłaby również służyć jako fundament budowy przyszłych algorytmów klastrujących.

2. Opis zastosowanych metodyk i algorytmów

W eksperymencie wykorzystane zostały różne podejścia definiowania kryterium dla grupowania zadanej sieci społecznościowej. Wszystkie z nich, za wyjątkiem modularności Q, wymagają określenia sposobu wyznaczania odległości pomiędzy dwoma węzłami sieci. Poniżej przedstawione zostaną zaimplementowane w projekcie sposoby obliczania odległości, a następnie zaimplementowane kryteria.

a. Zaimplementowane sposoby obliczania odległości

1) Adjacency Relation Distance (ARD)

2) Edge Path Distance

3) Neighbour Overlap Distance (NOD)

4) Pearson Correlation Distance (PCD)

b. Zaimplementowane kryteria

1) C Index Criteria

2) Davies Bouldin Criteria

3) Dunn Index

4) Modularity Q

5) PBM Criteria

6) Point Biserial Criteria

7) Silhouette Width Criteria (wariant SWC 2)

8) Variance Ratio Criteria

9) Z Statistics Criteria

c. Zaimplementowane sposoby wyliczania ewaluacji zewnętrznej

1) Jaccard Coefficient

d. Zastosowane sposoby obliczania korelacji wektorów

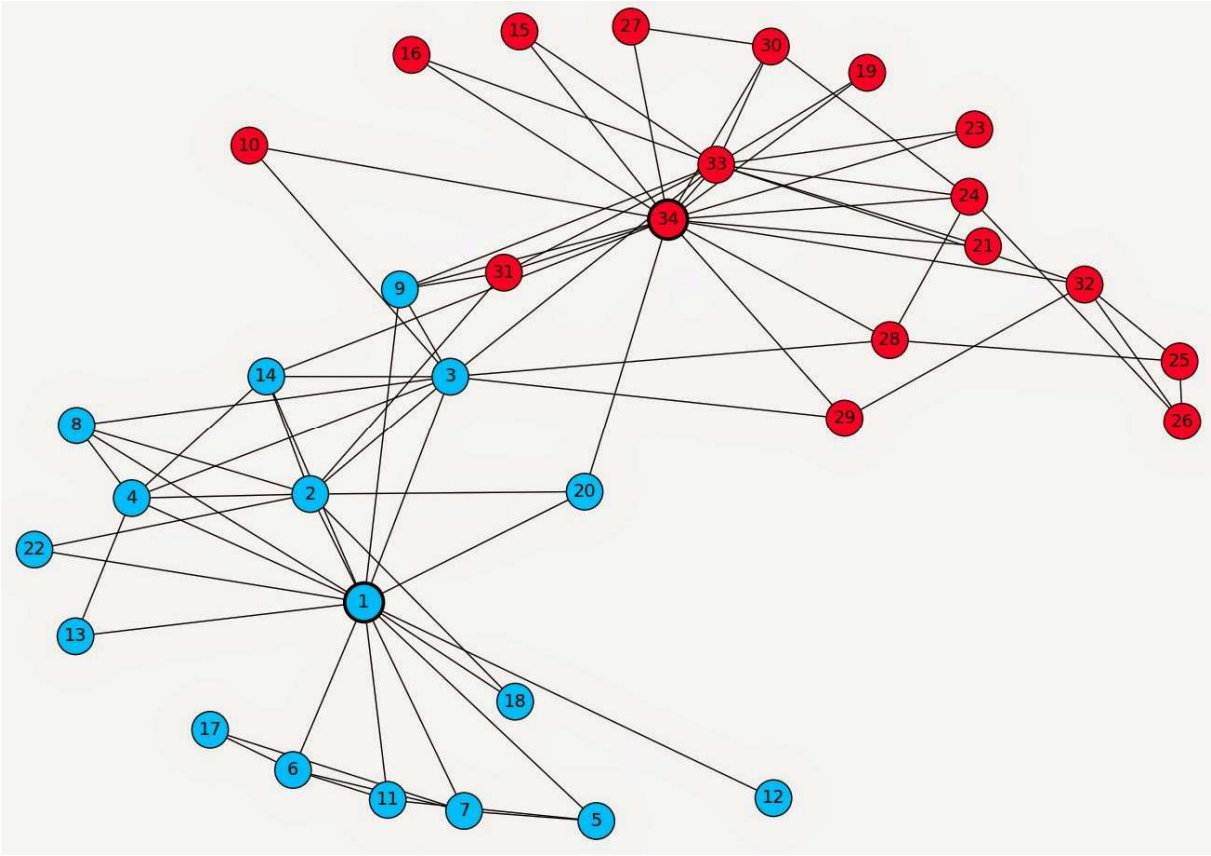
1) Spearman Correlation

2) Pearson Correlation

3. Opis wykorzystanych danych

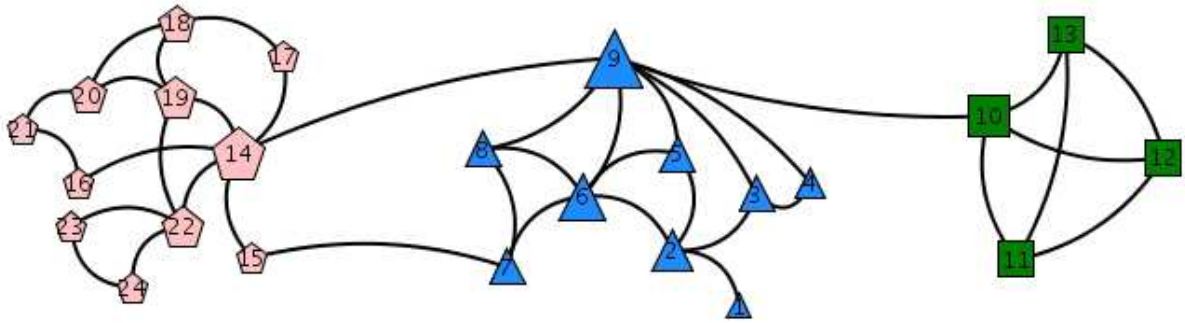
W eksperymencie zostały wykorzystane, analogicznie do bazowego artykułu, 3 sieci społecznościowe:

