Social media jako nové pole pro data mining

30. 5. 2013 Josef Šlerka

Social data profiling

řekni mi co lajkuješ a já to povím, kdo jsi

ČSSD Hostěradice
Political Party

OVV ČSSD Pardubice
Organization

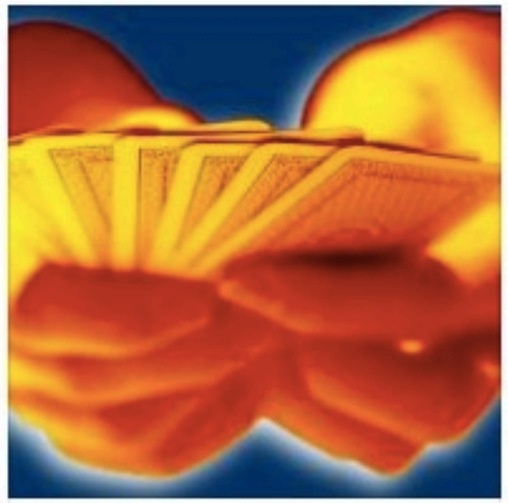
OVV ČSSD Strakonice
Political Party

Podporuji, aby prezidentem ...
Community

Vladimír Plaček
Politician

MO ČSSD Krnov
Political Party

ČSSD Moravská Ostrava a Pří...
Political Party



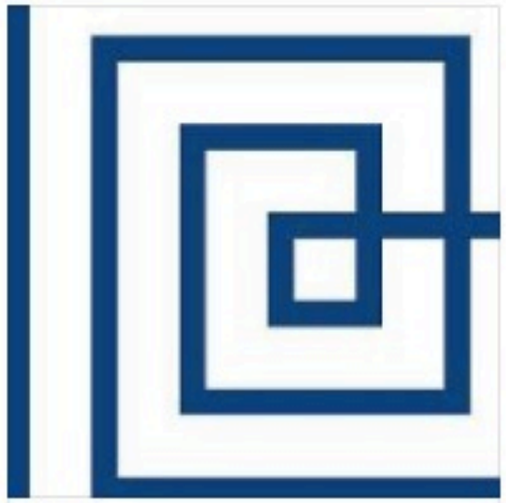
Výklad z karet
Community

Okresní organizace ČSSD Jičín
Political Party

Vinný sklípek u Kratochvílů
Restaurant

Deník Legie
Magazine

Adam Rykala
Politician



Centre Tchèque České Centr...
Non-Profit Organization

Kalhoty pro Václava Havla
Community

Bel Mondo
Media/News/Publishing

Izrael v České republice
Consulate & Embassy

Kamil Fila rulez
Community

Evropská observatoř žurnalis...
Media/News/Publishing

Woody Allen
Public Figure

TIME
Media/News/Publishing

100 let skautingu v českých ...
Community

causes.com
App

HNZprávy
Newspaper

David Lynch
Artist

Knihobežník
Games/Toys

UKáčko.cz – Portál studentů ...
News/Media Website

FFakt, časopis studentů Filoz...
Magazine

#procjsemtady
Community

Akční letenky
Travel/Leisure

Bloxxter
Local Business

brmlab
Education

Mládež pražského seniorátu ...
Community

ČCE Praha 10 – Vršovice
Religious Organization

qwerly
Product/Service

www.czechoriginalfashion.cz
Website

SNM.gug.cz
Community

Prezidentská volba

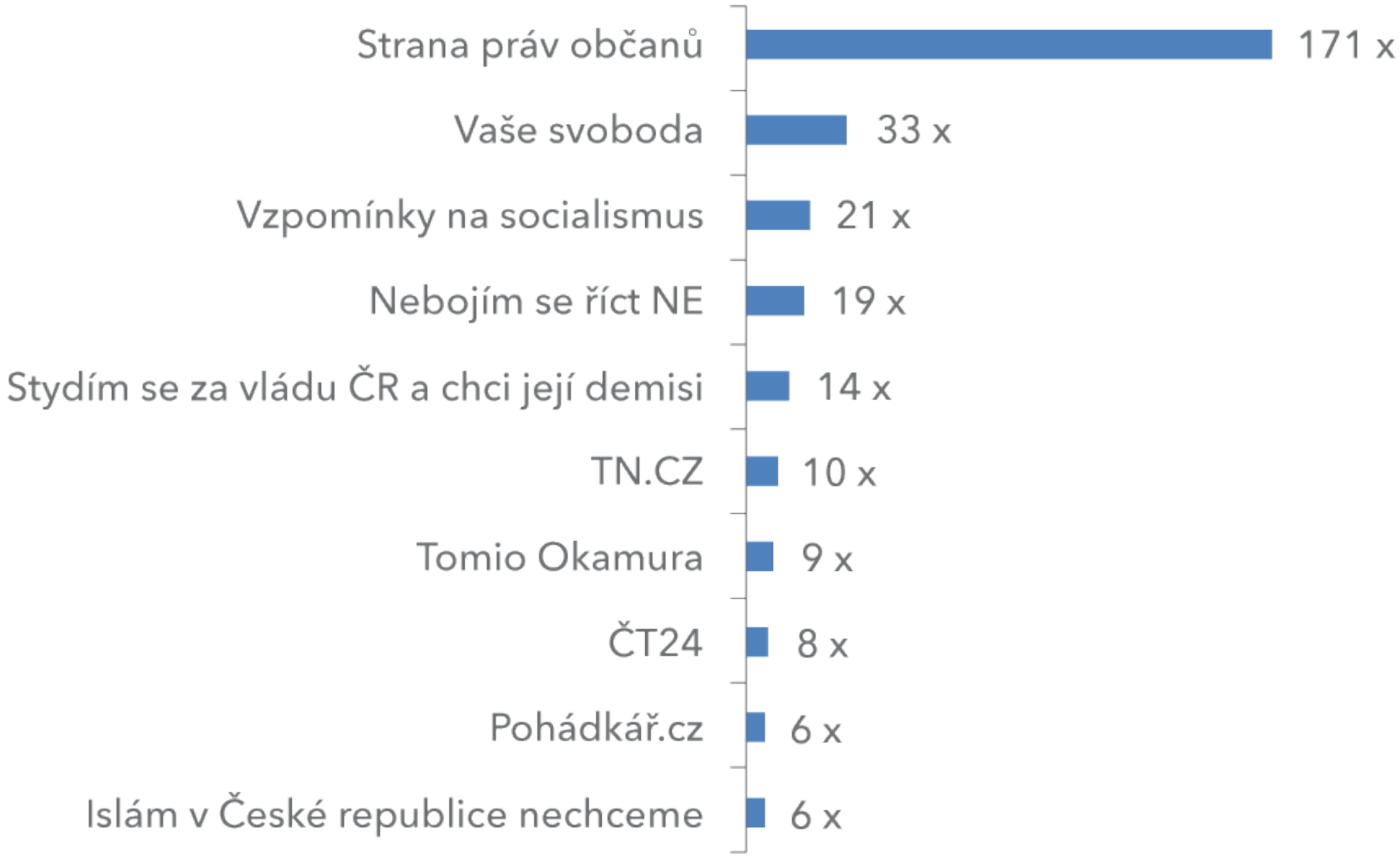
analýza fanoušků prezidentských kandidátů na Facebooku

analyzováno přes 1.600 fanoušků

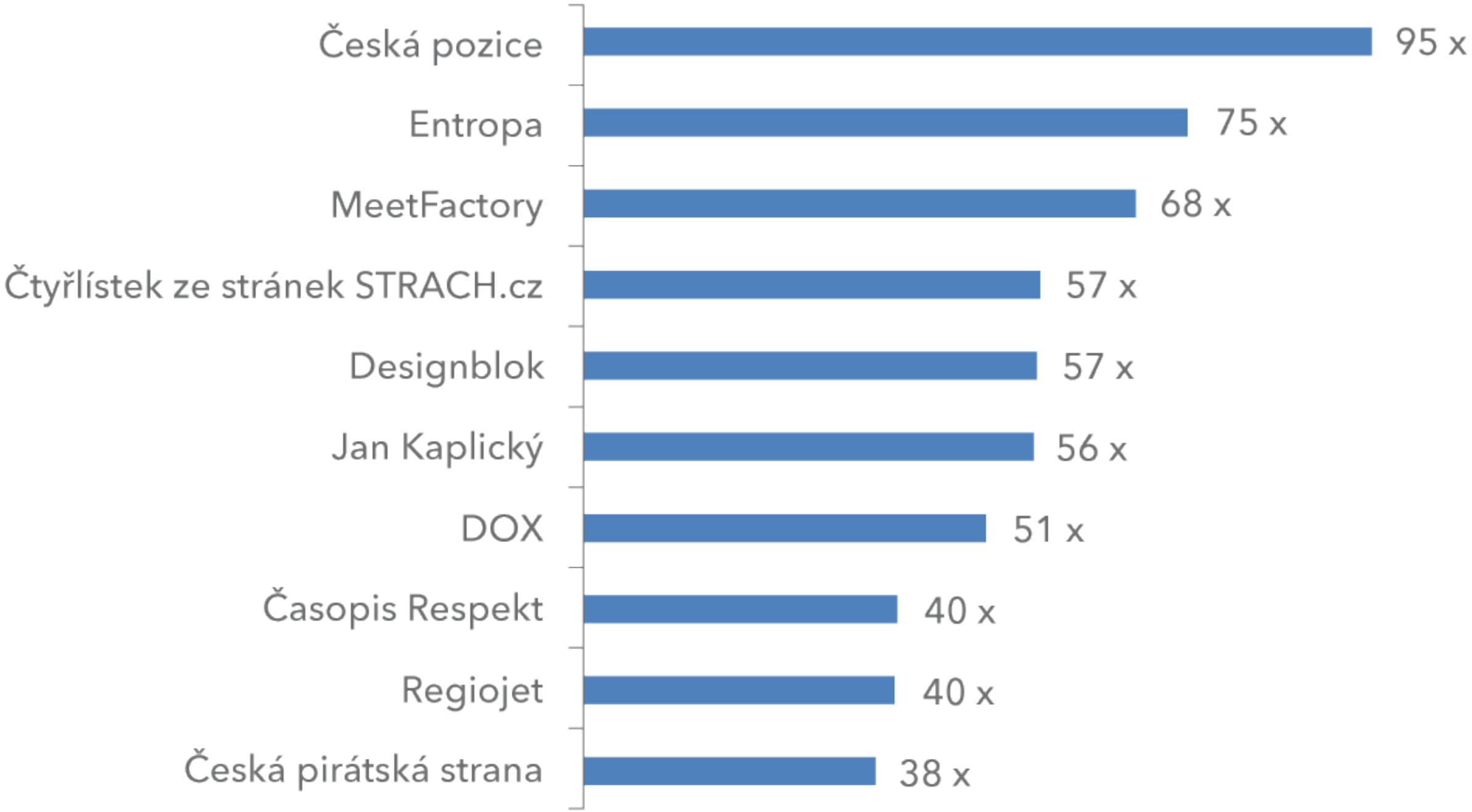
mapa charakteristických odchylek fanoušků od průměrného uživatele

společný projekt České televize a Studia nových médií a blízkých lidí (Josef Šlerka, Jan Schmid)

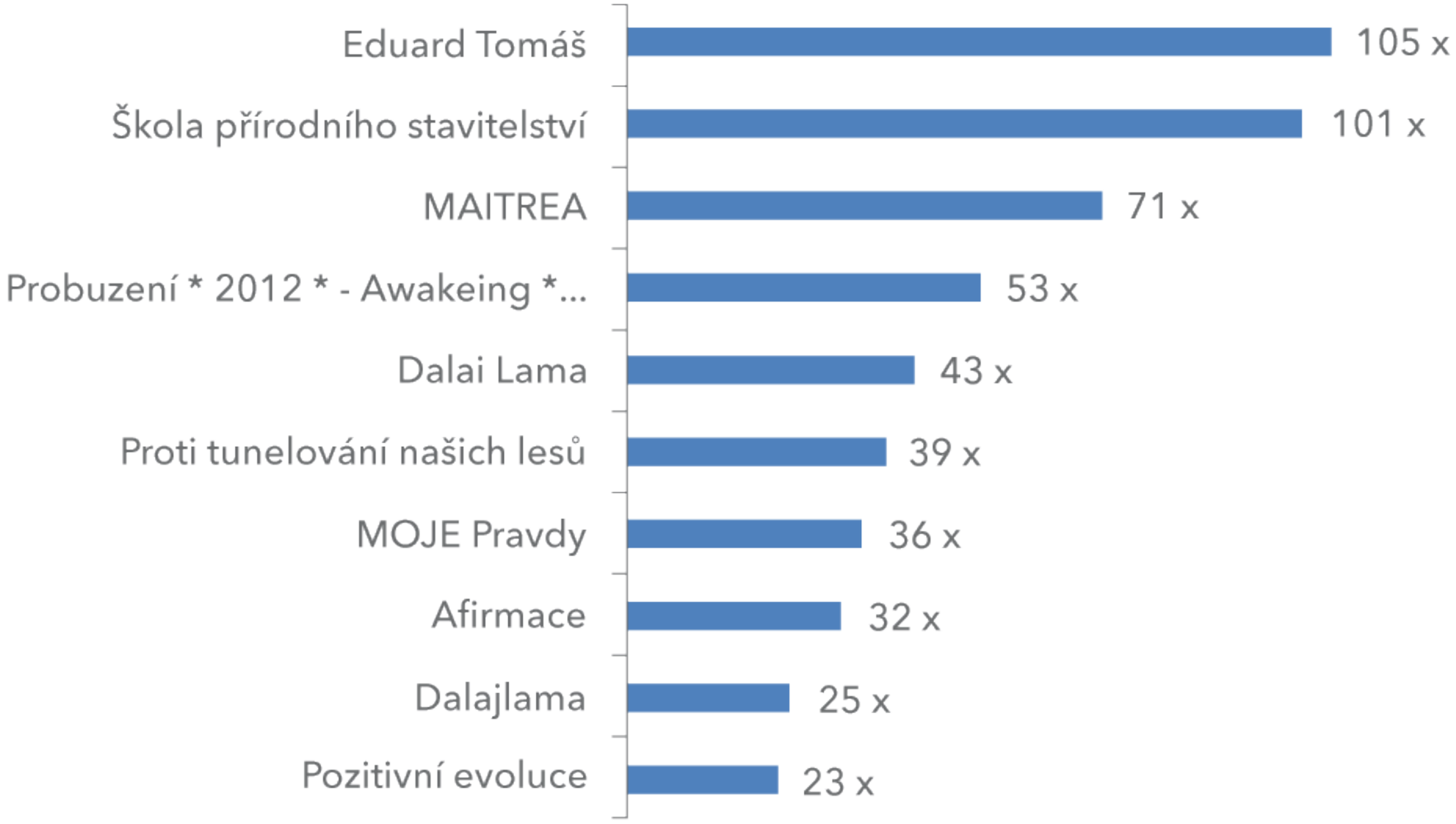
Miloš Zeman


Strana práv občanů
171 x
Vaše svoboda
33 x
Vzpomínky na socialismus
21 x
Nebojím se říct NE
19 x
Stydím se za vládu ČR a chci její demisi
14 x
TN.CZ
10 x
Tomio Okamura
9 x
ČT24
8 x
Pohádkář.cz
6 x
Islám v České republice nechceme
6 x

Vladimír Franz



Táňa Fišerová



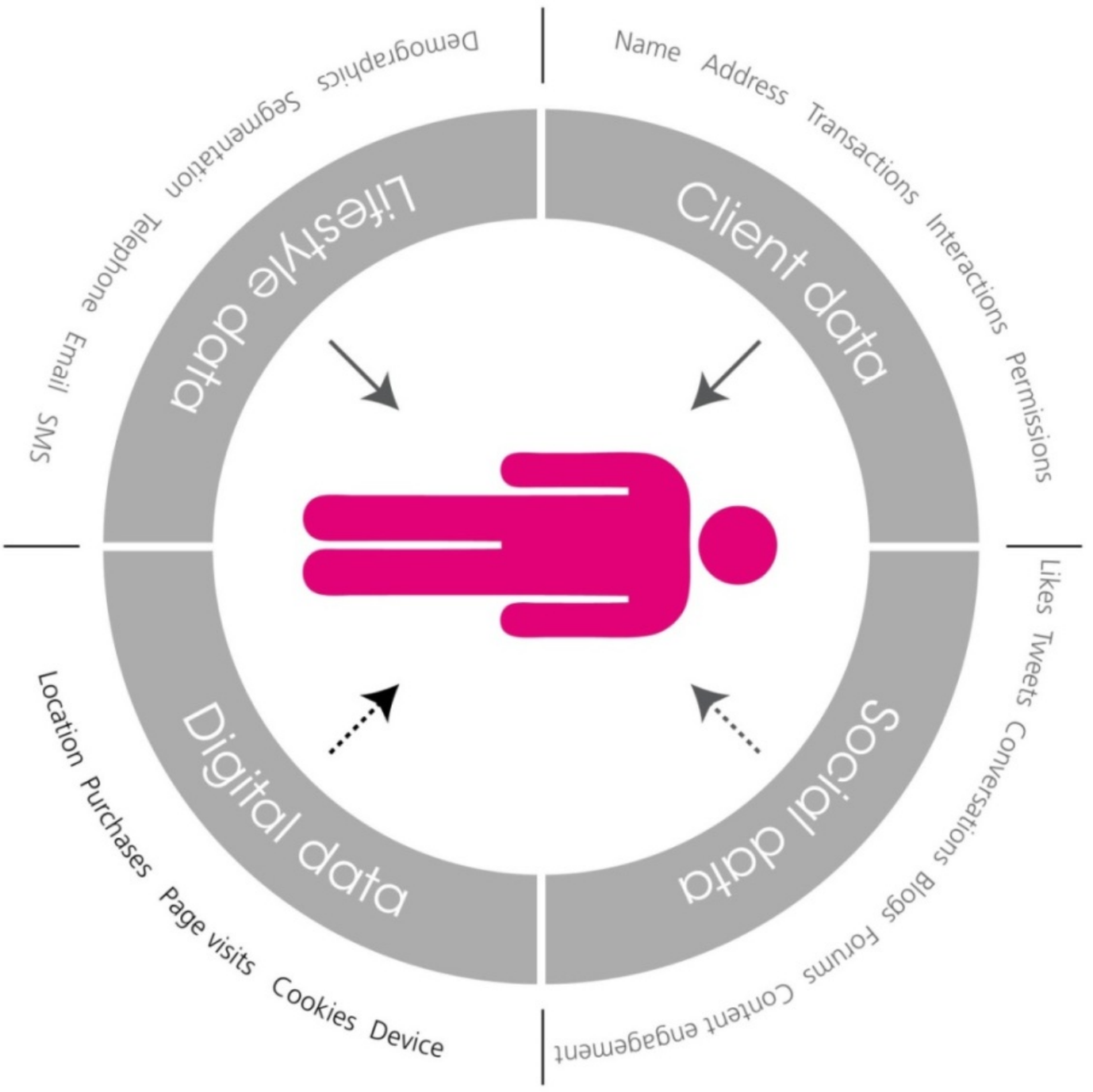
Pěkný, ale...

... příliš “big picture” pro každodenní business!

Social Insight Finder

ataxo

KBM GROUP
part of the wunderman network


Client data
Name Address Transactions Interactions Permissions
Lifestyle data
Demographics Segmentation Telephone Email SMS
Social data
Tweets Conversations Blogs Forums Content engagement
Digital data
Location Purchases Page visits Cookies Device

Základní principy

obchodní segmentace na tvrdých datech (transakční data, webová analytika aj.)

zájmové preference a psycho-demografie na datech ze sociálních sítí (Facebook, Twitter, blogy aj.)

Případová studie

velký klient z oblasti e-commerce

požadavek na segmentaci klientů a jejich charakteristiku

vhled do segmentu, který utrácí i do segmentu, který neutrácí

Postup

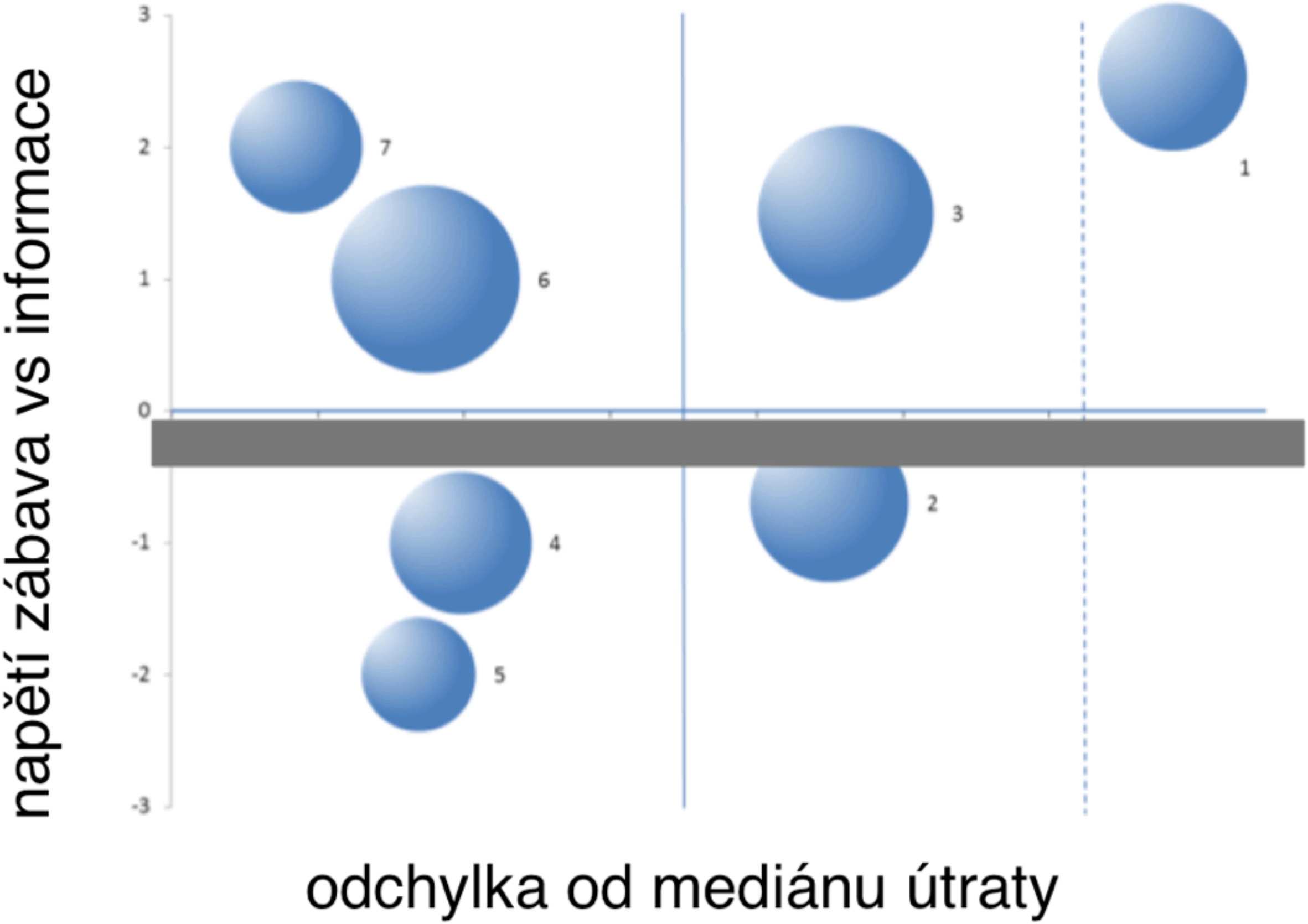
klasická segmentace z transkačních a dalších dat

obohacení o typické odchylky v preferencích na Facebooku od průměrné populace v jednotlivých segmentech

propojení obou analýz

Ukázky z výsledné analýzy (trochu)

7 obchodně aktivních segmentů



č. 2 URBAN YUPIIES

průměrné revenue

objednávají: goods, food, fun, méně travel

browsing: fun, food

na webu nadprůměrně aktivní ve všech kategoriích

pohlaví: nejvíce mužů z nakupujících segmentů

výrazně nadprůměrně platba kartou

věk: medián XX let

muži: městský starší teenager, po škole, single nebo nemá žádný vážný vztah, má peníze, chce se ukázat

price sensitive: Časopis dTest

technika: MobilMania.cz, Zive.cz

zábava: Fotbálek, Zelená, Show Jana Krause, Maxim

požitkáři: České pivo, JenProMuze.cz, Češky jsou nejkrásnější holky na světě etc.

ženy: studentky po nebo na VŠ (Fakulta roku), single nebo nemá žádný vážný vztah, má peníze pro sebe

chtějí se ukázat, řeší módu, značky – Starbucks, Botyk.cz, Fashion Days, CCC Boty, Zoot, Módní peklo

pózy – Vodu z vodovodu Zdarma, Potřebuji dovolenou

č. 8 Muži bez peněz

Nic nekupují, ale nějak se chovají a víme co mají rádi

č. 8 Muži bez peněz

insight: Hledám, co bych si koupil, ale nemám peníze

mix mužů bez peněz, hodně studenti (Státní maturity, Stáhněte si zadání) z menších měst (Brno, Ostrava)

pasivní zábava: Vyžeň nudu

sázení: Tipsport, Onlajny.cz, Chance

technické zájmy: Datart, Mobilmania, Asus, Škoda, Peugeot, Ford, Hyundai, Autosalon TV Prima

politika: Česká pirátská strana, Paroubek na Mars, Stydím se za vládu ČR

Facebook normalized distance ...

... příběh jedné hypotézy a problémy, které jí provázejí a provazely

Google distance

počítá se sémantická vzdálenost

autory jsou Rudi Cilibrasi a Paul M. B. Vitanyi

podobné věci sdílí stejné vlastnosti

tudíž se o nich mluví častěji dohromady

dvě reprezentace jsou si tím podobnější, čím méně složitých změn je třeba k převodu jedné v druhou

NGD je vyjádřená vzorcem:

m = log10(počet všech indexovaných stránek);

fx = log10(počet výsledků pro slovo X);

fy = log10(počet výsledků pro slovo Y);

fxy = log10(počet výsledků pro slovo X a Y);

GND = ((max(fx,fy) - fxy) / (m - min(fx,fy))

1	apple	2	microsoft	3	bmw	4	chrysler	5	toyota

Display

Additional keywords set (upto five, optional)

1		2		3		4		5	

Display

Restrict to domain (optional)

Domain or TLD

Display

Example keyword set: ***apple, microsoft, bmw, chrysler, toyota***

Result matrix

	apple	microsoft	bmw	chrysler	toyota
apple		0.11519974	0.77006889	0.85777333	0.54669557
microsoft	0.11519974		0.84528119	0.94983362	0.73886136
bmw	0.77006889	0.84528119		0.06765933	0.28139269
chrysler	0.85777333	0.94983362	0.06765933		0.42550257
toyota	0.54669557	0.73886136	0.28139269	0.42550257	

Lower numbers mean higher probability of keyword co-occurence

Tweet these results

http://www.mechanicalcinderella.com/

FND

Facebook normalized distance vychází z Google normalized distance

Lidé komentují na stránkách kandidátů, vůči kterým se především pozitivně, ale i negativně vymezují.

Pohybují se v určitých myšlenkovinách rovinách či diskursech

FND je vyjádřená vzorcem:

m = log10(počet všech českých účtů);

fx = log10(počet komentujících na stránce X);

fy = log10(počet komentujících na stránce Y);

fxy = log10(počet společných komentujících);

FND = ((max(fx,fy) - fxy) / (m - min(fx,fy))

Aplikace

buď objevování příbuzných stránek a témat

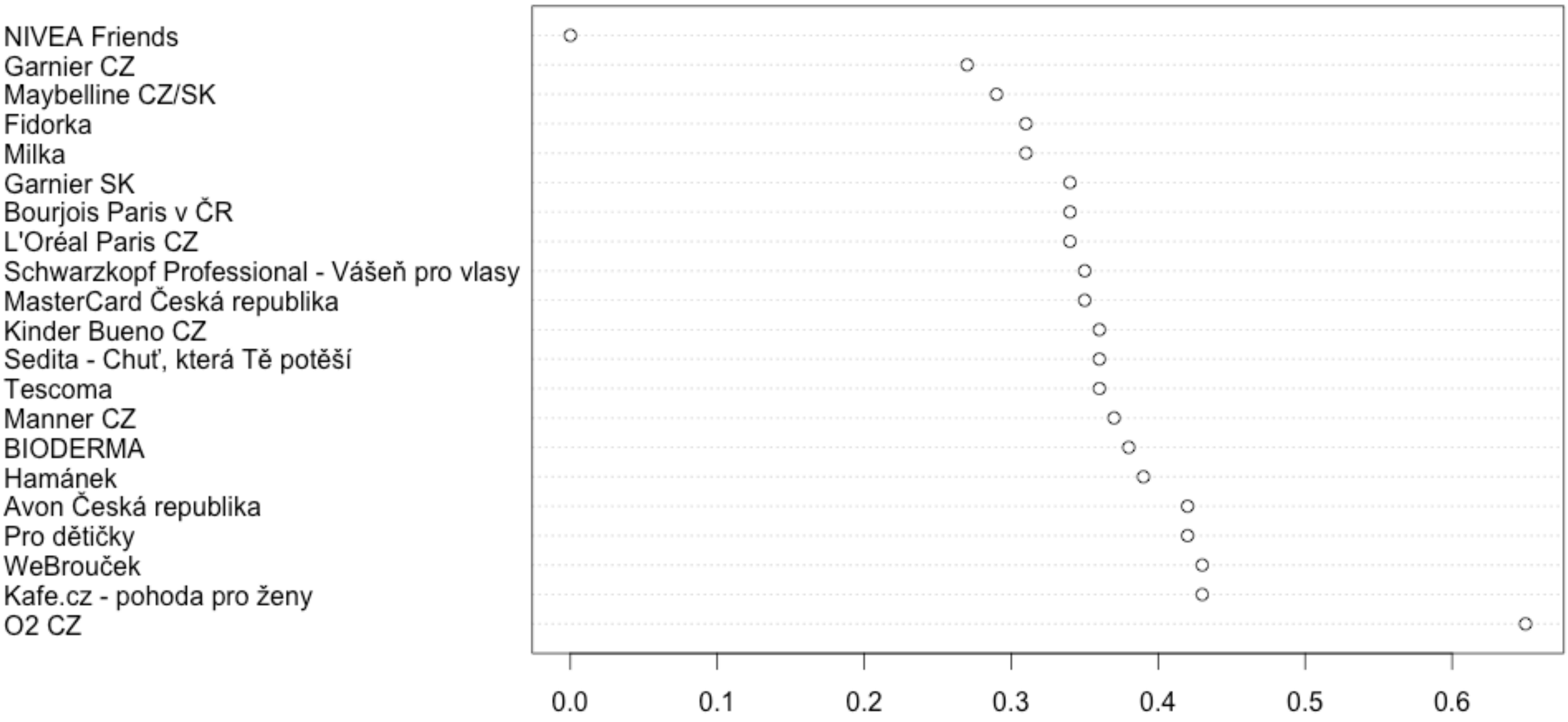
či k mapování vztahů mezi stránkami

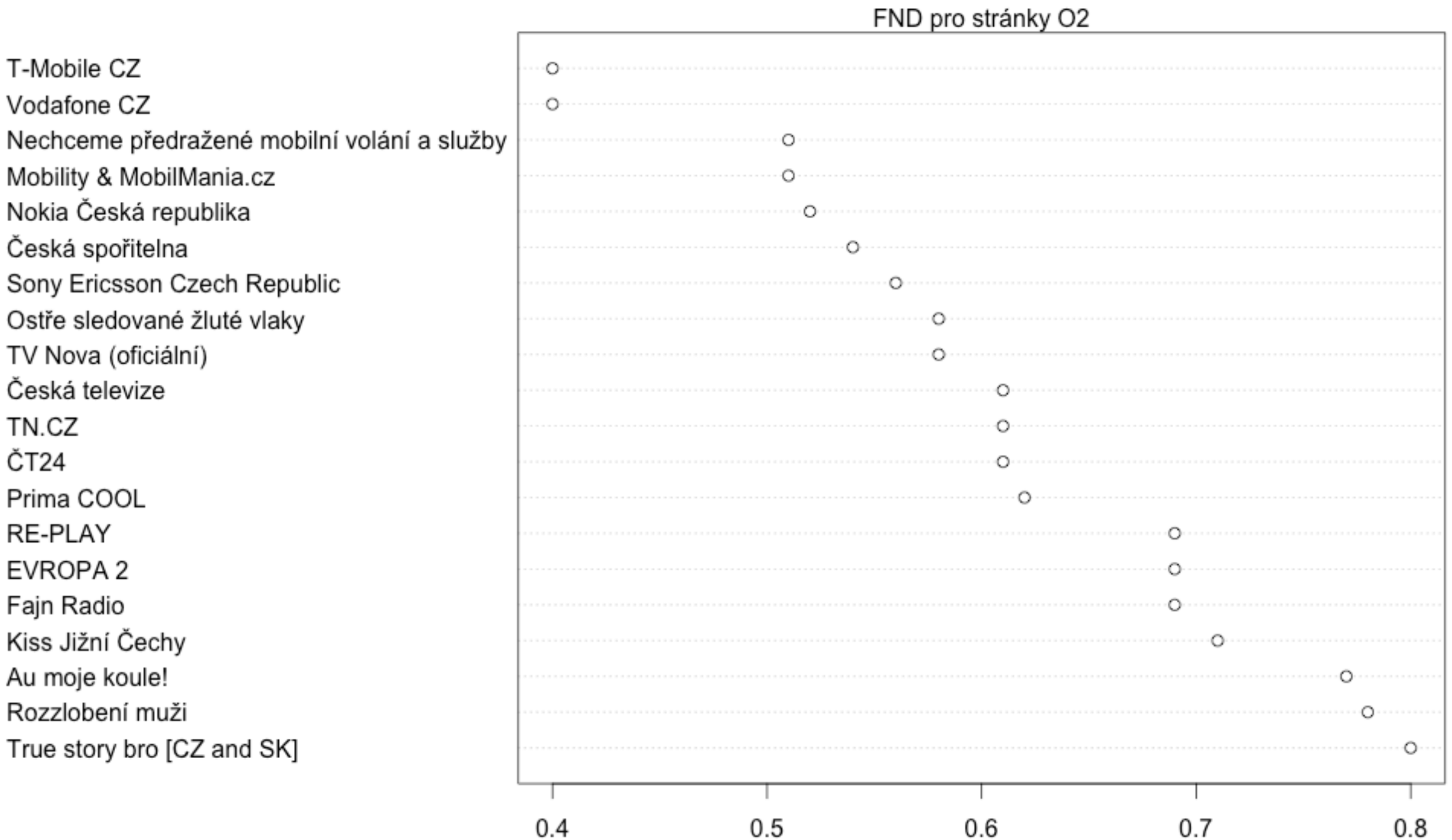
Vyhledávání

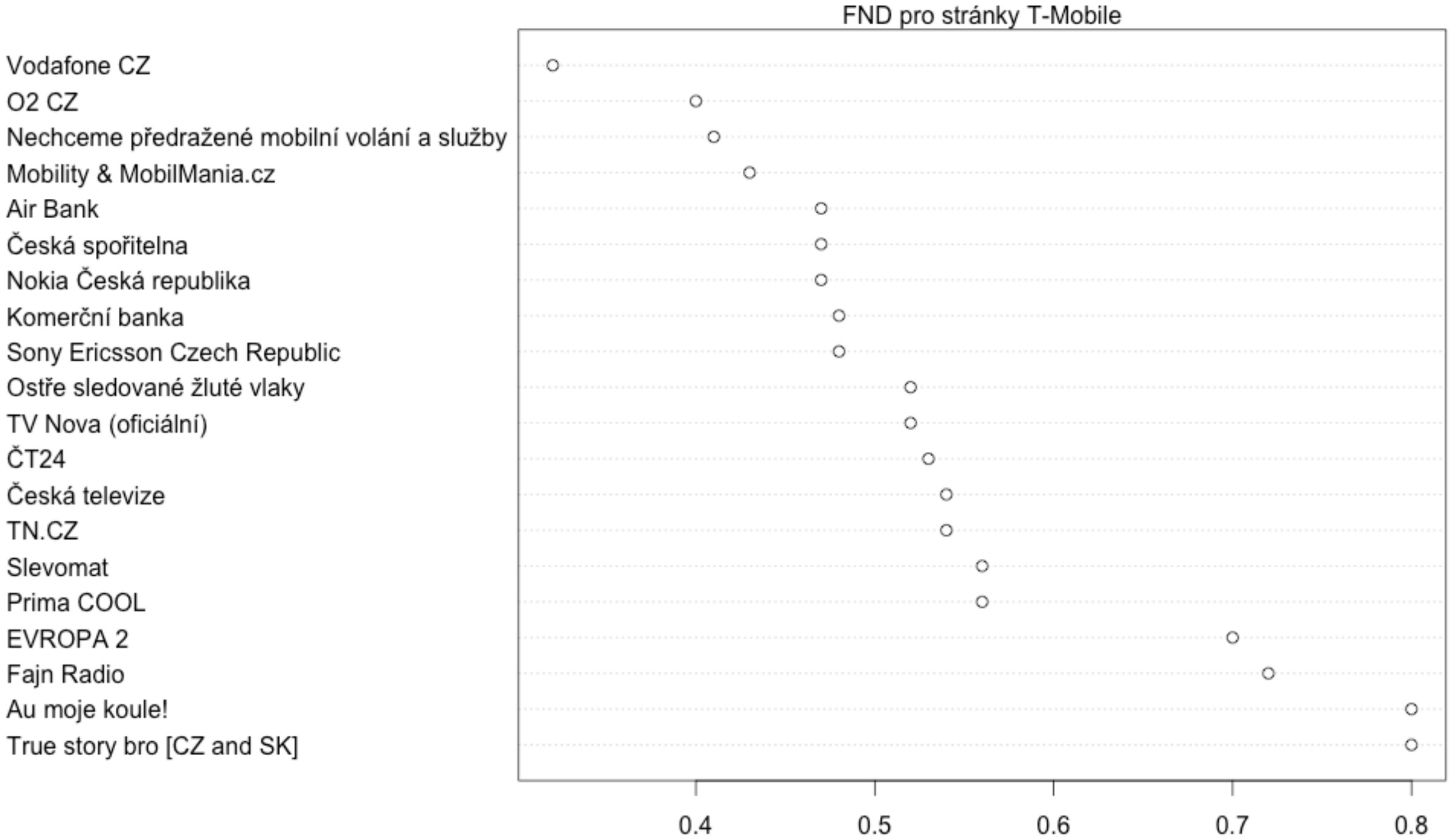
Najdi mi stránky, které mají blízko k mojí

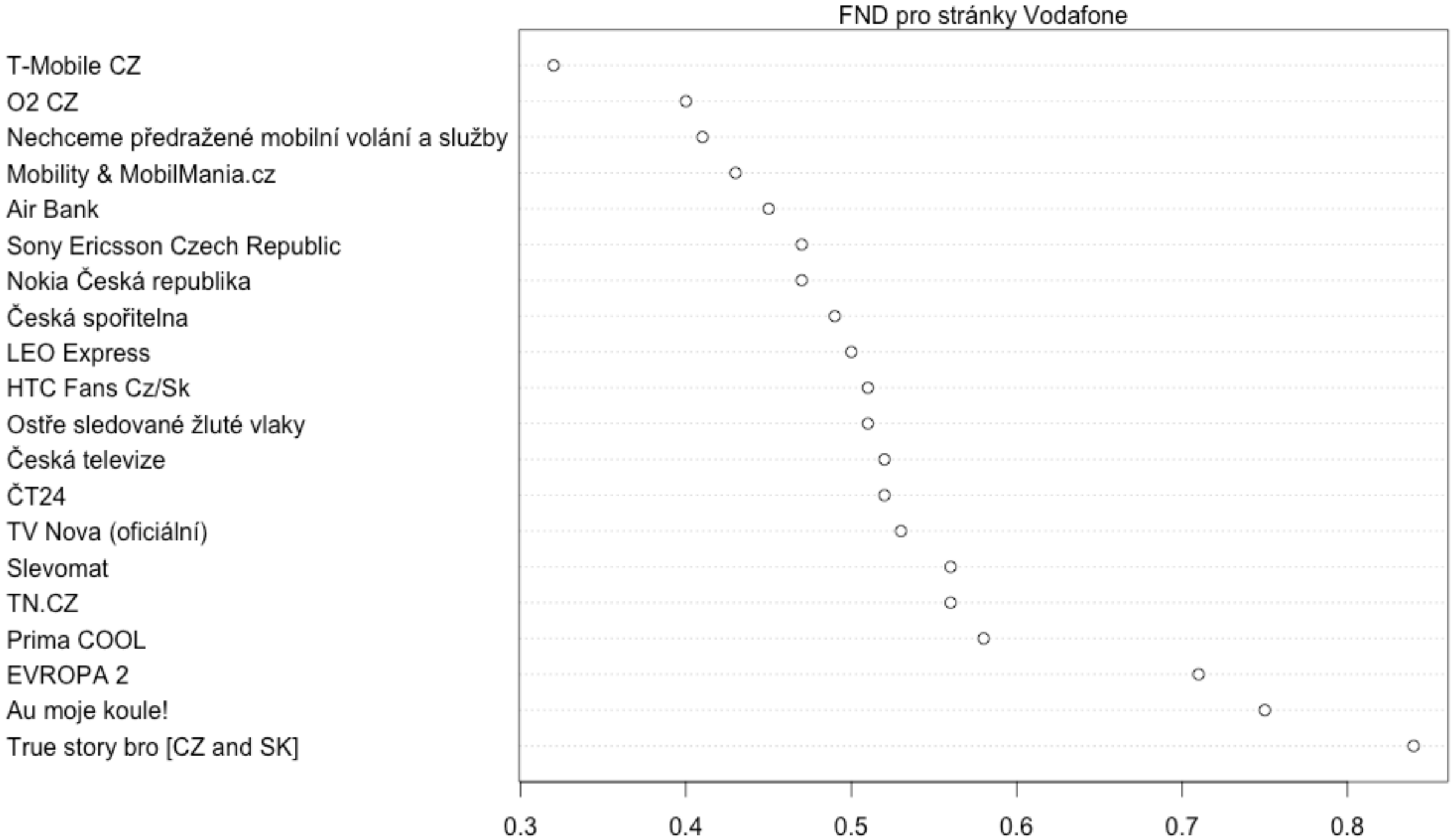
NGD	unique count	common count	common pct	page ID	page
	3377	3377	100.0	286483085526	T-Mobile CZ
0.32	4923	519	15.37	176688316811	Vodafone CZ
0.4	8560	515	15.25	120461304659141	O2 CZ
0.52	4327	116	3.44	253229733343	TV Nova (oficiální)
0.53	4436	112	3.32	137067469008	ČT24
0.41	677	105	3.11	275389913958	Nechceme předražené mobilní volání a služby
0.43	1069	104	3.08	115294422375	Mobility & MobilMania.cz
0.47	1454	90	2.67	123268337716736	Nokia Česká republika
0.54	3778	88	2.61	142069753862	TN.CZ
0.56	4133	87	2.58	97718196244	Prima COOL
0.47	1235	80	2.37	90594467955	Česká spořitelna
0.8	19710	78	2.31	352840478078551	True story bro [CZ and SK]
0.48	1257	78	2.31	273531189144	Sony Ericsson Czech Republic
0.54	3063	78	2.31	51828152685	Česká televize
0.47	1027	76	2.25	188676337835075	Air Bank
0.8	19219	74	2.19	109204525794230	"Au moje koule!"
0.7	8513	65	1.92	115991108439011	EVROPA 2
0.72	9153	63	1.87	111003257404	Fajn Radio
0.52	1662	63	1.87	191987746428	Ostře sledované žluté vlaky
0.56	2818	62	1.84	312917027851	Slevomat
0.48	712	56	1.66	361281313901258	Komerční banka

Blízkost stránek ke stránce NIVEA Friends

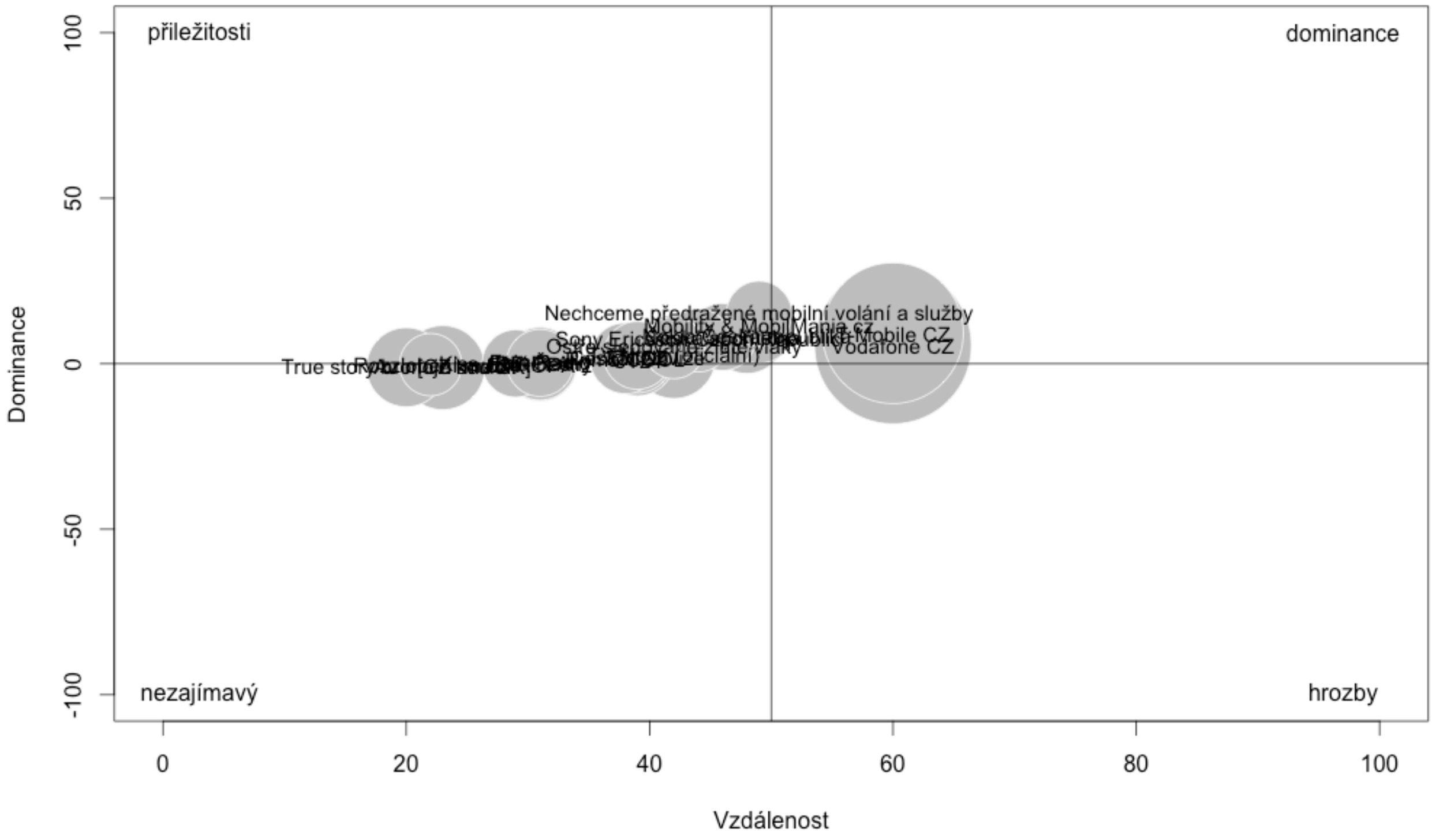



FND pro stránky O2
T-Mobile CZ
Vodafone CZ
Nechceme předražené mobilní volání a služby
Mobility & MobilMania.cz
Nokia Česká republika
Česká spořitelna
Sony Ericsson Czech Republic
Ostře sledované žluté vlaky
TV Nova (oficiální)
Česká televize
TN.CZ
ČT24
Prima COOL
RE-PLAY
EVROPA 2
Fajn Radio
Kiss Jižní Čechy
Au moje koule!
Rozzlobení muži
True story bro [CZ and SK]
0.4
0.5
0.6
0.7
0.8

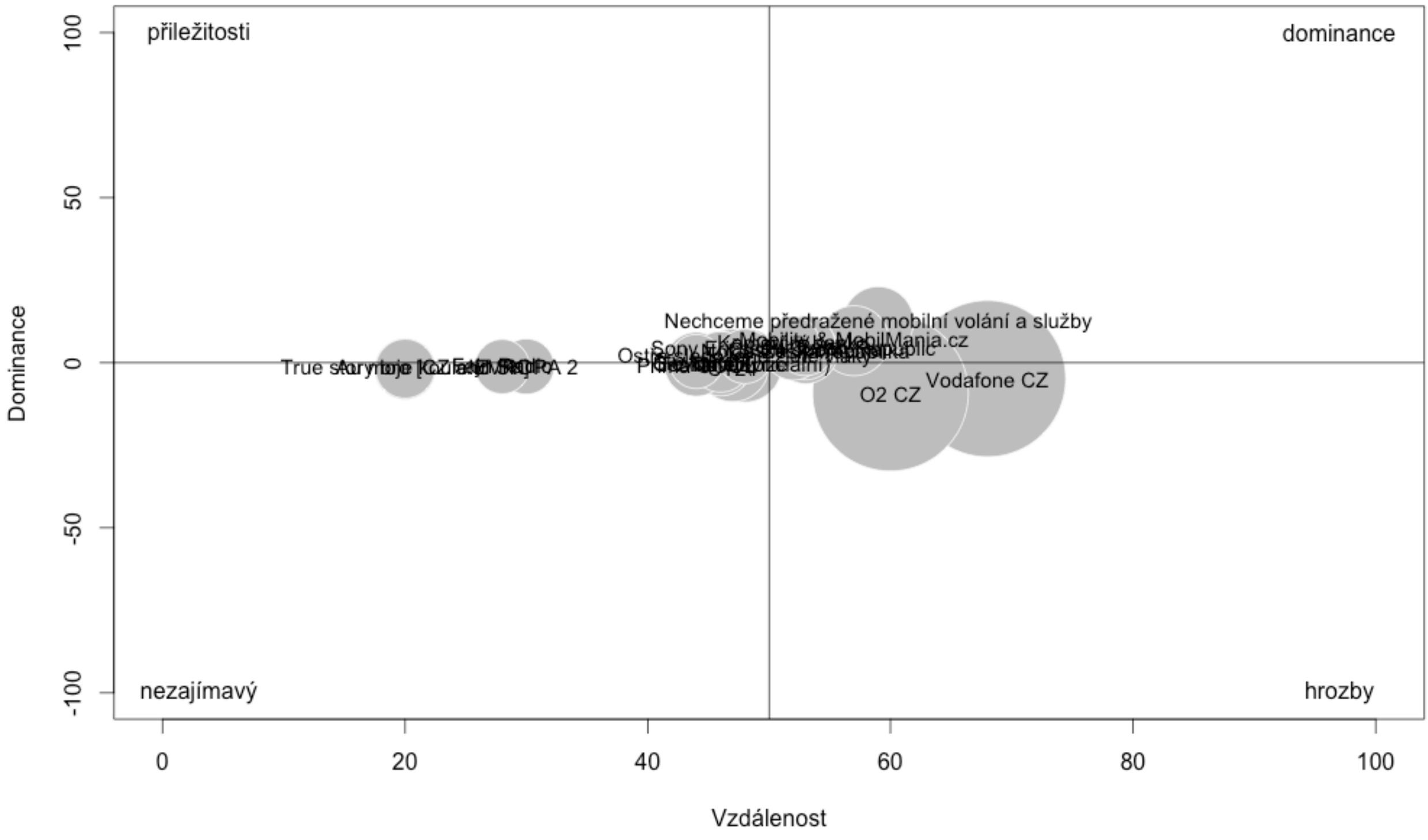

FND pro stránky T-Mobile
Vodafone CZ
O2 CZ
Nechceme předražené mobilní volání a služby
Mobility & MobilMania.cz
Air Bank
Česká spořitelna
Nokia Česká republika
Komerční banka
Sony Ericsson Czech Republic
Ostře sledované žluté vlaky
TV Nova (oficiální)
ČT24
Česká televize
TN.CZ
Slevomat
Prima COOL
EVROPA 2
Fajn Radio
Au moje koule!
True story bro [CZ and SK]
0.4
0.5
0.6
0.7
0.8


FND pro stránky Vodafone
T-Mobile CZ
O2 CZ
Nechceme předražené mobilní volání a služby
Mobility & MobilMania.cz
Air Bank
Sony Ericsson Czech Republic
Nokia Česká republika
Česká spořitelna
LEO Express
HTC Fans Cz/Sk
Ostře sledované žluté vlaky
Česká televize
ČT24
TV Nova (oficiální)
Slevomat
TN.CZ
Prima COOL
EVROPA 2
Au moje koule!
True story bro [CZ and SK]
0.3
0.4
0.5
0.6
0.7
0.8

Magic Quadrant for O2 page on Facebook



Magic Quadrant for T-Mobile page on Facebook

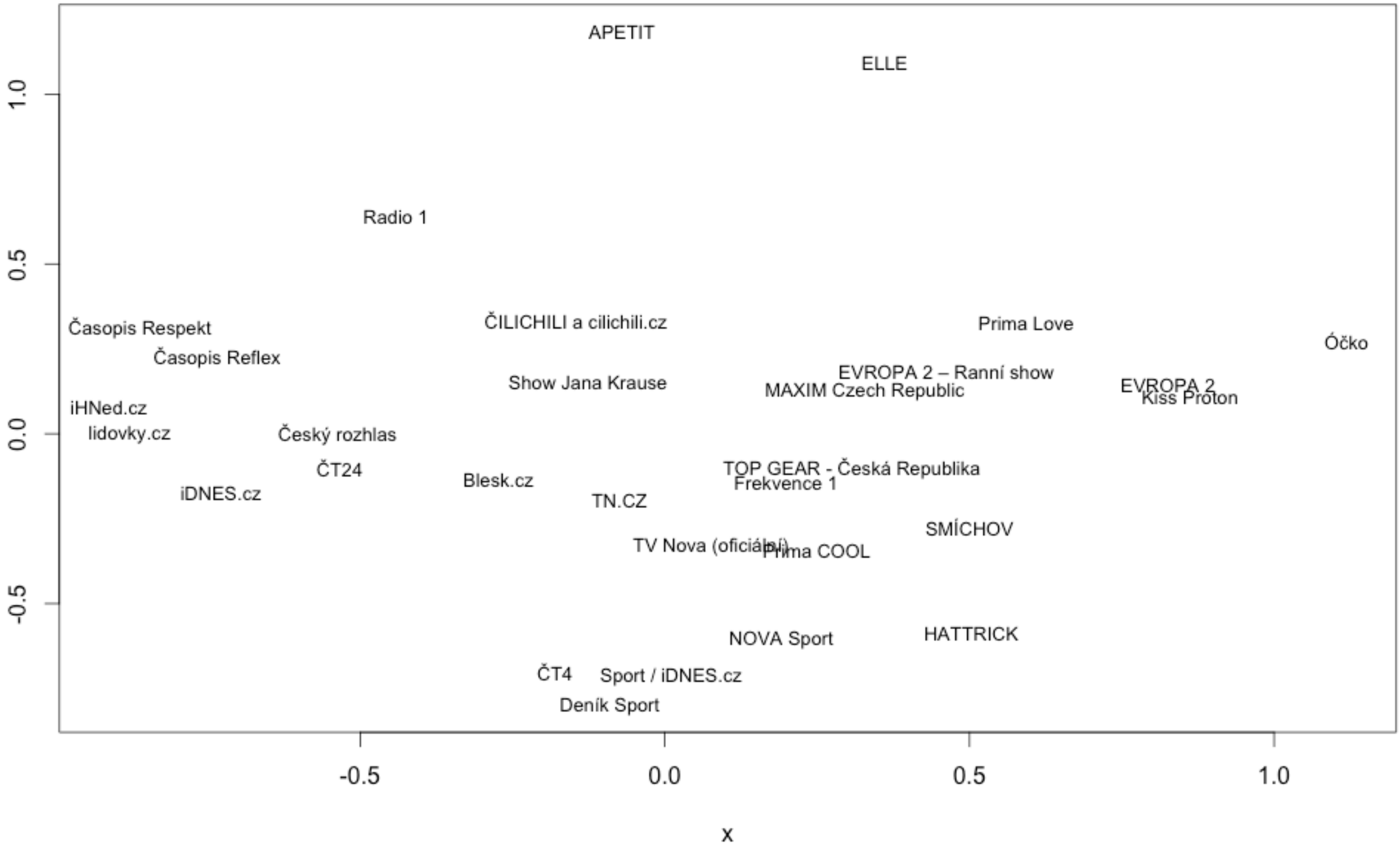


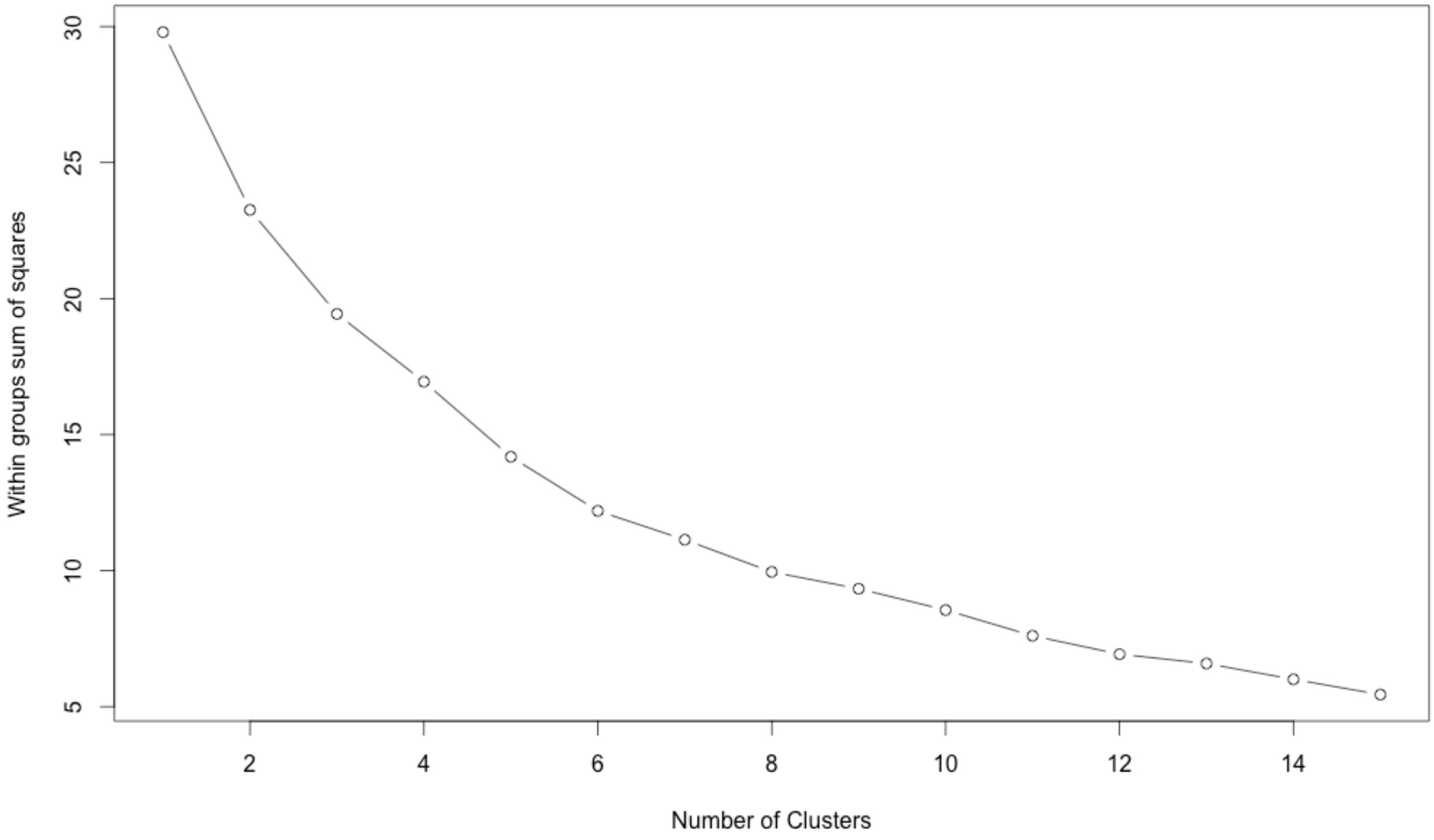
Vztahování

Kde se nachází moje stránka mezi ostatními

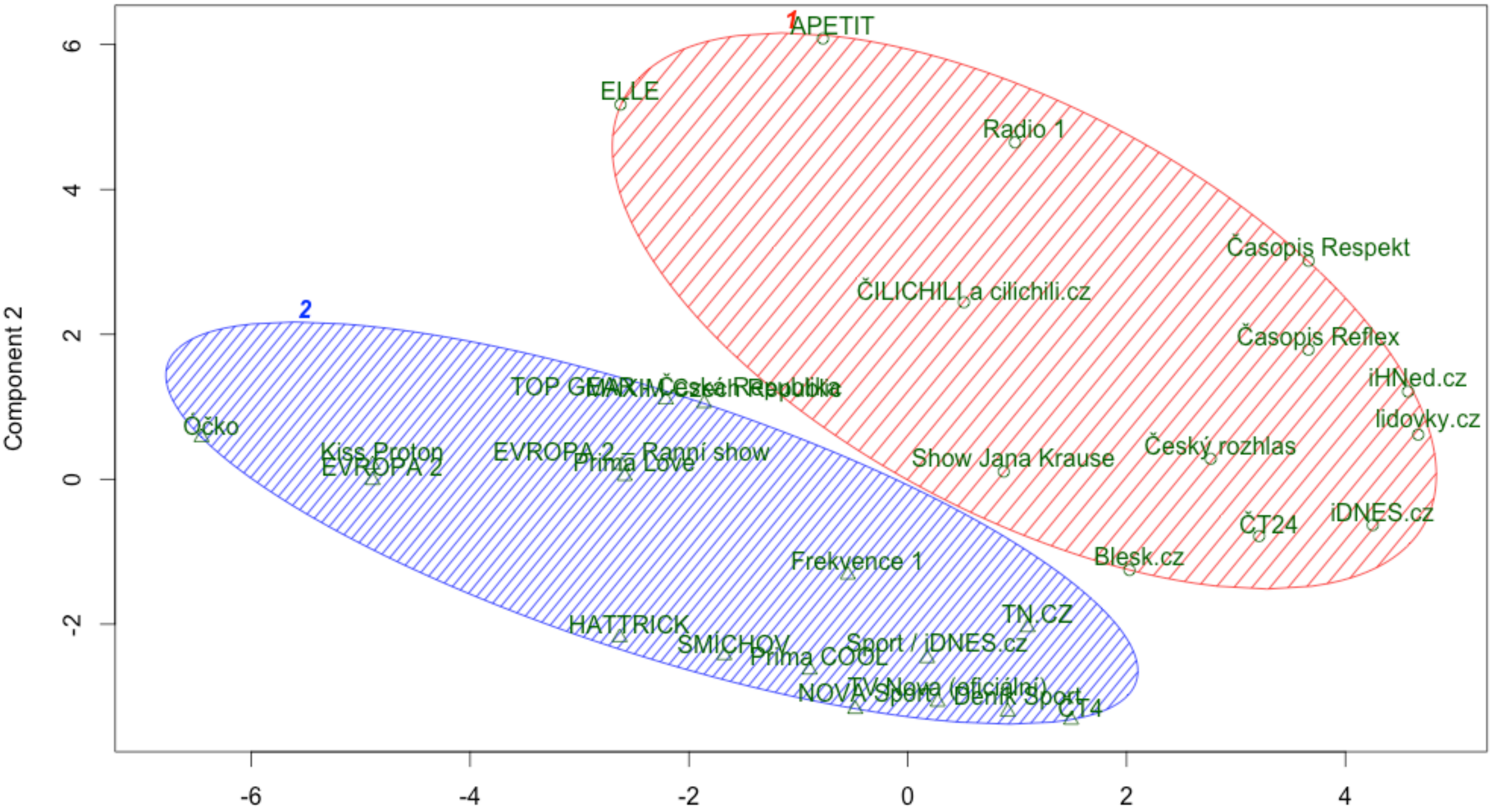
Mediální mapa FB

Jaké typy stránek máme mezi FB stránkami serveru

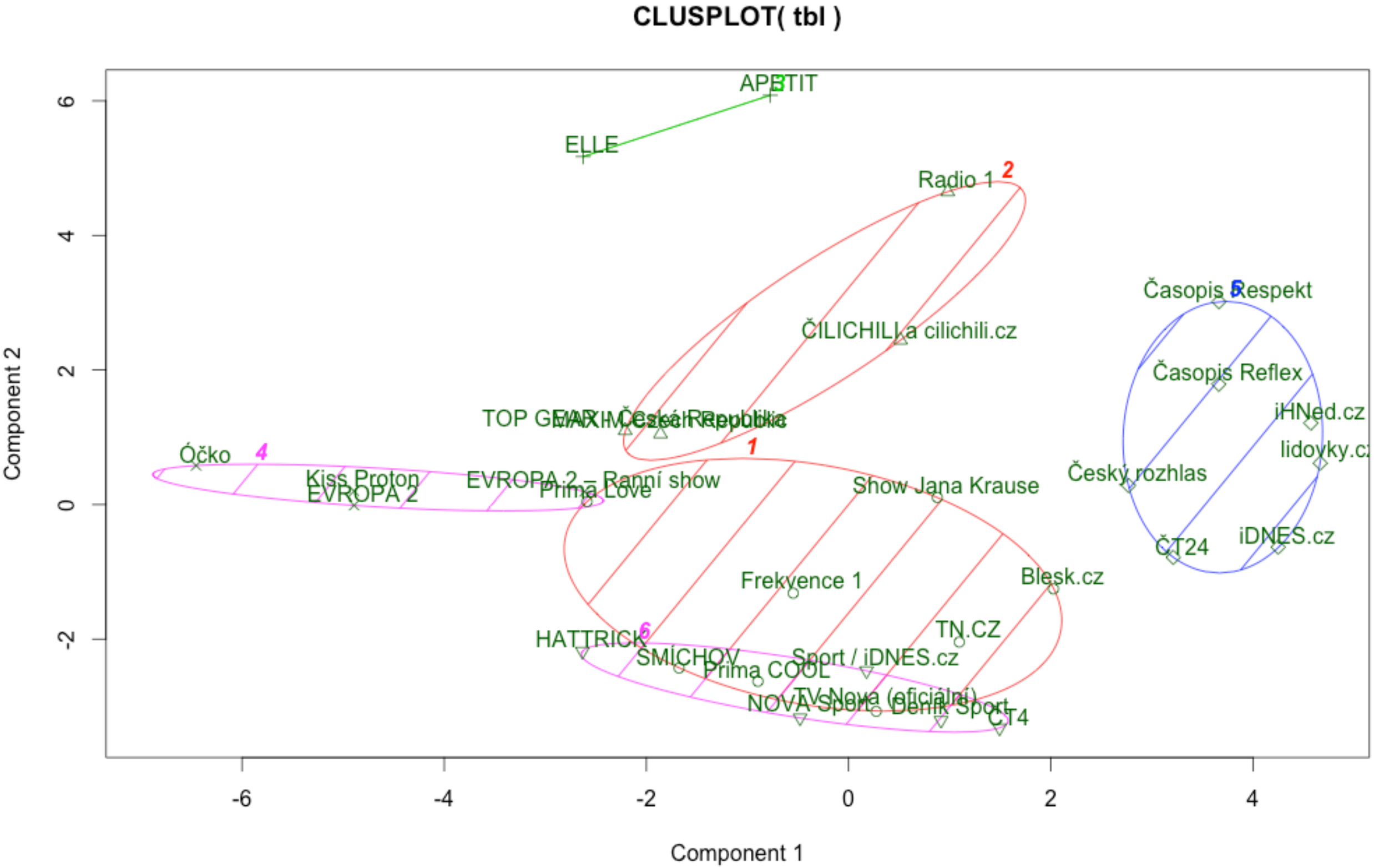

APETIT
ELLE
Radio 1
ČILICHILI a cilichili.cz
Prima Love
Óčko
Časopis Respekt
Časopis Reflex
Show Jana Krause
EVROPA 2 – Ranní show
MAXIM Czech Republic
EVROPA 2
Kiss Proton
iHNed.cz
lidovky.cz
Český rozhlas
ČT24
Blesk.cz
TOP GEAR - Česká Republika
Frekvence 1
iDNES.cz
TN.CZ
SMÍCHOV
TV Nova (oficiální)
Prima COOL
NOVA Sport
HATTRICK
ČT4
Sport / iDNES.cz
Deník Sport
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
x


Within groups sum of squares
30
25
20
15
10
5
2
4
6
8
10
12
14
Number of Clusters

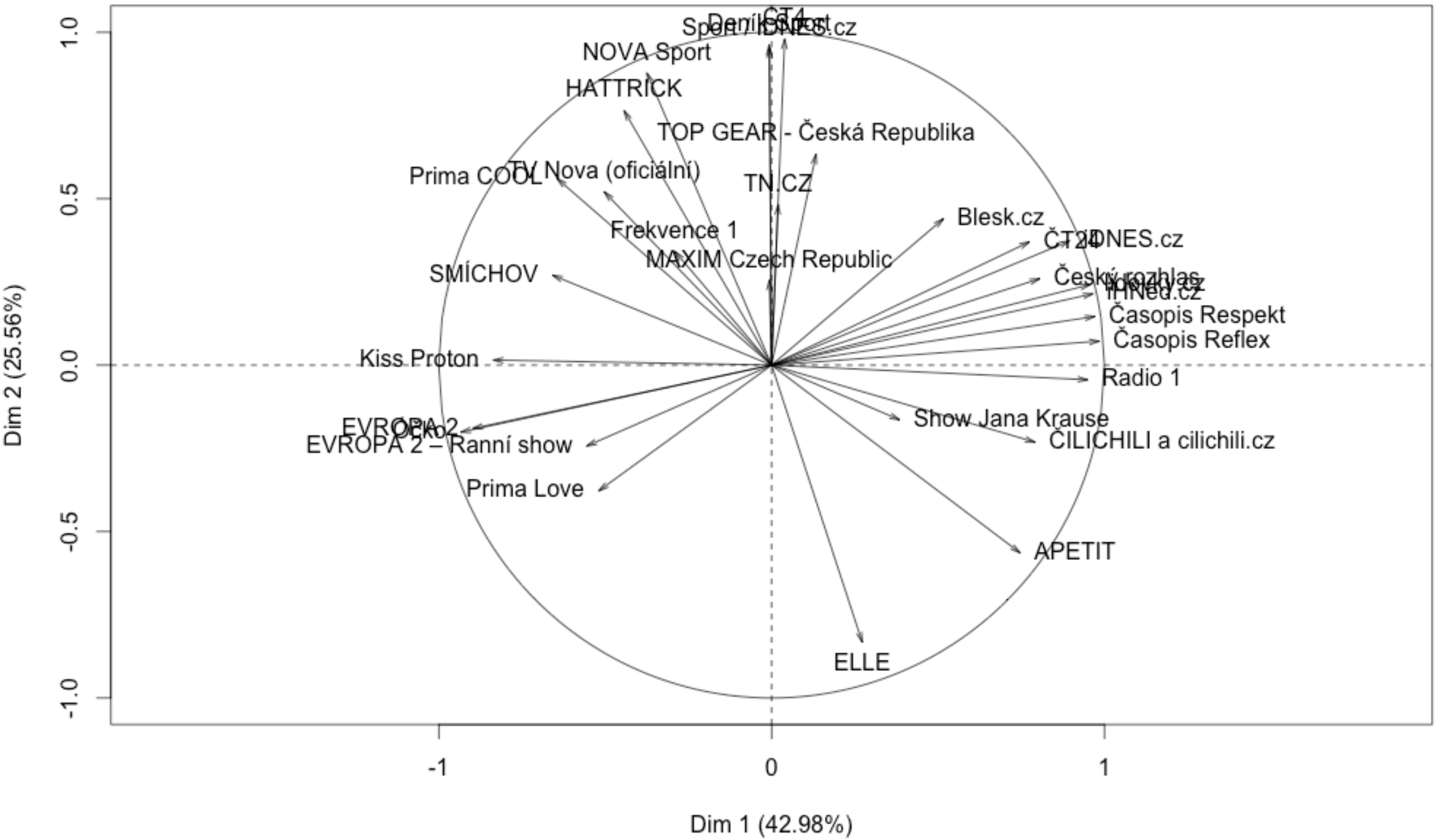
CLUSPLOT(tbl)

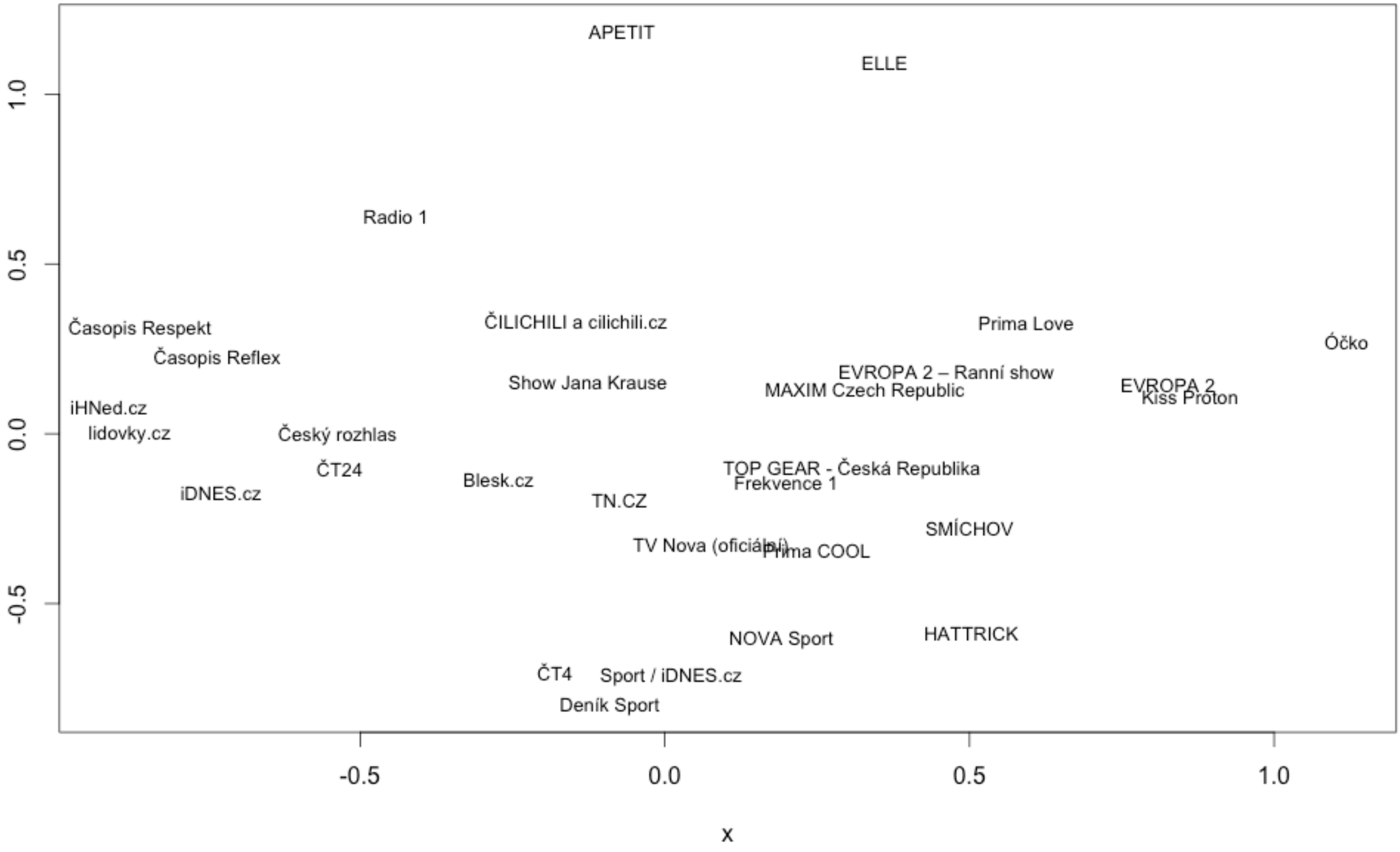


These two components explain 47.47 % of the point variability.



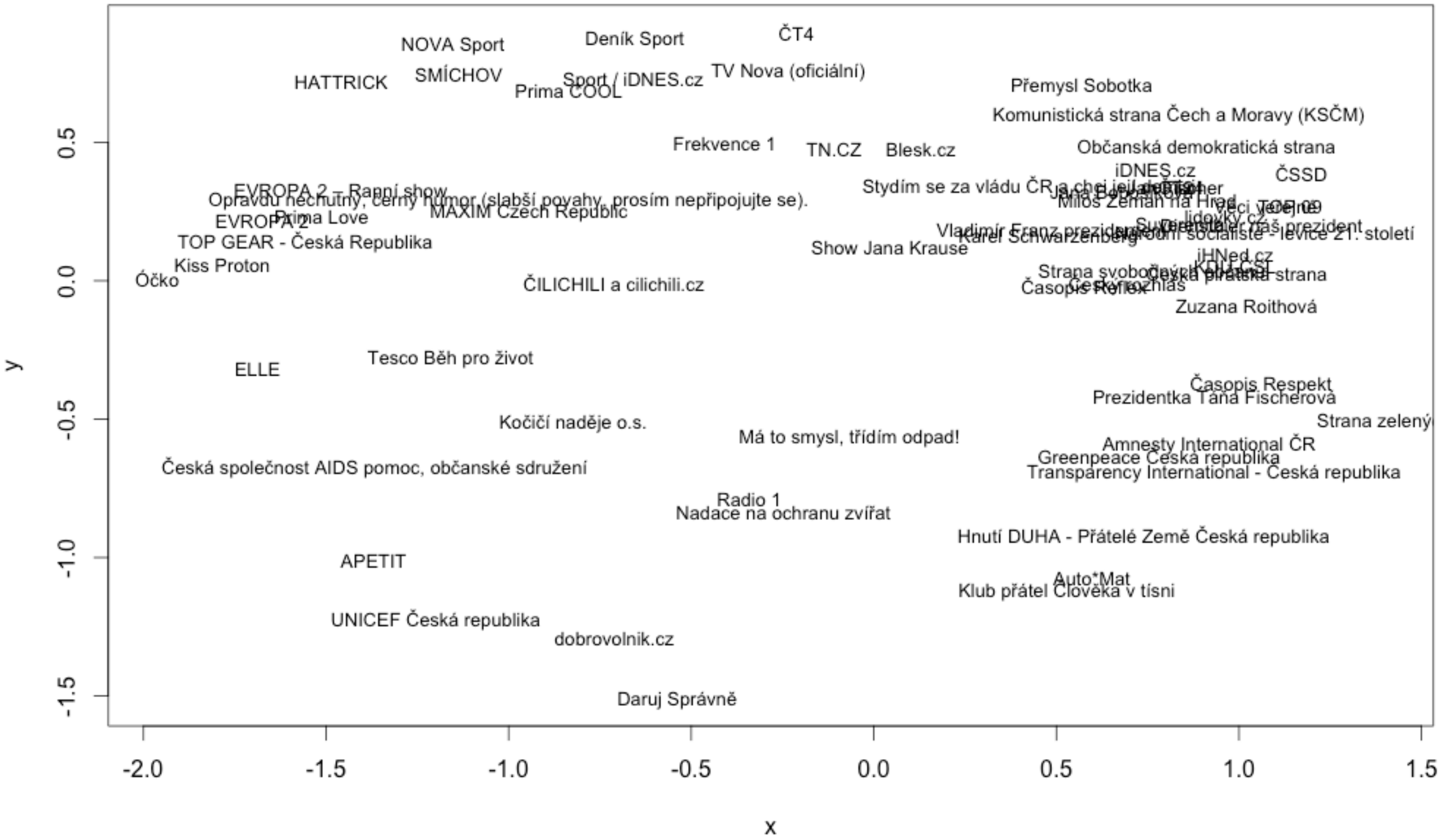
These two components explain 47.47 % of the point variability.

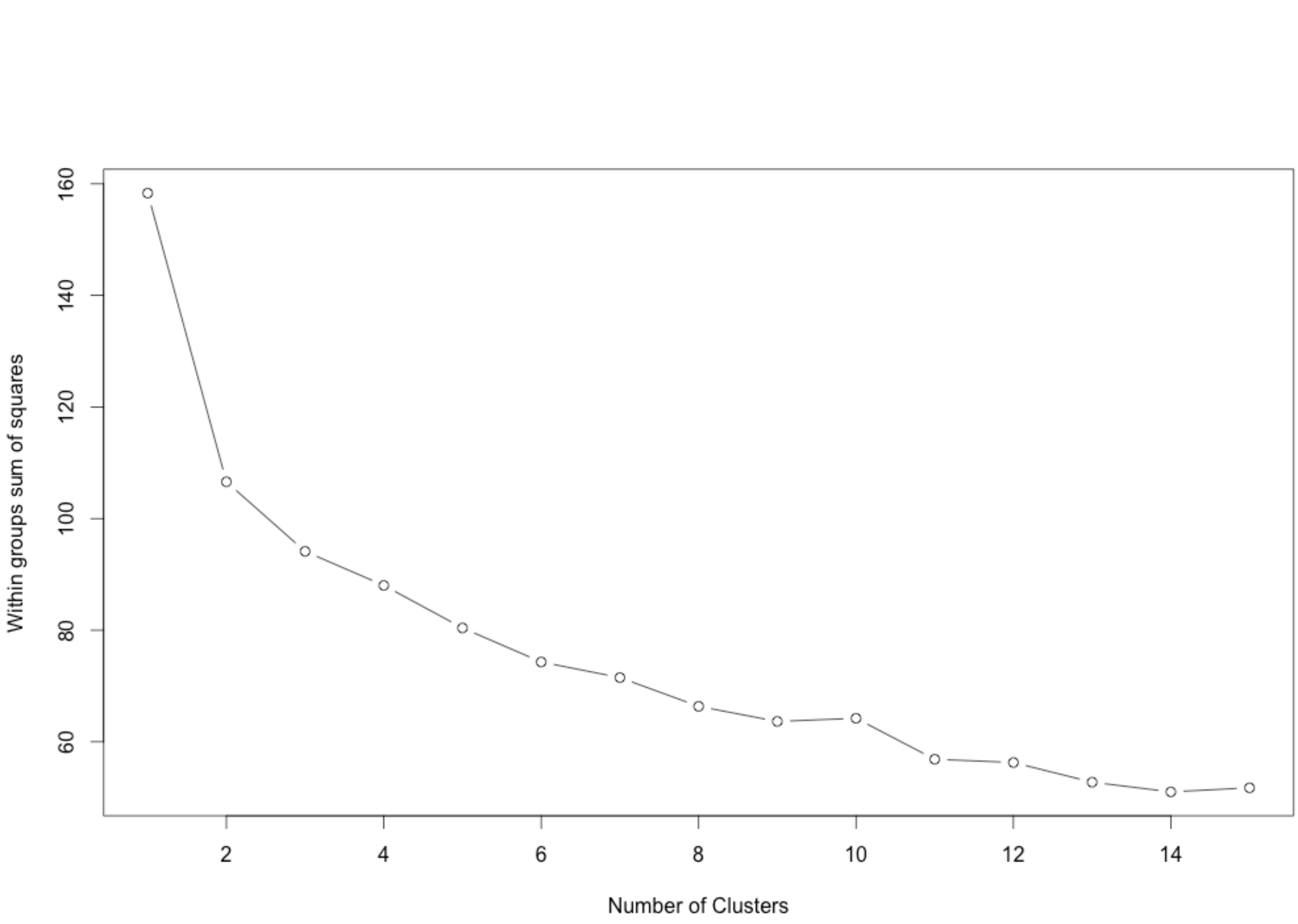

Variables factor map (PCA)
ČT4
Sport
iDNES.cz
NOVA Sport
HATTRICK
TOP GEAR - Česká Republika
Prima COOL
TV Nova (oficiální)
TN.CZ
Blesk.cz
Frekvence 1
ČT24
iDNES.cz
MAXIM Czech Republic
SMÍCHOV
Český rozhlas
Časopis Respekt
Časopis Reflex
Kiss Proton
Radio 1
Show Jana Krause
EVROPA 2
EVROPA 2 – Ranní show
ČILICHILI a cilichili.cz
Prima Love
APETIT
ELLE
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Dim 2 (25.56%)
Dim 1 (42.98%)


APETIT
ELLE
Radio 1
ČILICHILI a cilichili.cz
Prima Love
Óčko
Časopis Respekt
Časopis Reflex
Show Jana Krause
EVROPA 2 – Ranní show
MAXIM Czech Republic
EVROPA 2
Kiss Proton
iHNed.cz
lidovky.cz
Český rozhlas
ČT24
Blesk.cz
TOP GEAR - Česká Republika
Frekvence 1
iDNES.cz
TN.CZ
SMÍCHOV
TV Nova (oficiální)
Prima COOL
NOVA Sport
HATTRICK
ČT4
Sport / iDNES.cz
Deník Sport
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
x

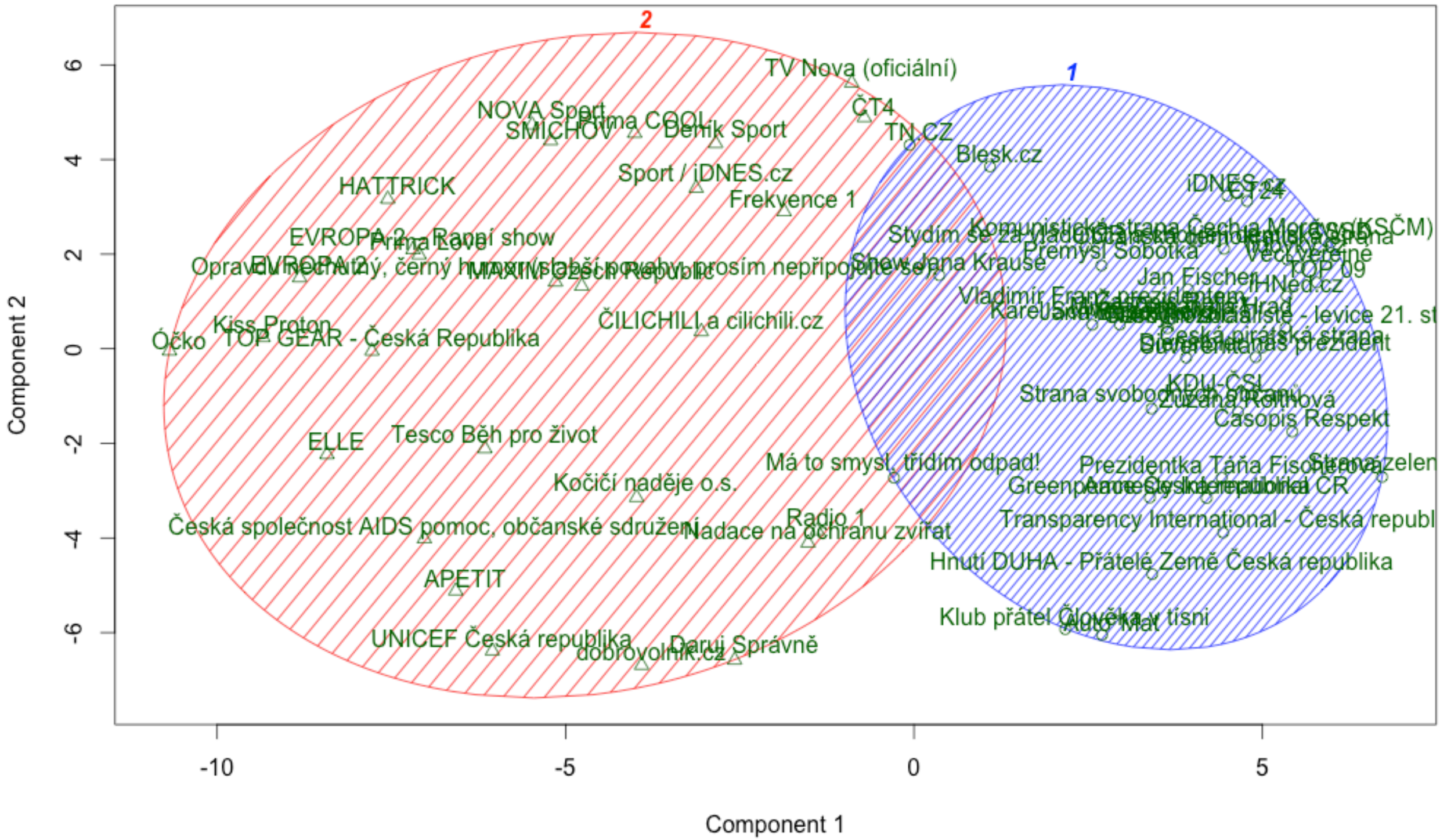
Velká mapa stránek

Neziskovky, strany, média....


NOVA Sport
Deník Sport
ČT4
HATTRICK
SMÍCHOV
Sport / iDNES.cz
TV Nova (oficiální)
Prima COOL
Přemysl Sobotka
Komunistická strana Čech a Moravy (KSČM)
Frekvence 1
TN.CZ
Blesk.cz
Občanská demokratická strana
iDNES.cz
ČSSD
Stydím se za vládu ČR a chci jejíl
Jana Bobošíková
Miloš Zeman na Hrad
EVROPA 2 – Ranní show
Opravdu nechutný, černý humor (slabší povahy, prosím nepřipojujte se).
MAXIM Czech Republic
Prima Love
EVROPA 2
Vladimír Franz prezident
Karel Schwarzenberg
TOP GEAR - Česká Republika
Show Jana Krause
Kiss Proton
Óčko
ČILICHILI a cilichili.cz
Strana svobodných občanů
Česká pirátská strana
Časopis Reflex
Zuzana Roithová
Tesco Běh pro život
ELLE
Časopis Respekt
Prezidentka Táňa Fischerová
Strana zelených
Kočičí naděje o.s.
Má to smysl, třídím odpad!
Amnesty International ČR
Greenpeace Česká republika
Česká společnost AIDS pomoc, občanské sdružení
Transparency International - Česká republika
Radio 1
Nadace na ochranu zvířat
Hnutí DUHA - Přátelé Země Česká republika
APETIT
Auto*Mat
Klub přátel Člověka v tísni
UNICEF Česká republika
dobrovolnik.cz
Daruj Správně
0.5
0.0
-0.5
-1.0
-1.5
y
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x

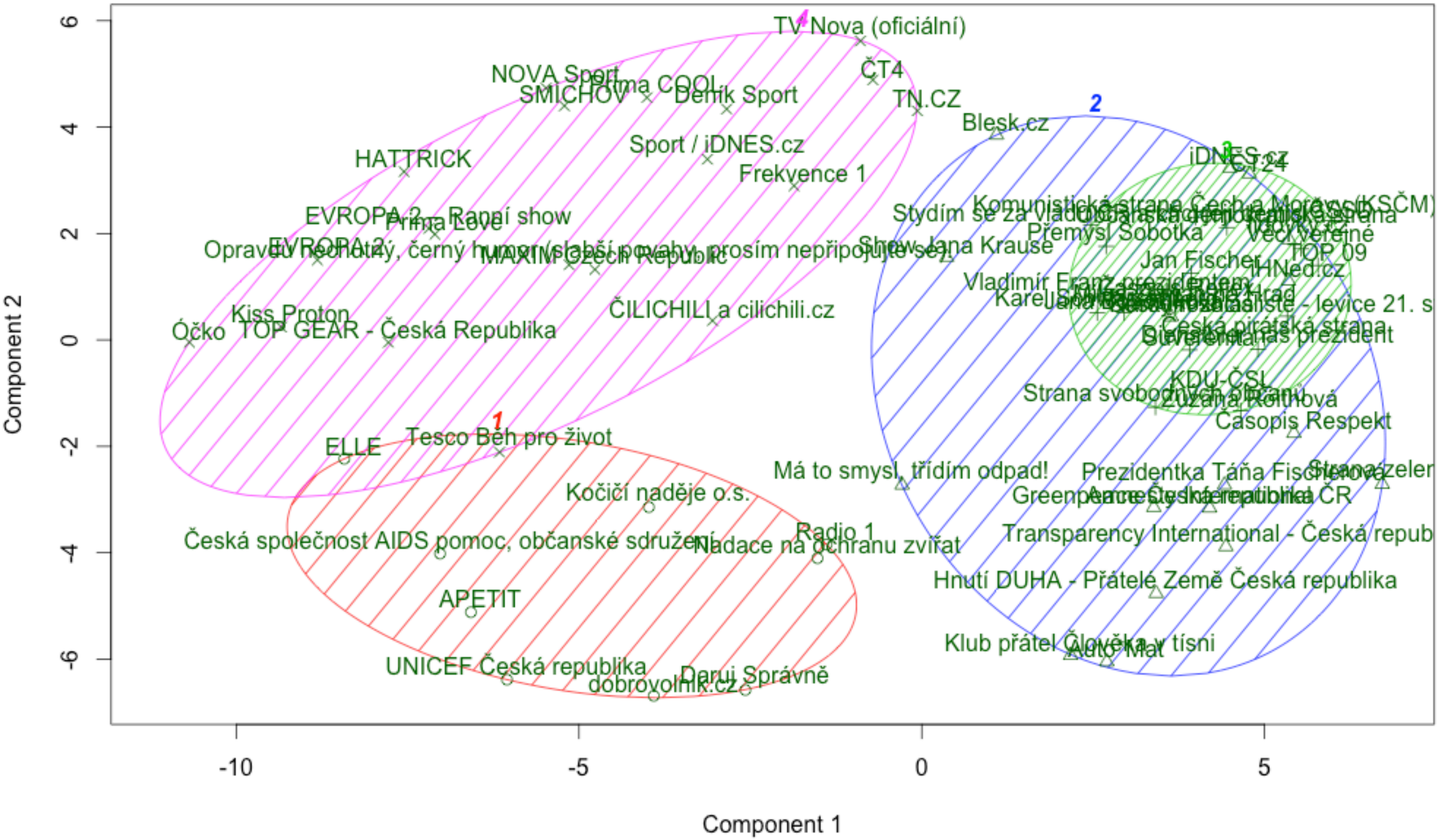

Within groups sum of squares
160
140
120
100
80
60
2
4
6
8
10
12
14
Number of Clusters

CLUSPLOT(tbl)



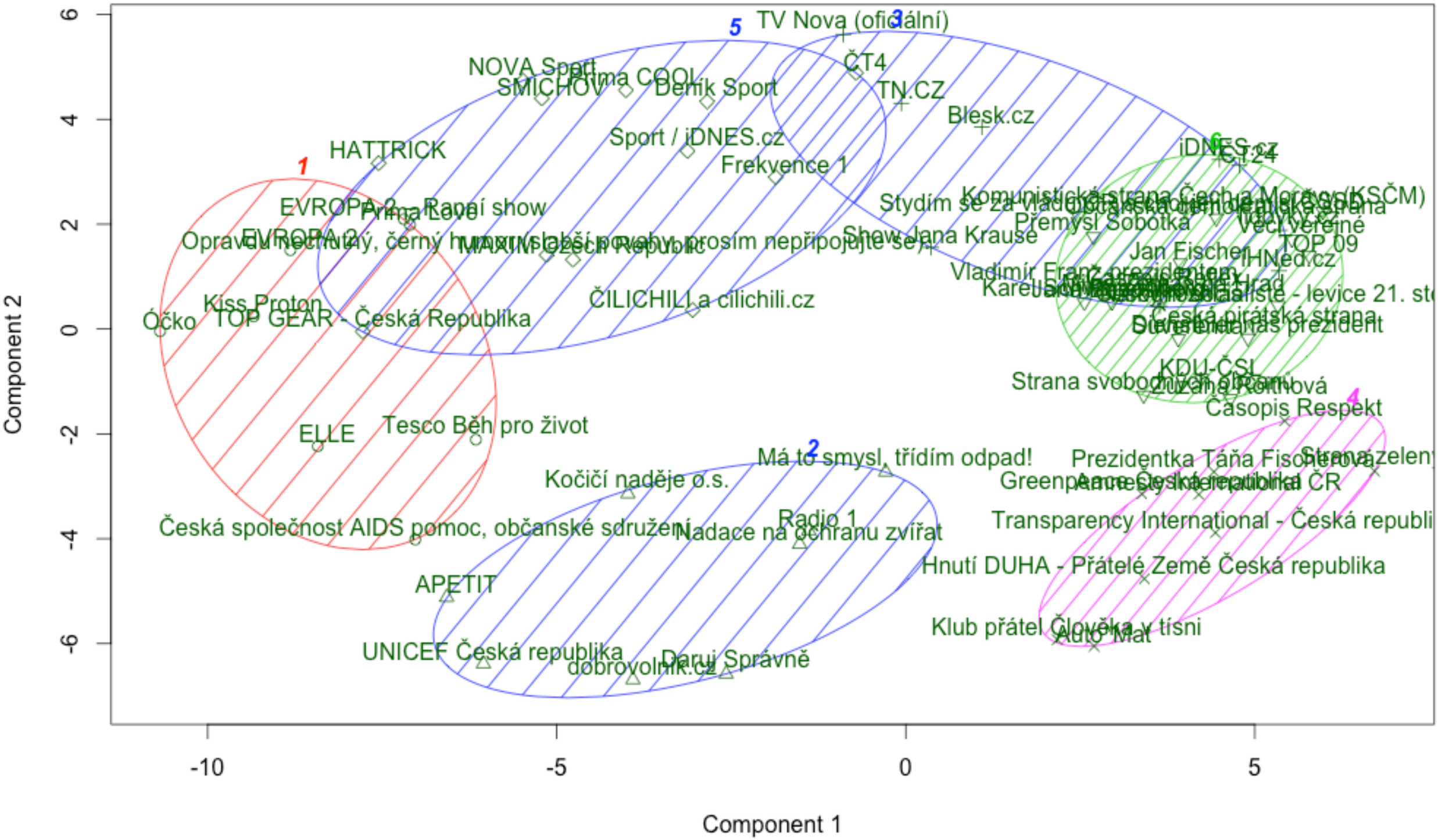
These two components explain 51.1 % of the point variability.

CLUSPLOT(tbl)



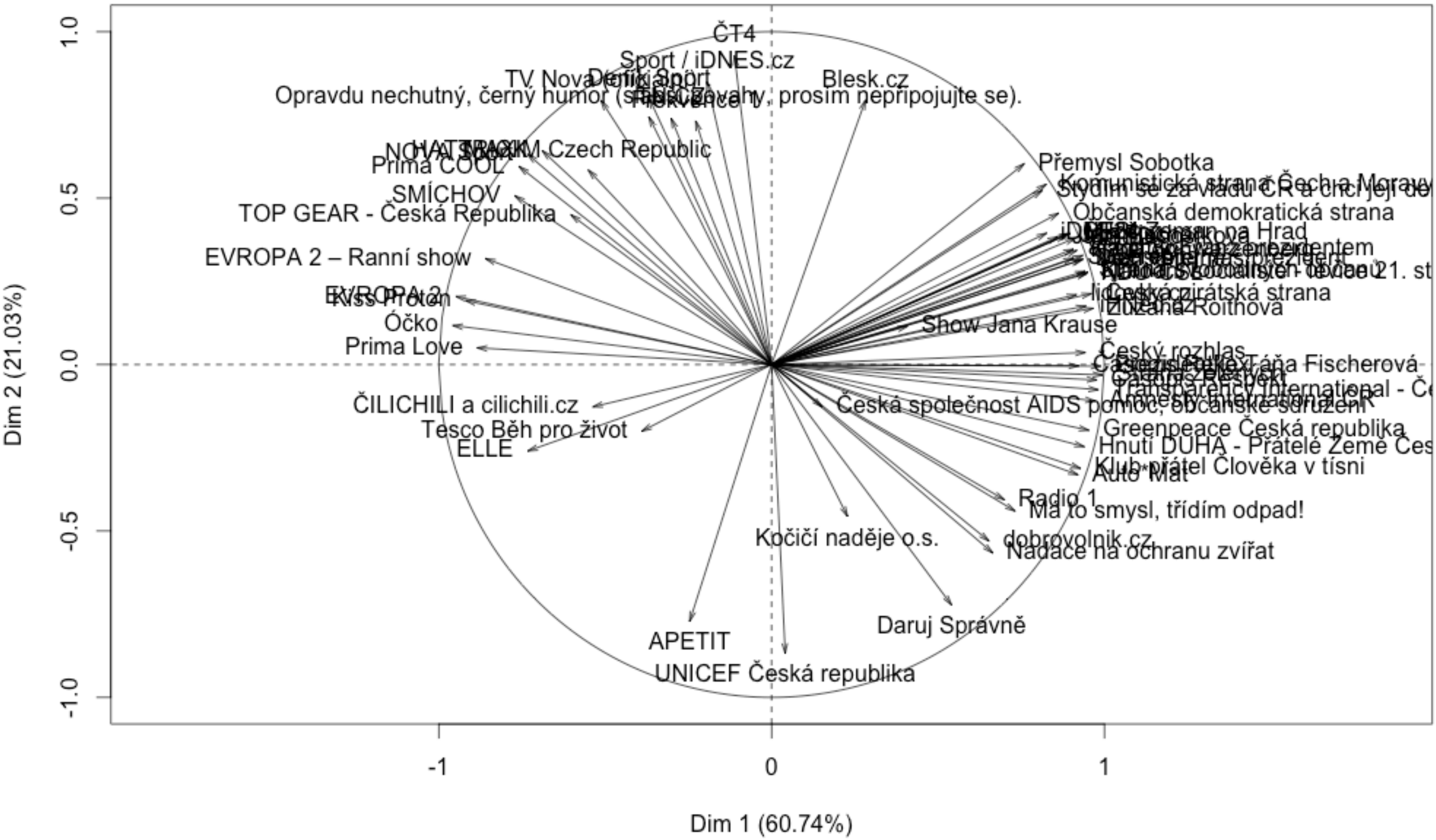
These two components explain 51.1 % of the point variability.

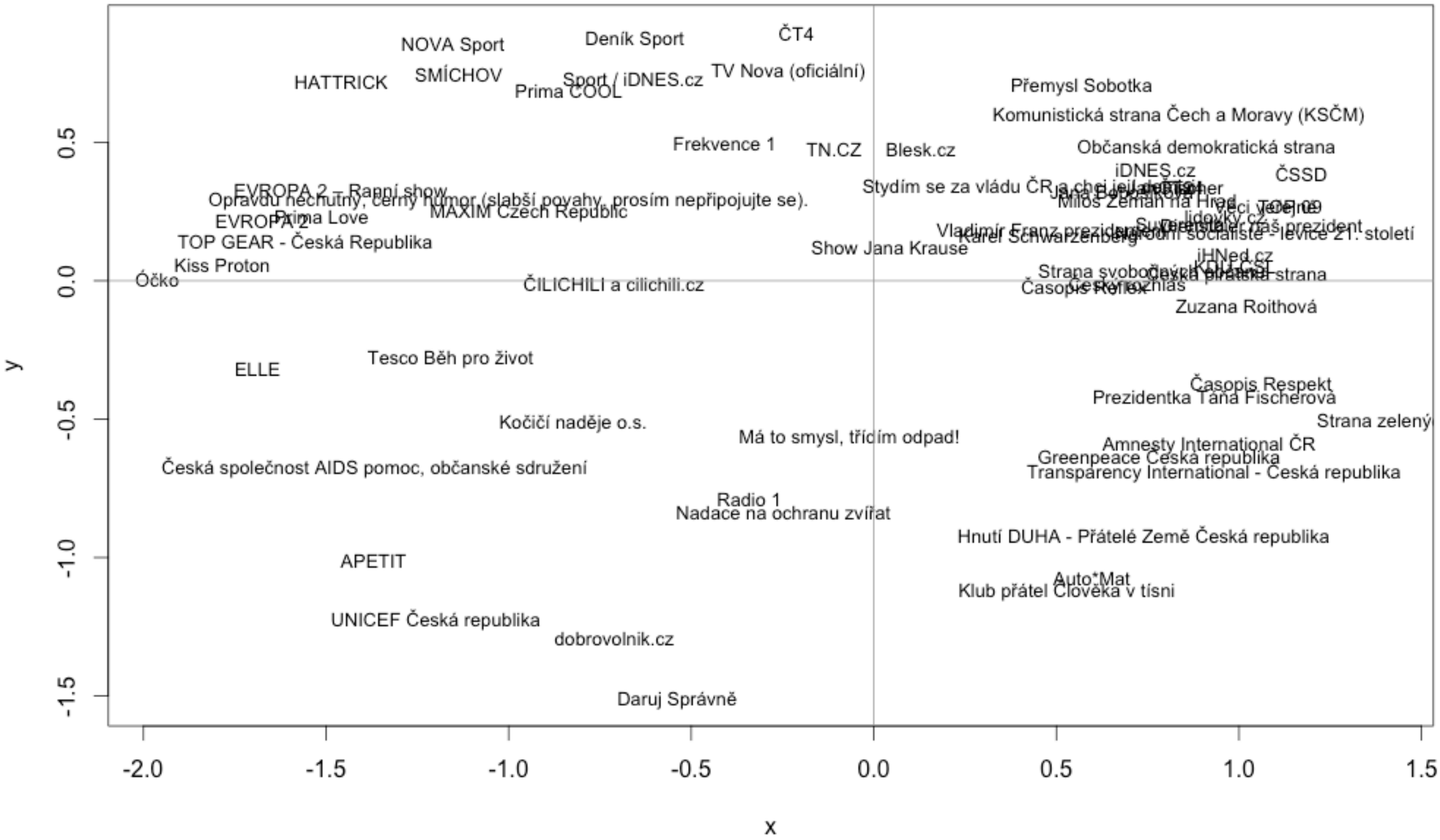
CLUSPLOT(tbl)



These two components explain 51.1 % of the point variability.

Variables factor map (PCA)

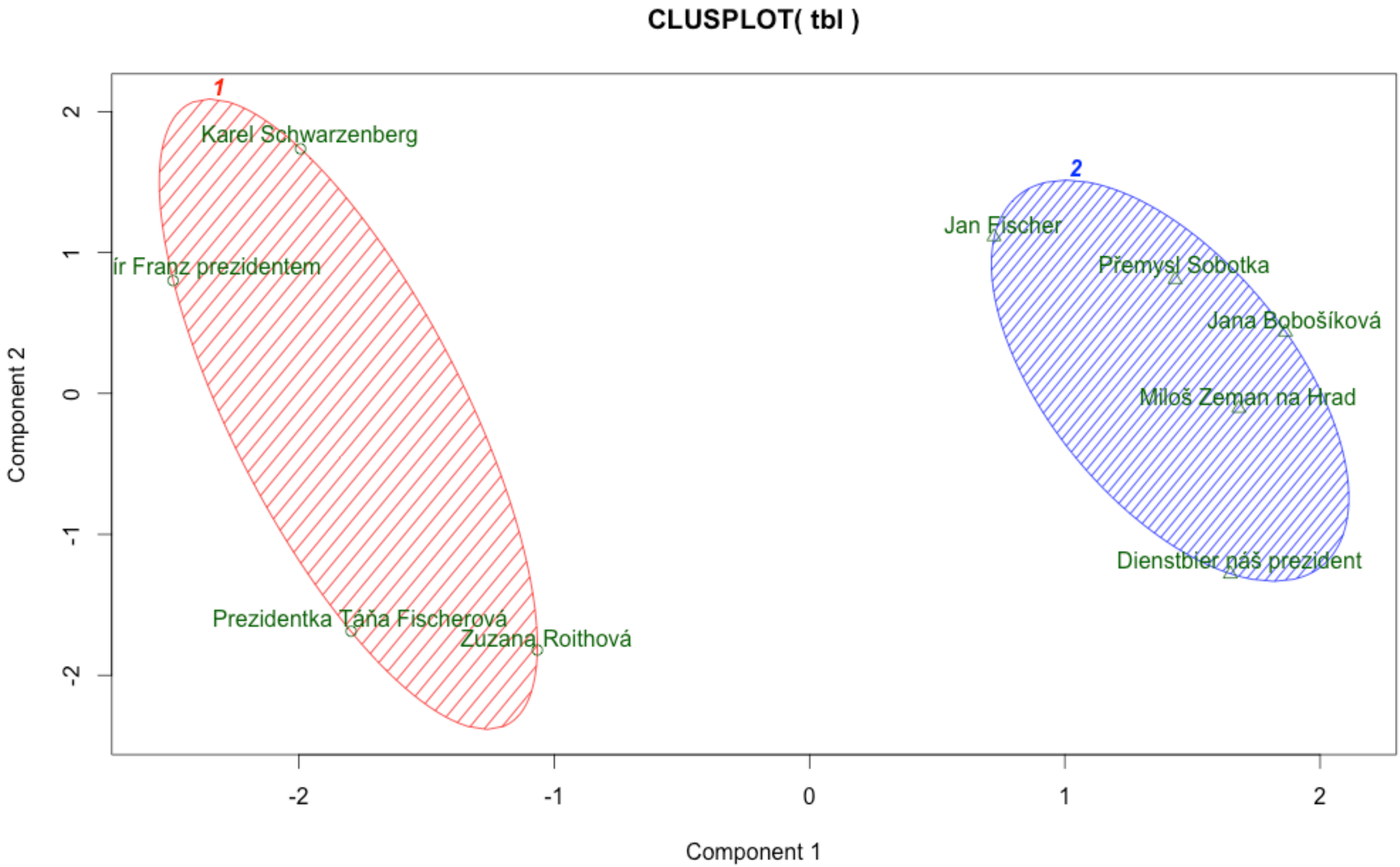



NOVA Sport
Deník Sport
ČT4
HATTRICK
SMÍCHOV
Sport / iDNES.cz
TV Nova (oficiální)
Prima COOL
Přemysl Sobotka
Komunistická strana Čech a Moravy (KSČM)
Frekvence 1
TN.CZ
Blesk.cz
Občanská demokratická strana
iDNES.cz
ČSSD
Stydím se za vládu ČR a chci jejíl
Jana Bobošíková
Miloš Zeman na Hrad
EVROPA 2 – Ranní show
Opravdu nechutný, černý humor (slabší povahy, prosím nepřipojujte se).
MAXIM Czech Republic
Prima Love
EVROPA 2
Vladimír Franz prezident
Karel Schwarzenberg
TOP GEAR - Česká Republika
Show Jana Krause
Kiss Proton
Óčko
ČILICHILI a cilichili.cz
Strana svobodných občanů
Česká pirátská strana
Časopis Reflex
Zuzana Roithová
Tesco Běh pro život
ELLE
Časopis Respekt
Prezidentka Táňa Fischerová
Strana zelených
Kočičí naděje o.s.
Má to smysl, třídím odpad!
Amnesty International ČR
Greenpeace Česká republika
Česká společnost AIDS pomoc, občanské sdružení
Transparency International - Česká republika
Radio 1
Nadace na ochranu zvířat
Hnutí DUHA - Přátelé Země Česká republika
APETIT
Auto*Mat
Klub přátel Člověka v tísni
UNICEF Česká republika
dobrovolnik.cz
Daruj Správně
0.5
0.0
-0.5
-1.0
-1.5
y
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x

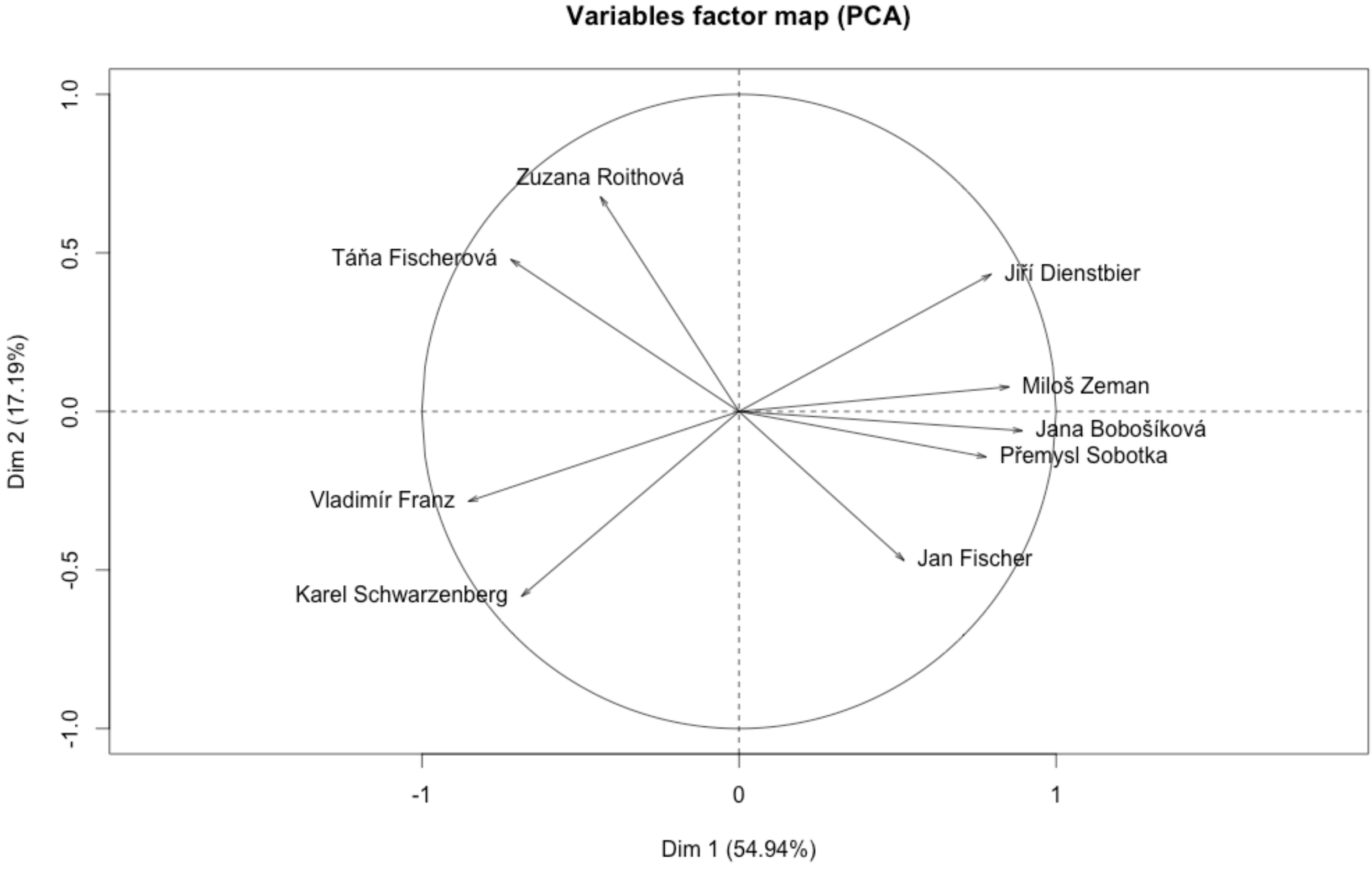
Případová studie

Prezidentské volby 2013

Mapa prezidentských kandidátů 14. 1. 2013



Mapa prezidentských kandidátů 14. 1. 2013



Hlavní úskalí

velikosti korpus a průniku (vzorec funguje “vždy”)

validace

úplnost korpusu při objevování

relativní drahost

vývoj v čase

Co nás čeká

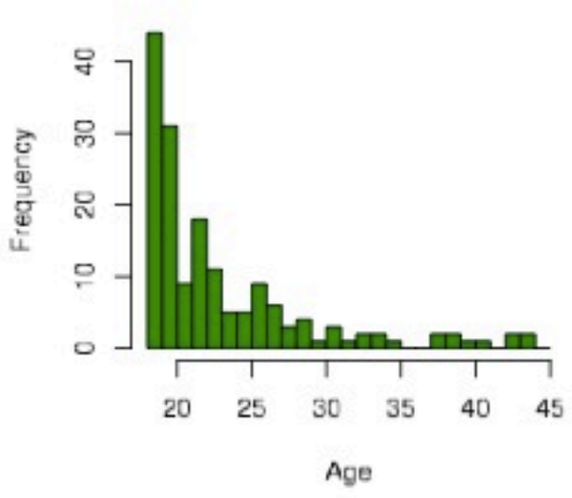
aneb na čem teď pracujeme

Tabuľka 1.3

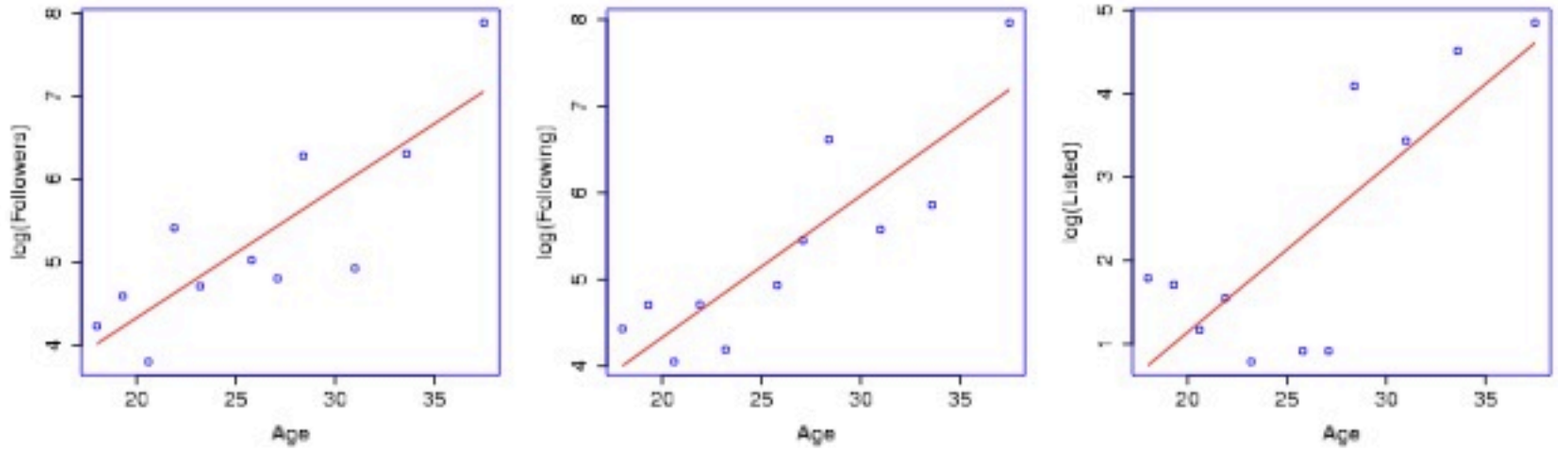
Prehľad charakteristík obsiahnutých v päťfaktorovom modeli osobnosti

Opis viazaný na vysoké skóre	Konvenčný názov a definícia faktoru	Opis viazaný na nízke skóre
spoločenský, aktívny, zhovorčivý, družný, priebojný, smelý	**Extraverzia (E; Extraversion)** Zisťuje kvantitu interpersonálnych interakcií, úroveň aktivácia a stimulácie.	uzavretý, plachý, málovravný, nesmelý, samotársky, tichý
dobrosrdečný, láskavý, znášanlivý, srdečný, skromný, zmierlivý	**Prívetivosť (A; Agreeableness)** Zisťuje kvalitu interpersonálnej orientácie na kontinuu od súcitenia po nepriateľskosť v myšlienkach, pocitoch i činoch.	panovačný, útočný, pomstychtivý, bezcitný, despotický, konfliktný
spoľahlivý, pracovitý, presný,	**Svedomitosť (C; Conscientiousness)** Zisťuje mieru motivácie a vytrvalosti na cieľ zameraného správania. Odlišuje spoľahlivých, na	bezcieľny, nedbalý, lenivý, nesvedomitý,

poriadkumilovný, zodpovedný, starostlivý	seba náročných ľudí od tých, ktorí sú ľahostajní a nedbalí.	chaotický, nevytrvalý
napätý, nepokojný, labilný, neistý, vznetlivý, popudlivý	**Emocionálna (ne)stabilita / Neuroticizmus (N; Neuroticism)** Odlišuje jedincov náchylných k psychickému vyčerpaniu a ťažko zvládajúcich psychickú záťaž od jedincov vyrovnaných a odolných voči psychickému vyčerpaniu.	pokojný, uvoľnený, vyrovnaný, stabilný, sebaistý, nezdolný
zvedavý, originálny, tvorivý, obrazotvorný, inteligentný, kultivovaný	**Otvorenosť voči skúsenosti / Intelekt / Kultúra (I; Intellect)** Zisťuje aktívne vyhľadávanie nových skúseností, zážitkov, toleranciu neznámeho.	konvenčný, pragmatický, realistický, neprispôsobivý, neinteligentný, nevzdelaný



(a) Distributions of Age.



(b) User activity varies with age.

Fig. 3. Age distribution and effects on Twitter activity

Trait	**Listeners** *log(Following)*	**Popular** *log(Followers)*	**Highly-read** *log(Listed)*	**Influential** *Klout*	**Influential** *log(TIME)*
O	0.05	0.05	**0.17***	0.13	0.00
C	0.08	0.10	0.02	0.01	**0.18*****
E	**0.13***	**0.15****	0.09	**0.15***	**0.25*****
A	0.07	0.02	0.03	-0.17	0.06
N	**-0.17****	**-0.19*****	-0.03	**-0.03***	**-0.20*****
log(Age)	**0.28***	**0.37***	0.13	0.05	**0.39***
Male	-0.05	-0.05	-0.05	-0.04	0.01

Děkuju za pozornost

@josefslerka