Analizy związków

Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

Często występujące wzorce

- Ang. Frequent Pattern
- Wzorce (głównie zbiory obiektów, ale też struktur), które występują często w zbiorach danych
  - Jakie produkty są często zamawiane wspólnie (wycieczka do Kairu + rejs statkiem)
  - Jakie usługi zamówi klient, jeżeli wykupi AllInc+?
  - Jakie grupy klientów są zainteresowani nową ofertą
  - **Jakie sklepy występują w sąsiednich lokalizacjach?**
- Zastosowania: modele biznesowe, markting, sprzedaż, **analizy przestrzenne zachowań**, analiza koszykowa

Co to jest transakcja?

ID	Items
1	{Al,Child,2Weeks,Cairo}
2	{Al,Child,1Week,Desert}
3	{Al,Child,2Weeks,WatPark}
4	{BB,Cairo,Luxor,Alexandr,2Weeks}
5	{Al,2Weeks,Luxor,Cairo,Desert}
...	...

Transakcje

Produkty turystyczne:
Al – All inclusive
Child – pakiet dla dzieci
2Week – dwa tygodnie
Cairo – wycieczka do..
WatPark – waterpark
....

{Al,Child} Przykład Frequent ItemSet
Child → Al Przykład reguły wiążącej (Association rule)

- Transakcje grupowane są w bazy transakcji
- Każda transakcja to **zbiór** elementów, co oznacza że elementy nie mogą się powatarzać

Elementy i zbiory elementów w transakcji

- **Item (element)** – element występujący przynajmniej w jednej transakcji w bazie transakcji, najczęściej oznaczony symbolem (A, „Al”, lub True, jeżeli baza transkakcji ma charakter zbioru binarnego
- **Itemset (zbiór elementów) *I*** – zbiór możliwych kombinacji elementów w transakcji: np. w transakcji {A,B,E}: {A}, {B}, {E}, {A,B}, {A,E}, {A,B,E}
- **Transakcja** – zbiór elementów występujących w jednym zdarzeniu (koszyk, grupa sąsiednich obiektów, rachunek)
- **Baza transakcji** – zbiór wszystkich analizowanych transakcji współdzielących te same elementy i zbiory elementów

Czynniki definiujące złożoność przeszukiwania

- Wielkość progu – im mniejszy tym więcej reguł jest tworzonych
- Wymiarowość – liczba pojedynczych items w zbiorze
- Liczba transakcji
- Wielkość transakcji – iczba items w pojednyczej transakcji

Sposób przechowywania transakcji w bazie

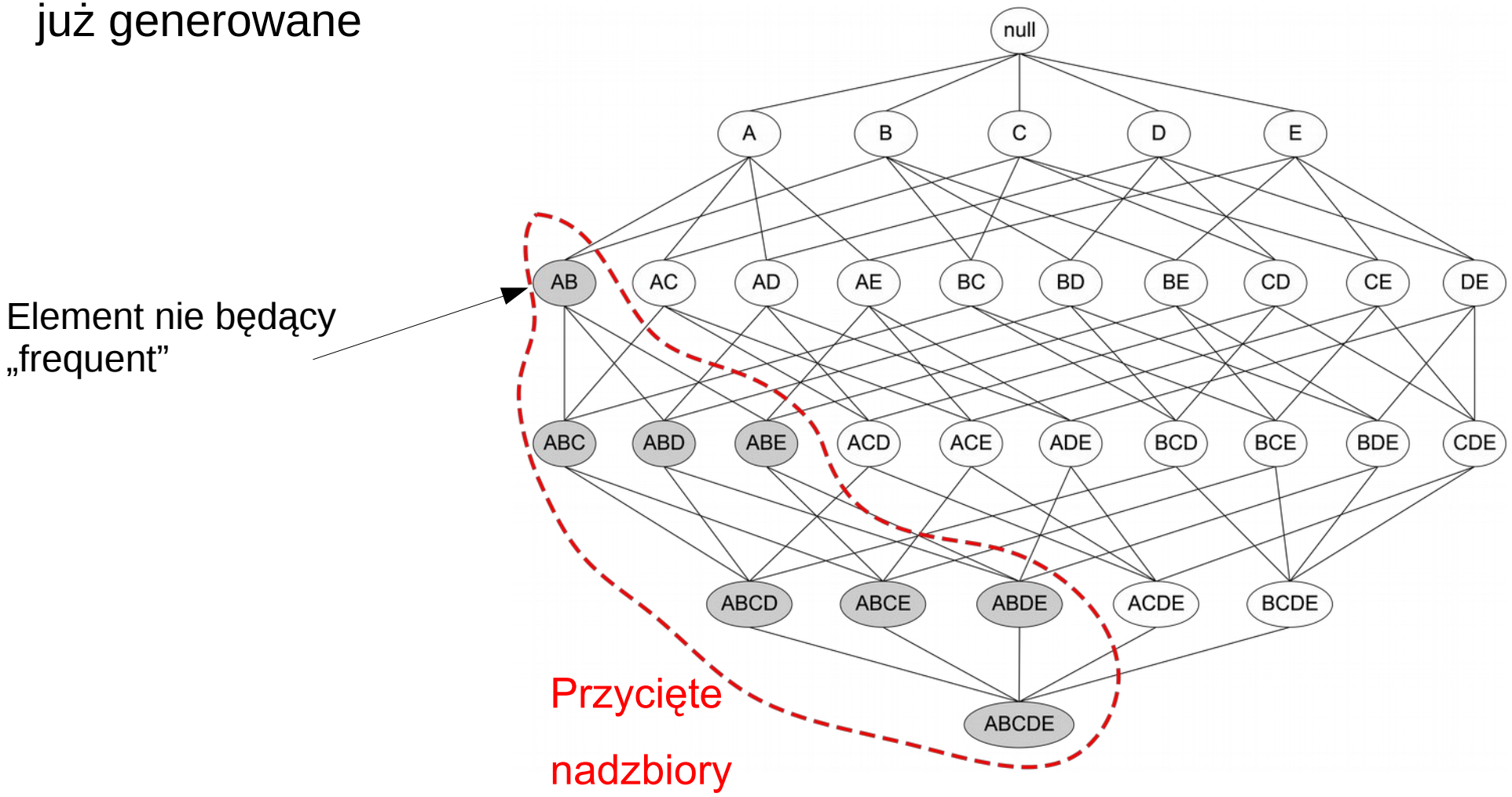
- Jako zbiór – każda transakcja jest zapisana jako zbiór (unikalnych) elementów
- Jako tablica binarna (bitmapa) gdzie kolumny to wszystkie możliwe elementy (Items) a wiersze transakcje

ID	Items
1	{AI,Child,2Weeks,Cairo}
2	{AI,Child,1Week,Desert}
3	{AI,Child,2Weeks,WatPark}
4	{BB,Cairo,Luxor,Alexandr,2Weeks}
5	{AI,2Weeks,Luxor,Cairo,Desert}
...	...

ID	AI	BB	Ch	2W	1W	Ca	Lx	Ax	Wp	Ds
1	1	0	1	1	0	1	0	0	0	0
2	1	0	1	0	1	0	0	0	0	1
3	1	0	0	1	0	0	0	0	1	0
4	0	1	0	1	0	1	1	1	0	0
5	1	0	0	1	0	1	1	0	0	1
..										

Generowanie reguł

- Reguły generowane są od najprostrzych
- Jeżeli prosta reguła nie jest „frequent” jej następniki też nie są i nie są już generowane



Co to znaczy że itemsert jest „frequent”

- Support – częstotliwość z jaką dany itemset pojawia się w bazie danych, liczba transakcji zawierających danych itemset do wszystkich transakcji
- „AI” pojawia się zarówno w 1 jak i 2, i 3
- **Frequent itemset** to itemset, którego support jest większy niż parametr min_support

ID	Items	support
1	{AI}	0.6
2	{AI,Child}	0.4
3	{AI,Child,2Weeks}	0.2
4	{Cairo,Luxor,Alexandr}	0.2
5	{Cairo}	0.4
...	...	...