Zahary's Weightened Karate Dataset



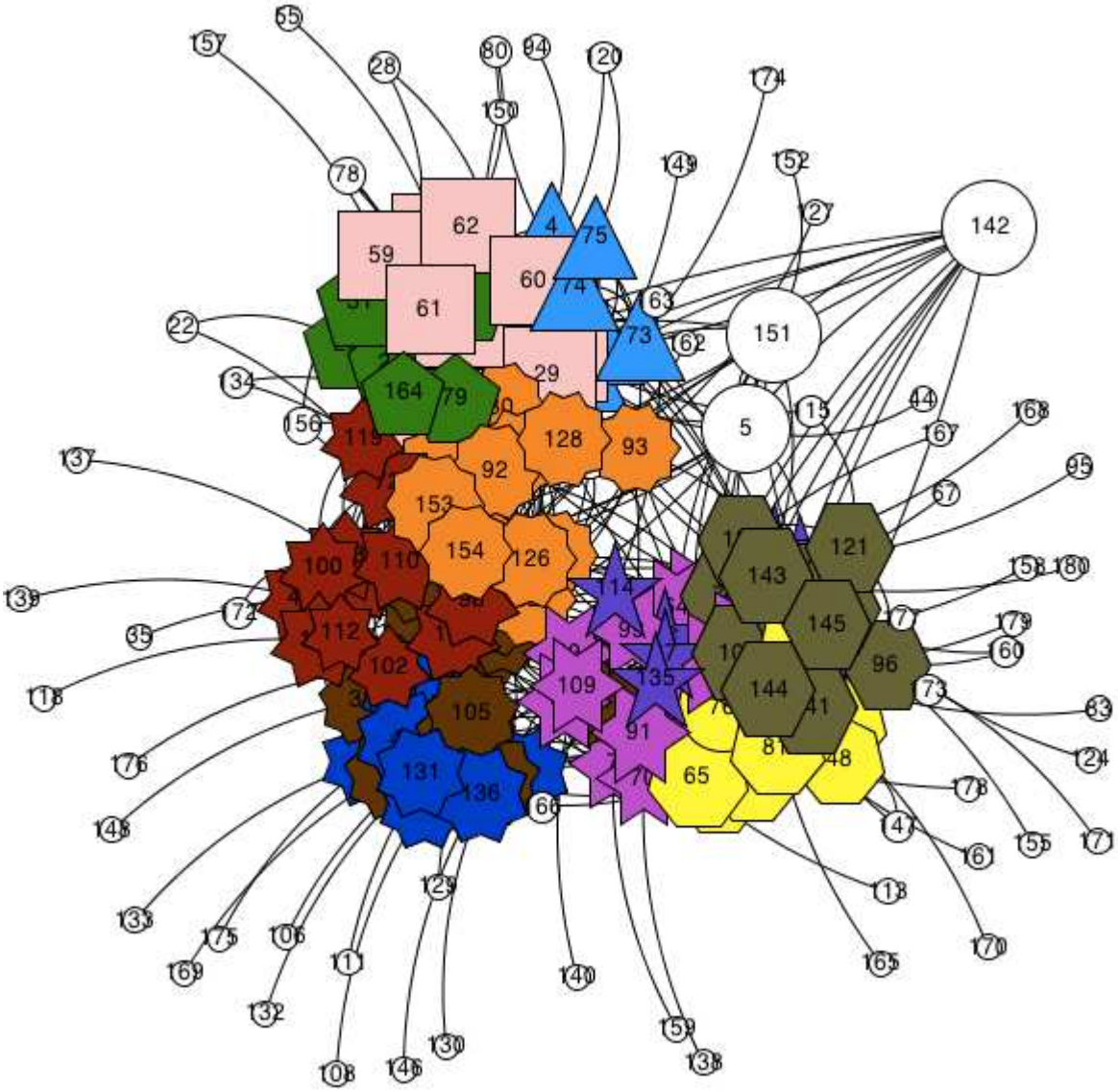
http://spaghetti-os.blogspot.com/2014/05/zacharys-karate-club.html

Sawmill Strike Dataset



http://webdocs.cs.ualberta.ca/~rabbanyk/TopLeader/strike.jpg

NCAA Football Bowl Subdivision Network



http://webdocs.cs.ualberta.ca/~rabbanyk/TopLeader/footbal.png

4. Przeprowadzone eksperymenty oraz wyniki

a. Eksperyment porównawczy kryterium

Eksperyment został przeprowadzony w analogiczny sposób jak w artykule bazowym z zastosowanymi modyfikacjami w zakresie generowania zestawu domyślnych klastrowań. Proces eksperymentu rozpoczynał się od doboru testowanych zbiorów danych. Wykorzystane zostały zbiory przedstawione w punkcie 3. Dla każdego ze zbiorów generowany był zbiór N losowych klastrowań tworzony jako modyfikacja stopnia K rzeczywistego pogrupowania sieci ("ground truth"). Następnie dla każdego z testowanych kryteriów, w każdym wariancie sposobu obliczania odległości - jeśli było to konieczne - obliczana była jego wartość dla poszczególnego z losowych pogrupowań. W ten sposób powstawał pierwszy wektor danych eksperymentalnych. Następnie obliczany był wektor korelacji tworzony na podstawie wartości ewaluacji zewnętrznej każdego z losowych klastrowań z rzeczywistym klastrowaniem ("ground truth"). W ten sposób dla każdego z kryterium uzyskiwane były 2 wektory: E - wektor ewaluacji zewnętrznej oraz wektor I - wektor wartości danego kryterium dla wszystkich losowych klastrowań. Następnie celem wyznaczenia poprawności zachowania kryterium obliczana była korelacja obu tych wektorów i w ten sposób uzyskiwana została ocena końcowa dla zadanego kryterium.

