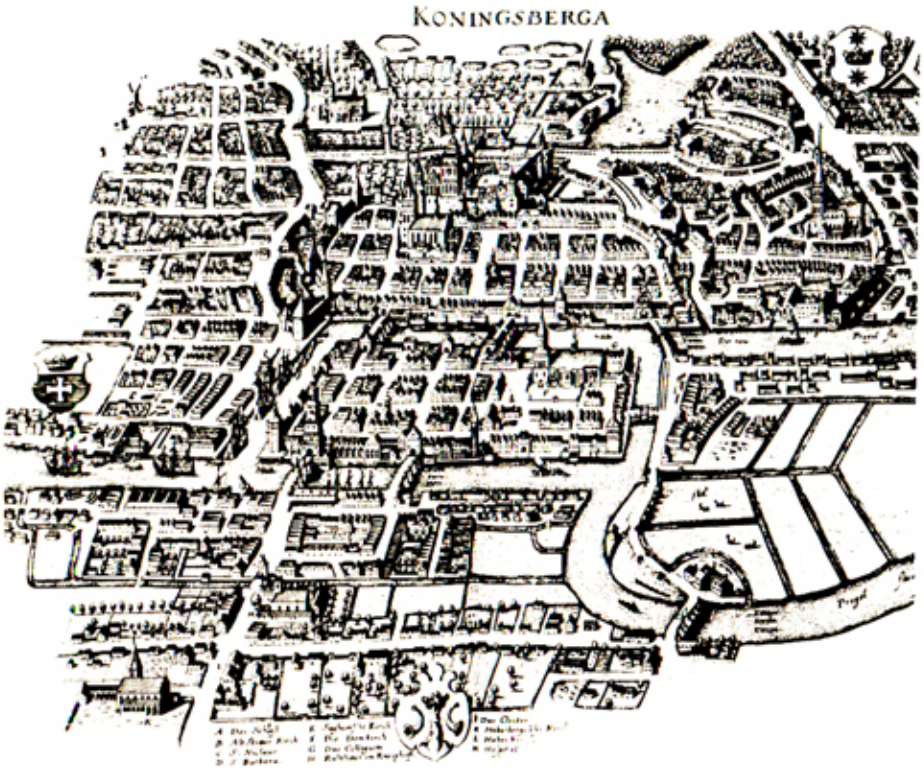


Univerza v Ljubljani
podiplomski študij statistike

Analiza omrežij
3. Zgradba omrežij: podomrežja

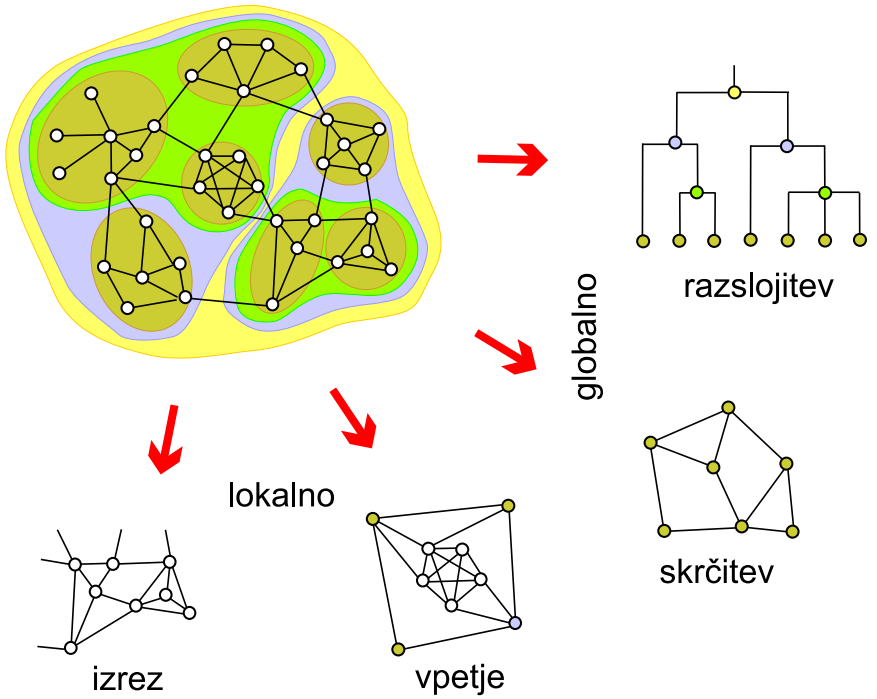
Vladimir Batagelj

Univerza v Ljubljani

Ljubljana, 17. november 2006 / 10. in 17. november 2003

Kazalo

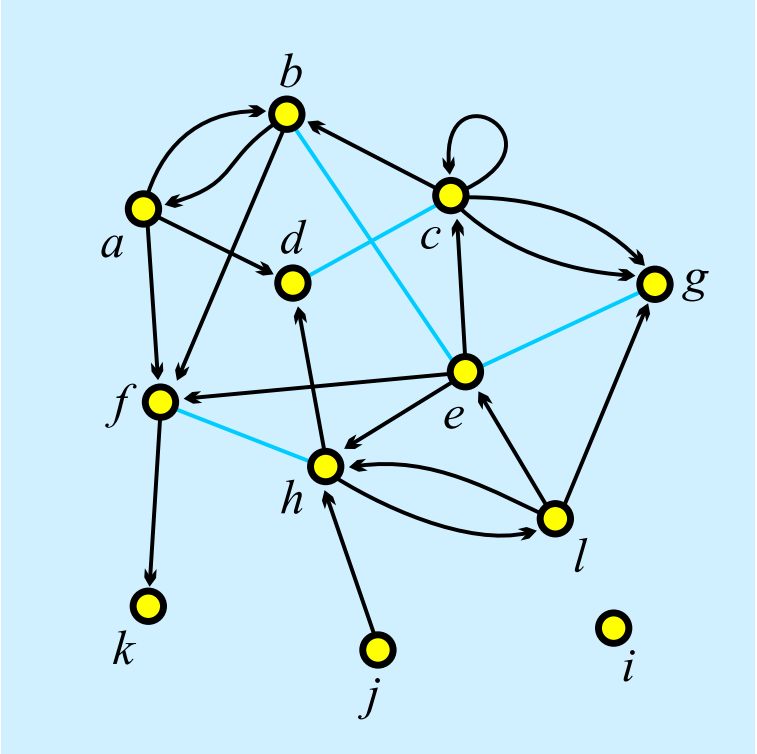
Pristopi k velikim omrežjem



Pri *velikih* omrežjih (več tisoč ali milijonov točk, omrežje je mogoče shraniti v pomnilniku) se moramo odpovedati celoviti sliki, uporabni so le redki postopki.

Za analizo velikih omrežij lahko uporabimo statistiko ali pa izpeljana mala in srednja (pod)omrežja.