Reguły asociacji

- Association rules.
- Reguły asocjacji to implikacje (wynikania), gdzie $X \rightarrow Y$, i gdzie X i Y są zbiorami rozłącznymi
- Lewa strona (lhs, antecedant) reguły zawiera dowolną ilość elementów, a strona prawa (rhs, consequent) zawiera jeden element, nie występujący po stronie prawej:

{Cairo,Alexandr} => {Luxor}

Wskaźniki wartościowych reguł

- **Support** – nie jest definiowany dla reguły ale dla itemset i dzieli się na trzy metryki: Support poprzednika (A - anecedant), Support następnika (C - consequent), Support (A => C) = support (A u C)
  - Jeśli zbyt mały – tracimy interesujące ale rzadkie items – np. drogie produkty
  - Jeśli zbyt duży – duża liczba itemsets i długi czas obliczeń
- **Condifence** – liczba transakcji zawierających A i C, przez liczbę transakcji zawierających A, jak często występowanie danego zbioru elementów (itemset) spowoduje pojawienie się elementu C. Wartość 1 oznacza że zbiór A zawsze będzie generował zbiór C (np. Al i Child będzie zawsze oznaczało 2Week)
- **Lift** – support A i C do support A * support C, miara jak często współwystępowanie A i C występują wspólnie względem ilości ich współwystępowania gdyby były statystycznie niezależne, w takiej sytuacji **Lift = 1**. Wartość lift poniżej 1 oznacza że dany wariant występuje rzadziej niż można się spodziewać

Jak rozumieć wskaźniki?

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

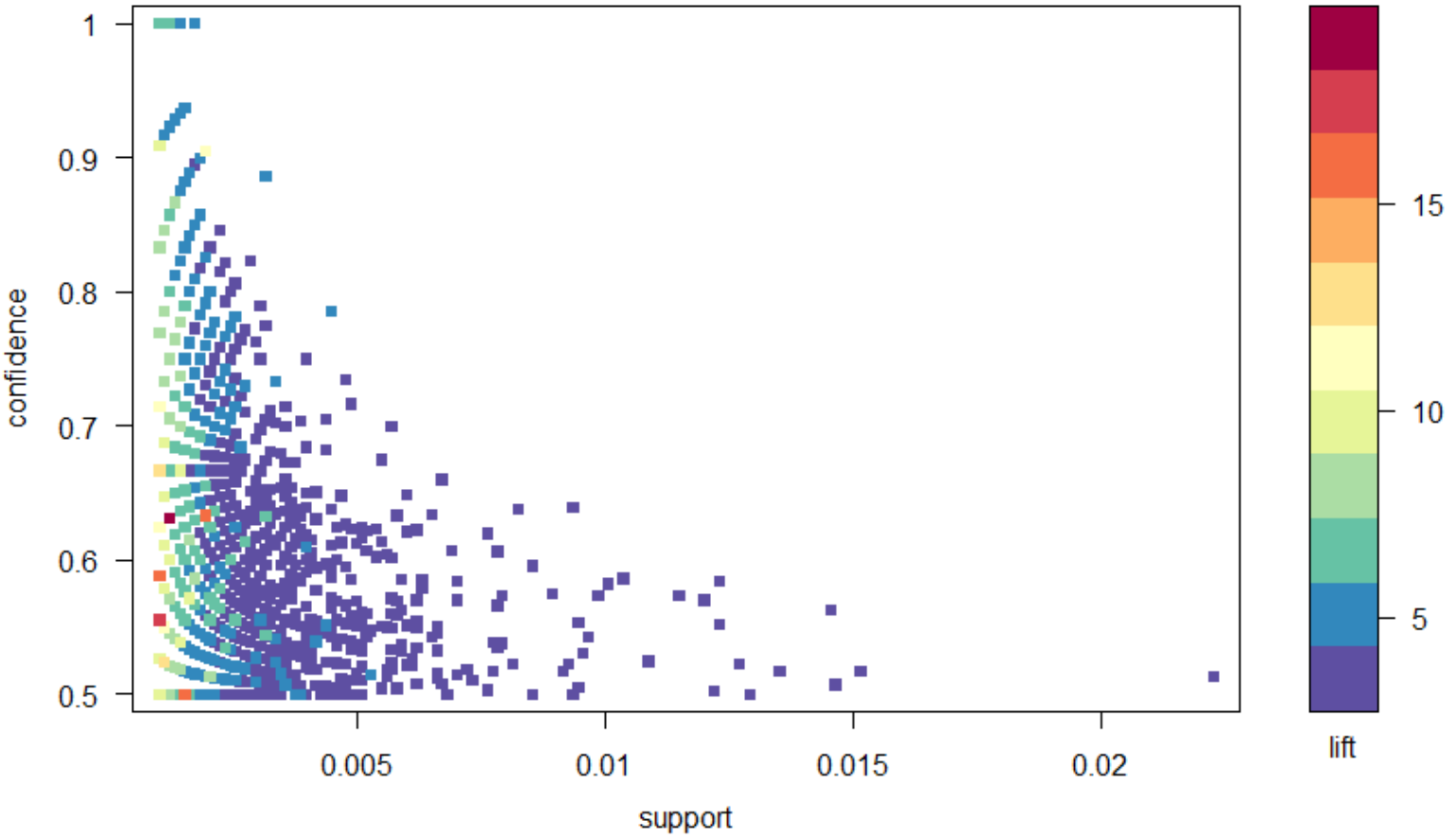
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

W 100 transakcjach kupiono:

- 50 razy chleb
- 5 razy masło
- Masło kupowano tylko wtedy gdy kupowano chleb

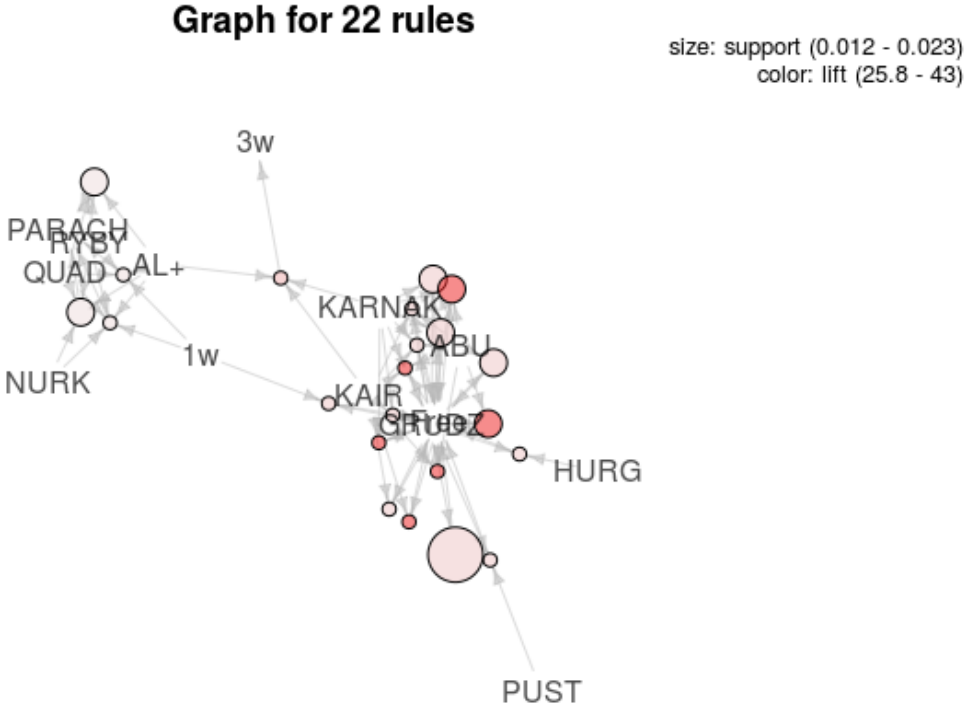
- Support chleb => masło = 5/10 = 0.05
- Confidence chleb => masło = 5/50 = 0.1
- Lift chleb => masło = 0.05/0.005 = 10