Istotne było wydzielenie 3 różnych przypadków losowego klastrowania zbiorów względem rzeczywistego pogrupowania: pesymistyczne, optymistyczne oraz średnie. Przypadki określane były średnią wartością ewaluacji zewnętrznej dla wszystkich losowych klastrowań składowych. W czasie eksperymentu można było to uzyskiwać odpowiednio modyfikując stopnień K modyfikacji rzeczywistego pogrupowania sieci (definiując na przykład, że co drugi węzeł z klastra był losowo przerzucany do innego).

Zaimplementowany zestaw 9 kryterium oraz 4 metryk tworzą zestaw testowy 33 permutacji (ponieważ modularność nie wykorzystuje metryki) kryterium, które były testowane podczas eksperymentu.

Poniżej przedstawione zostały wyniki eksperymentów dla trzech, hermetycznych przypadków. Ewaluacja zewnętrzna została wyliczona za pomocą formuły Jaccard'a. Korelacja między wektorami testowymi była liczona według wzorów Spearman'a i Pearson'a.

Przypadek pesymistyczny					
Spearman			Pearson		
1	C Index	Adjacency Relation Distance	1	C Index	Adjacency Relation Distance
2	C Index	Pearson Correlation Distance	2	C Index	Pearson Correlation Distance
3	SWC 2	Pearson Correlation Distance	3	C Index	Neighbour Overlap Distance
4	C Index	Neighbour Overlap Distance	4	SWC 2	Pearson Correlation Distance
5	Modularity Q		5	Modularity Q	
6	SWC 2	Neighbour Overlap Distance	6	PBM	Neighbour Overlap Distance
7	PBM	Pearson Correlation Distance	7	PBM	Pearson Correlation Distance
8	C Index	Edge Path Distance	8	C Index	Edge Path Distance
9	Z Statistics	Edge Path Distance	9	Z Statistics	Edge Path Distance
10	PBM	Neighbour Overlap Distance	10	SWC 2	Neighbour Overlap Distance
11	Z Statistics	Neighbour Overlap Distance	11	SWC 2	Adjacency Relation Distance
12	PBM	Adjacency Relation Distance	12	Z Statistics	Neighbour Overlap Distance
13	Variance Ratio	Edge Path Distance	13	PBM	Adjacency Relation Distance
14	SWC 2	Adjacency Relation Distance	14	Dunn Index	Neighbour Overlap Distance
15	Z Statistics	Pearson Correlation Distance	15	Point Biserial	Edge Path Distance
16	Variance Ratio	Adjacency Relation Distance	16	Dunn Index	Adjacency Relation Distance
17	Davies Bouldin	Adjacency Relation Distance	17	Davies Bouldin	Pearson Correlation Distance
18	PBM	Edge Path Distance	18	Davies Bouldin	Neighbour Overlap Distance
19	Point Biserial	Adjacency Relation Distance	19	Variance Ratio	Edge Path Distance
20	Davies Bouldin	Pearson Correlation Distance	20	Dunn Index	Pearson Correlation Distance
21	Point Biserial	Edge Path Distance	21	Z Statistics	Pearson Correlation Distance
22	Point Biserial	Neighbour Overlap Distance	22	Point Biserial	Adjacency Relation Distance
23	Dunn Index	Neighbour Overlap Distance	23	Variance Ratio	Pearson Correlation Distance
24	Davies Bouldin	Edge Path Distance	24	Variance Ratio	Adjacency Relation Distance
25	Dunn Index	Adjacency Relation Distance	25	Variance Ratio	Neighbour Overlap Distance
26	Variance Ratio	Pearson Correlation Distance	26	Point Biserial	Neighbour Overlap Distance
27	Point Biserial	Pearson Correlation Distance	27	Point Biserial	Pearson Correlation Distance
28	Davies Bouldin	Neighbour Overlap Distance	28	Davies Bouldin	Adjacency Relation Distance
29	Variance Ratio	Neighbour Overlap Distance	29	Z Statistics	Adjacency Relation Distance
30	Z Statistics	Adjacency Relation Distance	30	SWC 2	Edge Path Distance
31	Dunn Index	Pearson Correlation Distance	31	PBM	Edge Path Distance
32	SWC 2	Edge Path Distance	32	Davies Bouldin	Edge Path Distance
33	Dunn Index	Edge Path Distance	33	Dunn Index	Edge Path Distance