Stopnje grafa



stopnja točke v, $\deg(v) =$ je število povezav, ki imajo točko v za krajišče;
vhodna stopnja točke v, $\text{indeg}(v) =$ je število povezav, ki imajo točko v za konec (krajišče neusmerjene povezave je hkrati njen začetek in konec);
izhodna stopnja točke v, $\text{outdeg}(v) =$ je število povezav, ki imajo točko v za začetek.

$$n = 12,\ m = 23,\ \text{indeg}(e) = 3,\ \text{outdeg}(e) = 5,\ \deg(e) = 6$$

$$\sum_{v\in\mathcal{V}} \text{indeg}(v) = \sum_{v\in\mathcal{V}} \text{outdeg}(v) = |\mathcal{A}| + 2|\mathcal{E}|,\ \sum_{v\in\mathcal{V}} \deg(v) = 2|\mathcal{L}| - |\mathcal{L}_0|$$

Statistika

Vhodni podatki o točkah

- številski → **`vector`**
- urejenostni → **`permutation`**
- imenski → **`clustering`** (razbitje)

Izračunane lastnosti točk

globalne: število točk, usmerjenih/neusmerjenih povezav, komponent; največje središčno število, ...

local: stopnje, središčna števila, indksi (vmesnost, dostopnost, viri in kazala, ...)

pregledi: razbitja, vektorji, vrednosti povezav, ...

Analiza povezanosti med izračunanimi (strukturnimi) lastnostmi in vhodnimi (izmerjenimi) lastnostmi.

... Statistika

Globalne lastnosti izpišejo **Pajek**ovi ukazi v poročilo; največ jih je dosegljivih v izbiri **Info**. Pri uporabi *ponavljajočih* ukazov se shranijo v vektorje.

Lokalne lastnosti izračunajo razni **Pajek**ovi ukazi in jih shranijo v vektorje ali razbitja. Njihove vrednosti / porazdelitev si lahko ogledamo v izbiri **Info**.

Za primer si oglejmo omrežje The Edinburgh Associative Thesaurus. EAT je omrežje asociacij med besedami zbranih na študentski populaciji. Točke so besede. Povezave (X, Y) pa so določene z vprašanjem: Katera beseda Y vam pride prva na misel, ko slišite besedo X ? Utež povezave pove, kolikokrat je bila izbrana.

```
File/Network/Read  eatRS.net
Info/Network/General
```

Ima 23219 točk in 325624 usmerjenih povezav (564 zank); 227481 povezav ima utež 1.

… Statistika

Točke z največjimi stopnjami dobimo takole:

```
Net/Partitions/Degree/All
Partition/Make vector
Info/Vector  +10
```

V EAT so to:

	vertex	deg	label
1	12720	1108	ME
2	12459	1074	MAN
3	8878	878	GOOD
4	18122	875	SEX
5	13793	803	NO
6	13181	799	MONEY
7	23136	732	YES
8	15080	723	PEOPLE
9	13948	720	NOTHING
10	22973	716	WORK

Pajek in R

Pajek 0.89 (in kasnejši) omogoča uporabo statističnega programa R in tudi drugih programov kot orodij (izbira `Tools`).

V programu **Pajek** določimo stopnje točk in jih 'podtaknemo' R-ju

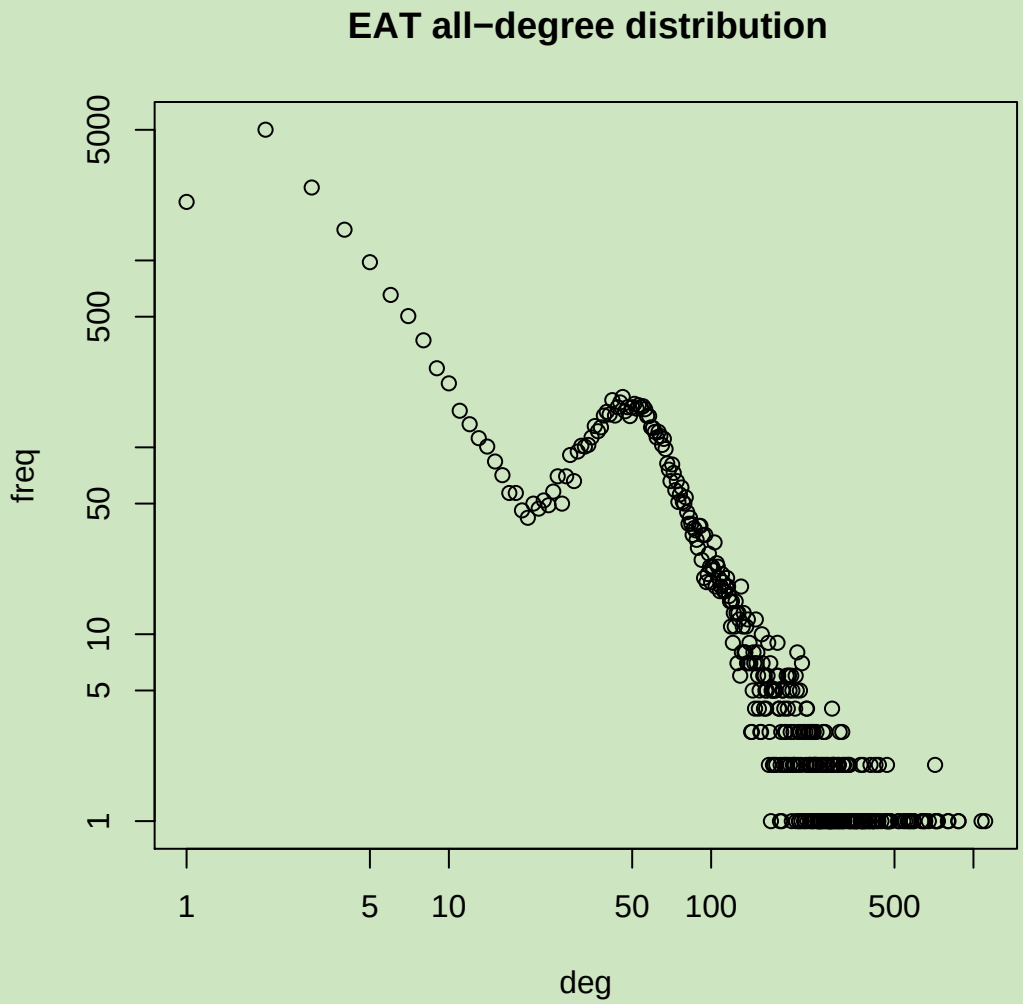
```
info/network/general
Net/Partitions/Degree/All
Partition/Make Vector
Tools/Program R/Send to R/Current Vector
```

Tu določimo porazdelitev stopenj in jo narišemo

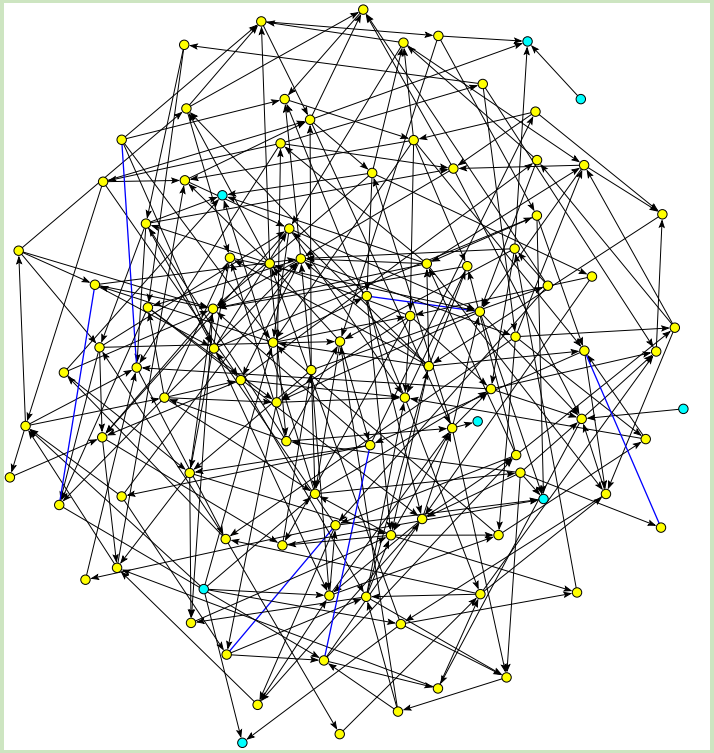
```
summary(v2)
t <- tabulate(v2)
c <- t[t>0]
i <- (1:length(t))[t>0]
plot(i,c,log='xy',main='degree distribution',
  xlab='deg',ylab='freq')
```

Pozor! Če obstajajo točke stopnje 0, jih `tabulate` ne upošteva.

