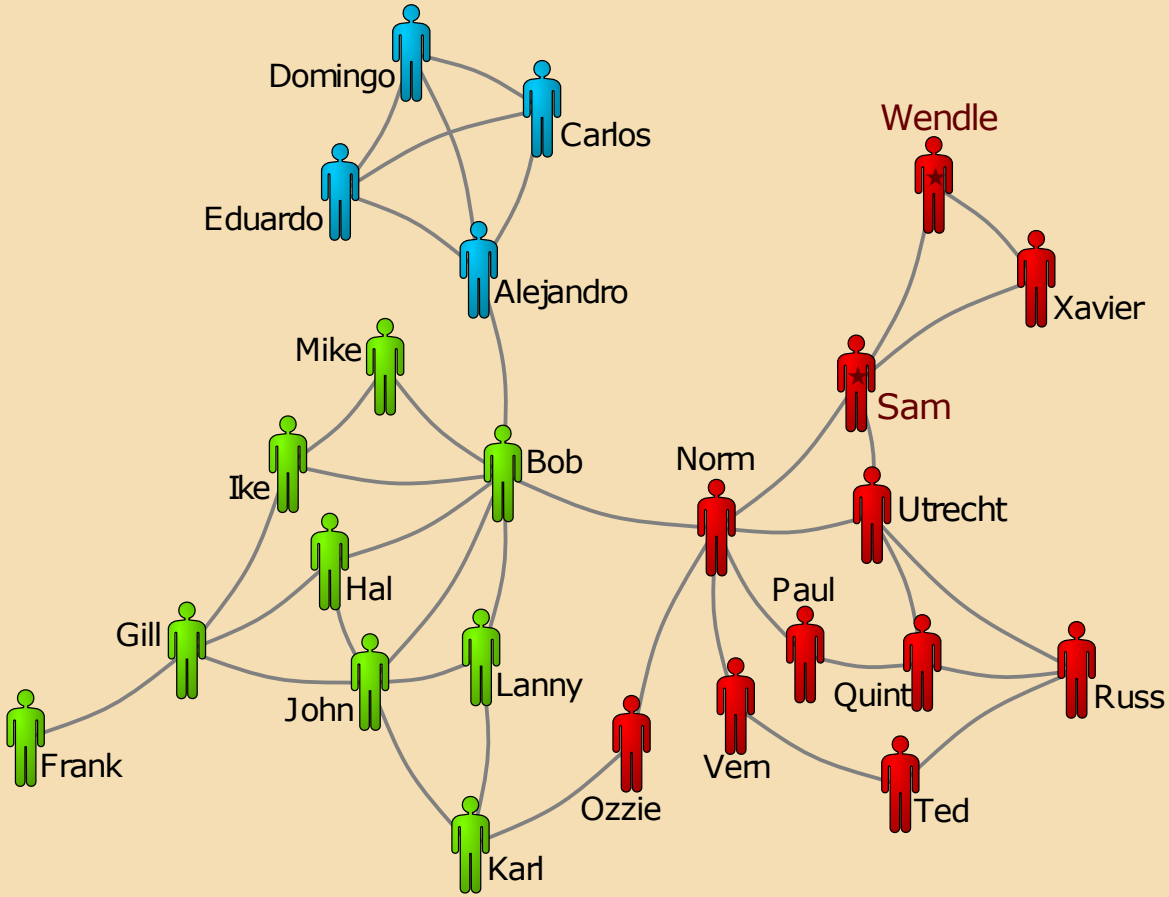
# Dodatne metode za analizo velikih omrežij




**Andrej Mrvar**

# Nakopičenost (Clustering Coefficient)...

## Definicija...



**Nakopičenost** v točki omrežja je določena kot razmerje med številom vseh povezav med sosedi dane točke in največjim možnim številom povezav med sosedi. Največje možno število povezav med vsemi sosedi dobimo, če so vsi sosedi povezani med sabo.
Naj bo $deg_v$ stopnja točke $v$, $|E(G^1(v))|$ število povezav med točkmi v 1-okolici točke $v$, $\Delta$ največja stopnja točke v omrežu in $|E(G^2(v))|$ število povezav med točkami v 2-okolici točke $v$.