Przypadek średni					
Spearman			Pearson		
1	C Index	Adjacency Relation Distance	1	C Index	Adjacency Relation Distance
2	C Index	Pearson Correlation Distance	2	C Index	Pearson Correlation Distance
3	C Index	Neighbour Overlap Distance	3	C Index	Neighbour Overlap Distance
4	Modularity Q		4	Modularity Q	
5	PBM	Pearson Correlation Distance	5	Z Statistics	Edge Path Distance
6	PBM	Neighbour Overlap Distance	6	C Index	Edge Path Distance
7	Z Statistics	Edge Path Distance	7	PBM	Pearson Correlation Distance
8	C Index	Edge Path Distance	8	PBM	Neighbour Overlap Distance
9	SWC 2	Pearson Correlation Distance	9	Davies Bouldin	Neighbour Overlap Distance
10	Davies Bouldin	Neighbour Overlap Distance	10	SWC 2	Pearson Correlation Distance
11	Point Biserial	Edge Path Distance	11	Z Statistics	Pearson Correlation Distance
12	Z Statistics	Pearson Correlation Distance	12	SWC 2	Edge Path Distance
13	Davies Bouldin	Pearson Correlation Distance	13	Point Biserial	Edge Path Distance
14	Z Statistics	Adjacency Relation Distance	14	Davies Bouldin	Pearson Correlation Distance
15	SWC 2	Neighbour Overlap Distance	15	Z Statistics	Adjacency Relation Distance
16	SWC 2	Edge Path Distance	16	Variance Ratio	Pearson Correlation Distance
17	SWC 2	Adjacency Relation Distance	17	Dunn Index	Neighbour Overlap Distance
18	Variance Ratio	Pearson Correlation Distance	18	SWC 2	Adjacency Relation Distance
19	Dunn Index	Neighbour Overlap Distance	19	Point Biserial	Adjacency Relation Distance
20	Dunn Index	Adjacency Relation Distance	20	Dunn Index	Adjacency Relation Distance
21	Point Biserial	Adjacency Relation Distance	21	Z Statistics	Neighbour Overlap Distance
22	Variance Ratio	Edge Path Distance	22	SWC 2	Neighbour Overlap Distance
23	PBM	Adjacency Relation Distance	23	Dunn Index	Pearson Correlation Distance
24	Z Statistics	Neighbour Overlap Distance	24	Point Biserial	Pearson Correlation Distance
25	Davies Bouldin	Edge Path Distance	25	Variance Ratio	Neighbour Overlap Distance
26	Point Biserial	Pearson Correlation Distance	26	Variance Ratio	Edge Path Distance
27	Variance Ratio	Neighbour Overlap Distance	27	Point Biserial	Neighbour Overlap Distance
28	Davies Bouldin	Adjacency Relation Distance	28	PBM	Adjacency Relation Distance
29	Dunn Index	Pearson Correlation Distance	29	Variance Ratio	Adjacency Relation Distance
30	PBM	Edge Path Distance	30	Davies Bouldin	Adjacency Relation Distance
31	Point Biserial	Neighbour Overlap Distance	31	PBM	Edge Path Distance
32	Variance Ratio	Adjacency Relation Distance	32	Davies Bouldin	Edge Path Distance
33	Dunn Index	Edge Path Distance	33	Dunn Index	Edge Path Distance

Przypadek optymistyczny					
Spearman			Pearson		
1	C Index	Neighbour Overlap Distance	1	C Index	Neighbour Overlap Distance
2	PBM	Pearson Correlation Distance	2	PBM	Pearson Correlation Distance
3	PBM	Neighbour Overlap Distance	3	C Index	Adjacency Relation Distance
4	C Index	Pearson Correlation Distance	4	C Index	Pearson Correlation Distance
5	Modularity Q		5	Davies Bouldin	Pearson Correlation Distance
6	C Index	Adjacency Relation Distance	6	PBM	Neighbour Overlap Distance
7	Z Statistics	Edge Path Distance	7	Z Statistics	Edge Path Distance
8	C Index	Edge Path Distance	8	C Index	Edge Path Distance
9	Davies Bouldin	Pearson Correlation Distance	9	Modularity Q	
10	SWC 2	Neighbour Overlap Distance	10	SWC 2	Edge Path Distance
11	SWC 2	Edge Path Distance	11	SWC 2	Pearson Correlation Distance
12	PBM	Adjacency Relation Distance	12	SWC 2	Neighbour Overlap Distance
13	Dunn Index	Pearson Correlation Distance	13	Variance Ratio	Pearson Correlation Distance
14	SWC 2	Adjacency Relation Distance	14	SWC 2	Adjacency Relation Distance
15	Dunn Index	Adjacency Relation Distance	15	PBM	Adjacency Relation Distance
16	Davies Bouldin	Adjacency Relation Distance	16	Dunn Index	Adjacency Relation Distance
17	Variance Ratio	Pearson Correlation Distance	17	Point Biserial	Edge Path Distance
18	Variance Ratio	Adjacency Relation Distance	18	Variance Ratio	Adjacency Relation Distance
19	Davies Bouldin	Neighbour Overlap Distance	19	Davies Bouldin	Neighbour Overlap Distance
20	SWC 2	Pearson Correlation Distance	20	Z Statistics	Pearson Correlation Distance
21	Point Biserial	Edge Path Distance	21	Dunn Index	Pearson Correlation Distance
22	Z Statistics	Pearson Correlation Distance	22	Davies Bouldin	Adjacency Relation Distance
23	Dunn Index	Neighbour Overlap Distance	23	Z Statistics	Neighbour Overlap Distance
24	Z Statistics	Neighbour Overlap Distance	24	Dunn Index	Neighbour Overlap Distance
25	Variance Ratio	Neighbour Overlap Distance	25	Z Statistics	Adjacency Relation Distance
26	Z Statistics	Adjacency Relation Distance	26	Variance Ratio	Neighbour Overlap Distance
27	Variance Ratio	Edge Path Distance	27	Point Biserial	Adjacency Relation Distance
28	PBM	Edge Path Distance	28	Point Biserial	Pearson Correlation Distance
29	Point Biserial	Adjacency Relation Distance	29	Point Biserial	Neighbour Overlap Distance
30	Point Biserial	Pearson Correlation Distance	30	PBM	Edge Path Distance
31	Point Biserial	Neighbour Overlap Distance	31	Variance Ratio	Edge Path Distance
32	Davies Bouldin	Edge Path Distance	32	Davies Bouldin	Edge Path Distance
33	Dunn Index	Edge Path Distance	33	Dunn Index	Edge Path Distance

b. Eksperyment porównawczy algorytmów klastrujących

Eksperyment porównawczy algorytmów klastrujących został utworzony analogicznie jak eksperyment porównania kryterium z modyfikacją etapu tworzenia klastrowań - tym razem tworzone było N klastrowań za pomocą danego algorytmu, a nie losowego przerzucania węzłów pomiędzy klastrami względem rzeczywistego klastrowania. Podczas prowadzonego eksperymentu część z kryterium. Proces był prowadzony dla jednego ze zbiorów danych - Karate.

Porównane zostały następujące algorytmy klastrujące:
Z biblioteki Jung: Bicomponent Clustering, Edge Betweenness Clustering, Voltage Clustering
Z biblioteki JavaML: Density Based Spatial Clustering, KNode Clustering, Self Organizing Maps Clustering

Ze wszystkich uzyskanych danych, jako kryterium decydujące został wybrany C-Index - metryka, która wypadła najlepiej w poprzednim eksperymencie.