Diagram Lift – confidence - support



Grafy zależności

- Są jednym z najlepszych narzędzi wizualizacyjnych dla poszukiwania związków. Pokazują support każdej reguły (częstotliwość występowania) i lift (nieprzypadkowość związków)
- Items stanowią węzły grafu, graf jest skierowany łuki łączą items z regułami
- Pozwala wykryć grupy (clusters) klientów zainteresowanych określonymi produktami. Z tego powodu jest to metoda nienadzorowana

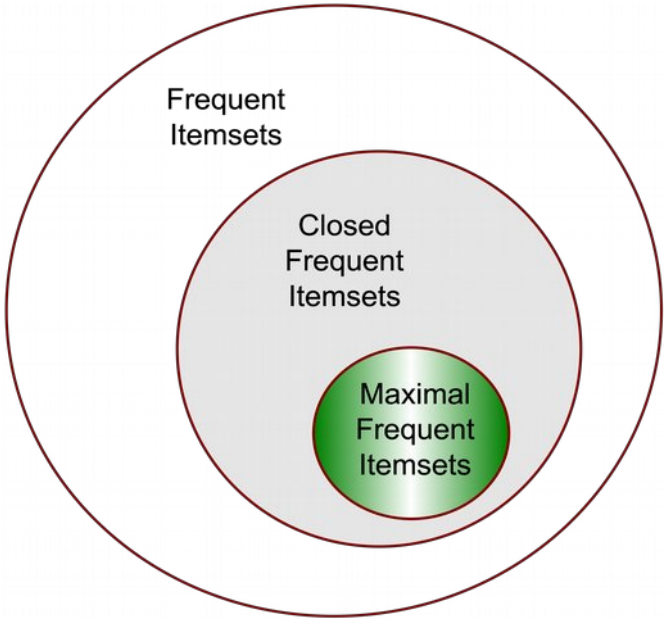


Ocena ważności reguł

- Szukanie związków generuje wiele reguł, z których wiele jest nieinteresujących lub nadmiarowych
- Nadmiarowość oznacza, że reguły mają taki sam support i confidence:
  $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$
- Miary ważności:
  - **Obiektywne** (21 miar ważności związków mn. Support, Gini, entropia itp.
  - **Subiektywne**: reguła spotyka się z oczekiwaniem użytkownika lub reguła jest użyteczna

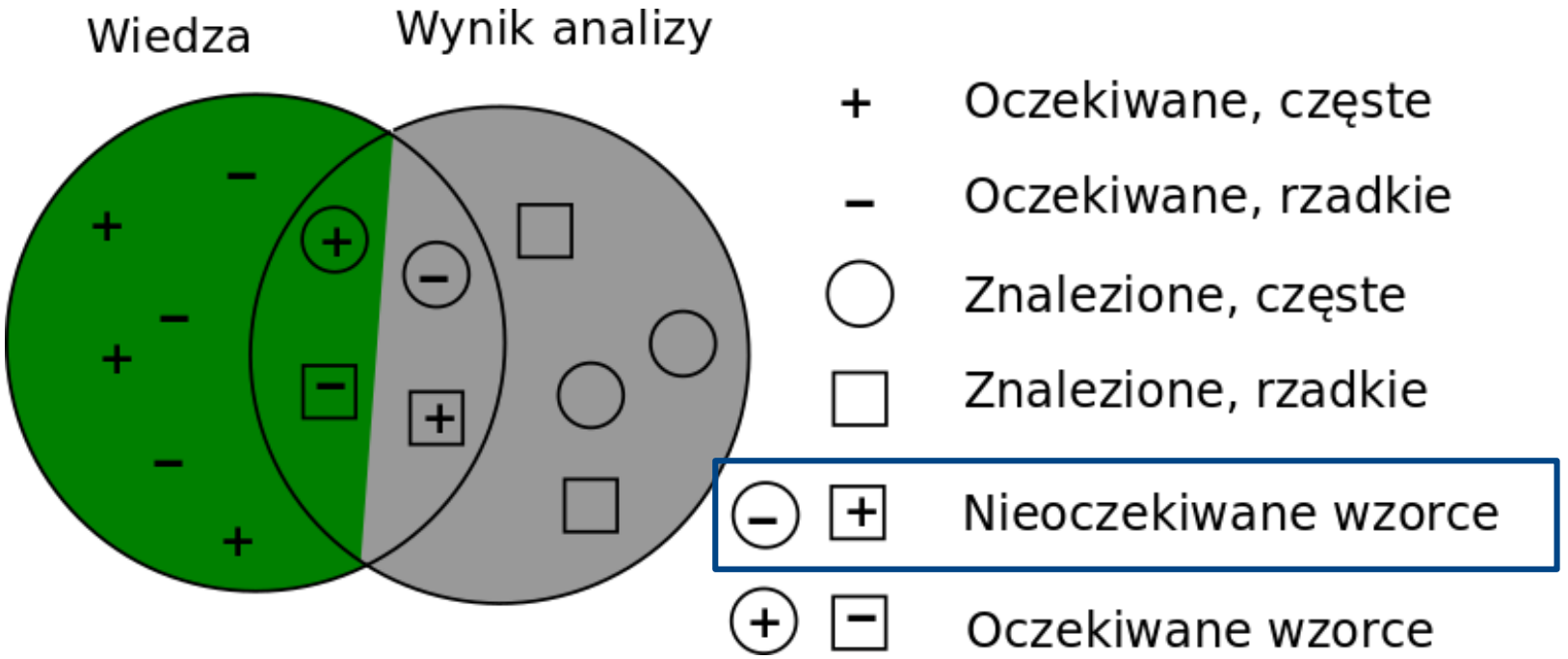
ważne frequent itemsets

- **Zamknięte** (closed) itemsets – to takie FI, których żaden z nadzbiorów nie ma takiego samego support jak dane itemset – nie jest nadmiarowo generowany przez inny itemset
- **Maksymalne** (maximal) – to takie closed itemsets, gdzie żaden z nadzbiorów nie jest frequent – wskazuje na interesujące zestawienie items



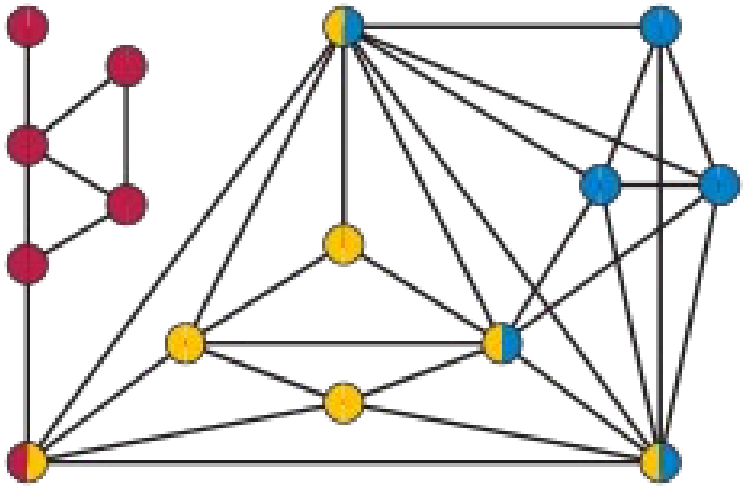
Reguły interesujące vs. zaskakujące

- Wymaga połączenia oczekiwań użytkownika wynikających z wiedzy z wynikami analizy