- $CC_1$ – mere, ki upoštevajo le 1-okolico točke:

$$CC_1(v) = \frac{2|E(G^1(v))|}{deg_v(deg_v - 1)}$$

$$CC_1'(v) = \frac{deg_v}{\Delta} CC_1(v)$$

$CC_1(v)$ je *Clustering Coefficient*, kot sta ga definirala Watts in Strogatz, 1998.

# ...Nakopičenost...

## Definicija...



- $CC_2$ – mere, ki upoštevajo 2-okolico točke:

$$CC_2(v) = \frac{|E(G^1(v))|}{|E(G^2(v))|}$$

$$CC_2'(v) = \frac{deg_v}{\Delta} CC_2(v)$$

Če je $deg_v \leq 1$, zavzamejo vsi koeficienti za točko $v$ vrednost 999999998 (manjkajoča vrednost).

V analizah bomo uporabljali samo prvi koeficient ($CC_1$).

---

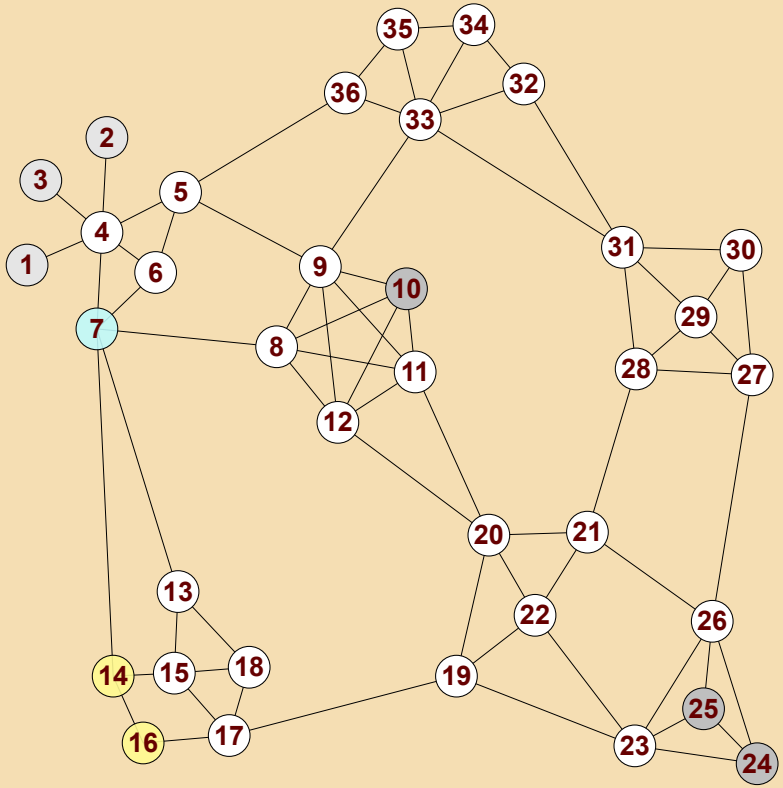
**Network / Create Vector / Clustering Coefficients**

V oknu Report se izpišeta tudi *Watts-Strogatz Clustering Coefficient* in *Network Clustering Coefficient (Transitivity)*. Prvi koeficient je neuteženo, drugi pa uteženo povprečje lokalnih nakopičenosti. V nadaljevanju bomo uporabljali le drugega: *Nakopičenost* ali *tranzitivnost* omrežja je delež vseh poti na dveh točkah v omrežju, ki so zaprte s tretjo povezavo.

# ...Nakopičenost...

## Primer






V zgornjem omrežju (shr1.net) imajo točke $v_{10}$, $v_{24}$, in $v_{25}$ najvišjo možno nakopičenost ($CC_1 = 1$, vsi njihovi sosedi so povezani med sabo); točki $v_{14}$, in $v_{16}$ imata najnižjo nakopičenost, $CC_1 = 0$ (nobene povezave med sosedi; točka $v_7$ ima naslednjo najmanjšo nakopičenost $CC_1 = 0.1$ (1 povezava med 5 točkami z 10 možnimi povezavami); za točke $v_1$, $v_2$, in $v_3$ pa $CC_1$ ni mogoče izračunati.

# ...Nakopičenost

## Tranzitivnost



Povprečno nakopičenost vseh lokalnih nakopičenosti (brez manjkajočih vrednosti) izračunamo z **Vector** / **Info**. V našem primeru dobimo za rezultat 0.451. To je neutežena nakopičenost ali *Watts-Strogatz Clustering Coefficient*.