EAT – porazdelitev stopenj



Slučajni grafi



Erdős in Rényi sta definirala *slučajni graf* takole: vsako mogočo povezavo vključimo v slučajni graf z dano verjetnostjo p.

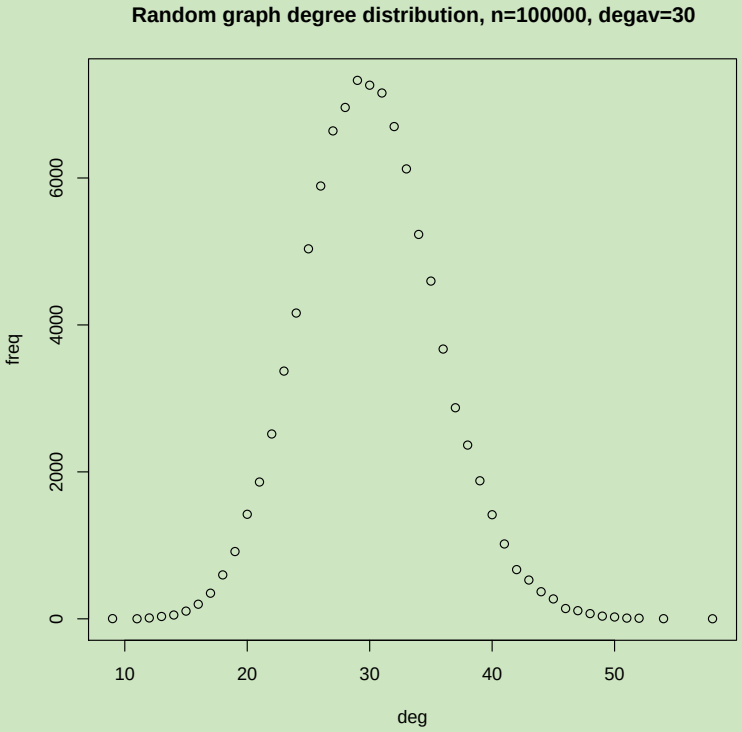
V programu **Pajek** (`Net / Random Network / Erdos-Renyi`) uporabljamo namesto verjetnosti p povprečno stopnjo

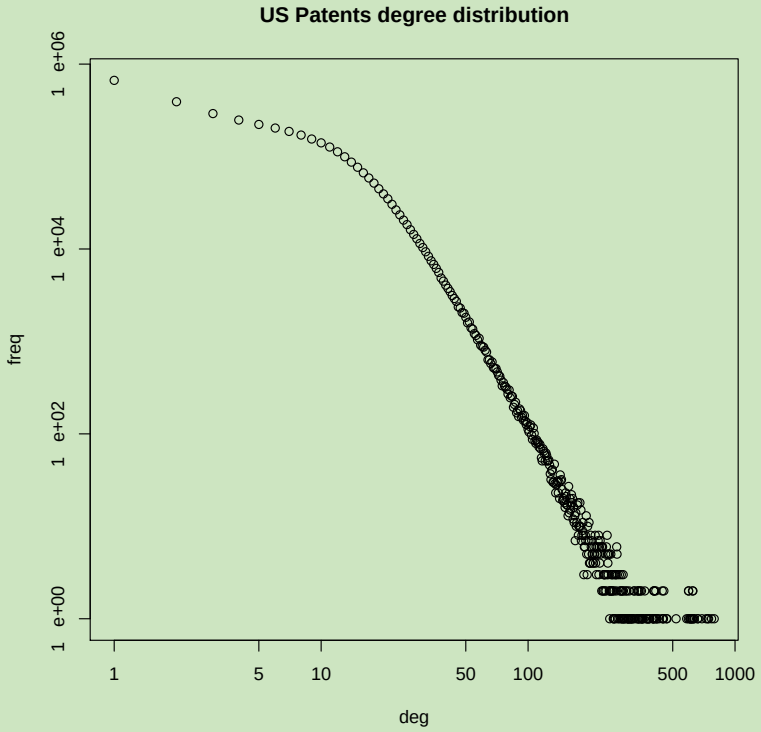
$$\overline{\deg} = \frac{1}{n} \sum_{v \in \mathcal{V}} \deg(v)$$

Velja $p = \frac{m}{m_{max}}$ in, za enostavne grafe, še $\overline{\deg} = \frac{2m}{n}$.

Na sliki je prikazan slučajni graf na 100 točkah s povprečno stopnjo 3.

Porazdelitve stopenj





Dejanska omrežja so vse prej kot slučajna. Analiza porazdelitev je dala nov pogled na zgradbo dejanskih omrežij – Watts (Small worlds), Barabási (nd/networks, Linked).

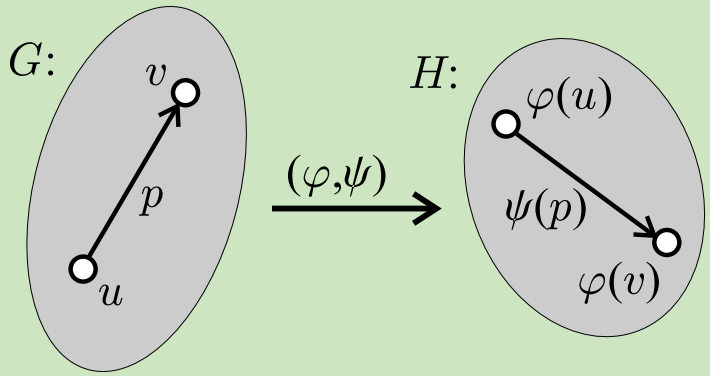
Povezanosti med grafi

Preslikavi (φ, ψ), $\varphi: \mathcal{V} \to \mathcal{V}'$ in $\psi: \mathcal{L} \to \mathcal{L}'$ določata *šibki homomorfizem* grafa $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ v graf $\mathcal{H} = (\mathcal{V}', \mathcal{L}')$ ntk velja:

$$\forall u, v \in \mathcal{V} \forall p \in \mathcal{L} : (p(u : v) \Rightarrow \psi(p)(\varphi(u) : \varphi(v)))$$

in določa *(krepki) homomorfizem* grafa $\mathcal{G}$ v graf $\mathcal{H}$ ntk velja:

$$\forall u, v \in \mathcal{V} \forall p \in \mathcal{L} : (p(u, v) \Rightarrow \psi(p)(\varphi(u), \varphi(v)))$$