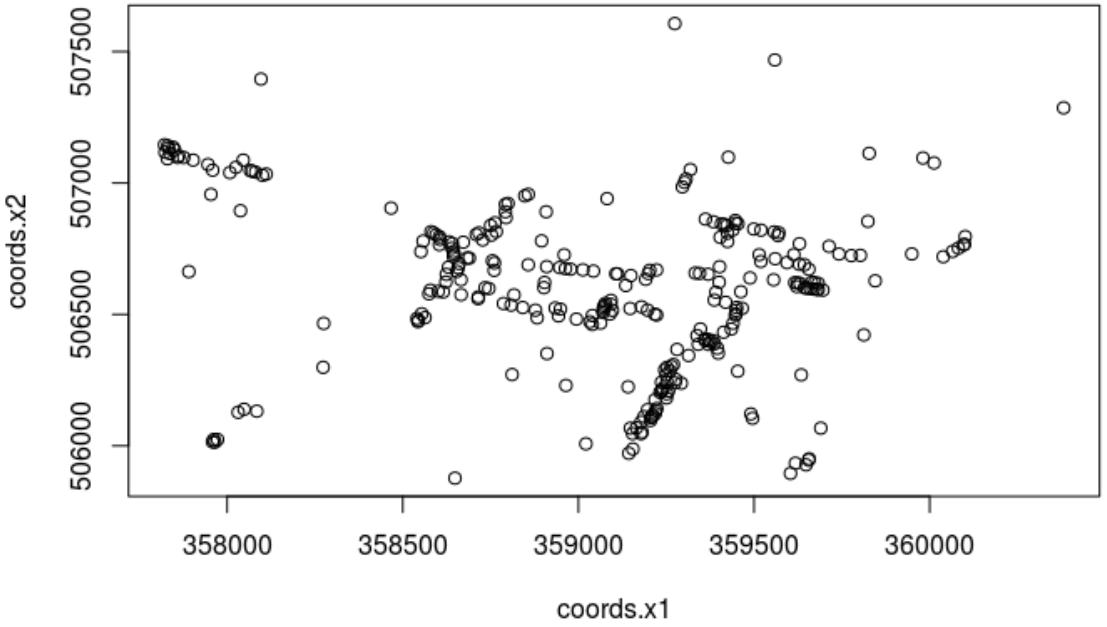
FI w analizie geoprzestrzennej

- Wykrywanie związków współwystępowania określonych obiektów blisko siebie
- Wymaga podania progu odległości i wyznaczenia klik (wszystkich obiektów znajdujących się w względem siebie bliżej niż założona wartość progowa
- Każdy obiekt może należeć do więcej niż jednej kliki

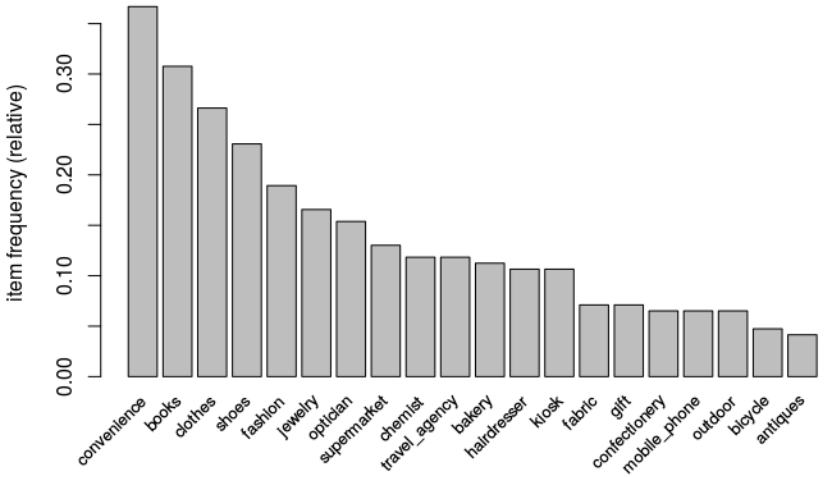


Przykład

- Współwystępowanie sklepów na pewnym obszarze



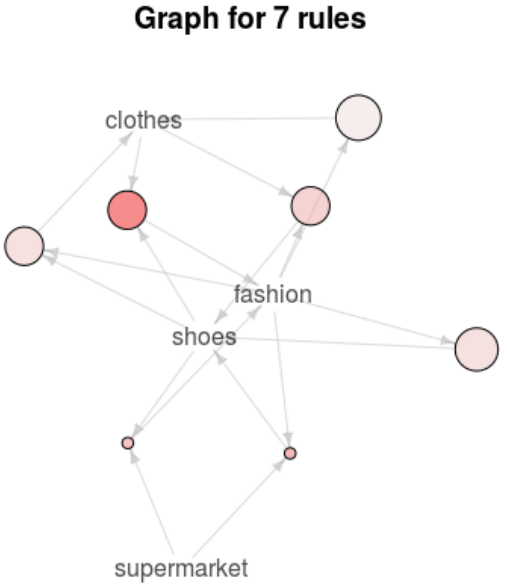

```
[5]   {kiosk}
[6]   {convenience}
[7]   {bicycle}
[8]   {books,chemist,convenience}
[9]   {deli}
[10]  {craft}
[11]  {books}
[12]  {supermarket}
[13]  {chemist,convenience,jewelry}
[14]  {chemist,clothes,fashion,houseware,shoes,supermarket}
[15]  {photo,shoes}
[16]  {clothes,shoes}
[17]  {bicycle,books,clothes,fashion,shoes}
[18]  {bicycle,books,clothes,fashion,jewelry,shoes}
[19]  {books,clothes,fashion,jewelry,shoes}
[20]  {books,clothes,fashion,jewelry,shoes}
[21]  {clothes,fashion,jewelry,shoes,supermarket}
```

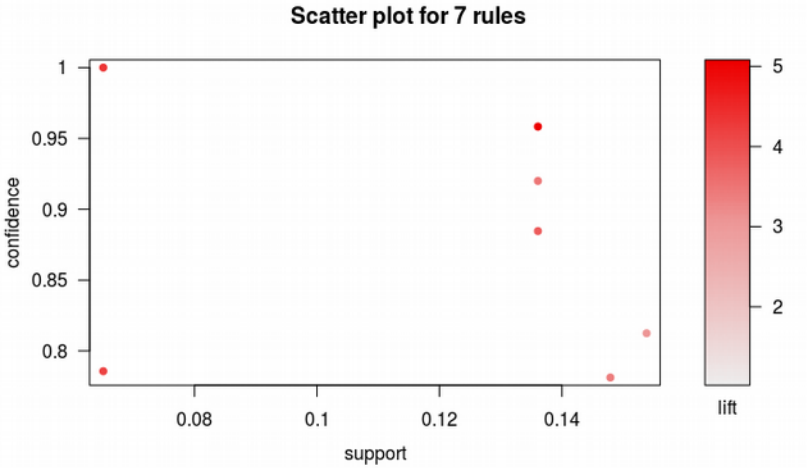


Wyniki analizy

```
    lhs                     rhs       support    confidence lift     count
[1] {fashion}            => {shoes}   0.14792899 0.7812500  3.385417 25
[2] {fashion}            => {clothes} 0.15384615 0.8125000  3.051389 26
[3] {fashion,supermarket} => {shoes}   0.06508876 1.0000000  4.333333 11
[4] {shoes,supermarket}   => {fashion} 0.06508876 0.7857143  4.149554 11
[5] {fashion,shoes}       => {clothes} 0.13609467 0.9200000  3.455111 23
[6] {clothes,fashion}     => {shoes}   0.13609467 0.8846154  3.833333 23
[7] {clothes,shoes}       => {fashion} 0.13609467 0.9583333  5.061198 23
```