O tranzitivnosti smo govorili že, ko smo predstavili spisek triad (*triadic census*). V zgornjem primeru imamo 30 tranzitivnih in 129 netranzitivnih triad. Uteženo nakopičenost, ki jo imenujemo tudi *Network Clustering Coefficient* (ali kratko *Tranzitivnost*) izračunamo takole (polne triade moramo šteti trikratno):

$$\frac{30 * 3}{30 * 3 + 129} = 0.410$$

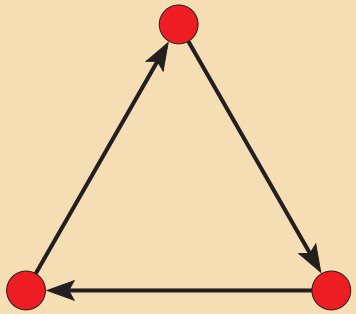
V nadaljevanju bomo uporabljali samo ta koeficient (*Network Clustering Coefficient* ali *Tranzitivnost*).

# Kratki cikli...

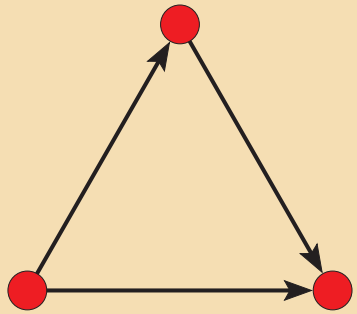
## Trikotniki (3-obroči)

Za vsako povezavo preštejemo koliko trikotnikom pripada. Števila se shranijo kot vrednosti na povezavah.
**Network / Create New Network / with Ring Counts stored as Line Values / 3-Rings**

- **Undirected** – preštejemo število trikotnikov v neusmerjenem omrežju.
- **Directed** – preštejemo ciklične, tranzitivne ali vse trikotnike v usmerjenem omrežju, ali pa kolikokrat neka povezava nastopa kot bližnjica (shortcut).



ciklični



tranzitivni

*Primera:* flow2.net (usmerjeno), shr1.net (neusmerjeno).

# ...Kratki cikli

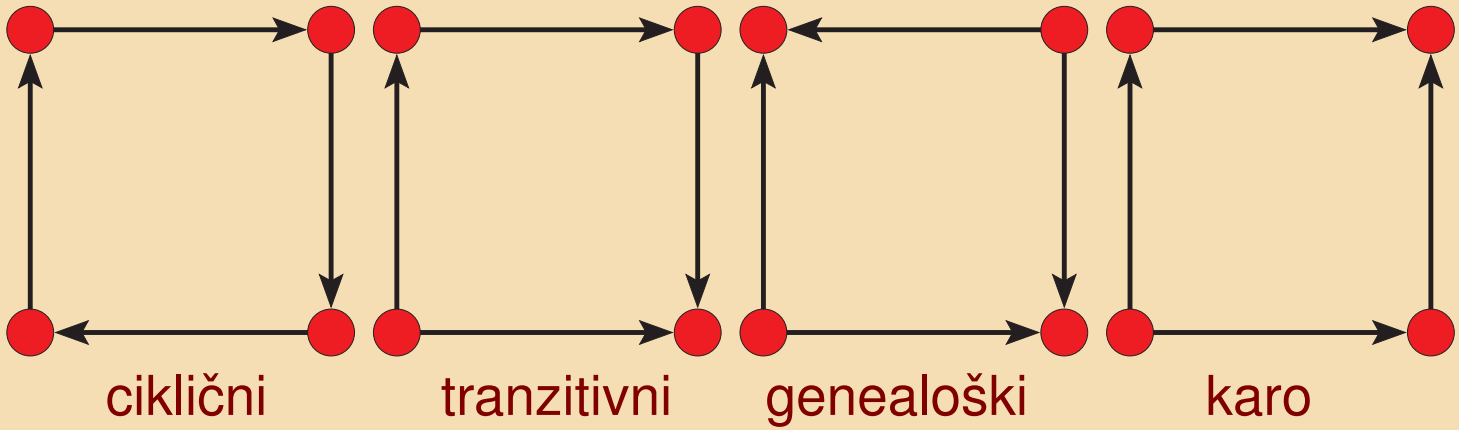
## 4-obroči



Za vsako povezavo preštejemo koliko štirikotnikom pripada. Števila se shranijo kot vrednosti na povezavah.