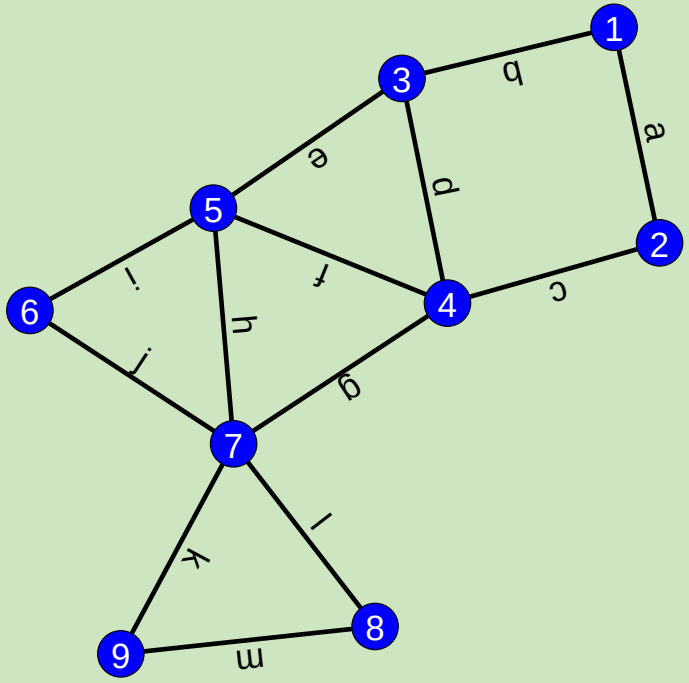


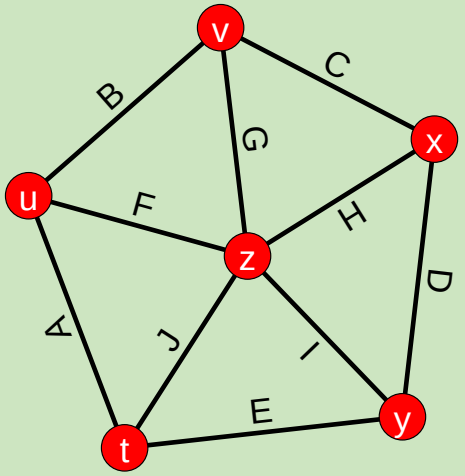
Ko sta φ in ψ bijekciji in ustrezni pogoj velja v obe smeri, govorimo o *izomorfizmu* grafov $\mathcal{G}$ in $\mathcal{H}$. Da sta grafa šibko izomorfna zapišemo $\mathcal{G} \sim \mathcal{H}$; da sta (krepko) izomorfna pa $\mathcal{G} \approx \mathcal{H}$. Velja $\approx \subset \sim$.

Stalnica ali *invarianta* grafa imenujemo vsako grafu prirejeno število, ki je enako za vse med seboj izomorfne grafe. .

EulerGT

Homomorfizem

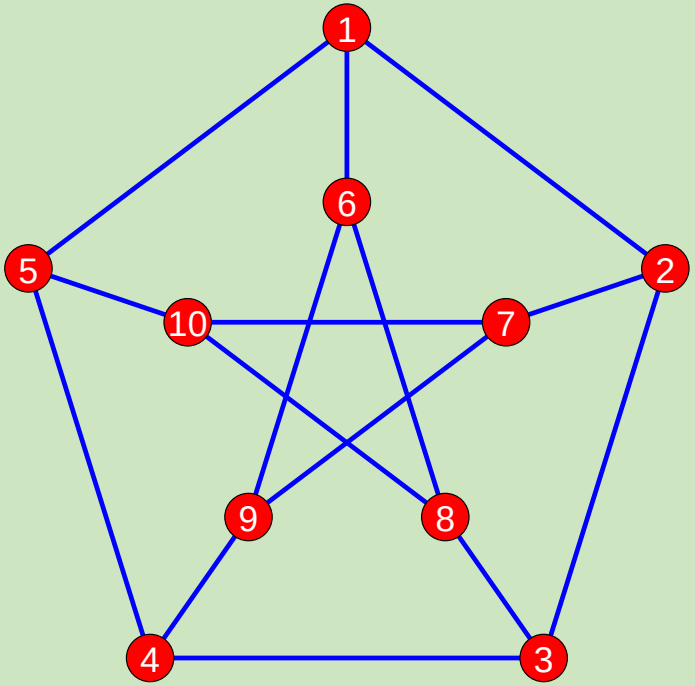


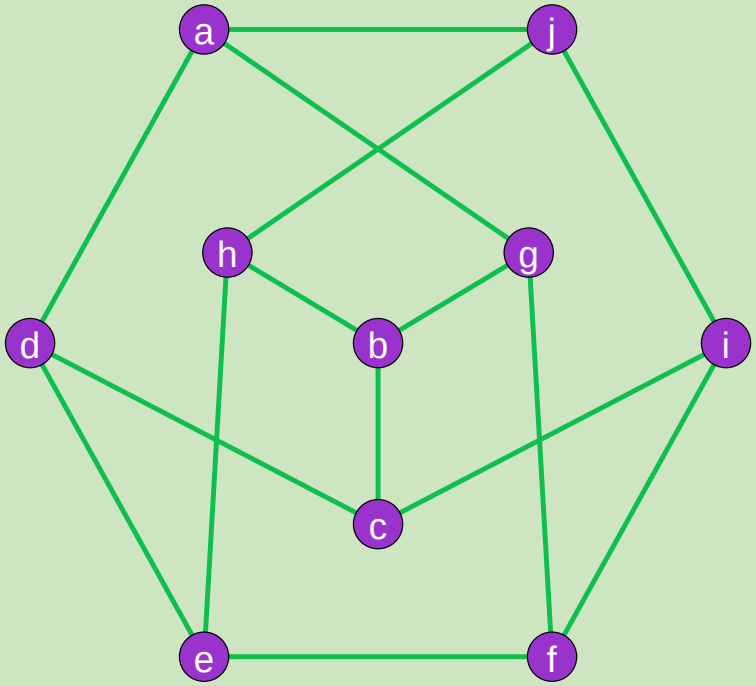


φ	1	2	3	4	5	6	7	8	9
	x	y	t	z	x	z	t	z	x

ψ	a	b	c	d	e	f	g	h	i	j	k	l	m
	D	E	H	J	E	I	J	E	I	J	E	J	I

Izomorfna grafa





φ	1	2	3	4	5	6	7	8	9	10
	b	h	j	a	g	c	e	i	d	f

Skupine, razvrstitve, razbitja, razslojitve

Neprazno podmnožico $C \subseteq \mathcal{V}$ imenujemo *skupina*. Neprazna množica skupin $\mathbf{C} = \{C_i\}$ je *razvrstitev*.

Razvrstitev $\mathbf{C} = \{C_i\}$ je *razbitje* ntk

$$\cup\mathbf{C} = \bigcup_i C_i = \mathcal{V} \quad \text{in} \quad i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

Razvrstitev $\mathbf{C} = \{C_i\}$ je *razslojitev* ali *hierarhija* ntk

$$C_i \cap C_j \in \{\emptyset, C_i, C_j\}$$

Razslojitev $\mathbf{C} = \{C_i\}$ je *polna*, če je $\cup\mathbf{C} = \mathcal{V}$; in je *osnovna*, če je za vsak $v \in \cup\mathbf{C}$ tudi $\{v\} \in \mathbf{C}$.