Name	Jaccard Ex. Evaluation	C-Index NOD	C-Index ARD	C-Index PCD
Jung Bicomponent	0.4605089949506484	0.18064609786610283	0.4985120814096288	0.2258478458692724
Jung Edge Betweenness	0.4866310160427806	0.3	0.4285714285714287	-0.2000000000000001
Jung Voltage Clustering	0.5187105584220729	0.2475895522660684	0.43187310694120346	0.2789870794384364
JavaML Density Based Spatial	0.4526717087868168	0.262039093595844	0.558564512088685	0.3547832539308114
JavaML Knode	0.495616880650531	0.16444965781240056	0.13367560477810442	0.19711628953958574
JavaML Self Organizing Maps	0.6724051154863459	0.22555020106813498	0.3239879888281027	0.24271250242535564

Poniżej przedstawione zostały całościowe wyniki eksperymentu:

Jung Bicomponent Clusterer		
[Jaccard] Jaccard Coefficient External Evaluation: 0.4605089949506484		
Modularity Q Criteria	Neighbour Overlap Distance	0.10046756994808943
Modularity Q Criteria	Edge Path Distance	0.10046756994808943
Modularity Q Criteria	Adjacency Relation Distance	0.10046756994808943
Modularity Q Criteria	Pearson Correlation Distance	0.10046756994808943
C Index Criteria	Neighbour Overlap Distance	0.18064609786610283
C Index Criteria	Edge Path Distance	0.916919457375983
C Index Criteria	Adjacency Relation Distance	0.4985120814096288
C Index Criteria	Pearson Correlation Distance	0.2258478458692724
Davies Bouldin Criteria	Neighbour Overlap Distance	1.1022267881136072
Davies Bouldin Criteria	Edge Path Distance	Infinity
Davies Bouldin Criteria	Adjacency Relation Distance	1.4570405169471183
Davies Bouldin Criteria	Pearson Correlation Distance	0.9654637132956236
Silhouette Width Criteria 2	Neighbour Overlap Distance	-0.027344210837022036
Silhouette Width Criteria 2	Edge Path Distance	-0.9399403239556694
Silhouette Width Criteria 2	Adjacency Relation Distance	-0.15745426903474244
Silhouette Width Criteria 2	Pearson Correlation Distance	-0.09474537669292894
Variance Ratio Criteria	Neighbour Overlap Distance	19.239870629476638
Variance Ratio Criteria	Edge Path Distance	0.5
Variance Ratio Criteria	Adjacency Relation Distance	3.4789342341615765
Variance Ratio Criteria	Pearson Correlation Distance	12.6025350107332
PBM Criteria	Neighbour Overlap Distance	0.014039332551747304
PBM Criteria	Edge Path Distance	0.016666666666666666
PBM Criteria	Adjacency Relation Distance	0.013773725298722669
PBM Criteria	Pearson Correlation Distance	0.01756247213035195
Dunn Index Criteria	Neighbour Overlap Distance	0.7574074074074079
Dunn Index Criteria	Edge Path Distance	5.562684646268003E-309
Dunn Index Criteria	Adjacency Relation Distance	0.2548235957188129
Dunn Index Criteria	Pearson Correlation Distance	0.3292723778722483
Z-Statistics Criteria	Neighbour Overlap Distance	390.23776422909236
Z-Statistics Criteria	Edge Path Distance	32.89135968974835
Z-Statistics Criteria	Adjacency Relation Distance	316.6988678293565
Z-Statistics Criteria	Pearson Correlation Distance	194.52756369360802
Point Biserial Criteria	Neighbour Overlap Distance	55.63821204182208
Point Biserial Criteria	Edge Path Distance	50.91231568412392
Point Biserial Criteria	Adjacency Relation Distance	59.291096242048255
Point Biserial Criteria	Pearson Correlation Distance	46.487156918183686

Jung Edge Betweenness Clusterer		
[Jaccard] Jaccard Coefficient External Evaluation: 0.4866310160427806		
Modularity Q Criteria	Neighbour Overlap Distance	0.025552369708213844
Modularity Q Criteria	Edge Path Distance	0.025552369708213844
Modularity Q Criteria	Adjacency Relation Distance	0.025552369708213844
Modularity Q Criteria	Pearson Correlation Distance	0.025552369708213844
C Index Criteria	Neighbour Overlap Distance	0.3
C Index Criteria	Edge Path Distance	0.5
C Index Criteria	Adjacency Relation Distance	0.4285714285714287
C Index Criteria	Pearson Correlation Distance	-0.2000000000000001
Davies Bouldin Criteria	Neighbour Overlap Distance	4.9E-324
Davies Bouldin Criteria	Edge Path Distance	4.9E-324
Davies Bouldin Criteria	Adjacency Relation Distance	4.9E-324
Davies Bouldin Criteria	Pearson Correlation Distance	4.9E-324
Silhouette Width Criteria 2	Neighbour Overlap Distance	0.029411764705882356
Silhouette Width Criteria 2	Edge Path Distance	0.029411764705882356
Silhouette Width Criteria 2	Adjacency Relation Distance	0.029411764705882356
Silhouette Width Criteria 2	Pearson Correlation Distance	0.029411764705882356
Variance Ratio Criteria	Neighbour Overlap Distance	FAILED
Variance Ratio Criteria	Edge Path Distance	FAILED
Variance Ratio Criteria	Adjacency Relation Distance	FAILED
Variance Ratio Criteria	Pearson Correlation Distance	FAILED
PBM Criteria	Neighbour Overlap Distance	0.0
PBM Criteria	Edge Path Distance	1.668805393880401E-308
PBM Criteria	Adjacency Relation Distance	0.0
PBM Criteria	Pearson Correlation Distance	0.0
Dunn Index Criteria	Neighbour Overlap Distance	FAILED
Dunn Index Criteria	Edge Path Distance	FAILED
Dunn Index Criteria	Adjacency Relation Distance	FAILED
Dunn Index Criteria	Pearson Correlation Distance	FAILED
Z-Statistics Criteria	Neighbour Overlap Distance	609.0603919263048
Z-Statistics Criteria	Edge Path Distance	36.07588943941051
Z-Statistics Criteria	Adjacency Relation Distance	479.01752235192373
Z-Statistics Criteria	Pearson Correlation Distance	315.0580627947995
Point Biserial Criteria	Neighbour Overlap Distance	0.0
Point Biserial Criteria	Edge Path Distance	0.0
Point Biserial Criteria	Adjacency Relation Distance	0.0
Point Biserial Criteria	Pearson Correlation Distance	0.0