**Network / Create New Network / with Ring Counts stored as Line Values / 4-Rings**

- **Undirected** – preštejemo število štirikotnikov v neusmerjenem omrežju.
- **Directed** – preštejemo ciklične, tranzitivne, genealoške, 'karo', ali vse štirikotnike v usmerjenem omrežju, ali pa kolikokrat neka povezava nastopa kot bližnjica (shortcut).

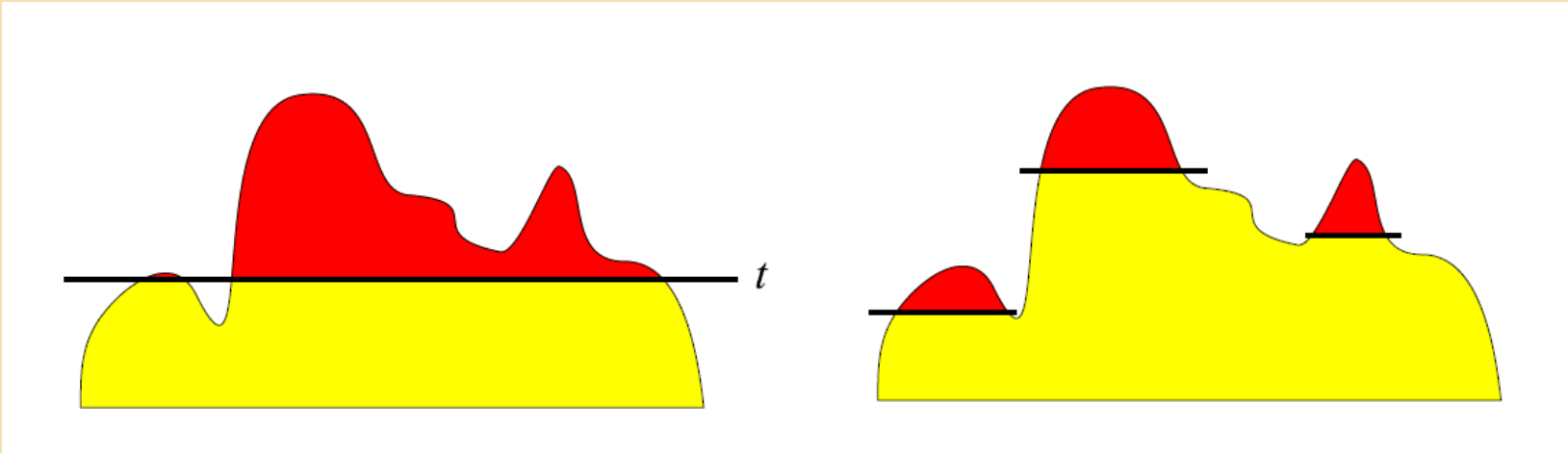



V dvovrstnem omrežju lahko obstajajo samo štirikotniki, trikotniki pa ne.
*Primeri:* flow2 (usmerjeno), shr1 (neusmerjeno), Davis (dvovrstno).

# Otoki...



Če dano lastnost/utež točk/povezav predstavimo kot njihovo višino, nam naše omrežje določa nekakšno pokrajino s hribi in dolinami. Če to pokrajino potopimo v vodo do izbrane višine, dobimo kot izrez otoke. S spreminjanjem višine vode dobivamo različne otoke. V uporabah nas običajno zanimajo ne preveliki in ne premajhni otoki - le otoki velikosti med izbranima k in K. Postopek 'otoki' začne s pokrajino popolnoma potopljeno v vodo. Nato znižujemo višino, dokler se ne pojavi otok prave velikosti...



# ...Otoki

**Otoki na povezavah**:
**Network / Create Partition / Islands / Line Weights**

Za te otoke potrebujemo uteženo omrežje (omrežje z vrednostmi na povezavah). Če omrežje ni uteženo, ga lahko naredimo uteženega npr. s štetjem 3 ali 4 ciklov, ali pa (v primeru dvovrstnih omrežij) pretvorimo dvovrstno omrežje v dve uteženi enovrstni omrežji.
*Primera:* shr1.net, Davis.net (dvovrstno omrežje)

---

**Otoki na točkah**:
**Operations / Network + Vector / Islands / Vertex Weights**

Kot vhod rabimo poleg omrežja še vektor, ki določa vrednosti v točkah.