Primer razbitja in razslojitve

$$
\begin{aligned}
\mathcal{V} &= \{a, b, c, d, e, f, g\} \\
\mathbf{C} &= \{\{a, b, e\}, \{c, g\}, \{d, f\}\} \\
C_2 &= \{c, g\} \\
\mathbf{H} &= \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}, \\
&\quad \{a, e\}, \{c, g\}, \{d, f\}, \{a, b, e\}, \\
&\quad \{c, d, f, g\}, \{a, b, c, d, e, f, g\}\}
\end{aligned}
$$

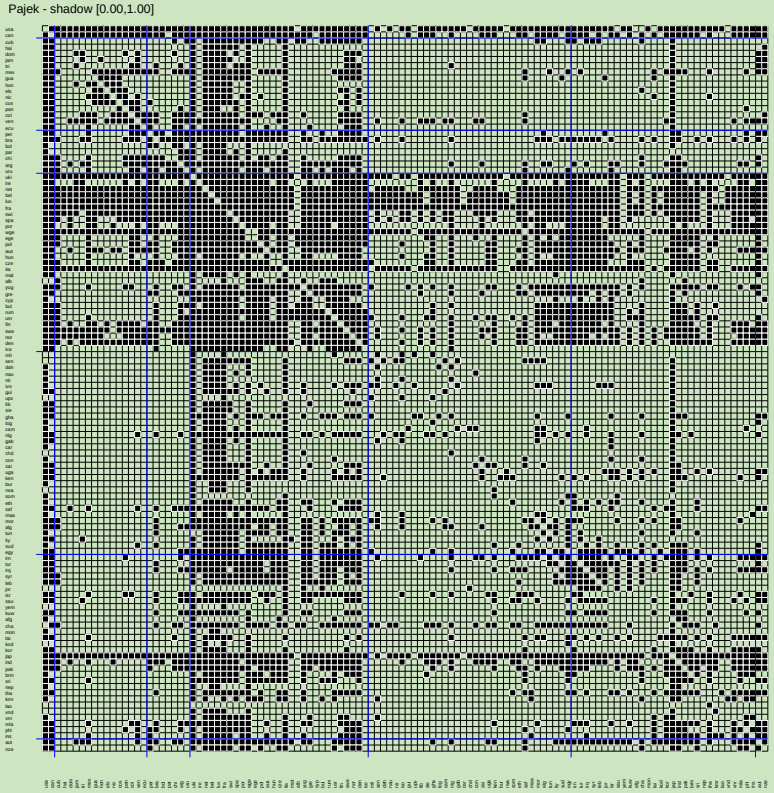
Skrčitev skupine

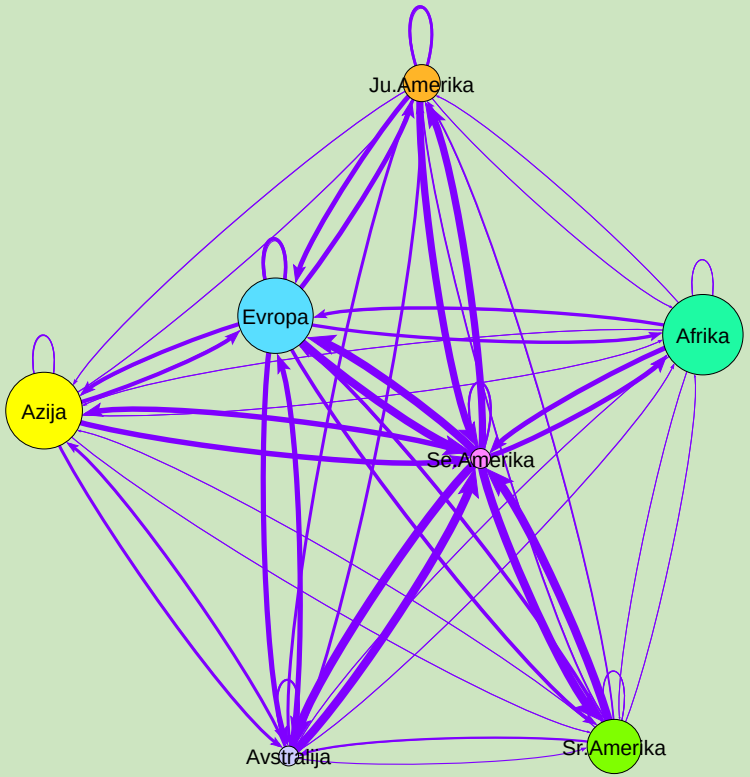
Skrčitev skupine C imenujemo graf $\mathcal{G}/C$, ki ga dobimo tako da vse točke skupine C zamenjamo z eno točko, recimo c. Natančneje $\mathcal{G}/C = (\mathcal{V}', \mathcal{L}')$, kjer je $\mathcal{V}' = (\mathcal{V} \setminus C) \cup \{c\}$ in $\mathcal{L}'$ sestavljajo povezave iz $\mathcal{L}$, ki imajo obe krajišči v $\mathcal{V} \setminus C$. Poleg teh pa še 'zvezda' z vrhom c in krakom (v, c), če $\exists p \in \mathcal{L}, u \in C : p(v, u)$, oziroma krakom (c, v), če $\exists p \in \mathcal{L}, u \in C : p(u, v)$. V točki c je zanka (c, c), če $\exists p \in \mathcal{L}, u, v \in C : p(u, v)$.

V omrežju nad grafom $\mathcal{G}$ moramo povedati še, kako so določene vrednosti/uteži v skrčenem delu. Običajno kar kot vsota ali maksimum/minimum izvornih vrednosti.

`Operations / Shrink Natwork / Partition`

Skrčitev skupin – trgovanje med državami





Snyder in Kickovi podatki o trgovanju med državami. Matrični prikaz gostih omrežij.

$$w(C_i, C_j) = \frac{n(C_i, C_j)}{n(C_i) \cdot n(C_j)}$$

Izračun uteži w

```
File / Pajek Project File / Read [SKtrade.paj]
Operations / Shrink Network / Partition [1][0]
```

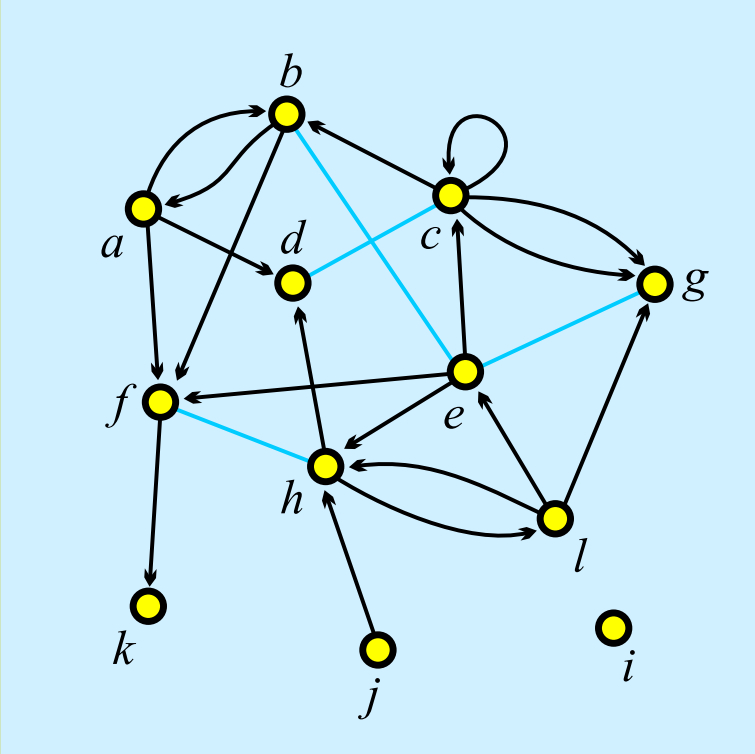
		1	2	3	4	5	6	7
#usa	1.	2	30	13	56	42	45	4
#cub	2.	30	74	25	196	20	37	12
#per	3.	12	28	33	124	16	36	5
#uki	4.	55	217	130	695	427	483	41
#mli	5.	42	8	14	406	122	117	11
#irn	6.	43	37	43	444	142	307	30
#aut	7.	4	4	5	39	9	30	2
count		2	15	7	29	33	30	2