Min support =0.06

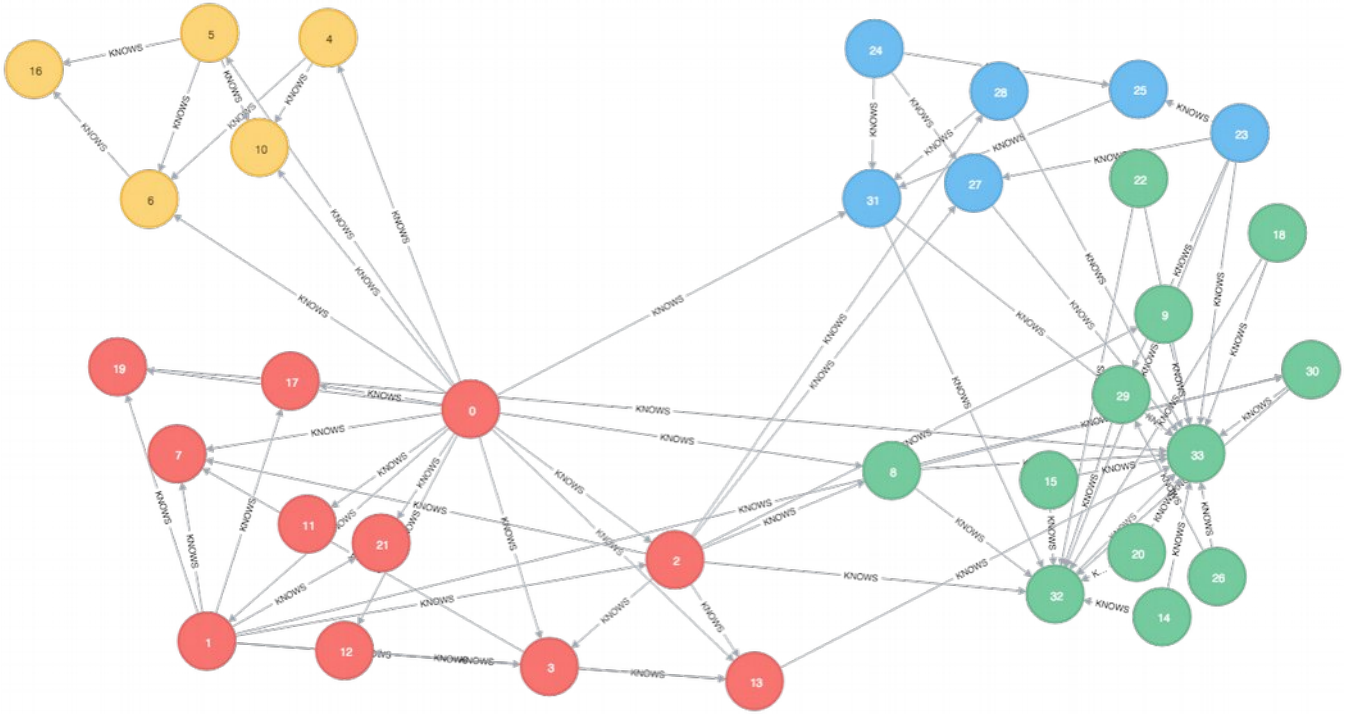




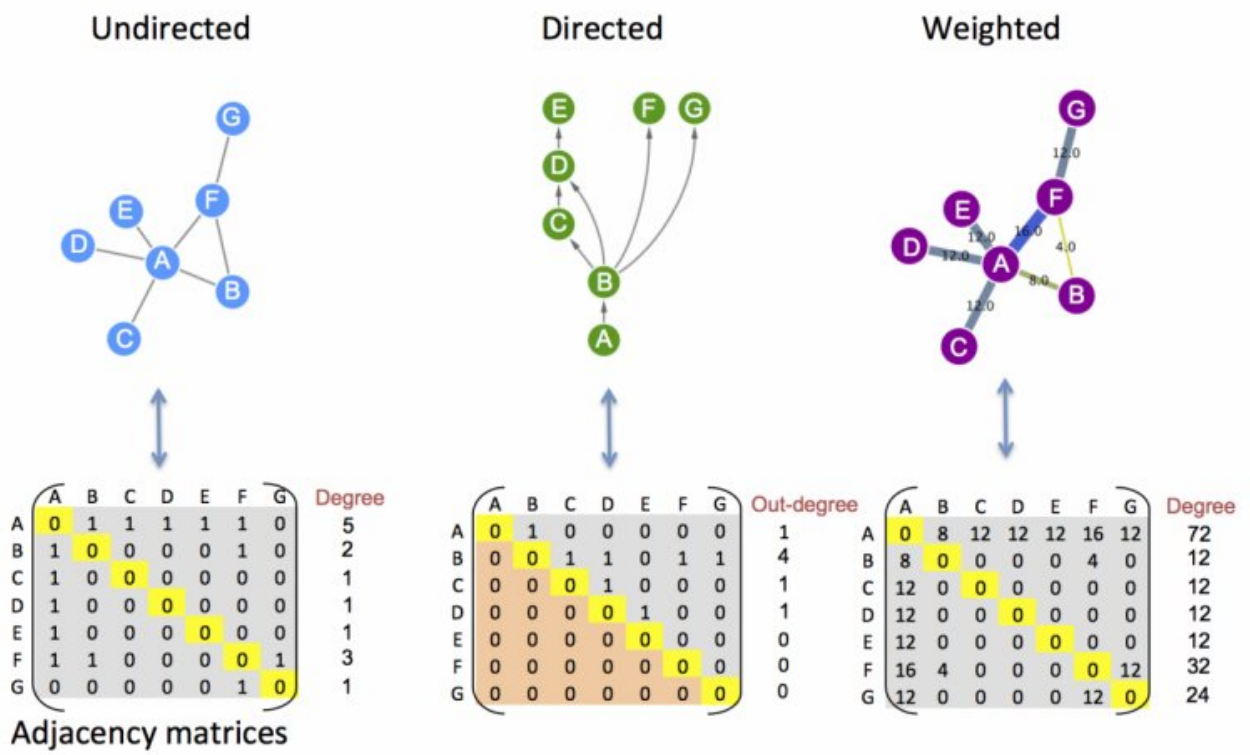
- Znalezienie reguły: sklepy z odzieżą powodują pojawienie się sklepów z butami i odwrotnie
- Supermarkety powodują pojawienie się sklepów z modą i sklepów z butami

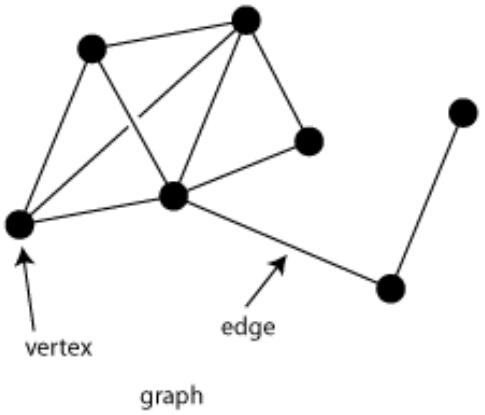
Wyszukiwanie informacji w sieciach

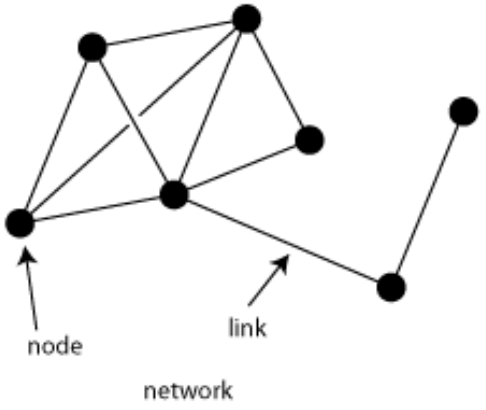
- Grupowanie na podstawie **połączeń między obiektami** a nie obiektami (podobnie jak w analizie asocjacyjnej)
- Połączenia między ludźmi i inne skomplikowane relacje
- Każdy obiekt może należeć do wielu społeczności
- Wykrywanie obiektów centralnych dla społeczności i łączących społeczności