Jung Voltage Clusterer		
[Jaccard] Jaccard Coefficient External Evaluation: 0.5187105584220729		
Modularity Q Criteria	Neighbour Overlap Distance	0.12809729952587093
Modularity Q Criteria	Edge Path Distance	0.12809729952587093
Modularity Q Criteria	Adjacency Relation Distance	0.12809729952587093
Modularity Q Criteria	Pearson Correlation Distance	0.12809729952587093
C Index Criteria	Neighbour Overlap Distance	0.2475895522660684
C Index Criteria	Edge Path Distance	0.8263149947000463
C Index Criteria	Adjacency Relation Distance	0.43187310694120346
C Index Criteria	Pearson Correlation Distance	0.2789870794384364
Davies Bouldin Criteria	Neighbour Overlap Distance	1.392261098430386
Davies Bouldin Criteria	Edge Path Distance	Infinity
Davies Bouldin Criteria	Adjacency Relation Distance	2.307484603443498
Davies Bouldin Criteria	Pearson Correlation Distance	1.1875274297517227
Silhouette Width Criteria 2	Neighbour Overlap Distance	-0.004260816262939314
Silhouette Width Criteria 2	Edge Path Distance	-0.7537472707243592
Silhouette Width Criteria 2	Adjacency Relation Distance	-0.008712130142009744
Silhouette Width Criteria 2	Pearson Correlation Distance	-0.024243677206292433
Variance Ratio Criteria	Neighbour Overlap Distance	28.65923300477113
Variance Ratio Criteria	Edge Path Distance	8.274223223783693E-306
Variance Ratio Criteria	Adjacency Relation Distance	10.490549192511583
Variance Ratio Criteria	Pearson Correlation Distance	28.44695419805719
PBM Criteria	Neighbour Overlap Distance	0.019660098204090756
PBM Criteria	Edge Path Distance	3.80249229034177E-309
PBM Criteria	Adjacency Relation Distance	0.013509172937091585
PBM Criteria	Pearson Correlation Distance	0.02323102775825919
Dunn Index Criteria	Neighbour Overlap Distance	0.5061823361823361
Dunn Index Criteria	Edge Path Distance	5.562684646268003E-309
Dunn Index Criteria	Adjacency Relation Distance	0.15921612870199006
Dunn Index Criteria	Pearson Correlation Distance	0.18740978296128236
Z-Statistics Criteria	Neighbour Overlap Distance	346.8773762962795
Z-Statistics Criteria	Edge Path Distance	29.52579699323996
Z-Statistics Criteria	Adjacency Relation Distance	272.93485473967434
Z-Statistics Criteria	Pearson Correlation Distance	172.19629018537836
Point Biserial Criteria	Neighbour Overlap Distance	29.00668096788241
Point Biserial Criteria	Edge Path Distance	40.113261178543624
Point Biserial Criteria	Adjacency Relation Distance	27.75077753608798
Point Biserial Criteria	Pearson Correlation Distance	20.57098804580455

JavaML Density Based Spatial Clustering		
[Jaccard] Jaccard Coefficient External Evaluation: 0.4526717087868168		
Modularity Q Criteria	Neighbour Overlap Distance	0.021675221126581313
Modularity Q Criteria	Edge Path Distance	0.021675221126581313
Modularity Q Criteria	Adjacency Relation Distance	0.021675221126581313
Modularity Q Criteria	Pearson Correlation Distance	0.021675221126581313
C Index Criteria	Neighbour Overlap Distance	0.262039093595844
C Index Criteria	Edge Path Distance	0.8492958328170821
C Index Criteria	Adjacency Relation Distance	0.558564512088685
C Index Criteria	Pearson Correlation Distance	0.3547832539308114
Davies Bouldin Criteria	Neighbour Overlap Distance	1.1127941772401195
Davies Bouldin Criteria	Edge Path Distance	Infinity
Davies Bouldin Criteria	Adjacency Relation Distance	1.2980264106305408
Davies Bouldin Criteria	Pearson Correlation Distance	0.8386866098842202
Silhouette Width Criteria 2	Neighbour Overlap Distance	-0.016609213100506857
Silhouette Width Criteria 2	Edge Path Distance	-0.7273075301279422
Silhouette Width Criteria 2	Adjacency Relation Distance	-0.03514983343082651
Silhouette Width Criteria 2	Pearson Correlation Distance	-0.04840170272804344
Variance Ratio Criteria	Neighbour Overlap Distance	5.0978434527518806
Variance Ratio Criteria	Edge Path Distance	5.81683752699527E-306
Variance Ratio Criteria	Adjacency Relation Distance	4.234576743879832
Variance Ratio Criteria	Pearson Correlation Distance	5.696712179536347
PBM Criteria	Neighbour Overlap Distance	0.019614075103632243
PBM Criteria	Edge Path Distance	1.650263111726175E-309
PBM Criteria	Adjacency Relation Distance	0.017460327313029173
PBM Criteria	Pearson Correlation Distance	0.023525812774654476
Dunn Index Criteria	Neighbour Overlap Distance	0.6032223871104515
Dunn Index Criteria	Edge Path Distance	9.3947562914749E-310
Dunn Index Criteria	Adjacency Relation Distance	0.24376003116884168
Dunn Index Criteria	Pearson Correlation Distance	0.21459624380632653
Z-Statistics Criteria	Neighbour Overlap Distance	503.46955871547294
Z-Statistics Criteria	Edge Path Distance	27.83821941138085
Z-Statistics Criteria	Adjacency Relation Distance	411.2798888028743
Z-Statistics Criteria	Pearson Correlation Distance	277.5012953846243
Point Biserial Criteria	Neighbour Overlap Distance	85.09624107865456
Point Biserial Criteria	Edge Path Distance	32.721512240961836
Point Biserial Criteria	Adjacency Relation Distance	81.6617565661445
Point Biserial Criteria	Pearson Correlation Distance	82.78286251258804