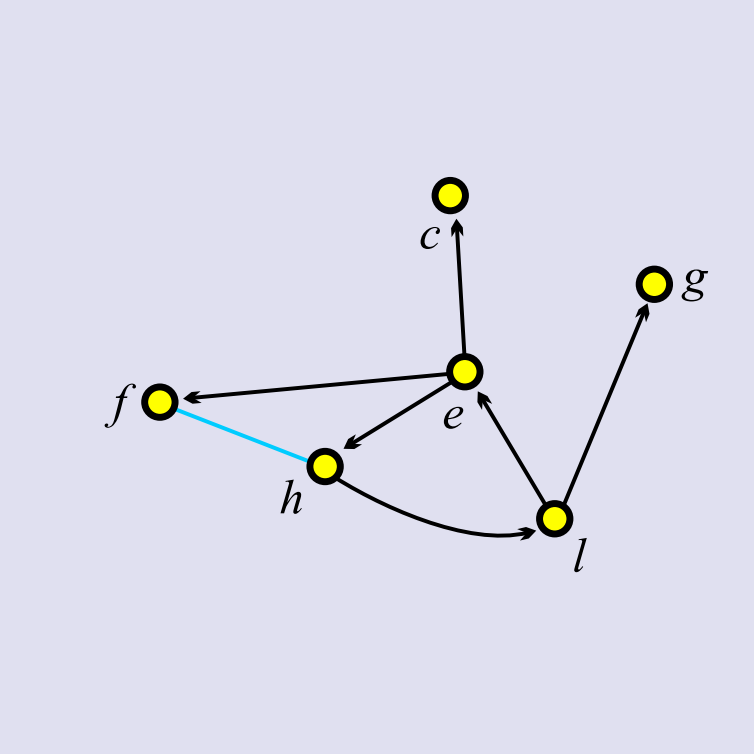
```
select partition SKtrade.clu
Partition / Count
Partition / Make Vector
Vector / Create Identity Vector [7]
select as second vector From partition ...
Vectors / Divide First by Second
select network Shrinking N? according to C?
Operations / Vector # Network / input
Operations / Vector # Network / output
```

		1	2	3	4	5	6	7
#usa	1.	0.50	1.00	0.93	0.97	0.64	0.75	1.00
#cub	2.	1.00	0.33	0.24	0.45	0.04	0.08	0.40
#per	3.	0.86	0.27	0.67	0.61	0.07	0.17	0.36
#uki	4.	0.95	0.50	0.64	0.83	0.45	0.56	0.71
#mli	5.	0.64	0.02	0.06	0.42	0.11	0.12	0.17
#irn	6.	0.72	0.08	0.20	0.51	0.14	0.34	0.50
#aut	7.	1.00	0.13	0.36	0.67	0.14	0.50	0.50

Ker so na diagonali 0, bi jih bilo smiselno pred izračunom postaviti na 1.