Budowanie grafu

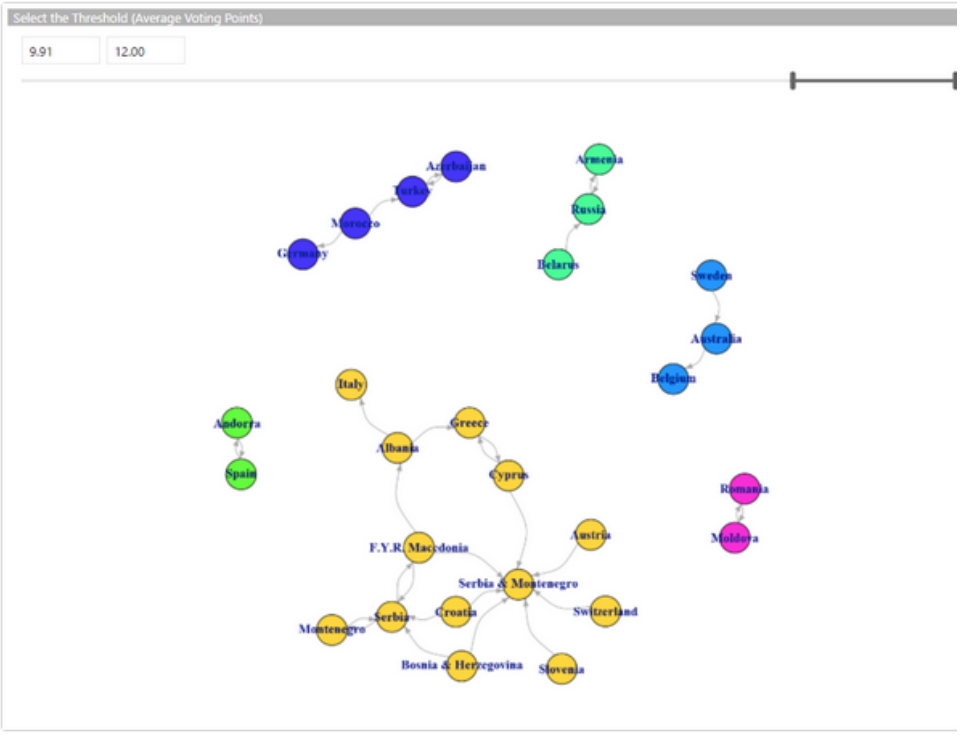




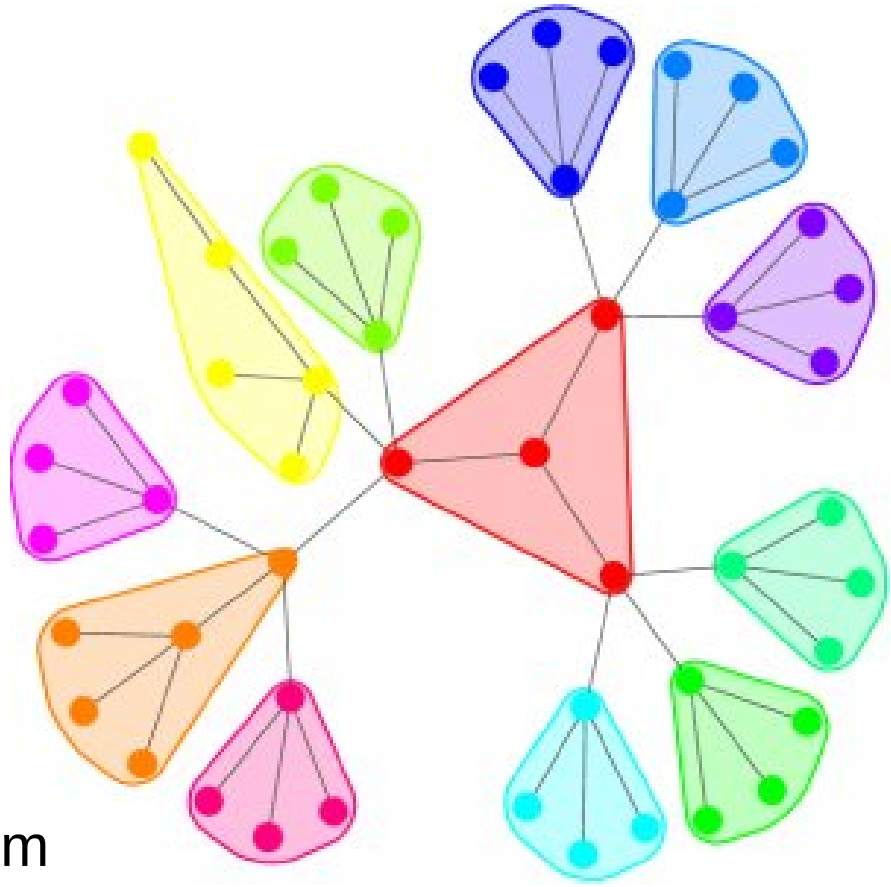


Grupowanie na podstawie połączeń

From country	ToCountry	AvgPoint
Albania	Andorra	1.50
Albania	Armenia	0.86
Albania	Australia	8.60
Albania	Austria	1.11
Albania	Azerbaijan	3.07
Albania	Belarus	0.50
Albania	Belgium	1.14
Albania	Bosnia & Herzegovina	6.25
Albania	Bulgaria	4.30
Albania	Croatia	2.00
Albania	Cyprus	2.54
Albania	Czech Republic	0.00
Albania	Denmark	1.00
Albania	Estonia	0.50