JavaML KNode Clustering		
[Jaccard] Jaccard Coefficient External Evaluation: 0.495616880650531		
Modularity Q Criteria	Neighbour Overlap Distance	0.03703325880534378
Modularity Q Criteria	Edge Path Distance	0.03703325880534378
Modularity Q Criteria	Adjacency Relation Distance	0.03703325880534378
Modularity Q Criteria	Pearson Correlation Distance	0.03703325880534378
C Index Criteria	Neighbour Overlap Distance	0.16444965781240056
C Index Criteria	Edge Path Distance	0.5844549976285407
C Index Criteria	Adjacency Relation Distance	0.13367560477810442
C Index Criteria	Pearson Correlation Distance	0.19711628953958574
Davies Bouldin Criteria	Neighbour Overlap Distance	1.1335422983649406
Davies Bouldin Criteria	Edge Path Distance	Infinity
Davies Bouldin Criteria	Adjacency Relation Distance	1.049729747166226
Davies Bouldin Criteria	Pearson Correlation Distance	0.9536897948681037
Silhouette Width Criteria 2	Neighbour Overlap Distance	-0.004485760022268611
Silhouette Width Criteria 2	Edge Path Distance	-0.4214964398268393
Silhouette Width Criteria 2	Adjacency Relation Distance	0.007091764324836521
Silhouette Width Criteria 2	Pearson Correlation Distance	-0.011291886027816427
Variance Ratio Criteria	Neighbour Overlap Distance	5.963973383298846
Variance Ratio Criteria	Edge Path Distance	3.8400000000000003
Variance Ratio Criteria	Adjacency Relation Distance	5.849053892822862
Variance Ratio Criteria	Pearson Correlation Distance	7.24247488735247
PBM Criteria	Neighbour Overlap Distance	0.01903750785850099
PBM Criteria	Edge Path Distance	4.406110624662539E305
PBM Criteria	Adjacency Relation Distance	0.030778091073152398
PBM Criteria	Pearson Correlation Distance	0.02329879384851664
Dunn Index Criteria	Neighbour Overlap Distance	0.63421199445618
Dunn Index Criteria	Edge Path Distance	1.020708484299335E-309
Dunn Index Criteria	Adjacency Relation Distance	0.4724121677126524
Dunn Index Criteria	Pearson Correlation Distance	0.36767879767031264
Z-Statistics Criteria	Neighbour Overlap Distance	505.3776222990005
Z-Statistics Criteria	Edge Path Distance	20.73916072443412
Z-Statistics Criteria	Adjacency Relation Distance	373.87650297977484
Z-Statistics Criteria	Pearson Correlation Distance	272.46424185337514
Point Biserial Criteria	Neighbour Overlap Distance	72.45628572158397
Point Biserial Criteria	Edge Path Distance	10.995626375615204
Point Biserial Criteria	Adjacency Relation Distance	51.594122637732426
Point Biserial Criteria	Pearson Correlation Distance	65.78926192806186

JavaML Self Organizing Maps Clustering		
[Jaccard] Jaccard Coefficient External Evaluation: 0.6724051154863459		
Modularity Q Criteria	Neighbour Overlap Distance	0.14105094518991249
Modularity Q Criteria	Edge Path Distance	0.14105094518991249
Modularity Q Criteria	Adjacency Relation Distance	0.14105094518991249
Modularity Q Criteria	Pearson Correlation Distance	0.14105094518991249
C Index Criteria	Neighbour Overlap Distance	0.22555020106813498
C Index Criteria	Edge Path Distance	0.6250526576632385
C Index Criteria	Adjacency Relation Distance	0.3239879888281027
C Index Criteria	Pearson Correlation Distance	0.24271250242535564
Davies Bouldin Criteria	Neighbour Overlap Distance	1.4269828253455654
Davies Bouldin Criteria	Edge Path Distance	Infinity
Davies Bouldin Criteria	Adjacency Relation Distance	2.796286048908365
Davies Bouldin Criteria	Pearson Correlation Distance	2.5186522430574736
Silhouette Width Criteria 2	Neighbour Overlap Distance	-0.008997308056958845
Silhouette Width Criteria 2	Edge Path Distance	-0.6793731631175604
Silhouette Width Criteria 2	Adjacency Relation Distance	-0.07055159848826993
Silhouette Width Criteria 2	Pearson Correlation Distance	-0.03648795844706942
Variance Ratio Criteria	Neighbour Overlap Distance	15.23164284430627
Variance Ratio Criteria	Edge Path Distance	1.9887006968390017E-306
Variance Ratio Criteria	Adjacency Relation Distance	8.649157608140422
Variance Ratio Criteria	Pearson Correlation Distance	18.621652270917544
PBM Criteria	Neighbour Overlap Distance	0.016728882461919194
PBM Criteria	Edge Path Distance	0.057726063829787226
PBM Criteria	Adjacency Relation Distance	0.012264561334440144
PBM Criteria	Pearson Correlation Distance	0.02157046392335171
Dunn Index Criteria	Neighbour Overlap Distance	0.7024086106047137
Dunn Index Criteria	Edge Path Distance	1.140791835393377E-309
Dunn Index Criteria	Adjacency Relation Distance	0.23328003635899983
Dunn Index Criteria	Pearson Correlation Distance	0.3995453059532888
Z-Statistics Criteria	Neighbour Overlap Distance	271.73976122845016
Z-Statistics Criteria	Edge Path Distance	19.88964268316983
Z-Statistics Criteria	Adjacency Relation Distance	203.26084429280385
Z-Statistics Criteria	Pearson Correlation Distance	138.698551214181
Point Biserial Criteria	Neighbour Overlap Distance	-13.363267709961429
Point Biserial Criteria	Edge Path Distance	15.161422543982209
Point Biserial Criteria	Adjacency Relation Distance	-23.032266645955286
Point Biserial Criteria	Pearson Correlation Distance	-23.69162105436469