Podgraf

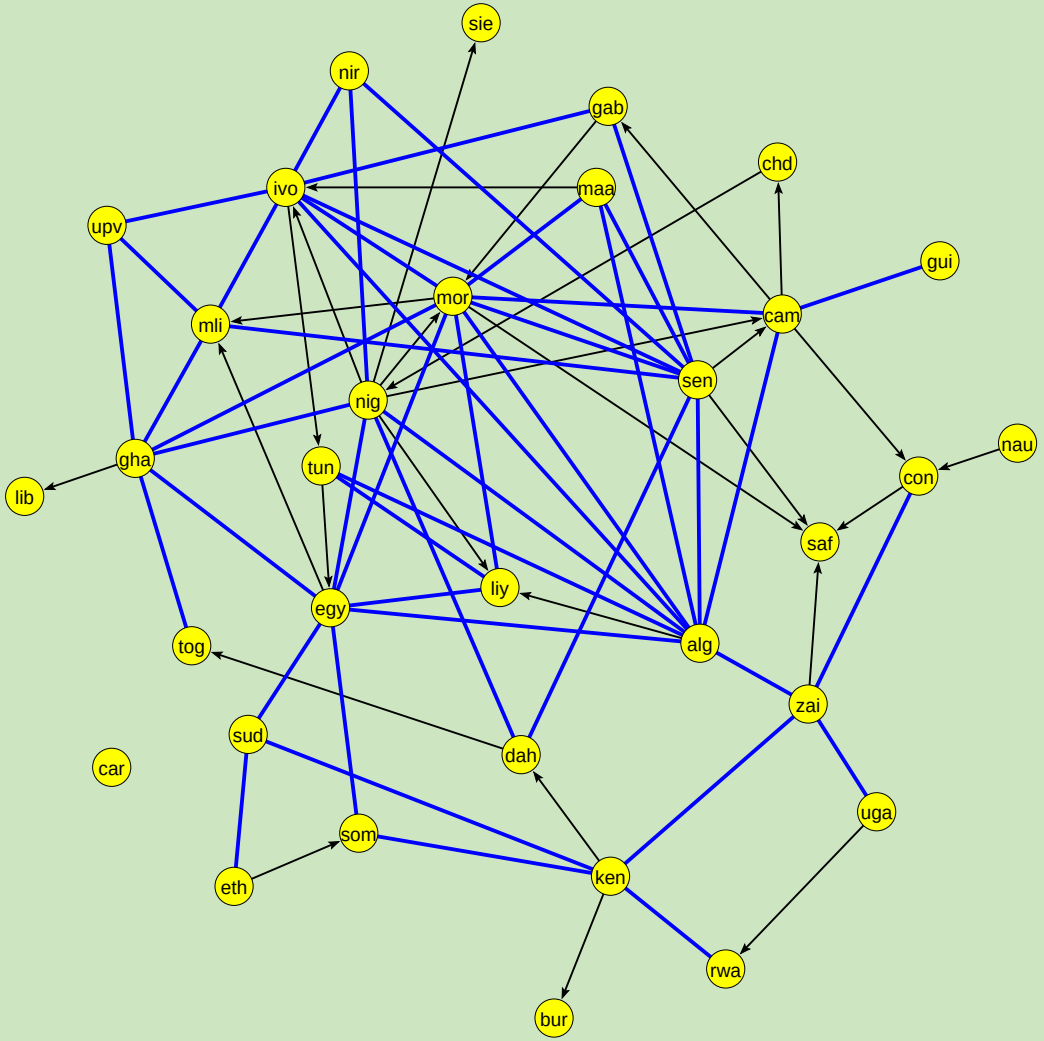




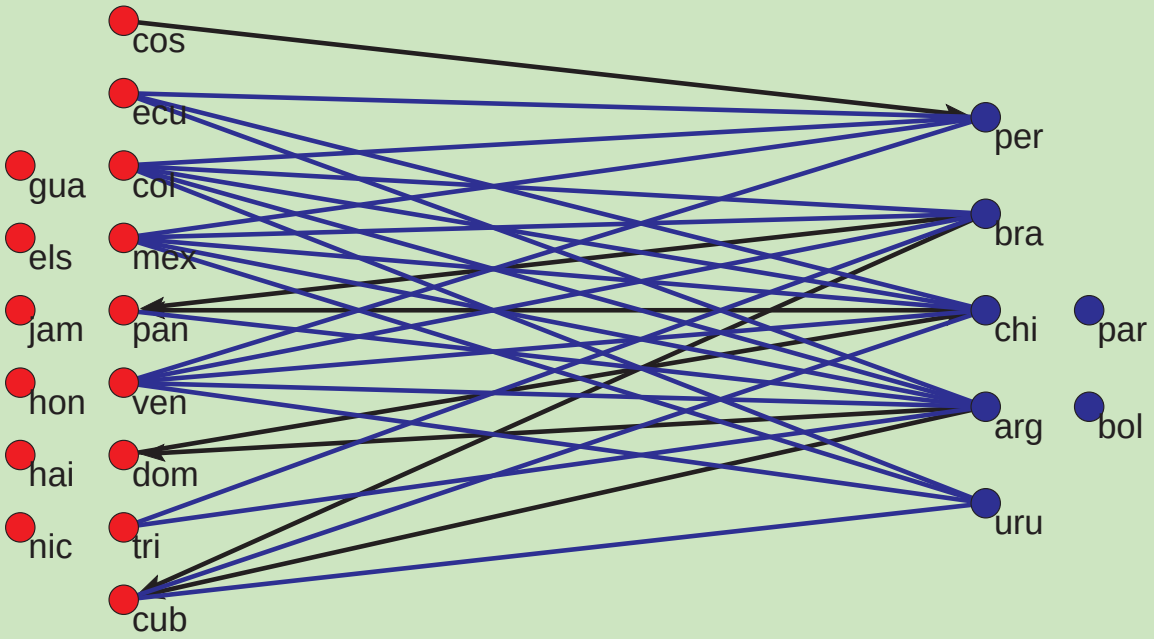
Podgraf $\mathcal{H} = (\mathcal{V}', \mathcal{L}')$ danega grafa $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ je graf, katerega povezave $\mathcal{L}'$ so vsebovane v povezavah grafa $\mathcal{G}$, $\mathcal{L}' \subseteq \mathcal{L}$, točke $\mathcal{V}'$ pa v točkah grafa $\mathcal{G}$, $\mathcal{V}' \subseteq \mathcal{V}$, in vsebujejo tudi vsa krajišča povezav $\mathcal{L}'$.

Podgraf je lahko *porojen* z dano podmnožico točk ali povezav. Podgraf je *vpet*, če je $\mathcal{V}' = \mathcal{V}$.

Izrez: Snyder in Kick – Afrika



Izrez: Snyder in Kick
Latinska Amerika : Južna Amerika




```
Operations/Extract from Network/Partition [5]

Operations/Extract from Network/Partition   [2,3]
Operations/Transform/Remove lines/Inside clusters   [2,3]

[Draw] Move/Grid
```

Prerezi

Točkovni prerez omrežja $\mathcal{N} = (\mathcal{V}, \mathcal{L}, p)$, $p : \mathcal{V} \rightarrow \mathbb{R}$, na *ravni* t je podomrežje $\mathcal{N}(t) = (\mathcal{V}', \mathcal{L}(\mathcal{V}'), p)$, določeno z množico točk

$$\mathcal{V}' = \{v \in \mathcal{V} : p(v) \geq t\}$$

kjer je $\mathcal{L}(\mathcal{V}')$ množica vseh povezav iz $\mathcal{L}$, ki imajo obe krajišči v $\mathcal{V}'$.

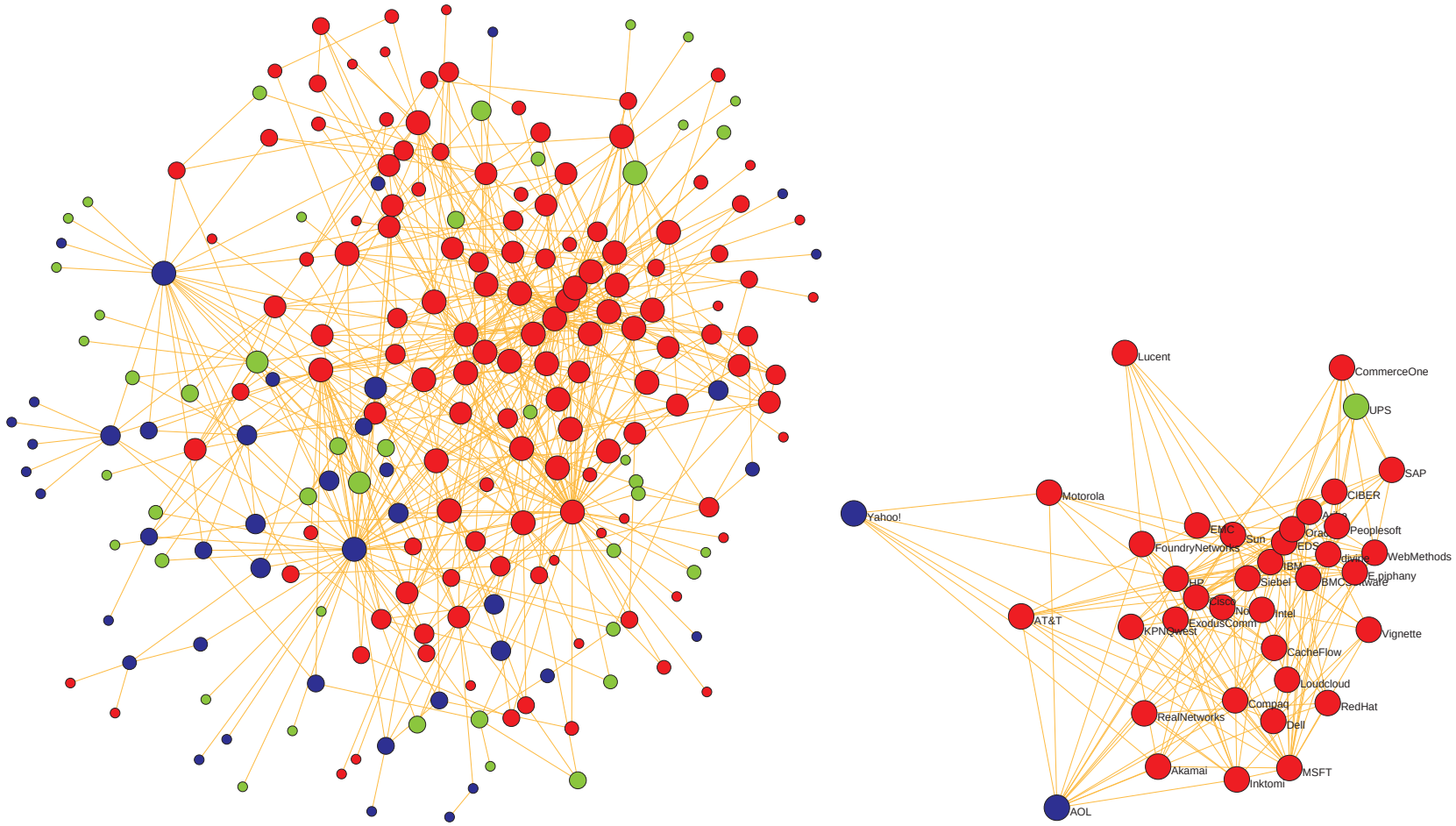
Povezavni prerez omrežja $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, $w : \mathcal{L} \rightarrow \mathbb{R}$, na *ravni* t je določeno z množico povezav