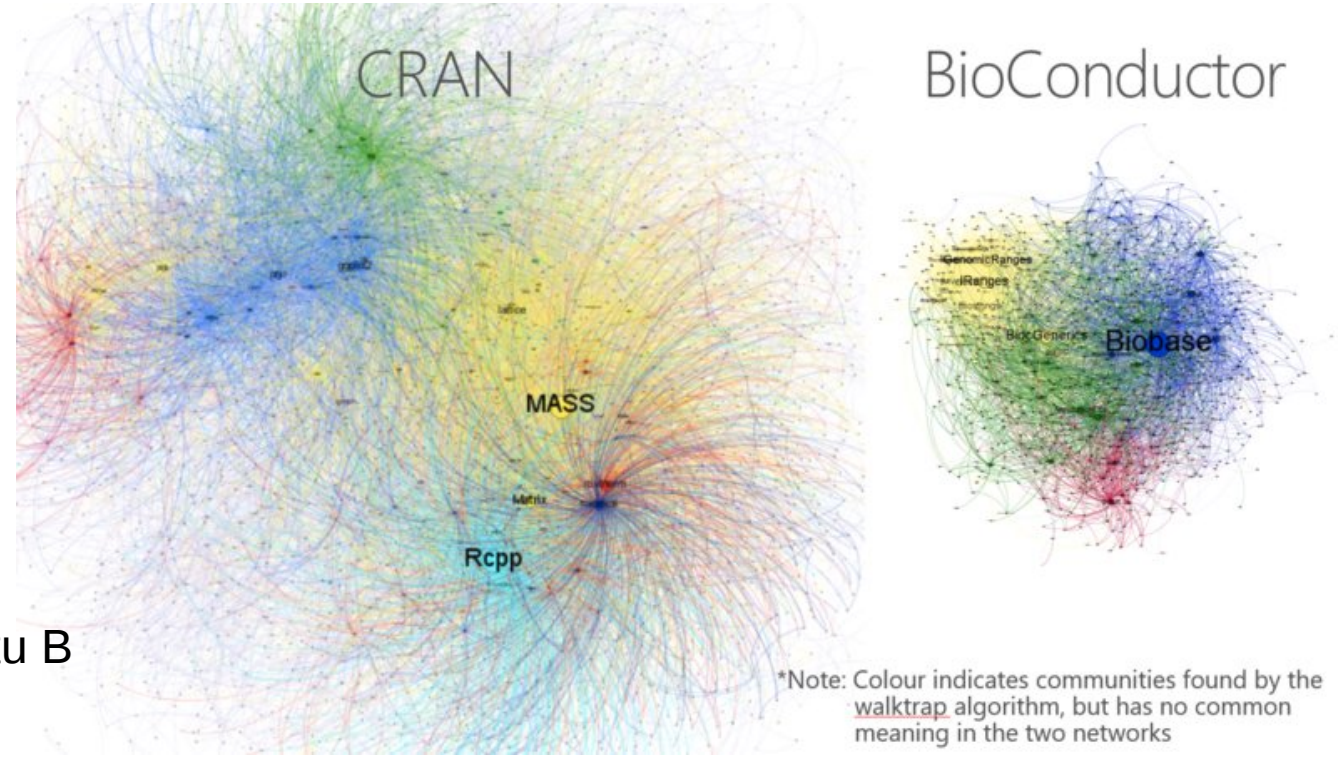


Budowanie communities



- Każdy vertex indywidualnym community
- Przemieszczanie węzłów do innych comm
- Jeśli nie można ulepszyć skupień - stop

Pakiety środowiska R

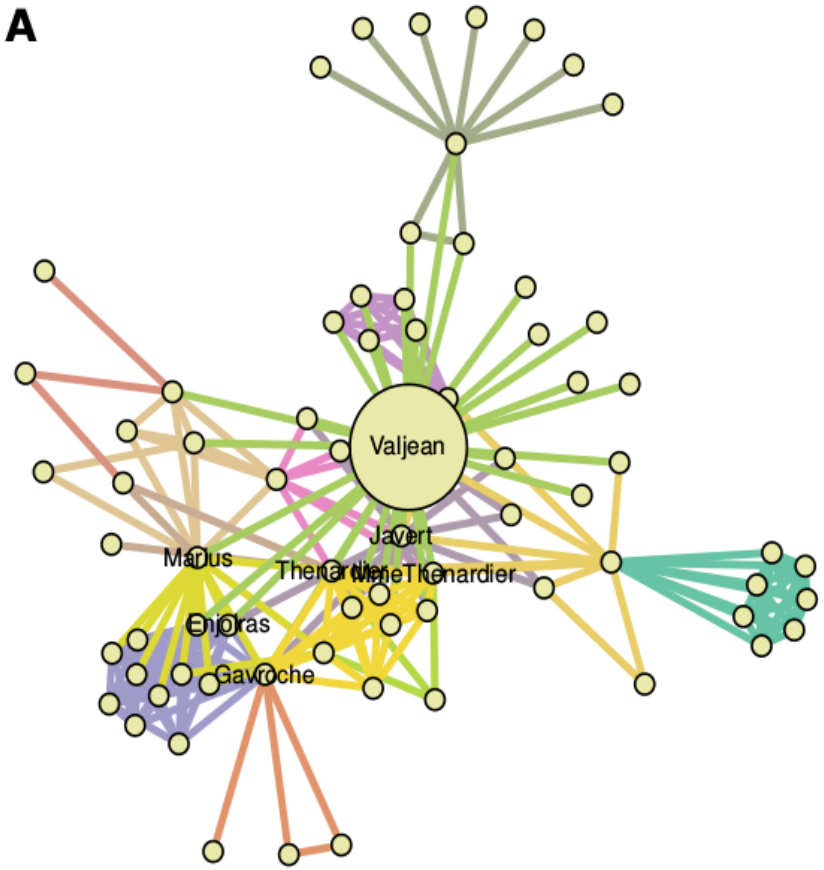


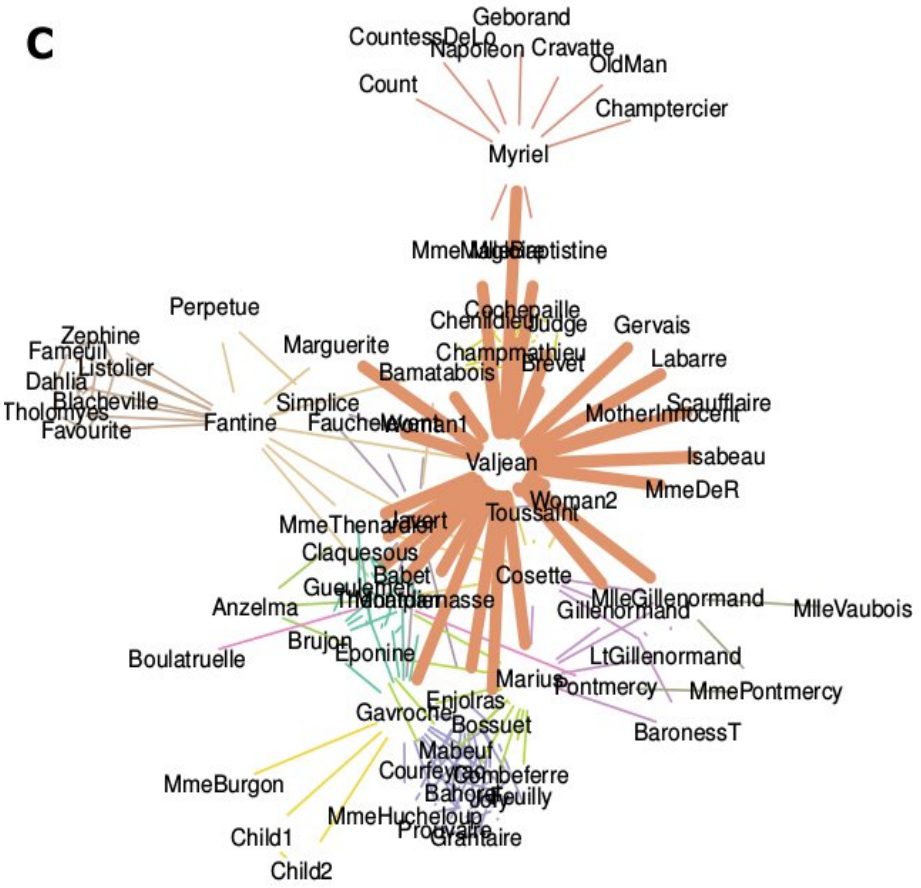
Zależności:
- pakiet A w zależnościach pakietu B
- pakiet B zależy od pakietu A