# Iskanje skupnosti - Community Detection...

Poiskati želimo take skupine, da bodo povezave znotraj skupin gostejše (in z večjimi vrednostmi), povezave med skupinami pa redkejše (in z manjšimi vrednostmi).
V Pajku sta na voljo dve metodi za iskanje skupin po metodah 'community detection' in sicer metoda **Louvain** in metoda **VOS Clustering**.
Po metodi **Louvain** iščemo razvrstitev v skupine, ki ima največjo **modularnost (modularity - Q)**:

$$Q = \frac{1}{2m} \sum_s (e_s - r * \frac{K_s^2}{2m})$$

- $m$ – skupno število povezav
- $s$ – skupina
- $e_s = \sum_{ij \in s} A_{ij}$ – dvakratno število povezav znotraj skupine $s$
- $K_s = \sum_{i \in s} k_i$ – vsota stopenj za točke v skupini $s$
- $r$ – *resolution parameter*, privzeta vrednost je 1, ki ustreza osnovni definiciji modularnosti

Podobna metoda je **VOS Clustering**, le da se namesto modularnosti uporabi *VOS quality function*.

# ...Iskanje skupnosti - Community Detection



Pajkov ukaz: **Network / Create Partition / Communities**
Na voljo je precej parametrov s katerimi usmerjamo iskanje skupin, ki pa jih za naše potrebe ni potrebno spreminjati (v poštev pridejo pri zelo velikih omrežjih).

Smiselno pa je preizkusiti različne vrednosti parametra **resolution**. Ta je ponavadi nastavljen na vrednost 1. Vrednost večja od 1 pomeni iskanje večjega števila (manjših) skupin, vrednost manjša od 1 pa iskanje manjšega števila (večjih) skupin.

Razlike med iskanjem *skupnosti* in iskanjem *otokov*:

- Metode za iskanje skupnosti (za razliko od iskanj otokov) lahko uporabimo tudi na neuteženih omrežjih.
- Kot rezultat iskanja skupnosti je vsaka točka dodeljena v eno skupino (skupnost).
- Pri iskanju otokov se samo 'dovolj visokim' točkam dodelijo skupine (otoki), ostale točke ostanejo nerazvrščene v skupini 0 (ne pripadajo nobenemu otoku).

*Primera:* shr1.net, Davis.net (dvovrstno omrežje)

# E-I Index

Obstaja enostavna mera za preverjanje kako dobro izračunane skupine razdelijo omrežje na kohezivne skupine. Imenuje se **E-I Index: External-Internal Index**. Kot je razvidno iz imena, ta indeks izračunamo tako, da od števila povezav med skupinami ($E$) odštejemo število povezav znotraj skupin ($I$) in razliko delimo s skupnim številom povezav ($m$). V poštev lahko vzamemo tudi vrednosti na povezavah.

$$Index_{E-I} = \frac{E - I}{m}$$

Rezultat je število med -1 in 1. Če je E-I Index enak -1, se vse povezave nahajajo znotraj skupin, če je 1 pa se vse povezave nahajajo med skupinami. Vrednost 0 dobimo, če je število povezav (ali vsota vrednosti na povezavah) med skupinami enako številu povezav (ali vsoti vrednosti na povezavah) znotraj skupin. Če razbitje dobro razdeli točke v kohezivne skupine, je večina povezav znotraj skupin, torej bo E-I Index negativen in blizu -1.

**Operations / Network + Partition / Info / E-I Index**

# Primerjave razbitij...

## Cramerjev koeficient



Ponavadi so si skupine dobljene po različnih metodah zelo podobne. Podobnost lahko preverimo s **Partitions** / **Info**, ki nam vrne **Cramerjev** in še nekatere druge koeficiente za primerjavo dveh razbitij.

Obstajajo številne mere za primerjavo dveh razbitij. Razbitja ponavadi predstavljajo nominalne spremenljivke, ki jih lahko predstavimo s kontingenčnimi tabelami.

**Partitions / Info / Cramer's V, Rajski, Adjusted Rand Index**

Za merjenje povezanosti poljubnih dveh spremenljivk (lahko tudi nominalnih) lahko uporabljamo Cramerjev koeficient (*Cramer's V*).