$$\mathcal{L}' = \{e \in \mathcal{L} : w(e) \geq t\}$$

To je podomrežje $\mathcal{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, kjer je $\mathcal{V}(\mathcal{L}')$ množica vseh krajišč povezav iz $\mathcal{L}'$.

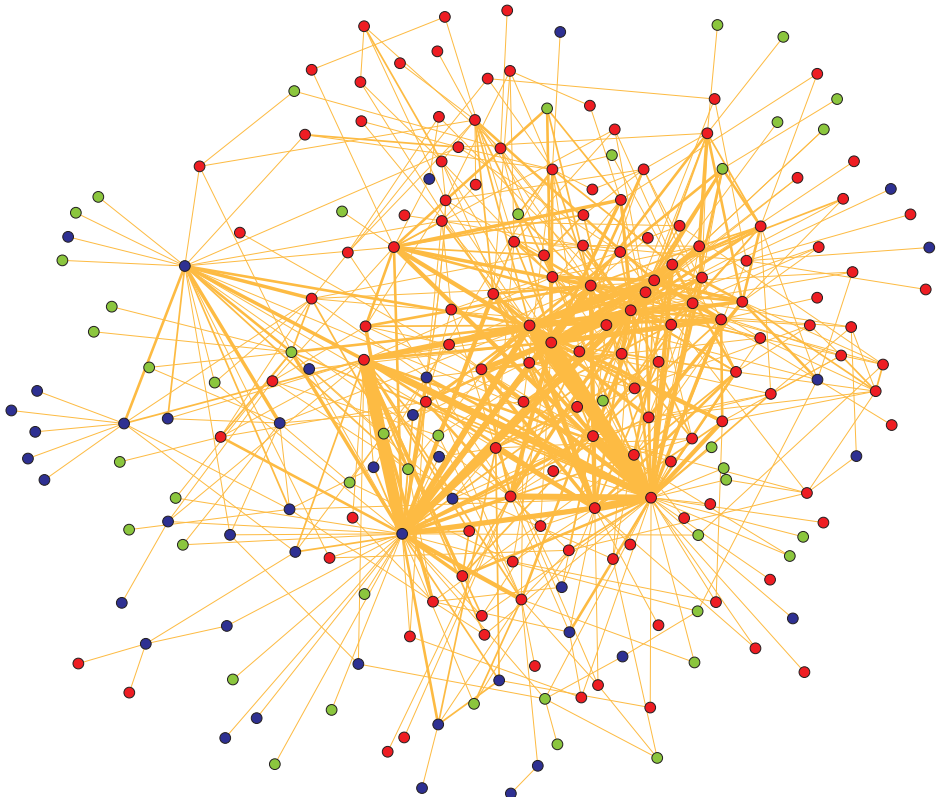
Vrednosti ravni t določimo na osnovi porazdelitve vrednosti funkcij w oziroma p. Običajno nas zanimajo komponente prereza, ki niso niti prevelike, niti premajhne.

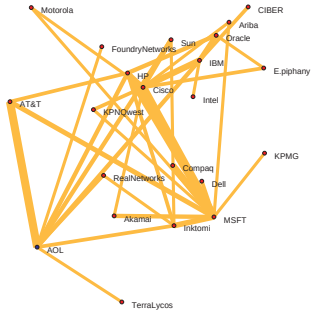
Točkovni prerez: Krebsova spletna podjetja, core=6

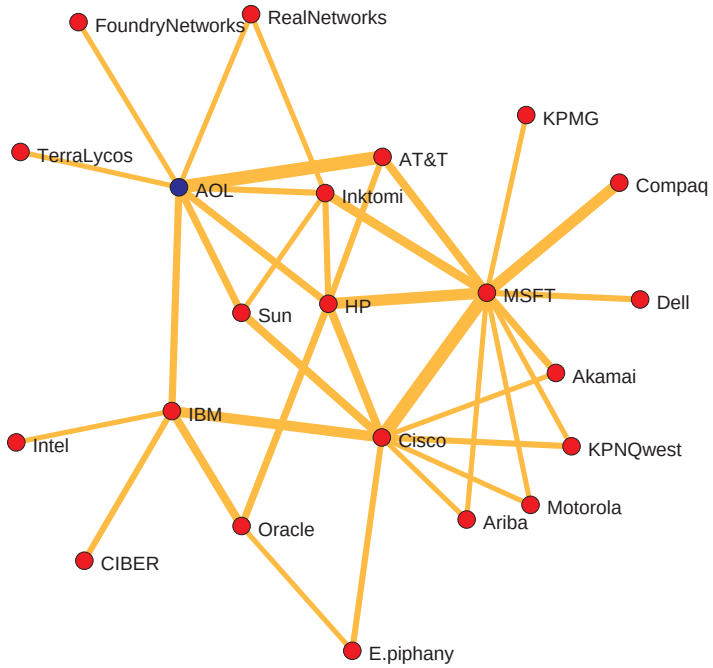


Vsaka točka predstavlja spletno podjetje dejavno v obdobju 1998 do 2001. $n = 219$, $m = 631$. rdeča – vsebina, modra – podporne storitve, zelena – trgovina. Podjetji sta povezani, če sta najavili skupni posel ali drugo obliko sodelovanja.

Povezavni prerez: Krebsova spletna podjetja, $w_3 \geq 5$







Povezavni prerez: EAT

```
File/Network/read  eatRS.net
Info/Network/Line values          ... >= 70
Net/Transform/Remove/Lines with Value/lower than   70
Net/Partitions/Degree/All
Operations/Extract from Network/Partition   1-*
Net/Components/Weak
Draw/Draw-Partition
```

Analiza omrežja s prerezi

Prerezi ponujajo enostaven pristop k analizi omrežij. Za izbrano lastnost/utež in raven t določimo pripadajoči prerez $\mathcal{N}(t)$. Pozornost posvetimo njegovim komponentam.

Število in velikost komponent je odvisna od ravni t. Pogosto se pojavi več majhnih komponent. Pri analizi nas običajno zanimajo le komponente 'prave' velikosti – vsaj k in ne večje kot K. Premajhne komponente zavržemo kot 'nezanimive'; prevelike komponente pa ponovno prerežemo na neki višji ravni.

Vrednost t, k in K določimo s pregledom porazdelitve vrednosti lastnosti/uteži in z upoštevanjem dodatnega vedenja o značilnostih omrežja in ciljih raziskave.

V program **`Pajek`** je vgrajenih nekaj novih, učinkovito izračunljivih lastnosti/uteži.