Rozproszony, liczne pakiety
bez połączeń
rozproszenie

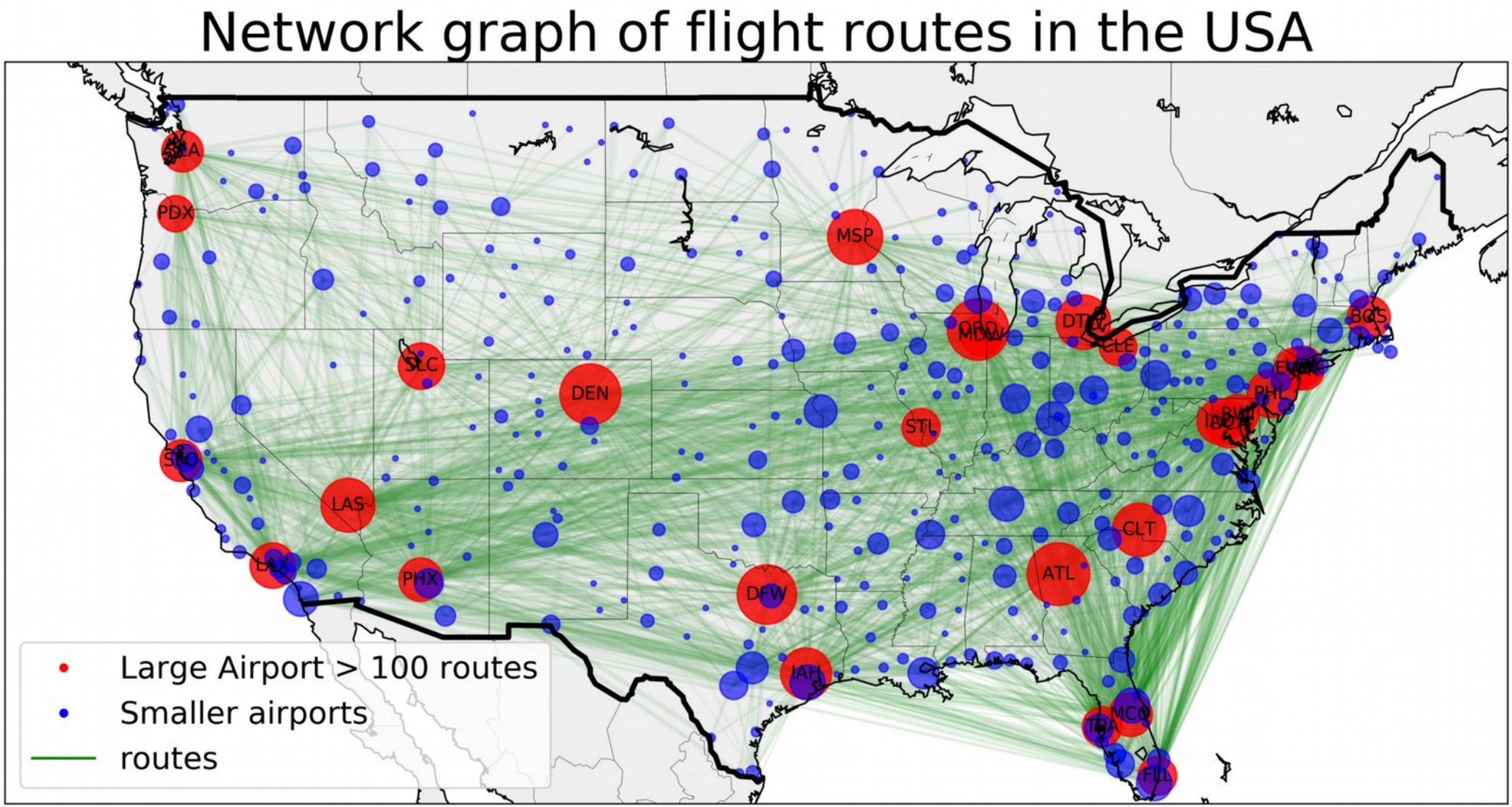
Zwarty, silenie połączony i
Pogrupowany
centralizacja

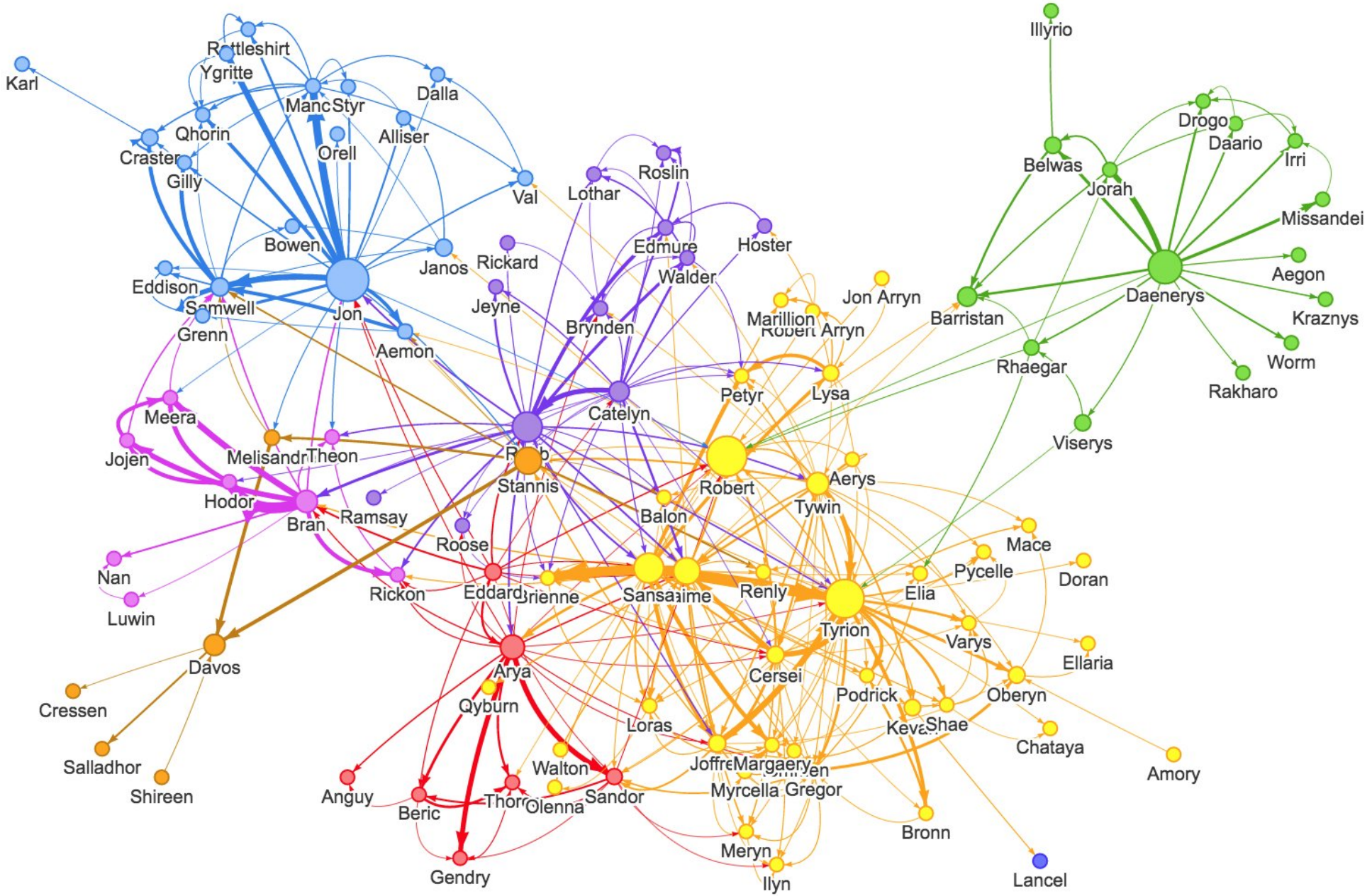
Społeczności vs obiekty centralne





Analiza geoprzestrzenna




Illyrio
Drogo
Daario
Irri
Belwas
Jorah
Missandei
Aegon
Daenerys
Kraznys
Barristan
Worm
Rhaegar
Rakharo
Viserys
Karl
Rattleshirt
Ygritte
Manc
Styr
Dalla
Qhorin
Alliser
Orell
Craster
Gilly
Val
Lothar
Roslin
Bowen
Janos
Rickard
Edmure
Hoster
Walder
Eddison
Samwell
Jon
Jeyne
Brynden
Jon Arryn
Marillion
Robert Arryn
Grenn
Aemon
Petyr
Lysa
Meera
Catelyn
Jojen
Melisandr
Theon
Stannis
Robert
Aerys
Tywin
Hodor
Bran
Ramsay
Balon
Roose
Mace
Pycelle
Doran
Nan
Rickon
Eddard
Brienne
Sansa
Jaime
Renly
Elia
Luwin
Tyrion
Varys
Davos
Arya
Cersei
Ellaria
Cressen
Qyburn
Podrick
Oberyn
Loras
Kevan
Shae
Chataya
Salladhor
Walton
Joffrey
Margaery
Amory
Shireen
Anguy
Beric
Thoros
Olenna
Sandor
Myrcella
Gregor
Bronn
Meryn
Gendry
Ilyn
Lancel