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$


# ...Primerjave razbitij...

## Koeficienti Rajskega...



Koeficient Rajskega (1964) je zgrajen na osnovi *entropije*:
Imejmo dve spremenjivki $X$ in $Y$. Spremenljivka $X$ naj zavzame $n$ različnih vrednosti, spremenljivka $Y$ pa $m$ različnih vrednosti.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$H(Y) = -\sum_{i=1}^{m} p(y_i) \log_2 p(y_i)$$

in

$$H(XY) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i, y_j)$$

# ...Primerjave razbitij...

## ...Koeficienti Rajskega...



*Informacija* med spremenljivkama $X$ in $Y$ je definirana takole:

$$I(X, Y) = H(X) + H(Y) - H(XY)$$

Informacija $I(X, Y)$ doseže vrednost 0, natanko takrat ko za vsak par $x_i$ in $y_j$ velja $p(x_i, y_j) = p(x_i)p(y_j)$, kar pomeni, da sta spremenljivki neodvisni.
Informacija $I(X, Y)$ doseže največjo vrednost, natanko takrat ko med spremenljivkama obstaja funkcijska zveza – v vsakem stolpcu in vsaki vrstici ustrezne kontingenčne tabele je največ en od nič različen element. Tedaj velja:

$$H(X) = H(Y) = H(XY) = I(X, Y)$$

Torej je informacija I(X,Y) mera funkcijske odvisnosti (določenosti) med spremenljivkama $X$ in $Y$.

# ...Primerjave razbitij...

## ...Koeficienti Rajskega...



Koeficienti Rajskega:

$$R(X \leftrightarrow Y) = \frac{I(X,Y)}{H(XY)}$$

$$R(X \rightarrow Y) = \frac{I(X,Y)}{H(Y)}$$

$$R(X \leftarrow Y) = \frac{I(X,Y)}{H(X)}$$

Vsi koeficienti zavzamejo vrednosti med 0 in 1. Vrednost 0 zavzamejo, ko sta spremenljivki neodvisni.
$R(X \rightarrow Y) = 1$, ko je $Y$ funkcija $X$,
$R(X \leftarrow Y) = 1$, ko je $X$ funkcija $Y$ in
$R(X \leftrightarrow Y) = 1$, ko obe spremenljivki ena drugo natanko določata.

# ...Primerjave razbitij...

## ...Koeficienti Rajskega...

**Primer 1:**

| | $y_1$ | $y_2$ | $y_3$ | Sum |
|---|---|---|---|---|
| $x_1$ | 2 | 2 | 1 | 5 |
| $x_2$ | 2 | 1 | 2 | 5 |
| Sum | 4 | 3 | 3 | 10 |

| $p(x_i, y_j)$ | $y_1$ | $y_2$ | $y_3$ | $p(x_i)$ |
|---|---|---|---|---|
| $x_1$ | 0.2 | 0.2 | 0.1 | 0.5 |
| $x_2$ | 0.2 | 0.1 | 0.2 | 0.5 |
| $p(y_j)$ | 0.4 | 0.3 | 0.3 | 1 |

$$R(X \leftrightarrow Y) = 0.0194$$

$$R(X \rightarrow Y) = 0.0312$$

$$R(X \leftarrow Y) = 0.0490$$

Vrednosti vseh treh koeficientov so majhne, na osnovi vrednosti ene spremenljivke ne moremo napovedati vrednosti druge spremenljivke.

**Primer 2:**

| | $y_1$ | $y_2$ | $y_3$ | Sum |
|---|---|---|---|---|
| $x_1$ | 0 | 3 | 0 | 3 |
| $x_2$ | 4 | 0 | 3 | 7 |
| Sum | 4 | 3 | 3 | 10 |

| $p(x_i, y_j)$ | $y_1$ | $y_2$ | $y_3$ | $p(x_i)$ |
|---|---|---|---|---|
| $x_1$ | 0 | 0.3 | 0 | 0.3 |
| $x_2$ | 0.4 | 0 | 0.3 | 0.7 |
| $p(y_j)$ | 0.4 | 0.3 | 0.3 | 1 |

$$R(X \leftrightarrow Y) = 0.5610$$

$$R(X \rightarrow Y) = 0.5610$$

$$R(X \leftarrow Y) = 1$$

Spremenljivka $X$ je funkcija $Y$: če poznamo vrednost spremenljivke $Y$ lahko natanko napovemo vrednost spremenljivke $X$, obratno pa ni res.

**Primer 3:**

| | $y_1$ | $y_2$ | $y_3$ | Sum |
|---|---|---|---|---|
| $x_1$ | 4 | 0 | 0 | 4 |
| $x_2$ | 0 | 0 | 3 | 3 |
| $x_3$ | 0 | 3 | 0 | 3 |
| Sum | 4 | 3 | 3 | 10 |

$$R(X \leftrightarrow Y) = 1$$

$$R(X \rightarrow Y) = 1$$

$$R(X \leftarrow Y) = 1$$

Spremenljivki $X$ in $Y$ ena drugo natanko določata.

**Primer 4:**

| | $y_1$ | $y_2$ | Sum |
|---|---|---|---|
| $x_1$ | 2 | 2 | 4 |
| $x_2$ | 2 | 2 | 4 |
| Sum | 4 | 4 | 8 |

$$R(X \leftrightarrow Y) = 0$$

$$R(X \rightarrow Y) = 0$$

$$R(X \leftarrow Y) = 0$$

Spremenljivki $X$ in $Y$ sta neodvisni


# ...Primerjave razbitij

## Adjusted Rand Index



### Adjusted Rand index [ edit ]

The adjusted Rand index is the corrected-for-chance version of the Rand index.[1][2][3] Such a correction for chance establishes a baseline by using the expected similarity of all pair-wise comparisons between clusterings specified by a random model. Traditionally, the Rand Index was corrected using the Permutation Model for clusterings (the number and size of clusters within a clustering are fixed, and all random clusterings are generated by shuffling the elements between the fixed clusters). However, the premises of the permutation model are frequently violated; in many clustering scenarios, either the number of clusters or the size distribution of those clusters vary drastically. For example, consider that in K-means the number of clusters is fixed by the practitioner, but the sizes of those clusters are inferred from the data. Variations of the adjusted Rand Index account for different models of random clusterings.[4]

Though the Rand Index may only yield a value between 0 and +1, the adjusted Rand index can yield negative values if the index is less than the expected index.[5]

#### The contingency table [ edit ]

Given a set $S$ of $n$ elements, and two groupings or partitions (*e.g.* clusterings) of these elements, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between $X$ and $Y$ can be summarized in a contingency table $[n_{ij}]$ where each entry $n_{ij}$ denotes the number of objects in common between $X_i$ and $Y_j$ : $n_{ij} = |X_i \cap Y_j|$.

| $X \backslash Y$ | $Y_1$ | $Y_2$ | $\cdots$ | $Y_s$ | sums |
|---|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1s}$ | $a_1$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2s}$ | $a_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $X_r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rs}$ | $a_r$ |
| sums | $b_1$ | $b_2$ | $\cdots$ | $b_s$ | |

#### Definition [ edit ]

The original Adjusted Rand Index using the Permutation Model is

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where $n_{ij}, a_i, b_j$ are values from the contingency table.

**https://en.wikipedia.org/wiki/Rand_index**


# Naloge...



Analizirajte omrežje **usair97.net**.

Vrednosti na povezavah v tem omrežju predstavljajo geografske razdalje med letališči. Teh vrednosti ne potrebujemo, oziroma z uporabo teh vrednosti bi dobili napačne rezultate, zato vrednosti na povezavah najprej odstranimo:
**Network / Create New Network / Transform / Line Values / Set All Line Values to 1**

1. **Skupnosti**
   1. Poiščite skupnosti po metodah *Louvain* in *VOS clustering* s parametrom resolucije enakim 1.
   2. Izračunajte kako dobro dobljene skupnosti razdelijo omrežje v kohezivne skupine (E-I index).
   3. Primerjajte razbitja dobljena po metodah Louvain in VOS (uporabite naslednje koeficiente Cramer, Rajski, Adjusted Rand).
   4. Ponovite nalogo še dvakrat: enkrat uporabite parameter resolucije večji od ena, enkrat pa manjši od 1.


# ...Naloge



2. **Otoki na povezavah**
   1. Določite neusmerjene *3-obroče* in v dobljenem omrežju poiščite otoke na povezavah.
   2. Določite neusmerjene *4-obroče* in v dobljenem omrežju poiščite otoke na povezavah.
   3. Primerjajte otoke dobljene na omrežju s 3 in 4-obroči (Cramer, Rajski, Adjusted Rand).