OBTEŽENA POVPREČJA IN PARADOKS PRIJATELJSTVA

BRIGITA FERČEC[1,2] IN NIKO TRATNIK[3]

[1]Fakulteta za energetiko, Univerza v Mariboru
[2]Center za uporabno matematiko in teoretično fiziko, Univerza v Mariboru
[3]Fakulteta za naravoslovje in matematiko, Univerza v Mariboru

Math. Subj. Class. (2010): 91D30

Članek obravnava zanimiv pojav, ki ga lahko opazimo na mnogih področjih življenja, tj. paradoks prijateljstva. Povezan je s posebno vrsto obteženih povprečij. Zato na začetku opišemo koncept obteženih povprečij skupaj s primeri situacij, v katerih se pogostokrat pojavi, in si ogledamo del pripadajoče matematične teorije. V zadnjem delu je opisan paradoks prijateljstva v kontekstu družabnih omrežij. Navedena je povezava z obteženimi povprečji kot tudi povezave z nekaterimi drugimi področji.

WEIGHTED AVERAGES AND THE FRIENDSHIP PARADOX

The paper describes an iteresting phenomenon which appears in many areas of life and is known as the friendship paradox. The latter is connected with a special type of weighted averages in mathematics. Thus, in the beginning the concept of weighted averages is described as well as its applications and a mathematical interpretation. The last part describes the friendship paradox as it appears in the context of social networks. The connection with weighted averages and connections with some other areas are stated.

Uvod

Večina ljudi je seznanjena z idejo računanja povprečja oz. aritmetičnega povprečja neke množice števil. Preprosto seštejemo vse elemente v tej množici in jih delimo s številom elementov množice. Vendar to deluje samo tedaj, ko so vsi elementi množice enakovredni oz. obteženi enako. Kot primer vzemimo povprečje mesečnega računa za elektriko za prejšnje leto. Seštejemo vrednosti dvanajstih položnic za elektriko za prejšnje leto in dobljeno vrednost delimo z 12, saj so obračuni narejeni mesečno.

Sedaj pa recimo, da smo opravljali izpit pri predmetu Matematika, ki je sestavljen iz treh delov: pisnega dela izpita, domačih nalog in ustnega dela izpita. Pri večini šolskih predmetov ti trije deli različno prispevajo h končni oceni, zato je v tem primeru primerno uporabiti *obteženo povprečje*.

Obteženo povprečje lahko opišemo kot povprečje, kjer nekatere vrednosti prispevajo več kot druge. Pri navadnem aritmetičnem povprečju pa so, drugače kot pri obteženem, vse vrednosti enakovredne. Formula za obteženo povprečje se uporablja za izračun povprečne vrednosti določene množice

števil z različnimi stopnjami pomembnosti oz. relevance. Relevanca vsakega števila se imenuje *utež števila*.

Vzemimo preprost primer množice števil $\{1, 2, 3, 4\}$ in izračunajmo povprečje teh števil kot

$$\text{povprečje} = \frac{1+2+3+4}{4} = 2{,}5.$$

Če bi v tem primeru dali vsakemu številu utež, bi v zgornjem primeru vsak element množice $\{1, 2, 3, 4\}$ dobil utež 25% (0,25) in bi povprečje lahko izračunali kot

$$\text{povprečje} = \frac{0{,}25 \cdot 1 + 0{,}25 \cdot 2 + 0{,}25 \cdot 3 + 0{,}25 \cdot 4}{0{,}25 + 0{,}25 + 0{,}25 + 0{,}25} = 2{,}5.$$

Sedaj pa spremenimo uteži in recimo, da število 1 dobi utež 0,1, število 2 dobi 0,2, število 3 utež 0,15 in število 4 dobi utež 0,55. Vsota uteži je 1 in vrednost povprečja je

$$\text{povprečje} = \frac{0{,}1 \cdot 1 + 0{,}2 \cdot 2 + 0{,}15 \cdot 3 + 0{,}55 \cdot 4}{1} = 3{,}15.$$

Iz zgornjega zelo preprostega primera vidimo, da se tedaj, ko nekatere vrednosti dobijo večje uteži kot druge, spremeni povprečje, ki se približa vrednosti z večjo utežjo.

Koncept obteženih povprečij se veliko uporablja tudi v ekonomiji, še posebej v poslovni in finančni ekonomiji. Kot preprost primer lahko navedemo investitorja, ki bi rad določil dobiček treh investicij, ki jih imenujmo investicija A, investicija B in investicija C. Recimo, da vloži 25 % svojega denarja v investicijo A, 25 % v investicijo B in 50 % vloži v investicijo C. Stopnja dobička za investicijo A je 5 %, za investicijo B 6 % in za investicijo C je 2 %. Če sedaj izračunamo obteženo povprečje glede na navedene podatke, dobimo povprečni dobiček (izračunan v deležu vloženega denarja)

$$\frac{0{,}25 \cdot (5\ \%) + 0{,}25 \cdot (6\ \%) + 0{,}50 \cdot (2\ \%)}{0{,}25 + 0{,}25 + 0{,}50} = 3{,}75\ \%.$$

Če bi investitor uporabljal običajno aritmetično povprečje, potem bi bilo povprečje 4,33 %. Ta precejšnja razlika v izračunu obeh povprečij nam kaže, kako pomembno je uporabiti pravo formulo za natančno analizo v podjetjih, kjer je pomembno vedeti, kako donosne so investicije.

Obtežena povprečja pa so tista, ki se velikokrat skrivajo za različnimi matematičnimi paradoksi. Primer v statistiki zelo znanega paradoksa je Simpsonov paradoks, ki se pogosto pojavlja v družbenih in zdravstvenih vedah. Le-ta včasih povzroči, da podatki, ki jih gledamo po nekih skupinah,

kažejo popolnoma drugačen trend, kot če te podatke gledamo združene skupaj. Eden izmed bolj znanih primerov tega paradoksa se je zgodil leta 1973 na Univerzi v Berkeleyju. Ko so analizirali vpis na univerzo, so ugotovili, da so bili moški, ki so se prijavili na študij, sprejeti v 44 % primerov, medtem pa je bilo pri vpisu uspešnih le 35 % žensk. Univerza je bila deležna številnih obtožb, povezanih s spolno diskriminacijo, saj so podatki kazali, da imajo moški večje možnosti, da so sprejeti na študij. Ko so se lotili analize po posameznih oddelkih, pa so ugotovili, da na nobenem oddelku moški niso bili bistveno bolj uspešni. Ravno nasprotno, na večini oddelkov so bile pri vpisu malo bolj uspešne ženske. Bistvo se je skrivalo v tem, da različni oddelki oz. študiji niso bili enako priljubljeni. Izkazalo se je, da so se ženske v veliki večini prijavljale na zelo priljubljene študije (npr. na angleščino), moški pa večinoma na manj priljubljene (npr. tehnika in kemija), to pa je bil razlog, da so bili skupno pri vpisu bolj uspešni.

Takih primerov navideznih paradoksov, kjer so v ozadju obtežena povprečja, je še veliko. V drugem poglavju si bomo ogledali še en paradoks, ki se da pojasniti z obteženimi povprečji. To je paradoks prijateljstva. V predzadnjem poglavju pa se bomo seznanili še s posplošenim paradoksom prijateljstva. Še prej pa se na kratko seznanimo z matematično razlago obteženih povprečij.

Obteženo povprečje neprazne množice podatkov $\{x_1, x_2, \ldots, x_n\}$ je

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n},$$

kjer je w_i utež, ki pripada podatku x_i. Zato podatki z večjo utežjo prispevajo več k obteženemu povprečju kot pa podatki z manjšo utežjo. Uteži niso nikoli negativne, nekatere, vendar ne vse (zaradi deljenja z nič), so lahko nič. Formula je enostavnejša, če so uteži normalizirane, tako da je njihova vsota 1, tj. $\sum_{i=1}^{n} w_i = 1$. Za takšne normalizirane uteži je obteženo povprečje preprosto

$$\overline{x} = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n.$$

Opazimo, da lahko uteži vedno normaliziramo s transformacijo uteži $w_i' = \frac{w_i}{\sum_{j=1}^{n} w_j}$, saj je

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{j=1}^{n} w_j} = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} x_i = \sum_{i=1}^{n} w_i' x_i,$$

kar je navadno obteženo povprečje.

Zgoraj navedeno obteženo povprečje je posplošitev aritmetičnega povprečja in se zato imenuje tudi *aritmetično obteženo povprečje*. Pri aritmetičnem povprečju dobi vsak element enako utež. Če vzamemo neprazno množico $\{x_1, \ldots, x_n\}$ in za vsak element x_i utež $w_i = \frac{1}{n}$, dobimo

$$\overline{x} = \frac{\sum_{i=1}^{n} \frac{1}{n} x_i}{\sum_{i=1}^{n} \frac{1}{n}} = \frac{\frac{1}{n}(x_1 + x_2 + \cdots + x_n)}{1} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

kar je znana formula za aritmetično povprečje.

Obstajata pa tudi *obteženo geometrijsko povprečje* in *obteženo harmonično povprečje*, ki vsako zase izhajata iz geometrijskega povprečja in harmoničnega povprečja.

Paradoks prijateljstva

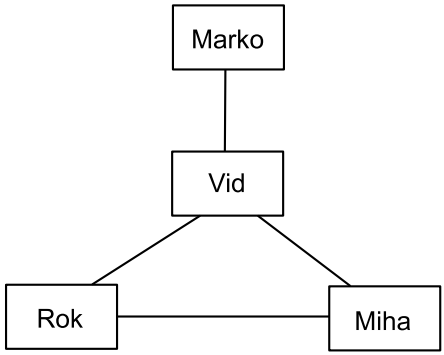
Dandanes ljudje veliko svojega časa namenimo družabnim omrežjem in povezovanju z ljudmi preko le-teh. Velika večina ljudi pa lahko opazi, da ima na teh omrežjih manj prijateljev kot večina njihovih prijateljev. Če ste med njimi tudi sami, potem ne skrbite, saj enako velja tudi za večino vaših prijateljev. To na primer potrjuje tudi obsežna študija Facebooka [3], kjer so raziskovalci ugotovili naslednji zanimiv rezultat. Najprej so pogledali, koliko ljudi ima manj prijateljev kot povprečno njegovi/njeni prijatelji. In izkazalo se je, da to velja za veliko večino uporabikov oziroma za kar 93 odstotkov vseh uporabnikov Facebooka. Merili pa so tudi povprečja na celotnem Facebooku in ugotovili, da imajo uporabniki v povprečju 190 prijateljev, medtem ko imajo njihovi prijatelji v povprečju 635 prijateljev. Kako točno so izračunali povprečje prijateljev od prijateljev, bomo videli v zgledu v nadaljevanju.

Tudi raziskave nevirtualnih družabnih omrežij kažejo enak trend. Ta pojav je namreč že leta 1991 odkril sociolog Scott L. Feld, ko internetna družabna omrežja še niso obstajala. Tako imamo tudi v nevirtualnem svetu večinoma manj prijateljev kot naši prijatelji. To pa seveda nima nikakršne povezave z osebnostmi, temveč sledi iz matematike. Za katerokoli omrežje, kjer ima nekaj ljudi več prijateljev kot drugi, velja, da je povprečno število prijateljev od prijateljev vedno večje kot povprečno število prijateljev. To trditev bomo v nadaljevanju strogo dokazali. Seveda bomo vedno predpostavljali, da so prijateljstva vzajemna.

Opisani pojav so poimenovali »paradoks prijateljstva« (v angleščini »the friendship paradox«). Njegova razlaga temelji na posebni vrsti obteženih povprečij, ki povzročajo različne navidezne paradokse tudi v mnogih drugih situacijah.

Da bomo lažje razmišljali, si za začetek zamislimo zelo preprost primer družabnega omrežja, ki ga sestavljajo samo štiri osebe. Dajmo jim nasle-

dnja imena: Marko, Vid, Rok in Miha. Recimo, da ima Marko samo enega prijatelja – Vida. Vid naj bo prijatelj z vsemi preostalimi, Rok in Miha pa naj bosta prijatelja še med seboj. Tako dobimo družabno omrežje, predstavljeno na sliki 1.



Slika 1. Primer družabnega omrežja.

Sedaj za vsakega posebej zapišimo, koliko prijateljev ima in koliko prijateljev imajo njegovi prijatelji.

Oseba	Število prijateljev	Število prijateljev od prijateljev	Povprečno število prijateljev od prijateljev
Marko	1	3	3
Vid	3	1; 2; 2	1,67
Rok	2	3; 2	2,5
Miha	2	3; 2	2,5

Takoj opazimo, da ima večina (Marko, Rok in Miha) manj prijateljev, kot imajo v povprečju prijateljev njegovi prijatelji. Le Vid, ki je bolj »priljubljen«, ima več prijateljev od svojih prijateljev.

Da bomo lahko to v splošnem pojasnili, označimo z A povprečno število prijateljev ljudi v omrežju (povprečje števil v drugem stolpcu tabele) in z B povprečno število prijateljev od prijateljev (povprečje števil v tretjem stolpcu tabele). Izračunajmo povprečji A in B za dani primer.

$$A = \frac{1+3+2+2}{4} = 2$$

$$B = \frac{3+(1+2+2)+(3+2)+(3+2)}{8} = 2{,}25$$

Opazimo, da za dani primer družabnega omrežja velja $A < B$. V nadaljevanju pa bomo dokazali, da ta neenakost velja za čisto vsako družabno

omrežje, v katerem nimajo vsi enakega števila prijateljev. Dejstvo, da je povprečno število prijateljev strogo manjše od povprečja prijateljev od prijateljev, je razlog za nastanek omenjenega pojava, saj ima zaradi tega večina ljudi manj prijateljev kot povprečno njihovi prijatelji.

Preden pa se lotimo strogega dokaza, poskušajmo zgornjo neenakost razložiti intuitivno. V ta namen zapišimo povprečje B nekoliko drugače. Ker ima Vid 3 prijatelje, ga bodo tudi trije omenili kot prijatelja in zato se bo v števcu števila B trikrat pojavilo število 3, torej $3 \cdot 3 = 3^2$. Podobno ima Rok 2 prijatelja, zato bosta tudi dva omenila število 2, ko bosta naštevala, koliko prijateljev imajo njuni prijatelji, in tako se bo v števcu pojavil člen 2^2. Podobno pa bo tudi Miha prispeval 2^2 in Marko 1^2. Tako je

$$B = \frac{3^2 + 2^2 + 2^2 + 1^2}{8}.$$

V povprečju B števila prijateljev pred seštevanjem še kvadriramo, s tem pa damo dodatno težo velikim številom, in zato je $B > A$. Povprečje B je torej obteženo povprečje z utežmi, ki so kar enake vrednostim, katerih povprečje računamo, saj velja

$$B = \frac{3 \cdot 3 + 2 \cdot 2 + 2 \cdot 2 + 1 \cdot 1}{3 + 2 + 2 + 1}.$$

Takoj opazimo še, da je v imenovalcu števila B vsota števila prijateljev vseh oseb (v našem primeru $1 + 3 + 2 + 2$ – vsota števil v prvem stolpcu). Očitno bo to vedno res.

Končno se lotimo še splošnega primera, ko imamo v družabnem omrežju n ljudi. Ugotovitev zapišimo kot izrek.

Izrek 1. *V poljubnem družabnem omrežju, v katerem nimajo vsi enakega števila prijateljev, označimo z A povprečno število prijateljev, z B pa povprečno število prijateljev od prijateljev. Potem velja $0 < A < B$.*

Dokaz. Naj ima družabno omrežje n ljudi. Prvi naj ima x_1 prijateljev, drugi x_2 prijateljev in tako naprej vse do zadnjega, ki ima x_n prijateljev. Povprečje prijateljev A v splošnem primeru zlahka izračunamo in dobimo

$$A = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

S pomočjo že znanih razmislekov pa ugotovimo tudi, da je povprečno število prijateljev od prijateljev

$$B = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{x_1 + x_2 + \cdots + x_n}.$$

Seveda je $A > 0$ in $B > 0$, saj je povprečje pozitivnih števil vedno pozitivno. Zapišimo naslednji račun.

$$\begin{aligned}\frac{(x_1 - A)^2 + (x_2 - A)^2 + \cdots + (x_n - A)^2}{n} &= \\ = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - 2A \cdot \frac{x_1 + x_2 + \cdots + x_n}{n} + A^2 &= \\ = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - A^2.\end{aligned}$$

Če izraz $\frac{(x_1-A)^2+(x_2-A)^2+\cdots+(x_n-A)^2}{n}$, ki ga v statistiki imenujemo varianca, označimo z $\mathrm{Var}(x)$, dobimo

$$\frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} = A^2 + \mathrm{Var}(x).$$

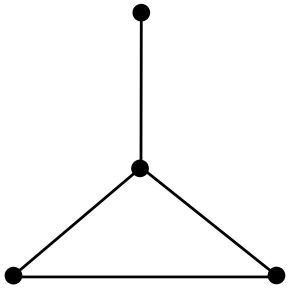
To enakost delimo z A in dobimo

$$B = A + \frac{\mathrm{Var}(x)}{A}.$$

Ker je vedno $\mathrm{Var}(x) \geq 0$ (in $\mathrm{Var}(x) = 0$ samo tedaj, ko je $x_1 = x_2 = \cdots = x_n$), za vsako družabno omrežje, kjer nimajo vsi enakega števila prijateljev, velja $A < B$. ∎

Opazimo, da lahko paradoks prijateljstva formuliramo na dveh nivojih: za posameznika in za družabno omrežje. Paradoks prijateljstva za družabno omrežje smo zapisali v zgornjem izreku. Na nivoju posameznika pa paradoks velja, če ima posameznik manj prijateljev kot povprečno njegovi/njeni prijatelji. Omenili pa smo že, da na nivoju posameznikov paradoks velja za veliko večino članov omrežja.

Seveda si lahko vsako družabno omrežje naravno predstavimo z grafom, kjer vozlišča grafa (točke) predstavljajo ljudi, pri tem pa sta dve vozlišči sosednji (to pomeni, da je med njima povezava), ko sta ustrezni osebi med seboj prijatelja. Graf družabnega omrežja, ki smo ga obravnavali prej, vidimo na sliki 2. Pri tem je očitno, da stopnja vozlišča (število sosedov vozlišča) pomeni število prijateljev ustrezne osebe. Paradoks prijateljstva v jeziku teorije grafov torej pove, da je povprečna stopnja vozlišč v grafu, v katerem nimajo vsa vozlišča iste stopnje, vedno manjša kot povprečna stopnja njihovih sosedov. Tudi raziskovalci, ki so proučevali Facebook, so opazovali lastnosti njegovega grafa. Ugotovili so na primer tudi, da kar 99,91 odstotka njegovih vozlišč (ljudi) pripada isti povezani komponenti. To pomeni, da lahko med temi za poljubna dva najdemo pot v grafu, ki ju povezuje.



Slika 2. Graf, ki prikazuje družabno omrežje s slike 1.

Posplošeni paradoks prijateljstva

Paradoks prijateljstva torej obravnava eno značilnost posameznikov, to je število njihovih prijateljev, oziroma stopnjo vozlišča v ustreznem grafu. Vendar pa imajo posamezniki tudi druge karakteristike, kot so na primer spol, starost, poklic ipd. Zato so v članku [1] paradoks prijateljstva posplošili tako, da ga lahko formuliramo za poljubno značilnost vozlišč, ki se da izraziti s številom. Kadar za značilnost izberemo stopnjo vozlišča, pa kot poseben primer dobimo paradoks prijateljstva. To posplošitev so poimenovali *posplošeni paradoks prijateljstva*. Nato so proučevali še mrežo znanstvenih člankov in prišli do podobnih rezultatov kot pri običajnem paradoksu prijateljstva. Ugotovili so, da imajo na primer vaši soavtorji zelo verjetno več soavtorjev, več citatov in tudi več objav kot vi. Oglejmo si posplošeni paradoks prijateljstva bolj natančno.

Vozlišča v grafu bodo označena z naravnimi števili, karakteristika vozlišča i naj bo x_i, njegova stopnja pa d_i. Posplošeni paradoks prijateljstva bomo zdaj obravnavali na nivoju posameznika in ne več na nivoju omrežja. Pravimo, da posplošeni paradoks prijateljstva velja za vozlišče i, če je izpolnjen naslednji pogoj:

$$x_i < \frac{\sum_{j \in N(i)} x_j}{d_i}, \tag{1}$$

kjer je $N(i)$ množica vseh sosedov vozlišča i. Takoj opazimo, da če izberemo $x_i = d_i$, posplošeni paradoks prijateljstva postane običajni paradoks prijateljstva.

V nadaljevanju bomo na kratko pogledali verjetnost in statistiko v omrežju soavtorstev, kot so to naredili v članku [1]. V ta namen bomo s $P(d, x)$ označili verjetnost, da vozlišče s stopnjo d in karakteristiko x zadošča enačbi (1). Seveda velja, da se pri fiksnem d z večanjem vrednosti x verjetnost $P(d, x)$ manjša. Raziskovalci so proučevali dve informacijski bazi: *Physical*

Review journals (PR) in *Google Scholar profile dataset of network scientists* (GS). Za vozlišča v grafu so vzeli vse avtorje, pri tem pa med dvema avtorjema obstaja povezava, če sta skupaj napisala kakšen članek. Omrežje PR je vsebovalo 242592 vozlišč, omrežje GS pa 29968. Pri tem so opazovali naslednje karakteristike vozlišč: število soavtorjev, število citatov, število objav in povprečno število citatov na objavo.

Raziskovalci so podatke obdelali statistično, pri tem pa so med drugim računali, kolikšna je povprečna verjetnost H, da posplošeni paradoks prijateljstva velja (pri tem so torej upoštevane vse verjetnosti $P(d, x)$). Ugotovili so, da je za vsako izmed proučevanih karakteristik ta verjetnost zelo velika, kar pomeni, da posplošeni paradoks prijateljstva velja za veliko večino vozlišč v omrežju. Na primer za število soavtorjev je ta verjetnost 0,934, za število citatov je 0,921, za število objav pa 0,912. Le za povprečno število citatov na objavo je ta verjetnost nekoliko manjša, in sicer 0,720. S tem so torej ugotovili, da kot pri običajnem paradoksu prijateljstva, tudi za druge karakteristike velja, da imajo pri veliki večini vozlišč manjšo vrednost kot pri njihovih sosedih.

Uporaba v praksi

Kot mnoge matematične ideje je tudi ta paradoks pripeljal do zanimivih praktičnih aplikacij. Nedavno je vzpodbudil sistem zgodnjega opozarjanja za odkrivanje izbruhov nalezljivih bolezni. V študiji, ki so jo opravili na Harvardu v času pandemične gripe leta 2009, sta znanstvenika Nicholas Christakis in James Fowler spremljala status gripe v veliki skupini naključno izbranih študentov in njihovih prijateljev. Nenavadno, prijatelji so zboleli dva tedna pred naključno izbranimi študenti, domnevno zato, ker so bili na splošno bolj povezani znotraj družabne mreže, kar tudi pričakujemo iz paradoksa prijateljstva. V drugih okoliščinah je lahko dva tedna dolgo prehodno obdobje, kot je bilo to, zelo koristno, da organi za javno zdravje načrtujejo odziv na okužbe, preden le-te napadejo množice.

LITERATURA

[1] Young-Ho Eom, Hang-Hyun Jo, *Generalized friendship paradox in complex networks: The case of scientific collaboration*, Scientific Reports **4**, 4603 (2014).

[2] Scott L. Feld, *Why your friends have more friends than you do?*, American Journal of Sociology **96** 6, (1991) 1464–1477.

[3] J. Ugander et al., *The Anatomy of the Facebook Social Graph*, arXiv:1111.4503v1 (2011).

[4] S. Strogatz, *Friends You Can Count On*, The New York Times (2012), `http://opinionator.blogs.nytimes.com/2012/09/17/friends-you-can-count-on/`, ogled: 28. 1. 2016.

[5] *Friendship paradox*, Wikipedia, `https://en.wikipedia.org/wiki/Friendship_paradox`, ogled: 28. 1. 2016.