5. Wnioski oraz podobieństwa z artykułem źródłowym

Podsumowując dane przedstawione w punkcie 4. dla eksperymentu porównawczego kryterium klastrowania sieci społecznościowych zauważalne są następujące wnioski:

- najbardziej uniwersalnym kryterium dla wszystkich skrajnych przypadków jest C Index, korzystający z metryki Neighbour Overlap Distance, Pearson Correlation Distance czy Adjacency Relationship Distance do obliczania odległości pomiędzy węzłami;
- klasyczna metryka, modularność, zajęła również wysoką pozycję. Tendencja ta mogłaby zostać zaburzona w przypadku, gdyby wykorzystane zostały sieci społecznościowe o dużej ilości węzłów - wtedy to widoczna byłaby słabość modularności;
- aktualnie najczęściej stosowaną metryką określania jakości pogrupowania jest modularność. Powyższy eksperyment wykazał, że w przyszłości korzystniejsze może okazać się implementowanie heurystyk służących do znajdowania grup w sieciach społecznych bazujących na optymalizacji innych kryterium - na przykład C Index.

Jeżeli chodzi o wyniki uzyskane w eksperymencie, są one dość zbliżone do rezultatów uzyskanych w bazowym artykule w przypadku bliskim przypadkowi optymistycznemu. Zauważalne jest natomiast pewne odchylenie w średnim i skrajnie pesymistycznym przypadku pogrupowania sieci. Po części może to wynikać z innego sposobu losowego modyfikowania rzeczywistego klastrowania niż miało to miejsce w artykule, na którym wzorowany był całościowy eksperyment.

Near Optimal Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	Q	0.736±0.266	5	5	2
2	CIndex PCD	0.72±0.326	1	1	3
3	SWC2 SPD	0.718±0.389	3	3	4
4	CIndex SPD	0.716±0.14	4	4	1
5	SWC2 ICD	0.713±0.396	2	2	5
6	ASWC2 ICD	0.687±0.334	11	10	7
Medium Far Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	CIndex PCD	0.608±0.202	8	18	1
2	CIndex NOD	0.58±0.053	39	13	2
3	CIndex ARD	0.513±0.313	26	62	5
4	Dunn01 ICD	0.457±0.173	58	83	8
5	SWC2 NOD	0.447±0.19	5	9	3
6	ASWC2 PCD	0.446±0.191	7	3	9
7	SWC2 PCD	0.446±0.19	6	2	10
8	Dunn03 ICD	0.439±0.109	43	37	11
9	Dunn31 SPD	0.437±0.177	56	47	15
10	Dunn01 SPD	0.434±0.205	29	67	7
11	Q	0.409±0.353	4	7	16
12	DB ICD	0.405±0.072	40	38	18
Far Far Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	SWC2 NOD	0.634±0.217	3	13	1
2	ASWC2 NOD	0.583±0.191	5	21	2
3	Q	0.498±0.179	4	38	5
4	CIndex PCD	0.493±0.282	2	4	13
5	CIndex SPD	0.437±0.291	1	11	4
6	SWC3 NOD	0.436±0.344	8	2	25

Wyniki uzyskane w artykule bazowym.

Dla eksperymentu porównywawczego zestawu algorytmów klastrujących, którego wyniki zostały przedstawione w punkcie 4. - najlepszym algorytmem dla klastrowanego zbioru według najlepiej wypadającego kryterium C-Index jest algorytm K Node zaimplementowany w bibliotece JavaML. Co ciekawe zauważalna jest rozbieżność w rankingu tworzonym przez wartości kryterium C-Index, a rzeczywistym stosunkiem do klastrowania ground-truth.

6. Podsumowanie

Zaimplementowany i przeprowadzony w projekcie eksperyment porównawczy kryterium klastrujących został wykonany dla wszystkich kryterium opisanych w artykule bazowym oraz części zawartych tam metryk obliczania odległości pomiędzy węzłami sieci (tylko 4). Wszystkie kryteria w raz z metrykami implementowane były od podstaw bazując na formułach zawartych w artykule bazowym jak i innych artykułach, zawartych w źródłach tej dokumentacji. Wykorzystane zostały wszystkie 3 sieci społecznościowe użyte w źródłowej publikacji. Proces eksperymentu przeprowadzony został dla skrajnych przypadków celem wykrycia najbardziej uniwersalnego kryterium klastrującego. Uzyskane wyniki są zadowalające, jednak w pewnym stopniu różnią się od bazowego artykułu - szczególnie dla pesymistycznego przypadku. Najlepsze kryterium według pierwszego eksperymentu - C-Index zostało wykorzystane jako kryterium rozstrzygające, który z algorytmów grupujących testowanych w drugiej części projektu był najlepszy.

7. Źródła

- http://webdocs.cs.ualberta.ca/~zaiane/postscript/SNA-Encyclopedia.pdf
- http://webdocs.cs.ualberta.ca/~rabbanyk/criteriaComparison/ASONAM/cameraReadyVersion/asonam12rabbanyk.pdf
- *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*, Nguyen Xuan Vinh http://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pdf
- http://spaghetti-os.blogspot.com/2014/05/zacharys-karate-club.html
- http://hal.elte.hu/~lanna/Publications/GraphEPLFinal_6o.pdf
- http://webdocs.cs.ualberta.ca/~rabbanyk/TopLeader/
- http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf