UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY




# HOMOFÍLIA V SOCIÁLNYCH SIEŤACH

DIPLOMOVÁ PRÁCA

2019 Bc. Katarína Semančíková

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# HOMOFÍLIA V SOCIÁLNYCH SIEŤACH

## DIPLOMOVÁ PRÁCA

| | |
|---|---|
| Študijný program: | Ekonomicko-finančná matematika a modelovanie |
| Študijný odbor: | 1114 Aplikovaná matematika |
| Školiace pracovisko: | Katedra aplikovanej matematiky a štatistiky |
| Vedúci práce: | doc. RNDr. Mgr. Beáta Stehlíková, PhD. |

Bratislava 2019 **Bc. Katarína Semančíková**






Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Bc. Katarína Semančíková
**Študijný program:** ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
**Študijný odbor:** aplikovaná matematika
**Typ záverečnej práce:** diplomová
**Jazyk záverečnej práce:** slovenský
**Sekundárny jazyk:** anglický

**Názov:** Homofília v sociálnych sieťach.
*Homophily in social networks.*

**Anotácia:** Homofília už podľa názvu - homo=rovnaký, prípona philia=mať niečo rád) vyjadruje známe príslovie 'vrana k vrane sadá'. V kontexte sociálnych sietí to znamená, že napríklad pravdepodobnosť, že dvaja ľudia sú priateľmi na Facebooku je väčšia, ak majú spoločné určité charakteristiky. V diplomovej práci sa vysvetľuje, ako toto matematicky merať a ako testovať prítomnosť tohto efektu v danej sieti. Obsahuje prehľad literatúry zaoberajúcej sa touto otázkou, ako aj samostatné analýzy na zvolených sieťach.

**Vedúci:** doc. RNDr. Mgr. Beáta Stehlíková, PhD.
**Katedra:** FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
**Vedúci katedry:** prof. RNDr. Daniel Ševčovič, DrSc.

**Dátum zadania:** 24.01.2018

**Dátum schválenia:** 25.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

**Poďakovanie**

Chcela by som poďakovať v prvom rade svojej vedúcej diplomovej práce doc. RNDr. Mgr. Beáte Stehlíkovej, RhD. za čas, ochotu, veľkú podporu a odborné rady, ktoré mi pomohli pri písaní práce. V druhom rade by som chcela poďakovať rodine a to hlavne mojej mame za starostlivosť. Zároveň by som chcela poďakovať dedkovi za motiváciu a všetkým priateľom za ochotu.

# Abstrakt v štátnom jazyku

SEMANČÍKOVÁ, Katarína: Homofília v sociálnych sieťach [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Mgr. Beáta Stehlíková, PhD., Bratislava, 2019, 106 s.

Práca sa venuje analýze sietí, ktorá spadá pod teóriu grafov. Sleduje sa efekt homofílie vo vzorke ľudí na sociálnej sieti. V práci sa opisuje, ako toto matematicky merať. Homofília v kontexte diplomovej práce predstavuje pravdepodobnosť, že dvaja ľudia budú kamaráti na základe ich spoločných charakteristík a vlastností. Na analýzu sietí sa využívajú ERG modely. V úvode práce sa najprv priblíži teória sietí, ktorú ERG modely využívajú. Následne sa priblížia ERG modely, ktoré sa v nasledovných kapitolách aplikujú na dve skupiny dát. Opisujú sa jednotlivé modely ako aj spätné testovanie kvality modelov. V závere práce je zhodnotená efektivita modelov na jednotlivé dáta a zistené výsledky homofílie v sieťach.

**Kľúčové slová:** Siete, ERGM, Goodness of Fit, Facebook, Homofília

# Abstract

SEMANČÍKOVÁ, Katarína: Homophily in social networks [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Mgr. Beáta Stehlíková, PhD., Bratislava, 2019, 106 p.

The thesis uses network analysis as based on the graph theory. More specifically, the homophily effect is assessed in a sample of people within a social network. The thesis describes ways how to measure the effect mathematically. In the context of this master's thesis, Homophily is defined as the probability that two people have mutual friends based on their shared interests and attributes. To analyse networks are used ERG models. The introduction describes the network theory used by ERG models. Then ERG are defined that are subsequently applied to two groups of data. Individual models are described as well as the recursive testing of model quality. The conclusion evaluates efficiency of models and results of homophily in networks.

**Keywords:** network, ERGM, Goodness of Fit, Facebook, Homophily

# Obsah

**Zoznam obrázkov** **13**

**Zoznam tabuliek** **15**

**Úvod** **17**

**1 Teória sietí** **19**

**2 ERG modely** **22**

2.1 Definícia modelov . . . . . 22

2.2 Vysvetľujúce premenné . . . . . 24

2.3 Informačné kritériá . . . . . 27

**3 Faktorová analýza** **28**

**4 Aplikácia modelov na dáta o žiakoch zo strednej školy v Glasgowe** **30**

4.1 Dáta . . . . . 30

4.2 Modely s charakteristikami vrcholov . . . . . 32

4.2.1 Bernoulliho model . . . . . 33

4.2.2 Modely sledujúce základné osobné údaje . . . . . 34

4.2.3 Model sledujúci hudobné preferencie . . . . . 39

4.2.4 Modely sledujúce fajčenie . . . . . 43

4.2.5 Modely sledujúce voľnočasové aktivity . . . . . 45

4.2.6 Spojenie modelov . . . . . 48

4.3 Testovanie modelov . . . . . 51

4.4 Modely s charakteristikami siete . . . . . 54

**5 Aplikácia modelov na dáta od priateľov na Facebooku** **60**

5.1 Dáta . . . . . 60

5.2 Modely s charakteristikami vrcholov . . . . . 62

5.2.1 Bernoulliho model . . . . . 62

5.2.2 Modely sledujúce základné údaje . . . . . 63

5.2.3 Modely sledujúce hudobné preferencie . . . . . 68

5.2.4 Modely sledujúce voľnočasové aktivity . . . . . 71
5.2.5 Modely sledujúce rôzne druhy športu . . . . . 73
5.2.6 Spojenie modelov . . . . . 77
5.3 Testovanie modelu . . . . . 80
5.4 Modely s charakteristikami siete . . . . . 82

**Záver** **89**

**Zoznam použitej literatúry** **91**

**Príloha A - dáta z Glasgowu** **94**

**A.1 Výstupy pre modely s vlastnosťami vrcholov** **94**

**A.2 Testovanie modelu s vlastnosťami vrcholov** **96**

**A.3 Výstupy pre modely s charakteristikami sietí** **97**

**A.4 Testovanie modelu so zahrnutou charakteristikou siete** **99**

**Príloha B - dáta z Facebooku** **101**

**B.1 Výstupy pre modely s vlastnosťami vrcholov** **101**

**B.2 Testovanie modelu s vlastnosťami vrcholov** **103**

**B.3 Výstupy pre modely s charakteristikami sietí** **105**

**B.4 Testovanie modelu so zahrnutou charakteriskou siete** **105**

# Zoznam obrázkov


1 Príklady neorientovanej a orientovanej siete. . . . . 19
2 Príklady počtu spoločných susedov pre hranu medzi vrcholmi A a B. . . . 21
3 Markovova hypotéza závislosti (prebrané z článku [29]). . . . . 25
4 Vytvorená sieť na základe dát z Glasgowu. . . . . 31
5 Histogram počtu žiakov v závislosti od veku. . . . . 37
6 Pomer odpovedí o počúvaní hudobných štýlov. . . . . 40
7 Počty hrán medzi žiakmi, ktorí odpovedali rovnako o počúvaní hudobných štýlov. . . . . 41
8 Korelácia hudobných štýlov medzi sebou. . . . . 41
9 Korelácie vytvorených faktorov s hudobnými štýlmi. . . . . 42
10 Pomer odpovedí o fajčení. . . . . 44
11 Počty hrán medzi žiakmi, ktorí odpovedali rovnako o fajčení. . . . . 44
12 Pomer odpovedí pri voľnočasových aktivitách. . . . . 46
13 Počty hrán medzi žiakmi, ktorí odpovedali rovnako pri voľnočasových aktivitách. . . . . 46
14 Počty hrán medzi žiakmi, ktorí odpovedali rovnako na sledované premenné. 48
15 Rozdelenie stupňov vrcholov v sieťach. . . . . 52
16 Rozdelenie počtu spoločných susedov v sieťach. . . . . 52
17 Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi v sieťach. . . . . 53
18 Testovanie modelu s charakteristikou. . . . . 58
19 Vytvorená sieť na základe dát z Facebooku. . . . . 61
20 Histogram počtu osôb vzhľadom na vek. . . . . 64
21 Pomer odpovedí pri jednotlivých premenných. . . . . 66
22 Pomer odpovedí na počúvanie hudobných štýlov. . . . . 69
23 Počty hrán medzi osobami, ktoré odpovedali rovnako pri počúvaní hudobných štýlov. . . . . 69
24 Korelácia hudobných štýlov medzi sebou. . . . . 70
25 Korelácia hudobných štýlov s faktormi. . . . . 70
26 Pomer odpovedí pri voľnočasových aktivitách. . . . . 71

27 Počty hrán medzi osobami, ktoré odpovedali rovnako pri voľnočasových aktivitách. . . . . 72
28 Pomer odpovedí pri športoch. . . . . 74
29 Počty hrán medzi osobami, ktoré odpovedali rovnako pri športoch. . . . . 75
30 Korelácie medzi jednotlivými športami. . . . . 75
31 Faktory a ich korelácie na základe športov. . . . . 76
32 Rozdelenie stupňov vrcholov v sieťach. . . . . 81
33 Rozdelenie počtu spoločných susedov v sieťach. . . . . 81
34 Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi v sieťach. . . . . 82
35 Hodnoty informačných kritérií pre modely s jednotlivými stupňami zobrazené v grafe. . . . . 83
36 Hodnoty informačných kritérií pre modely s počtami spoločných susedov. . 84
37 Testovanie modelu s charakteristikou. . . . . 86

# Zoznam tabuliek


1 Porovnanie AIC a BIC pre modely sledujúce pohlavie. . . . . 36
2 Porovnanie AIC a BIC pre modely sledujúce stav žiakov. . . . . 38
3 Porovnanie AIC a BIC pre modely sledujúce základné osobné údaje. . . . . 38
4 Porovnanie AIC a BIC pre modely spájajúce základné osobné údaje. . . . . 39
5 Porovnanie AIC a BIC pre modely sledujúce fajčenie. . . . . 45
6 Porovnanie AIC a BIC pre modely sledujúce voľnočasové aktivity. . . . . 47
7 Porovnanie AIC a BIC pre modely sledujúce stupne. . . . . 54
8 Porovnanie AIC a BIC pre modely sledujúce k-star. . . . . 55
9 Porovnanie AIC a BIC pre modely sledujúce rozdelenie počtu spoločných susedov. . . . . 56
10 Porovnanie AIC a BIC pre modely sledujúce pohlavie. . . . . 64
11 Porovnanie AIC a BIC pre modely sledujúce stav. . . . . 65
12 Porovnanie AIC a BIC pre modely sledujúce základné osobné údaje. . . . . 67
13 Porovnanie AIC a BIC pre modely spájajúce základné osobné údaje. . . . . 67
14 Porovnanie AIC a BIC pre modely sledujúce voľnočasové aktivity. . . . . 73
15 Porovnanie AIC a BIC pre modely spájajúce voľnočasové aktivity. . . . . 73
16 Porovnanie AIC a BIC pre modely spájajúce stupne. . . . . 83
17 Porovnanie AIC a BIC pre modely sledujúce gwesp s rôznymi hodnotami parametra $\tau$. . . . . 85

# Úvod

Človek už od svojho počiatku nie je samotár. Už v dávnej minulosti sa ľudia zoskupovali do skupín aby dosiahli väčšiu silu, kvôli sympatiám alebo iným dôvodom. V dnešnej dobe sa človek dostane za svoj život do kontaktu s veľkým množstvom ľudí. S niektorými len na pár minút, iní zostanú v kontakte po dlhú dobu. Ako išla doba dopredu vznikali rôzne prostriedky, aby komunikácia medzi ľuďmi bola ľahšia. Najprv sa posielali listy, neskôr maily a momentálne ľudia medzi sebou komunikujú aj pomocou sociálnych sietí. Jednou z najznámejších sociálnych sietí je Facebook. Každý používateľ má svoj profil. Vie si pridávať iné profily do listu kamarátov a tak ľahko zdieľať svoj život, zážitky, ale aj jednoduchšie komunikovať so svojimi kamarátmi. Prečo sú dvaja používatelia kamaráti a čo mohlo byť podnetom k tomuto kamarátstvu? Ovplyvnila priateľstvo na Facebooku niektorá z ich vlastností? Týmto sa budeme zaoberať v našej diplomovej práci.

Homofília predstavuje tendenciu ľudí vytvárať sociálne väzby s ľuďmi, s ktorými majú zhodné vlastnosti alebo spoločné charakteristiky. Sociálne väzby v tejto práci budú predstavovať jednotlivé kamarátstva. Na pracovanie s takýmito väzbami sa využíva teória sietí, ktorá spadá pod teóriu grafov. Sieť sa skladá z vrcholov a hrán. V našom prípade vrcholy budú predstavovať ľudí a hrana ich priateľstvo. Teóriu sietí a jednotlivé vlastnosti sietí predstavíme v prvej kapitole. Jedny zo skupiny modelov, ktoré pracujú so sieťami a vedia ich analyzovať sú ERG modely. V druhej kapitole uvádzame základný tvar ERG modelu. Pomocou týchto modelov budeme testovať, aké ľudské charakteristiky alebo vlastnosti mohli ovplyvniť kamarátstva v našich sieťach. Do modelov sa dajú zahrnúť, nie len vlastnosti vrcholov ale aj vlastnosti siete. Možnosti vlastností sietí a ich alternatívy priblížime taktiež v druhej kapitole. Po zahrnutí rôznych parametrov budeme potrebovať zistiť ako zlepšujú model. Na porovnanie modelov nám budú slúžiť informačné kritéria, ktoré taktiež v krátkosti priblížime.

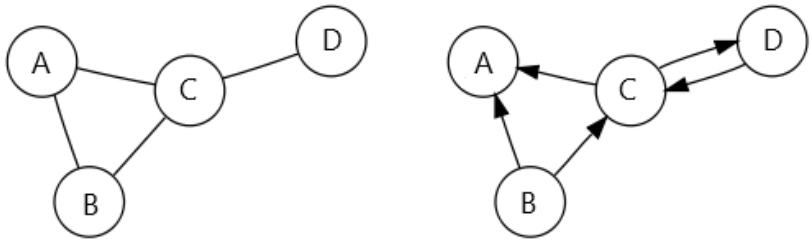
ERG modely aplikujeme najprv na dáta od žiakov zo školy v Glasgowe, ktoré sme našli na internete. Na vytváranie modelov použijeme knižnicu *statnet* v *R* [8, 11]. Nakoľko dáta pochádzajú z minulosti, sociálne siete ako ich poznáme teraz, neboli veľmi rozvinuté. V dátach sa žiaci označovali za kamarátov medzi sebou a tým sme dostali sieť. Vytvoríme rôzne modely sledujúce rôzne informácie a taktiež otestujeme modely ako dobre opisujú dáta. Následne v ďalšej kapitole vytvoríme modely na základe siete, ktorú sme vytvorili

z nami zozbieraných dát. Táto sieť bude pozostávať z priateľov na Facebooku, ktorým sme dali vyplniť dotazník. Tento dotazník sa pýtal na základné informácie, záujmy a preferencie. Taktiež zadefinujeme konkrétne modely a posnažíme sa nájsť konkrétne charakteristiky, ktoré ovplyvňujú existenciu hrán. Porovnanie našich modelov a ako dobre reprezentujú sieť popisujeme v priebehu kapitoly.

Ku koncu kapitol, v ktorých prezentujeme aplikáciu ERG modelov, zahrnieme do modelov aj charakteristiky sietí. Na záver kapitol uvedieme odhadnuté parametre modelov a zistené výsledky vplyvu homofílie.

# 1 Teória sietí

Teória sietí spadá pod teóriu grafov. Ako sme naznačili v úvode, siete sa skladajú z vrcholov a hrán. Vrcholy zväčša predstavujú veci, látky, osoby a hrany ich interakcie medzi sebou. Matematicky to vieme zapísať ako graf $G(V, E)$, kde $V$ predstavuje množinu vrcholov a $E$ množinu hrán [27]. Počet všetkých vrcholov označujeme $|V|$, alebo neskôr v práci $n$. Počet všetkých hrán označujeme $|E|$.




**(a)** Neorientovaná sieť **(b)** Orientovaná sieť

**Obr. 1:** Príklady neorientovanej a orientovanej siete.

Nech $u$ a $v$ sú vrcholy, medzi ktorými existuje hrana. Túto hranu zapíšeme ako dvojicu $(u, v) \in E$. Hrany môžu byť neorientované alebo orientované. V prípade orientovanej hrany dostávame usporiadanú dvojicu $(u, v)$. Ak všetky hrany v sieti sú neorientované, tak sieť nazývame neorientovanou. Na obrázku 1 uvádzame príklady sietí. V oboch prípadoch máme množinu vrcholov $V = \{A, B, C, D\}$.

Množiny hrán majú nasledovný tvar:

$$E_1 = \begin{pmatrix} A & B \\ A & C \\ C & B \\ B & D \end{pmatrix} = \begin{pmatrix} B & A \\ C & A \\ B & C \\ D & B \end{pmatrix}, \qquad E_2 = \begin{pmatrix} B & A \\ B & C \\ C & A \\ C & D \\ D & C \end{pmatrix}, \tag{1}$$

pre lepšiu prehľadnosť sme tieto množiny zapísali ako maticu s dvomi stĺpcami. Do riadkov matice sme zapísali jednotlivé dvojice vrcholov, medzi ktorými existuje hrana. Matica $E_1$ prislúcha neorientovanej sieti. Má rôzne možnosti zapísania, nakoľko dvojice sú neusporiadané. Matica $E_2$ prislúcha orientovanej sieti a tým pádom, má len jednu možnosť zapísania. Pritom uvažujeme, že hrana vychádza z prvého uvedeného vrcholu v dvojici.

Ďalším spôsobom zadania siete je štvorcovou maticou susednosti $A$ (z angl. *adjacency matrix*) [27]. Jej rozmery sú $|V| \times |V|$ a skladá sa z hodnôt 0 a 1, nasledovne:

$$A(u,v) = \begin{cases} 1, & \text{ak } (u,v) \in E, \\ 0, & \text{inak .} \end{cases} \tag{2}$$

Ak by matica $A$ bola symetrická dostali by sme neorientovanú sieť. Zároveň, ak by na diagonále matice bola jednotka, daný vrchol by mal hranu sám so sebou. Túto možnosť v našej práci neuvažujeme a v ďalších kapitolách budeme pracovať už len s neorientovanými sieťami. Siete z obrázku 1 majú tvar matice susednosti:

$$A_1 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{3}$$

pričom matica susednosti $A_1$ prislúcha k neorientovanej sieti a $A_2$ k orientovanej sieti.

Jednou zo základných vlastností siete je rozdelenie stupňov vrcholov v sieti (z angl. *degree distribution*). Stupeň vrcholu $k$ predstavuje počet vrcholov, s ktorými má daný vrchol hranu. Definuje sa hodnota $d_k(y)$, ktorá nám hovorí o počte vrcholov v sieti $y$, ktoré majú stupeň $k$. Predelením hodnôt $d_i(y)$ počtom všetkých vrcholov $n$ dostaneme hľadané rozdelenie v tvare $\frac{d_0(y)}{n}, \ldots, \frac{d_k(y)}{n}$ [13]. V neorientovanej sieti z obrázku 1, označíme si ju $z$, má vrchol $A$ stupeň 2, lebo má hranu s vrcholmi $B$ a $C$. Zároveň tento stupeň má aj vrchol $B$, tým pádom hodnota $d_2(z)$ je rovná 2. Vrchol $C$ má ako jediný stupeň 3, a preto $d_3(z) = 1$. Stupeň 1 má v sieti iba vrchol D, a teda $d_1(z) = 1$.

Alternatíva k rozdeleniu stupňov sieti je sledovanie k-star rozdelenia (z angl. *k-star statistics*) [13]. Definuje sa k-star ako skupina $k$ hrán, ktoré majú spoločný vrchol. Treba poznamenať, že vrchol ktorý má hranu s $i$ vrcholmi, je stredom $\binom{n}{i}$ k-star. Definujú sa hodnoty rozdelenia $s_1(y), \ldots, s_{n-1}(y)$, pričom $s_k(y)$ predstavuje počet jednotlivých k-star v sieti $y$. Matematicky hodnoty definujeme:

$$s_k(y) = \sum_{i=k}^{n-1} \binom{i}{k} d_i(y), \qquad 2 \leq k \leq n-1, \tag{4}$$

$$s_1(y) = \frac{1}{2} \sum_{i=1}^{n-1} i d_i(y). \tag{5}$$

Hodnota $s_1(y)$, predstavujúca počet 1-star, je synonymum hrany, preto má aj zvlášť definovaný predpis rovnicou 5. V balíku *statnet* v $R$, ktorý využívame v našej práci, je počet 1-star definovaný bez násobenia $\frac{1}{2}$, a teda počet 1-star sa neskôr v práci bude rovnať dvojnásobnej hodnote počtu hrán v sieti. Na základe balíku *statnet* sa v sieti $z$ na obrázku 1 $(a)$ hodnota $s_1(z)$ rovná 8. Počet 2-star v sieti je 5, preto $s_2(z) = 5$. Najväčšiu k-star máme v sieti jednu 3-star a tú tvoria hrany, ktoré majú spoločný vrchol $C$, tým pádom $s_3(z) = 1$.

Ďalšou vlastnosťou siete, ktorú využijeme neskôr v práci, je rozdelenie počtu spoločných susedov (z angl. *the edgewise shared partner distribution*). Definujú sa hodnoty $ep_k(y)$ ako počet hrán, ktoré majú $k$ spoločných susedov v sieti $y$. Ak vrcholy, ktoré prislúchajú k sledovanej hrane, majú oba hranu s ďalším vrcholom, potom tento vrchol nazývame spoločným susedom. Hodnoty nám hovoria o počte trojuholníkov v našej sieti. Výsledné hodnoty rozdelenia dostaneme predelením počtom hrán, respektíve na základe [13], alternatívou k počtu hrán a to počtom 1-star, definovaným rovnicou (5). Rozdelenie má tvar: $\frac{ep_0(y)}{s_1(y)}, \frac{ep_1(y)}{s_1(y)}, \ldots, \frac{ep_{n-2}(y)}{s_1(y)}$. Príklady trojuholníkov pre hranu medzi vrcholmi $A$ a $B$ vidíme na obrázku 2. Pozrime sa na sieť $z$, uvedenú na obrázku 1 $(a)$ na začiatku kapitoly. Hodnoty rozdelenia pre túto sieť sú: $ep_0(z) = 1$, $ep_1(z) = 3$ a $ep_2(z) = 0$.




**(a)** jeden spoločný sused **(b)** dvaja spoloční susedia

**Obr. 2:** Príklady počtu spoločných susedov pre hranu medzi vrcholmi A a B.

Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi (z angl. *the geodesic distance distribution*) bude naša posledná sledovaná vlastnosť siete. Vzdialenosť medzi dvomi vrcholmi definujeme ako dĺžku najkratšej cesty medzi vrcholmi po existujúcich hranách. Napríklad v sieti na obrázku 1 $(a)$, dĺžka najkratšej cesty medzi vrcholmi $A$ a $D$ je 2. Takéto cesty dĺžky 2 máme v sieti len dve. Ostatné cesty majú dĺžku 1.

# 2 ERG modely

V tejto kapitole sa zameriame na modely, ktoré využívajú siete, ako sme ich definovali v predchádzajúcej kapitole. Najštandardnejší model, ktorý pracuje s náhodnými sieťami je Bernoulliho model, respektíve nazývaný aj Erdős-Rényiho model po jeho autoroch. Tento model definovali v roku 1959. Model sleduje konečný počet vrcholov v sieti. Každá hrana medzi vrcholmi vzniká s pravdepodobnosťou $p$, pričom jej vznik je nezávislý od vzniku ostatných hrán v sieti [27]. V ERG modeloch sa uvažuje, že hrana môže závisieť aj od existencie iných hrán v sieti a zároveň aj od charakteristík jednotlivých vrcholov. V roku 1986 uviedli autori Frank a Strauss Markovovu závislosť. V štúdii uviedli podmienenú závislosť a to takú, že dve hrany môžu byť závislé, ak zdieľajú vrchol [3]. V tejto kapitole si predstavíme základný tvar ERG modelov a priblížime závislé vysvetľujúce premenné, ktoré budú vstupovať do modelu. Zároveň, aby sme neskôr pri odhadovaní modelov na konkrétnych sieťach vedeli modely porovnať, uvedieme informačné kritériá, ktoré budeme využívať.

## 2.1 Definícia modelov

ERG modely (z angl. *exponential random graph models*) patria do skupiny modelov, ktoré sa snažia identifikovať pravdepodobnosti jednotlivých hrán, ako aj celej siete. Uvádzame základný tvare ERG modelu z článku [11]. Označme si za $Y$ náhodnú $n \times n$ maticu, obsahujúcu hodnoty $\{1;0\}$, ktorá bude predstavovať našu maticu susednosti. Matica $Y$ nám bude definovať sieť s $n$ vrcholmi. Následne množinu všetkých možných $n \times n$ matíc označme $\Upsilon$. Nakoľko v tejto práci uvažujeme neorientované siete, matice budú symetrické. Taktiež neuvažujeme, že vrchol môže mať hranu sám so sebou, preto na diagonále týchto matíc budú nuly. Do model budú vstupovať vysvetľujúce premenné, ktoré priblížime v nasledujúcej podkapitole.

Vytvoríme pravdepodobnostný model pre danú realizáciu $y$ náhodnej premennej $Y$. Označme si za $g(y)$ vektor vysvetľujúcich premenných pre danú realizáciu náhodnej premennej $Y$. Parametre modelu, ktoré budú prislúchať k jednotlivým vysvetľujúcim premenným $g_i(y)$ si označíme $\theta_i$, respektíve ako vektor $\theta \in \mathbb{R}$.

Model je definovaný v tvare:

$$P_{\theta,\Upsilon}(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta, \Upsilon)}. \tag{6}$$

Člen v menovateli $\kappa(\theta, \Upsilon)$ predstavuje normalizačnú konštantu, ktorá je definovaná nasledovne:

$$\kappa(\theta, \Upsilon) = \sum_{z \in \Upsilon} \exp\{\theta^\top g(z)\}. \tag{7}$$

Tento člen zabezpečuje, že rovnica (6) korektne definuje pravdepodobnosť. Predstavuje hodnotu čitateľa pre všetky možné siete z množiny $\Upsilon$.

Na základe rovnice (6) vieme vypočítať pravdepodobnosť konkrétnej siete. Avšak niekedy nás zaujíma aj pravdepodobnosť hrany medzi dvomi vrcholmi. Definuje sa preto vektor zmeny vysvetľujúcich premenných $\delta_g(y)_{ij}$. Vektor je funkciou vysvetľujúcich premenných, konkrétnej realizácie $y$ náhodnej premennej $Y$ a konkrétnej dvojice vrcholov $(i, j)$. Má tvar:

$$\delta_g(y)_{ij} = g(y^+_{ij}) - g(y^-_{ij}), \tag{8}$$

pričom $y^+_{ij}$, $y^-_{ij}$ reprezentujú sieť s hodnotou $y_{ij} = 1$, respektíve $y_{ij} = 0$ v matici $y$. Pričom zvyšok matice zostane rovnaký. Inak povedané, $\delta_g(y)_{ij}$ nám hovorí, ako sa zmenili štatistiky $g(y)$ vzniknutím hrany medzi vrcholmi $i$ a $j$.

Pripomeňme si ako je definovaná *logit*-funkcia pre nejakú pravdepodobnosť $p$:

$$logit[p] = log(odds(p)) = log\left(\frac{p}{1-p}\right). \tag{9}$$

Ak použijeme pravdepodobnosť definovanú rovnicou (6), pričom sa pozrieme práve na efekt, ak $Y_{ij} = 1$. Pritom ostatné hodnoty zostanú nezmenené, čo nám definuje podmienená časť pravdepodobnosti $(Y^c_{ij} = y^c_{ij})$. Pár operáciami dostaneme z rovnice (9) tvar:

$$logit[p_{\theta,\Upsilon}(Y_{ij} = 1 | Y^c_{ij} = y^c_{ij})] = \theta^\top \delta_g(y)_{ij}. \tag{10}$$

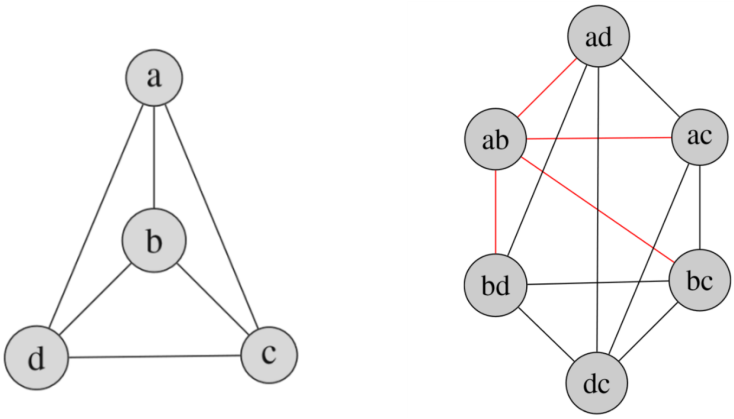
Ak by sme uvažovali, že do modelu budú vstupovať aj vysvetľujúce premenné, ktoré prislúchajú k jednotlivým vrcholom, tak by sme dostali tvar vektora vysvetľujúcich premenných $g(y, X)$. Práve premenná $X$ by predstavovala pridané informácie o vrcholoch. Taktiež aj zmenu vysvetľujúcich premenných by sme dostali závislú od $X$ a teda v tvare $\delta_g(y, X)_{ij}$.

## 2.2 Vysvetľujúce premenné

V prechádzajúcej podkapitole sme uviedli, ako je definovaný ERG model v základnom tvare. Do modelu vstupovali vysvetľujúce premenné, ktoré sa môžu vzťahovať k vrcholom alebo môžu zahŕňať závislosť medzi hranami navzájom. Takéto premenné nazývame aj štrukturálne premenné. Sledovanie rôznych vlastností vie zlepšiť odhadnutý model. Pokým sledovanie vlastností vrcholov predstavuje sledovanie nezávislých premenných, štrukturálne premenné patria medzi závislé premenné. Model len s nezávislými premennými sa odhaduje pomocou likelihood funkcie, a tým pádom opakovaním odhadov dostaneme rovnaké hodnoty. Na druhú stranu model, ktorý zahŕňa aj závislé premenné, sa odhaduje Markov chain Monte Carlo maximum likelihood odhadom, ktorý predstavuje stochastickú aproximáciu pre likelihood funkciu, tým pádom sa odhadnuté parametre môžu v malej miere líšiť [11, 13]. V tejto časti priblížime, ako sledovať závislosti medzi hranami a aj ich alternatívy pre degenerované modely.

Jednou z najznámejších hypotéz zaoberajúcich sa závislosťami hrán navzájom je Markovova hypotéza. Pri sledovaní existencie hrán sa autor článku [29] zamyslel napríklad aj nad scenárom, že ak dve firmy majú viacero spoločných partnerov, majú väčšiu tendenciu spolupracovať. Môže to byť dôsledkom na jednu stranu tým, že ak sa obe firmy zaujímajú o rovnaké veci, a teda pracujú na podobných veciach, majú aj spoločné záujmy. Na druhú stranu, ak firma hľadá spoluprácu, často sa pozerá na hodnotenia a určite viac dôveruje hodnoteniam od vlastných partnerov. Na ukážku predpokladajme, že máme 4 rôzne firmy. Ak sa pozrieme na nasledujúci obrázok 3, každý vrchol predstavuje jednu firmu a hrana ich spoluprácu. Zameriame sa na hranu medzi firmami $a$ a $b$. Hypotéza hovorí, že existencia tejto hrany závisí aj od hrán s firmami, s ktorými majú firmy $a$ a $b$ hranu. V našom prípade uvažujeme hrany s firmami $c$ a $d$. Jednotlivé hrany a ich možné závislosti sú zobrazené v grafe závislostí v časti $(b)$, pričom červenou farbou sú označené hrany, od ktorých môže závisieť hrana medzi firmami $a$ a $b$.

Túto hypotézu zahrnuli Frank a Strauss v Markovovom modeli v roku 1986 [3]. Do modelu zahrnuli sledovanie troch parametrov. Prvým parametrom je počet hrán v sieti. Následne zahrnuli parameter sledujúci 2-stars. Ak sa pozrieme na obrázok 3 na časť $(b)$, hrana medzi $ab$ a $ad$ predstavuje 2-star pre firmu $a$. Sledovanie jednotlivých 2-stars by priviedlo veľa parametrov do modelu. Predpokladá sa, že výskyt tejto vlastnosti má rov-




**(a)** Markovov graf **(b)** Graf závislostí

**Obr. 3:** Markovova hypotéza závislosti (prebrané z článku [29]).

nakú pravdepodobnosť pre všetky dvojice vrcholov, a preto sa sleduje celkový počet 2-star v sieti. Podobne aj pre 3-stars, 4-stars a vyššie k-stars. Tretí parameter v Markovovom modeli zahŕňa sledovanie trojuholníkov medzi firmami. Pre ilustráciu, ak sa pozrieme na firmy $b$, $c$, $d$, tak trojuholník dostaneme medzi vrcholmi $bc$, $bd$ a $cd$.

Model sa dá následne zapísať v tvare:

$$p(Y = y) = \frac{1}{\psi(\theta)} \exp(\theta_E \cdot v_E(y) + \theta_{S_2} \cdot v_{S_2} + \ldots \theta_{S_{n-1}} \cdot v_{S_{n-1}} + \theta_\Delta \cdot v_\Delta), \tag{11}$$

pričom $v_E(y)$ hovorí o počte hrán, $v_{S_i}$ o počte jednotlivých i-star a parameter $v_\Delta$ o počte trojuholníkov v sieti. Člen $\psi(\theta)$ predstavuje normalizačnú konštantu, podobne ako v základnom ERG modeli, definovanú:

$$\psi(\theta) = \sum_{y \in \Upsilon} \exp(\theta_E \cdot v_E(y) + \theta_{S_2} \cdot v_{S_2} + \ldots \theta_{S_{n-1}} \cdot v_{S_{n-1}} + \theta_\Delta \cdot v_\Delta). \tag{12}$$

Zistilo sa, že modely v takomto tvare sú náchylné na degeneráciu [29]. V článku [7] uvádzajú príklad degenerácie modelu s parametrom počtu hrán a počtu 2-star. Zároveň ukazujú, pre aké hodnoty parametrov v sieti so siedmimi vrcholmi nebude model degenerovaný. Jedna z mnohých degenerácií môže nastať dôsledkom toho, že ak pridáme jednu hranu do grafu, môžu sa zvýšiť aj hodnoty iných parametrov. Napríklad pridaním hrany sa môže zvýšiť počet 2-stars alebo sa môžu zmeniť na 3-stars. Zahrnutie všetkých možných podmnožín siete by mohlo model veľmi skomplikovať. Ukázalo sa, že takto zvolený model kladie väčšiu pravdepodobnosť na siete s vyššími stupňami. Na odstránenie týchto problémov sa pri odhadovaní modelov zaviedli nové parametre založené na čiastočnej

podmienenej závislosti [23]. Idea je založená na pridaní váhy k rozdeleniu stupňov v sieti. Nižším stupňom sa váha zvýši, a naopak s rastúcim stupňom sa bude váha zmenšovať. Túto myšlienku môžeme zapísať v tvare [29]:

$$u_{\tau}^{(d)} = \sum_{k=0}^{n-1} e^{-\tau k} d_k(y), \tag{13}$$

kde $d_k(y)$ predstavuje počet vrcholov stupňa $k$ a parameter $\tau$ predstavuje danú váhu. Túto premennú nazývame geometricky vážené rozdelenie stupňov (z angl. *geometrically weighted degree distribution*). Ak sa zamyslíme, rozdelenie stupňov vrcholov je alternatívne ku k-stars rozdeleniu, tým pádom vieme prepísať rovnicu (13) pomocou k-star na tvar [29]:

$$u_{\lambda}^{(S)} = S_2 - \frac{S_3}{\lambda} + \frac{S_4}{\lambda^2} - \ldots + (-1)^{n-2}\frac{S_{n-1}}{\lambda^{n-3}} = \sum_{k=2}^{n-1}(-1)^k \frac{S_k}{\lambda^{k-2}}, \tag{14}$$

pričom $S_k$ predstavuje počet k-stars a $\lambda$ je parameter. Túto metódu nazývame striedajúce sa k-stars (z angl. *alternating k-stars*). Rozdielom medzi týmito dvomi metódami je v meniacom sa znamienku.

Pre snahu zlepšiť sklon k degenerácii a zmenšiť počet parametrov sa zaviedla nová premenná aj pre rozdelenie počtu spoločných susedov. Nazýva sa geometricky vážené rozdelenie počtu spoločných susedov (z angl. *geometrically weighted edgewise shared partner*) a bola odvodená v [14]. Definovaná je v tvare:

$$ep^G(y, \tau) = e^{\tau} \sum_{i=1}^{n-2} \{1 - (1 - e^{-\tau})^i\} ep_i(y), \tag{15}$$

kde $\tau$ je pridaný parameter. V článku [13] sa autori pozreli na hodnoty $\tau$ z množiny $\{0{,}5; 1\}$. V prípade, že by sme nezadali parameter $\tau$ a nechali parameter modelom taktiež odhadnúť dostali by sme CERG modely (z angl. *curved exponentional random graph models*). V tejto práci sa pri aplikáciách modelov na siete pozrieme bližšie aj na tento parameter.

## 2.3 Informačné kritériá

Jednou z možností ako medzi sebou porovnávať modely, a tým vyradiť premenné, ktoré nám komplikujú model, pričom neprinášajú jeho výrazné zlepšenie, sú informačné kritériá. Najvyužívanejšie sú Akaikeho a Bayesovo informačné kritérium, ktoré si rozoberieme v tejto časti podľa [13].

Obe informačné kritériá sú definované ako funkcia od parametra $M$, ktorý predstavuje sledovaný model. Akaikeho informačné kritérium je definované v tvare:

$$AIC(M) = -2(\text{maximized loglikelihood under M}) + 2(\#\text{ of parameters in M}). \quad (16)$$

Cieľom je túto funkciu minimalizovať, avšak pre ERG modely hodnota likelihood funkcie nie je vždy vypočítateľná, a preto sa využíva jej aproximácia. Konkrétnejšie a detailnejšie informácie sa nachádzajú v [1, 13].

Podobne je definované Bayseovo informačné kritérium:

$$BIC(M) = -2(\text{maximized loglikelihood under M}) + logN(\#\text{ of parameters in M}), \quad (17)$$

pričom $N$ predstavuje rozsah výberu. Porovnanie týchto kritérií a aj matematické pozadie sa nachádza v článku [17]. V práci budeme neskôr tieto kritériá označovať AIC a BIC.

# 3 Faktorová analýza

V niektorých prípadoch, ak sledujeme vplyv vlastností vrcholov na hrany v sieti, dostaneme veľa parametrov, ktoré by sme mohli zahrnúť do modelu. Môže byť preto vhodné pokúsiť sa počet zmenšiť a jednotlivé možnosti zoskupiť. Tento prípad uvidíme neskôr v kapitolách pri sledovaní hudobných štýloch alebo športov. Jedným z riešení tohto problému je faktorová analýza, ktorá na základe korelácie medzi jednotlivými možnosťami vytvorí faktory.

Uvažujme, že náhodný vektor premenných $X = (X_1, \ldots, X_p)^T$ vie byť reprezentovaný $m$-rozmerným náhodným vektorom $F = (F_1, \ldots, F_m)^T$, kde $m$ je menšie číslo než $p$. Taktiež uvažujme, že vektory sú v nasledovnom vzťahu:

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + U_i, \qquad i = 1, \ldots p, \tag{18}$$

$$X = AF + U, \tag{19}$$

kde $A$ nazývame faktorové náklady (z angl. *factor loadings*) a $U$ špecifické faktory (z angl. *specific factors*) [9]. Neznámou hodnotou je matica $A$, ktorú budeme odhadovať. Na základe teoretických predpokladov a teórie hlavných komponentov, definovaných v [9], vieme zapísať korelačnú maticu vektora $X$ v tvare:

$$\Psi = AA^T + D, \tag{20}$$

kde hodnoty matice $A$ sú hodnoty korelácie medzi $X$ a $F$. Matica $D$ predstavuje odchýlky od konkrétnych korelácií, ktoré vzniknú po odhadnutí matice $A$. Chceme, aby tieto hodnoty boli minimálne. Jednou zo základných myšlienok faktorovej analýzy je, že modely $X = AF + U$ a $X = (AV)(V^T F) + U$, kde $V$ je $m \times m$ ortogonálna matica, sa nedajú rozlíšiť, nakoľko model v rovnici (19) je preparametrizovaný [9].

Dve najznámejšie metódy na odhadnutie matice faktorových nákladov $A$ sú metóda maximálnej vierohodnosti a metóda hlavných faktorov (z angl. *principal factors*). Priblížime si metódu hlavných faktorov [9]. Ak by sme poznali maticu $D = diag(d_1, \ldots, d_p)$ z rovnice (20), pričom $d_i = 1 - \sum_{j=1}^{m} a_{ij}^2$, vedeli by sme vypočítať redukovanú korelačnú maticu tvaru $\Psi - D = AA^T$. Túto maticu však nepoznáme, preto metóda hlavných faktorov využíva na odhad matice $\Psi$ odhadnutú korelačnú maticu $R$ na základe dát. Prvky matice

$D$ odhadneme na základe matice $R$ v nasledovnom tvare:

$$\hat{d}_i = \frac{1}{(R^{-1})_{ii}}. \tag{21}$$

Následne dostaneme odhad redukovanej korelačnej matice:

$$R^* = R - diag(\hat{d}_1, \ldots, \hat{d}_p). \tag{22}$$

Vypočítame vlastné hodnoty $\lambda_1, \ldots, \lambda_p$ a ortogonálne vlastné vektory $u_1, \ldots, u_p$. Zoberieme prvých $m$ nezáporných vlastných hodnôt, k nim prislúchajúce vlastné vektory a vytvoríme odhad matice $A$, v tvare:

$$\hat{A} = (\sqrt{\lambda_1}u_1, \ldots, \sqrt{\lambda_m}u_m). \tag{23}$$

Tento spôsob odhadu vieme opakovať s novým odhadom hodnôt vektora $D$, pričom $d_i = 1 - \sum_{j=1}^{m}(\hat{A})_{ij}^2, i = 1 \ldots p$, dokým nedokonvergujeme k rovnakému riešeniu. V literatúre [9] sa uvádza príklad pravidla na správne zvolenie počtu faktorov v metóde hlavných faktorov, pravidlo pre vlastné hodnoty v tvare:

$$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{i=1}^{p}|\lambda_i|} > 0{,}8. \tag{24}$$

Odhadnutú maticu $A$ vieme transformovať pomocou ortogonálnej matice $V$, ktorá nám vie uľahčiť interpretáciu vytvorených faktorov. Táto transformácia sa nazýva aj rotácia. V našej práci využijeme varimaxovú metódu na odhadnutie matice rotácie $V$. Metóda je definovaná pomocou funkcie od matice $V$, v tvare [9]:

$$f_v(V) = \frac{1}{m}\sum_{t=1}^{p}[\frac{1}{p}\sum_{j=1}^{p}((\hat{A}V)_{jt}^2 - \frac{1}{p}\sum_{k=1}^{p}(\hat{A}V)_{kt}^2)^2]. \tag{25}$$

Výslednú maticu $V$ dostaneme maximalizovaním funkcie $f_v(V)$.

# 4 Aplikácia modelov na dáta o žiakoch zo strednej školy v Glasgowe

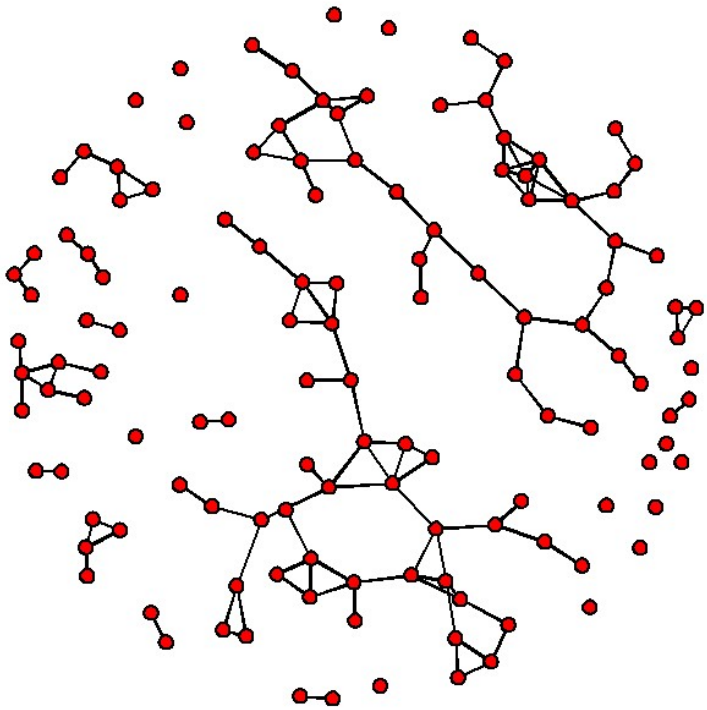
V tejto kapitole aplikujeme ERG modely na konkrétne dáta. Dáta pochádzajú od žiakov zo strednej školy v Glasgowe [22]. Žiakov sledovali niekoľko rokov a snažili sa odhaliť procesy, ktoré ovplyvňovali postoj žiakov k fajčeniu. V prvej podkapitole bližšie predstavíme pôvod dát a analýzy, ktoré boli s nimi už vytvorené. Zároveň uvedieme spôsob, akým sme si upravili tieto dáta. Následne v podkapitolách opíšeme základný Bernoulliho model a nami vytvorené modely, sledujúce vlastnosti žiakov. Ku koncu kapitoly tieto modely medzi sebou porovnáme. Z nich vyberieme podľa informačných kritérií najlepší model a ten otestujeme, na základe troch charakteristík siete. Na záver kapitoly pridáme k vybranému modelu charakteristiky sietí. Pozrieme sa na ich rôzne alternatívy a taktiež vytvoríme finálny model, na ktorom budeme prezentovať odhadnuté parametre a vplyv homofílie.

## 4.1 Dáta

V roku 1995 sa autori Patrick West a Lynn Michell štúdie *Peer pressure to smoke: the meaning depends on the method* z *University Glasgowe* rozhodli sledovať žiakov po dobu 2 rokov [21]. Ako sme už naznačili, snažili sa identifikovať vplyvy, ktoré ovplyvňujú postoj žiakov k fajčeniu v dospievaní. Prvý krát zozbierali dáta v februári 1995. Žiaci mali vtedy 13 rokov. Sledovanie pokračovalo ešte dvomi pozorovaniami, pričom posledné pozorovanie sa uskutočnilo v januári 1997 [22]. Každého sledovania sa zúčastnilo 160 žiakov, pričom 129 žiakov sa zúčastnilo všetkých pozorovaní. Úlohou žiakov bolo vymenovať 6 najbližších kamarátov. Žiak mal taktiež odpovedať na otázky ohľadne svojho života, hudobných preferencií, voľnočasových aktivít a užívaní návykových látok. Pre komplexnosť a rôzne zahrnuté informácie sa tieto dáta už využili na rôzne pozorovania. V článku *Girls, pecking order and smoking* v časopise *Social Science and Medicine* v roku 1997 sa autori L. Michell a A. Amos zamerali na rozdielnosť pravidelnosti fajčenia podľa pohlavia. Zároveň sa pozreli na súvislosť medzi fajčením žien a ich spoločenským postavením v skupine [20]. Dáta využili aj v roku 2010, autori Ch. Steglich, T.A.B. Snijders a M. Pearson v článku *Dynamic Networks and Behavior: Separating Selection from Influence* v časopise *Sociological Methodology* [24]. Prezentujú vyvinutú skupinu modelov, ktoré sa snažia od-

deliť efekt výberu priateľov danej osoby od účinku sociálneho vplyvu na tento výber. Na základe dát sledujú dôsledok výberu kamarátstiev a vzájomného ovplyvňovania na dynamiku priateľstiev a vplyv užívania návykových látok v sledovaných obdobiach. Kompletný opis dát a aj štúdie, ktoré ich využívali sa nachádzajú v [22].

V našej práci využívame prvé sledovanie a iba žiakov, ktorí sa zúčastnili vo všetkých pozorovaniach. Žiaci označovali svojich najlepších kamarátov, a tým pádom vznikla orientovaná sieť. Na účely nášho modelovania sme si zobrali len tie kamarátstva, kedy sa žiaci označili za kamarátov navzájom. Dostali sme tak neorientovanú sieť zobrazenú na nasledujúcom obrázku. Môžeme si všimnúť, že takouto úpravou, sme dostali malý počet hrán oproti počtu vrcholov.



**Obr. 4:** Vytvorená sieť na základe dát z Glasgowu.

Ku každému žiakovi sú v sade dát zahrnuté rôzne informácie. V našej práci využijeme nasledovné:

- pohlavie,
- vek,
- intenzita pitia alkoholu,
- intenzita fajčenia tabaku,
- intenzita fajčenia marihuany,
- stav,
- fajčenie v mieste bydliska,
- fajčenie rodičov,
- fajčenie súrodencov,
- hudobné štýly, ktoré žiak počúva,
- účasť na rôznych voľnočasových aktivitách.

Konkrétne informácie priblížime, keď ich budeme zahŕňať do modelov.

## 4.2 Modely s charakteristikami vrcholov

Ako prvé sa pozrieme na základný model, ktorý sa nazýva Bernoulliho model, alebo v niektorých literatúrach nazývaný Null model. Sleduje veľkosť siete, respektíve počet všetkých hrán. Následne do modelu postupne zahrnieme základné osobné údaje a pozrieme sa, ako sa model zlepšuje na základe informačných kritérií. V ďalšej časti sa pozrieme na koreláciu medzi hudobnými štýlmi, nakoľko máme v dátach veľký počet hudobných štýlov. Aby sme zmenšili počet parametrov a nevkladali jednotlivo každý hudobný štýl, vytvoríme faktory, ktoré budú reprezentovať skupiny štýlov na základe ich korelácie medzi sebou. Na základe článkov, ktoré sledovali a rôznymi spôsobmi testovali vplyv fajčenia, sme sa rozhodli taktiež vytvoriť modely, ktoré budú sledovať homofíliu vzhľadom na fajčenie a to nie len danej osoby, ale aj vplyv fajčenia v mieste bydliska alebo fajčenie rodinných

príslušníkov. Zo života vieme, že ľudia sa môžu zoznámiť aj pri rovnakých záľubách a voľnočasových aktivitách. Pri pravidelnom stretávaní sa osoby môžu skamarátiť, preto zahrnieme do modelov aj tieto informácie a pozrieme sa, či efekt homofílie vzhľadom na voľnočasové aktivity nám zlepší model. Po vytvorení modelov s jednotlivými parametrami sa pozrieme sa model, v ktorom tieto parametre spojíme. Vytvoríme tým pádom jeden model, s ktorým budeme v ďalších podkapitolách pracovať.

### 4.2.1 Bernoulliho model

Bernoulliho model, nazývaný aj Null alebo Erdős-Rényi model, sleduje ako vysvetľujúcu premennú počet hrán. V našom prípade máme v sieti 124 hrán. Tento model sa dá použiť aj na sledovanie, ako sa zložitejšie modely zlepšujú. Našou úlohou je odhadnúť parameter $\theta$ z rovnice (10):

$$logit(p(Y_{ij} = 1|Y^c_{ij} = y^c_{ij}) = \log(odds(Y_{ij} = 1)) = \theta \cdot \delta(y)_{ij}, \tag{26}$$

ktorú sme uviedli pri základnom tvare ERG modelu. Pre lepšiu prehľadnosť podmienenú časť pravdepodobnosti nebudeme ďalej v texte písať. Ako sme už spomenuli, $\delta(y)_{ij}$ hovorí o zmene vysvetľujúcich premenných na základe existencie hrany medzi vrcholmi $i$ a $j$. V prípade Bernoulliho modelu sa $\delta(y)_{ij}$ rovná 1, keďže naša vysvetľujúca premenná je iba počet hrán v sieti [28]. 0dhad parametra $\theta$ v Bernulliho modeli sa niekedy označuje aj ako intercept [18]. Pomocou programu *R* a knižnice *statnet* sme dostali výstup:

```
==========================
Summary of model fit
==========================

Formula:   net.sym ~ edges

Iterations:  7 out of 20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -4.18328    0.09048      0  -46.23   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 11445  on 8256  degrees of freedom
 Residual Deviance:  1287  on 8255  degrees of freedom

AIC: 1289    BIC: 1296    (Smaller is better.)
```

Hodnotu hľadaného parametra $\theta$ sme dostali $-4{,}18$. Prepísaním rovnice (26) a vyjadrením pravdepodobnosti dostaneme:

$$\begin{aligned} \log(odds(y_{ij}=1)) &= \theta \cdot \delta(y)_{ij} = \theta, && (27)\\ odds(y_{ij}=1) &= \frac{p(y_{ij}=1)}{1-p(y_{ij}=1)} = \exp(\theta), && (28)\\ p(y_{ij}=1) &= \frac{\exp(\theta)}{1+\exp(\theta)}. && (29) \end{aligned}$$

Dosadením parametra dostaneme:

$$p(y_{ij}=1) = \frac{\exp(-4{,}18)}{1+\exp(-4{,}18)} = 0{,}015. \quad (30)$$

Pravdepodobnosť ľubovoľnej hrany v sieti je 0,015. Výsledná hodnota je veľmi malá a to dôsledkom, ako sme už spomenuli, malého počtu hrán ku počtu vrcholov.

### 4.2.2 Modely sledujúce základné osobné údaje

O jednotlivých žiakoch máme rôzne informácie, pričom za základné osobné údaje budeme v tejto podkapitole považovať pohlavie, vek a stav. S každým osobným údajom vytvoríme samostatný model a pozrieme sa, ako sa modely zlepšili oproti Null modelu. Údaje si zoradíme podľa toho, ako sa modely zlepšili a následne budeme z nich vytvárať spoločný model. Posnažíme sa vytvoriť model, ktorý nebude veľmi zložitý, ale zároveň bude čo najlepšie opisovať dáta. Na rozhodovanie budeme využívať Akaikeho a Bayesovo informačné kritériá, ktoré sme definovali v kapitole 2.

Prvou informáciou, ktorú sledujeme v tejto časti, je pohlavie žiakov. Vytvoríme model, sledujúci kamarátstva medzi rovnakým pohlavím, pričom nerozlíšime kamarátstvo medzi dvomi chlapcami a dvomi dievčatami. V dátach máme 112 hrán medzi rovnaký pohlavím z celkových 124 hrán. Po odhadnutí modelu, ktorého výstup sa nachádza v prílohe A.1, sme dostali vektor odhadnutých parametrov v tvare:

$$\theta = \begin{pmatrix} -5{,}82 \\ 2{,}24 \end{pmatrix}. \quad (31)$$

Označme prvý koeficient vo vektore $\theta$ ako $\theta_1 = -5{,}82$ a druhý koeficient $\theta_2 = 2{,}24$. Takéto označovanie budeme využívať v celej práci pri opisovaní odhadnutých hodnôt parametrov. Parameter $\theta_1$ zodpovedá premennej sledujúcej počet hrán a druhý parameter $\theta_2$ práve

vplyvu homofílie vzhľadom na pohlavie. Na interpretáciu koeficientov sa môžeme pozrieť dvomi spôsobmi. Prvým spôsobom je pozrieť sa na jednotlivé možnosti, ktoré sa môžu v sieti vyskytnúť a vypočítať ich šance, respektíve pravdepodobnosti. Jednou z možností, ktorá sa môže vyskytnúť v sieti je hrana medzi osobami s rovnakým pohlavím. Na základe rovnice (10) vieme vypočítať logaritmus šancí a pravdepodobnosť hrany nasledovne:

$$\begin{aligned} \log(odds(y_{ij}=1)) &= -5{,}82 \cdot 1 + 2{,}24 \cdot 1 = -3{,}58, && (32)\\ odds(y_{ij}=1) &= \exp(-3{,}58) = 0{,}03, && (33)\\ p(y_{ij}=1) &= \frac{\exp(-3{,}58)}{1+\exp(-3{,}58)} = 0{,}027, && (34) \end{aligned}$$

pričom hodnoty $\delta_1(y)_{ij}$ aj $\delta_2(y)_{ij}$ sa rovnajú 1, nakoľko uvažujeme, že sa oba sledované parametre, počet hrán a počet hrán medzi rovnakým pohlavím, zväčšia o 1.

Obdobne vieme vypočítať pravdepodobnosť hrany medzi osobami s rozdielnym pohlavím:

$$\begin{aligned} \log(odds(y_{ij}=1)) &= -5{,}82 \cdot 1 + 2{,}24 \cdot 0 = -5{,}82, && (35)\\ odds(y_{ij}=1) &= \exp(-5{,}82) = 0{,}003, && (36)\\ p(y_{ij}=1) &= \frac{\exp(-5{,}82)}{1+\exp(-5{,}82)} = 0{,}0029. && (37) \end{aligned}$$

Druhým spôsobom, akým sa môžeme pozrieť na odhadnuté koeficienty a ich vplyv, je uvedený v [16]. V tomto prípade sa pozrieme ako parametre ovplyvňujú šance jednotlivo. Pritom sa predpokladá, že ostatné parametre zostanú zachované. Môžeme si situáciu predstaviť ako porovnanie dvoch dvojíc vrcholov, pričom budeme uvažovať, že tieto dvojice sa líšia iba v interpretovanom parametri. Takýto prístup k interpretácii budeme využívať aj neskôr v práci, keď už budeme mať zložitejšie modely s viacerými vysvetľujúcimi premennými. Pozrime sa na zmenu šancí:

$$\Delta odds(y_{ij}=1) = \exp(\theta_2) = \exp(2{,}24) = 9{,}39. \quad (38)$$

Dostali sme, že šance hrany medzi osobami s rovnakým pohlavím sa zvýšia faktorom 9,4 oproti hrane medzi osobami s rozdielnym pohlavím. Výsledok nám dáva informáciu, že existuje veľký vplyv homofílie vzhľadom na pohlavie v sieti.

Poďme sa ale pozrieť, či existuje rozdiel medzi kamarátstvami medzi dvomi chlapcami a dvomi dievčatami. Ak budeme takto rozlišovať pohlavie, v modeli vzniknú dve premenné namiesto pôvodnej jednej, ktorá sledovala pohlavie. V našej sieti máme 66 kamarátstiev medzi dvomi chlapcami a 46 kamarátstiev medzi dvomi dievčatami. Po odhadnutí modelu sme dostali vektor parametrov v tvare:

$$\theta = \begin{pmatrix} -5{,}82 \\ 2{,}17 \\ 2{,}35 \end{pmatrix}, \tag{39}$$

pričom $\theta_2 = 2{,}17$ sleduje kamarátstva medzi dvomi chlapcami a $\theta_3 = 2{,}35$ medzi dvomi dievčatami. Celý výstup s odhadnutými parametrami sa nachádza v prílohe A.1. Už na prvý pohľad vidíme, že odhadnuté parametre sa veľmi nelíšia, tým pádom ani výsledný vplyv na šance a pravdepodobnosti nebude veľmi rozdielny. Dostávame:

$$\Delta_1 odds(y_{ij} = 1) = \exp(\theta_2) = \exp(2{,}17) = 8{,}76, \tag{40}$$

$$\Delta_2 odds(y_{ij} = 1) = \exp(\theta_3) = \exp(2{,}35) = 10{,}49. \tag{41}$$

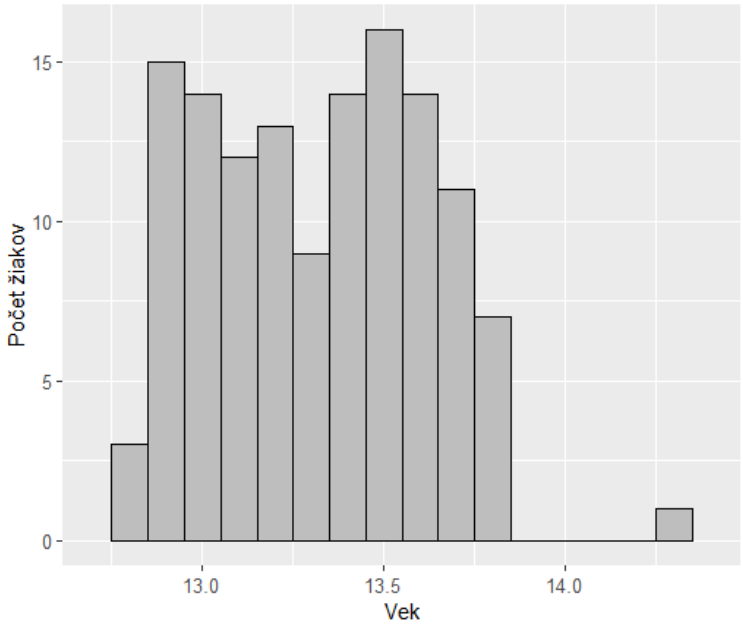
Vplyvu kamarátstva medzi dvomi chlapcami na šance prislúcha $\Delta_1$ a dvomi dievčatami $\Delta_2$. Vidíme, že zmeny šancí sú stále veľké ale nie je v nich veľký rozdiel.

Aby sme sa vedeli rozhodnúť, ktorý model je lepší, pozrime sa na informačné kritériá, ktoré sú zahrnuté vo výstupe v *R*. V tabuľke 1 vidíme, že rozlišovaním medzi pohlaviami sa model zhoršil, tým pádom budeme neskôr uvažovať iba prvý spomenutý model, ktorý nerozlišuje medzi pohlaviami.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model nerozlišujúci medzi pohlaviami | 1199 | 1213 |
| Model rozlišujúci medzi pohlaviami | 1200 | 1221 |

**Tabuľka 1:** Porovnanie AIC a BIC pre modely sledujúce pohlavie.

Druhou informáciou, ktorú sledujeme, je vek. Žiaci pochádzajú z jedného ročníka. Mohli by sme predpokladať, že ich vek nebude veľmi rôznorodý, avšak v dátach máme zadaný vek s presnosťou na jedno desatinné miesto. Najmladší žiak mal 12,8 rokov a naopak najstarší mal 14,3 rokov. Pre lepšiu predstavu na obrázku 5 uvádzame histogram počtu žiakov vzhľadom na vek. Vidíme, že iba jeden žiak je výraznejšie starší a žiaden konkrétny vek neprevláda. V sieti máme iba 11 kamarátstiev medzi žiakmi s rovnakým vekom.




**Obr. 5:** Histogram počtu žiakov v závislosti od veku.

Môžeme očakávať, že táto premenná nebude ovplyvňovať kamarátstva. Už aj z bežného života vieme, že malé vekové rozdiely, ako napríklad mesiac, neovplyvnia, či sa budeme s danou osobou kamarátiť. V prípade väčších vekových rozdielov, ako môže byť 5, 10 a viac rokov, by mohol byť tento efekt výraznejší. Z odhadnutého modelu sme dostali hodnoty parametrov:

$$\theta = \begin{pmatrix} -4{,}18 \\ -0{,}04 \end{pmatrix}. \tag{42}$$

Hodnota parametra $\theta_2$, sledujúca efekt existencie hrany medzi osobami s rovnakým vekom, je malá. Táto premenná by výsledné šance ovplyvňovala malým faktorom a to konkrétne faktorom $\exp(-0{,}04) = 0{,}96$. Poďme sa pozrieť na Akaikeho a Bayesovo informačné kritéria. Model vypočítal hodnotu Akaikeho kritéria na 1291 a Bayesovho kritéria na 1305.

Zo skúseností z predchádzajúcich modelov vieme, že je to veľké číslo oproti hodnotám kritérií uvedených doteraz.

Treťou informáciou a zároveň posledným základným osobným údajom, na ktorý sa zameriame, je stav. Žiak môže byť zadaný alebo slobodný. Najprv sa pozrieme na model, ktorý nebude rozlišovať, či sa kamarátia dvaja slobodní žiaci alebo obaja zadaní. Následne odhadneme model, ktorý tieto stavy bude rozlišovať a modely medzi sebou porovnáme. V našej sieti je 93 hrán medzi žiakmi s rovnakým stavom, z toho 79 hrán je medzi dvomi slobodnými a zvyšných 14 hrán je medzi zadanými žiakmi. Odhadnuté parametre pre modely sa nachádzajú v prílohe A.1. Podobne ako v prípade s modelom sledujúcim pohlavie sa pozrieme na Akaikeho a Bayesovo informačné kritérium. V tabuľke 2 vidíme, že hodnoty kritérií sa veľmi nelíšia. Zároveň však model, ktorý je v tomto smere jednoduchší, a teda model nerozlišujúci medzi stavmi, sa vidí potenciálne lepším.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model nerozlišujúci medzi stavmi | 1279 | 1293 |
| Model rozlišujúci medzi stavmi | 1280 | 1301 |

**Tabuľka 2:** Porovnanie AIC a BIC pre modely sledujúce stav žiakov.

Pozrime sa na všetky sledované údaje a ich informačné kritéria dokopy v tabuľke 3. Porovnaním hodnôt môžeme vybrať tie modely, ktoré najviac zlepšili Null model. Ukázalo sa, že najviac zlepšilo Null model sledovanie pohlavia. Ako sme očakávali, pre model sledujúci vek sa obe informačné kritériá zhoršili, preto už porovnáme len spojenie vysvetľujúcich premenných pohlavie a stav.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 1289 | 1296 |
| Model sledujúci pohlavia | 1199 | 1213 |
| Model sledujúci vek | 1291 | 1305 |
| Model sledujúci stav | 1279 | 1293 |

**Tabuľka 3:** Porovnanie AIC a BIC pre modely sledujúce základné osobné údaje.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 1289 | 1296 |
| Model sledujúci pohlavia | 1199 | 1213 |
| Model sledujúci pohlavie a stav | 1189 | 1210 |

**Tabuľka 4:** Porovnanie AIC a BIC pre modely spájajúce základné osobné údaje.

Z výsledných kritérií, uvedených v tabuľke 4, môžeme vybrať model sledujúci pohlavie a stav ako najlepší a budeme s ním pracovať v podkapitole 4.2.6. Model s odhadnutými parametrami sa nachádza v prílohe A.1.

### 4.2.3 Model sledujúci hudobné preferencie

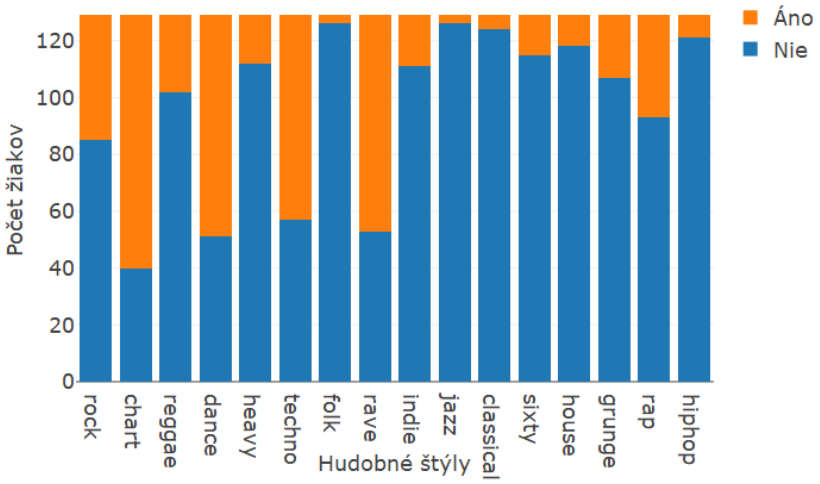
V tejto časti sme sa zamysleli nad ovplyvňovaním kamarátstiev počúvaním rovnakých hudobných štýlov. Inšpirovali sme sa článkom [25], v ktorom autori predstavujú aplikáciu programu SIENA na dáta. Program je využívaný na štatistické analýzy sietí a ich dynamiky, ktoré sa zameriavajú na sociálne siete. Predpokladá sa, že osoby, predstavujúce vrcholy, zohrávajú hlavnú úlohu vplyvu na existenciu hrán. V článku autori sledujú zmeny v kamarátstvach, počúvaní hudby a konzumácii alkoholu v priebehu sledovacieho obdobia. Z veľkého počtu sledovaných štýlov vytvorili tri skupiny, ktoré nazvali *rock, classical* a *techno*. Dospeli k záveru, že *rock* pozitívne vplýva na kamarátstva a podobne aj žiaci počúvajúci hudobné štýly zahrnuté v skupine *classical* mali tendenciu sa viac kamarátiť medzi sebou.

V dátach máme informáciu o 16 hudobných štýloch, a to:

- rock,
- chart,
- reggae,
- dance,
- heavy,
- techno,
- folk,
- rave,
- indie,
- jazz,
- classical,
- sixty,

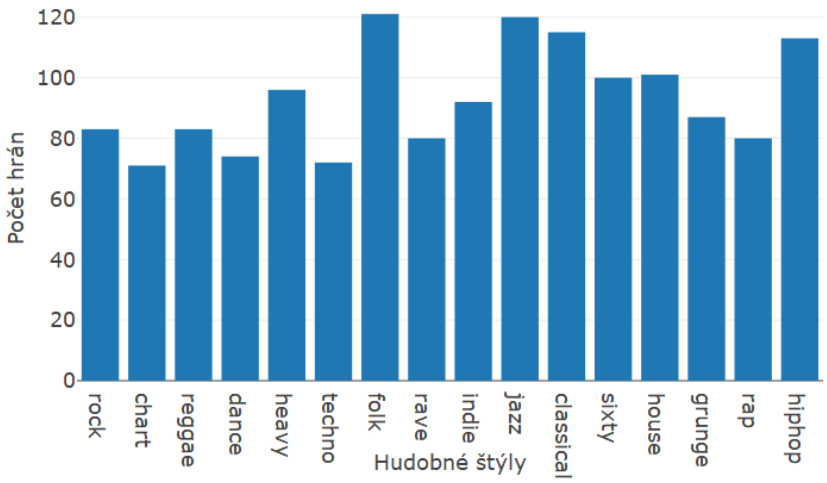
- house,
- grunge,
- rap,
- hip-hop.

Ku každému hudobnému štýlu žiak odpovedal, či daný štýl počúva. Na obrázku 6 vidíme, ako žiaci odpovedali na jednotlivé hudobné štýly. Medzi žiakmi prevláda počúvanie hudobných štýlov *dance, chart, rave* a *techno*. Na druhú stranu hudobné štýly *folk, jazz, classical* a *hiphop* počúva malá časť žiakov.

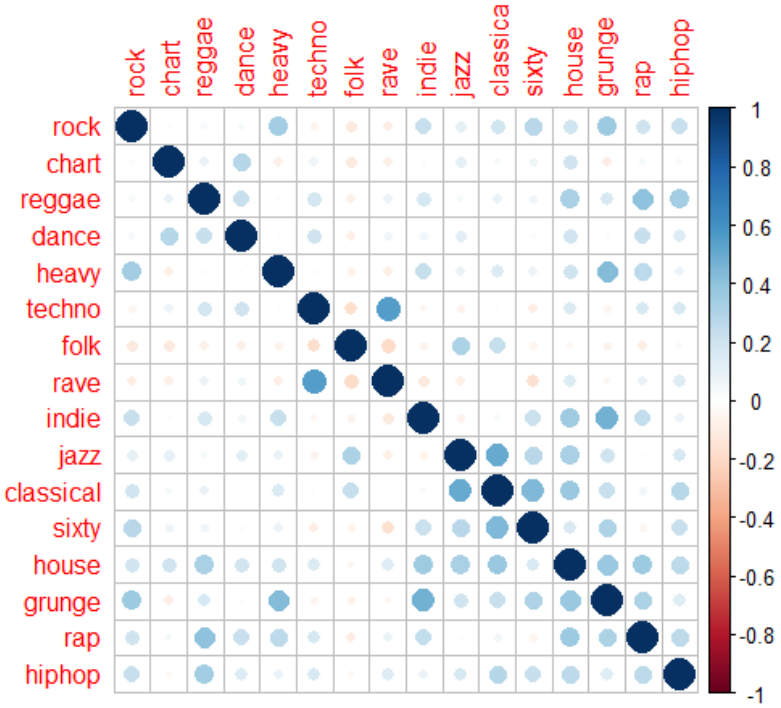



**Obr. 6:** Pomer odpovedí o počúvaní hudobných štýlov.

Pozrime sa na počty hrán medzi rovnakými odpoveďami o počúvaní jednotlivých štýlov. Graf s týmito počtami hrán je zobrazený na obrázku 7. Vidíme, že väčšina hrán v sieti je medzi osobami, ktoré rovnako počúvajú, respektíve nepočúvajú hudobné štýly *folk, jazz, classical* a *hiphop*. Z grafu taktiež vidíme, že pri každom hudobnom štýle je takmer polovica hrán medzi žiakmi, ktorí odpovedali rovnako. Poďme sa pozrieť na korelácie medzi jednotlivými štýlmi. Na obrázku 8 vidíme vykreslené jednotlivé korelácie. Čím máme väčší kruh pri dvoch štýloch, tým väčšia korelácia je v dátach. Môžeme si všimnúť, že hudobné štýly *techno* a *rave* spolu významnejšie korelujú. Ako by sme mohli očakávať kladne korelujú štýly *jazz, classical, sixty*.

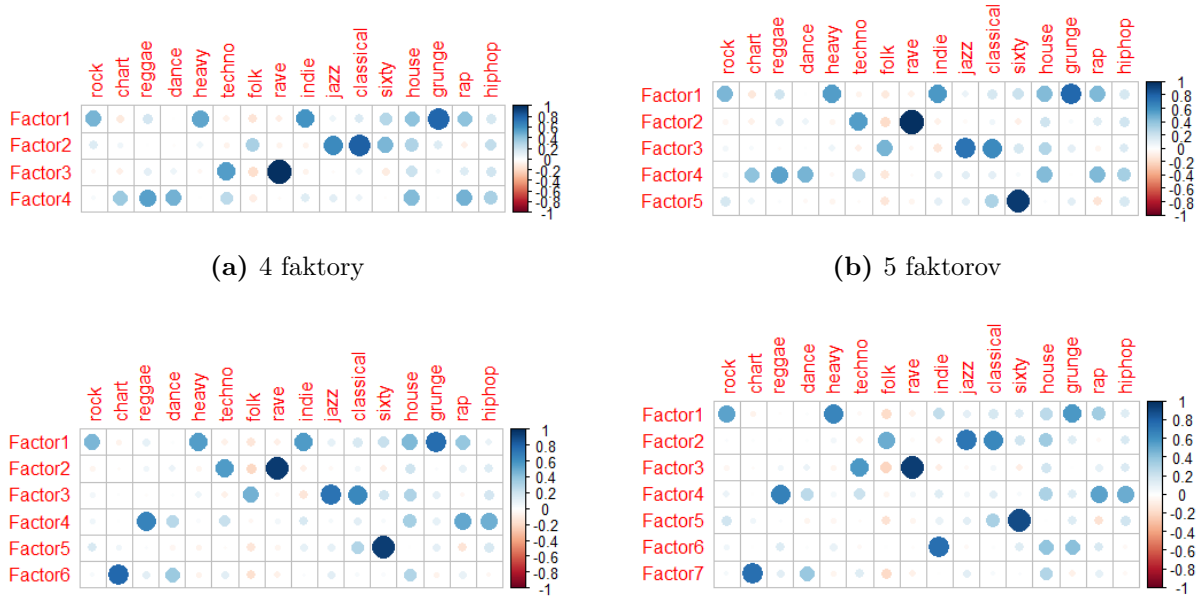



**Obr. 7:** Počty hrán medzi žiakmi, ktorí odpovedali rovnako o počúvaní hudobných štýlov.




**Obr. 8:** Korelácia hudobných štýlov medzi sebou.

Tieto korelácie nám slúžia na informatívny charakter. Bolo by ťažké vybrať, ktoré hudobné štýly spojiť. Aplikujeme preto na tieto premenné faktorovú analýzu, definovanú v kapitole 3. Využijeme varimaxovú rotáciu a pozrieme sa na hodnoty pre 4, 5, 6 a 7 faktorov. Na obrázku 9 uvádzame jednotlivé korelácie hudobných štýlov s vytvorenými faktormi.




**(a)** 4 faktory

**(b)** 5 faktorov

**(c)** 6 faktorov

**(d)** 7 faktorov

**Obr. 9:** Korelácie vytvorených faktorov s hudobnými štýlmi.

Pri rozhodovaní, ktorý počet faktorov je najvhodnejší, musíme zobrať do úvahy nie len počet faktorov, ale aj ako dané faktory vysvetľujú korelácie v dátach. Vo výstupe v *R* dostávame dve informácie na posúdenie. Jednou z nich je kumulovaná variancia a druhou je p-hodnota, ktorá nám hovorí o dostačujúcnosti počtu faktorov [26]. Vytvorené štyri faktory nám zahŕňajú iba 40% kumulovanej variancie a p-hodnota je pre nich 0,08, pri piatich faktoroch už nám hodnoty stúpnu na 46%, respektíve 0,40. Pri šiestich faktoroch dostaneme 50% kumulovanej variancie a p-hodnotu 0,75. Pri siedmich faktoroch už dostávame 53% a p-hodnotu 0,97. Chceme protichodné vlastnosti, a to menší počet faktorov, ale zároveň čo najväčšie percento kumulovanej variancie a čo najväčšiu p-hodnotu. Za zváženie nám teda stojí šesť a sedem faktorov. Ak by sme si zvolili prahovú hodnotu na koreláciu medzi faktorom a hudobným štýlom na hodnotu 0,5, dostali by sme nasledovné skupiny hudobných štýlov:

Rozdelenie pre 6 faktorov:

- 1. faktor: rock, heavy, indie,
- 2. faktor: techno, rave,
- 3. faktor: folk, jazz, classical,
- 4. faktor: reggae, rap, hip-hop,
- 5. faktor: sixty,
- 6. faktor: chart.

Rozdelenie pre 7 faktorov:

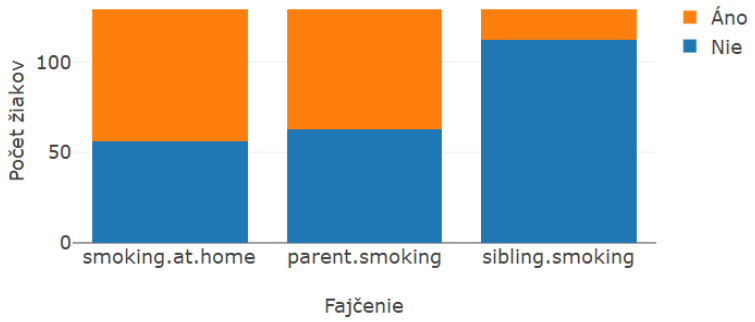
- 1. faktor: rock, heavy, grunge,
- 2. faktor: folk, jazz, classical,
- 3. faktor: techno, rave,
- 4. faktor: reggae, rap, hip-hop,
- 5. faktor: sixty,
- 6. faktor: indie,
- 7. faktor: chart.

Môžeme si všimnúť, že najväčší rozdiel medzi skupinami je v zaradení hudobného štýlu *indie*. Aby sme sa vedeli rozhodnúť, ktorý model je lepší, pozreli sme sa na informačné kritériá z výstupov, ktoré sú zahrnuté v prílohe A.1. Pre model so šiestimi faktormi sme dostali hodnoty 1286 a 1335. Obe kritériá sa pre model so siedmimi faktormi zhoršili, a to na hodnoty 1288 a 1344. Rozhodli sme sa preto zvoliť šesť faktorov, s ktorými budeme neskôr v práci vytvárať model.

### 4.2.4 Modely sledujúce fajčenie

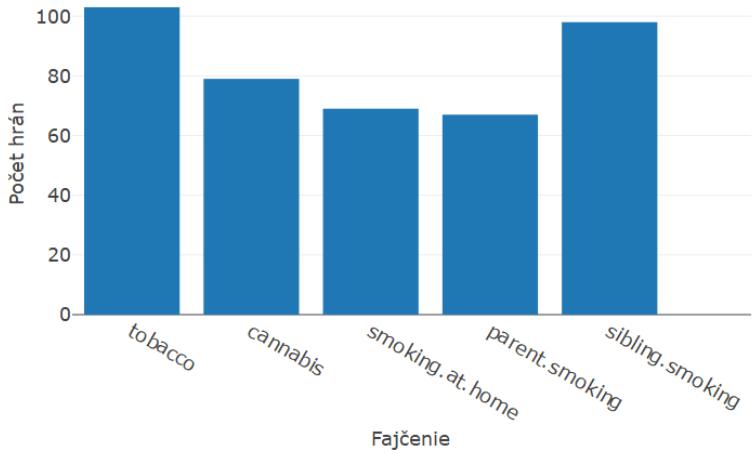
V tejto časti sa pozrieme na zlozvyky, respektíve užívanie návykových látok. Budeme sledovať ako fajčenie tabaku a fajčenie marihuany ovplyvňuje kamarátstva. Žiaci môžu tieto aktivity vykonávať spolu, a tým pádom sa viac kamarátiť. Taktiež zahrnieme do sledovania aj informácie, či sa u žiaka fajčí v domácnosti alebo aj informácie o fajčení rodičov a súrodencov.

Pri odpovediach ohľadne fajčenia tabaku si mali respondenti vybrať z troch rôznych možností a to nefajčím, fajčím príležitostne alebo pravidelne. V dátach máme 109 žiakov, ktorí nefajčia. Príležitostne fajčí 10 žiakov a zvyšných 10 žiakov fajčí pravidelne. Pri otázke o fajčení marihuany si respondenti vyberali z možností: nefajčím, vyskúšal som raz, fajčím príležitostne alebo pravidelne. Väčšina žiakov, konkrétne 94, nefajčí marihuanu. Marihuanu vyskúšalo raz 19 žiakov a 14 žiakov si ešte príležitostne zapáli. Pravidelne fajčia iba 2 žiaci. Pri otázkach ohľadne fajčenia v domácnosti, fajčenia rodičov alebo súrodencov mali možnosť odpovedať áno, respektíve nie. Na obrázku 10 vidíme pomer odpovedí žiakov pri jednotlivých možnostiach. Môžeme si všimnúť, že približne polovica rodičov žiakov fajčí a ešte vo viac domácnostiach sa celkovo fajčí. Môže to byť dôsledkom fajčenia iných rodinných príslušníkov, o ktorých informáciu nemáme.




**Obr. 10:** Pomer odpovedí o fajčení.

Pozrime sa na počty hrán medzi rovnakými odpoveďami. Na obrázku 11 vidíme jednotlivé počty. Veľký počet z celkových 124 hrán, ktoré máme v sieti, je medzi žiakmi, ktorí rovnako odpovedali ohľadne fajčenia tabaku a fajčenia súrodencov.




**Obr. 11:** Počty hrán medzi žiakmi, ktorí odpovedali rovnako o fajčení.

Podobne ako v predchádzajúcich častiach sme odhadli modely s parametrami jednotlivo. Celé výstupy aj s odhadnutými koeficientami sú zahrnuté v prílohe A.1. Hodnoty informačných kritérií vidíme v tabuľke 5. Ani jeden z parametrov výrazne nezlepšil informačné kritériá. Bayesovo kritérium, ktoré je prísnejšie a dáva väčšiu váhu na počet parametrov v modeli, sa zhoršilo vo všetkých modeloch. Môžeme skonštatovať, že tieto parametre by nám len skomplikovali model.

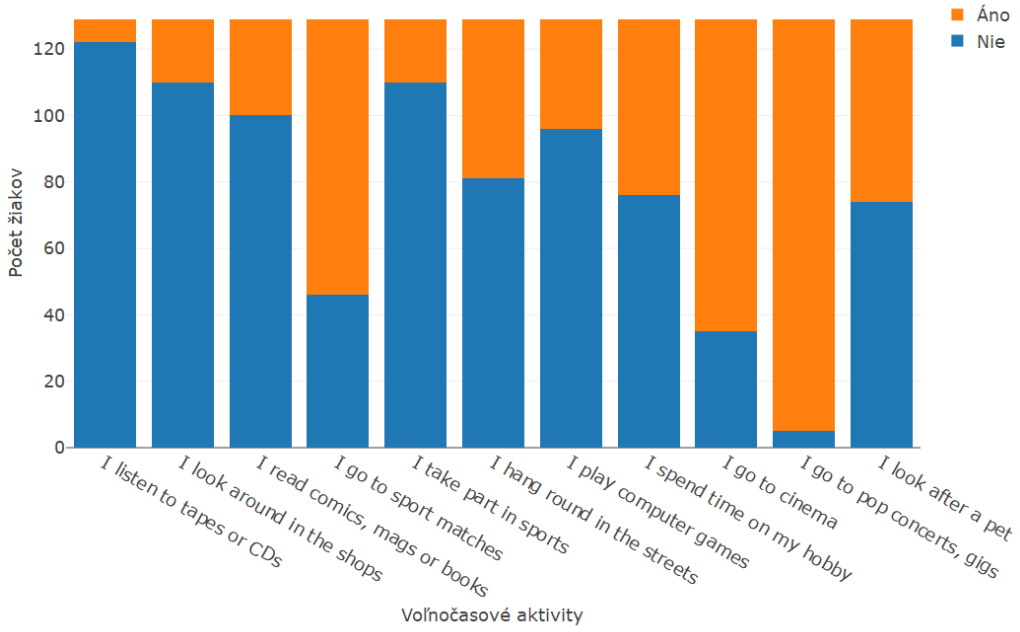
| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 1289 | 1296 |
| Model sledujúci fajčenie tabaku | 1283 | 1297 |
| Model sledujúci fajčenie marihuany | 1288 | 1302 |
| Model sledujúci fajčenie doma | 1290 | 1304 |
| Model sledujúci fajčenie rodičov | 1290 | 1304 |
| Model sledujúci fajčenie súrodencov | 1291 | 1305 |

**Tabuľka 5:** Porovnanie AIC a BIC pre modely sledujúce fajčenie.

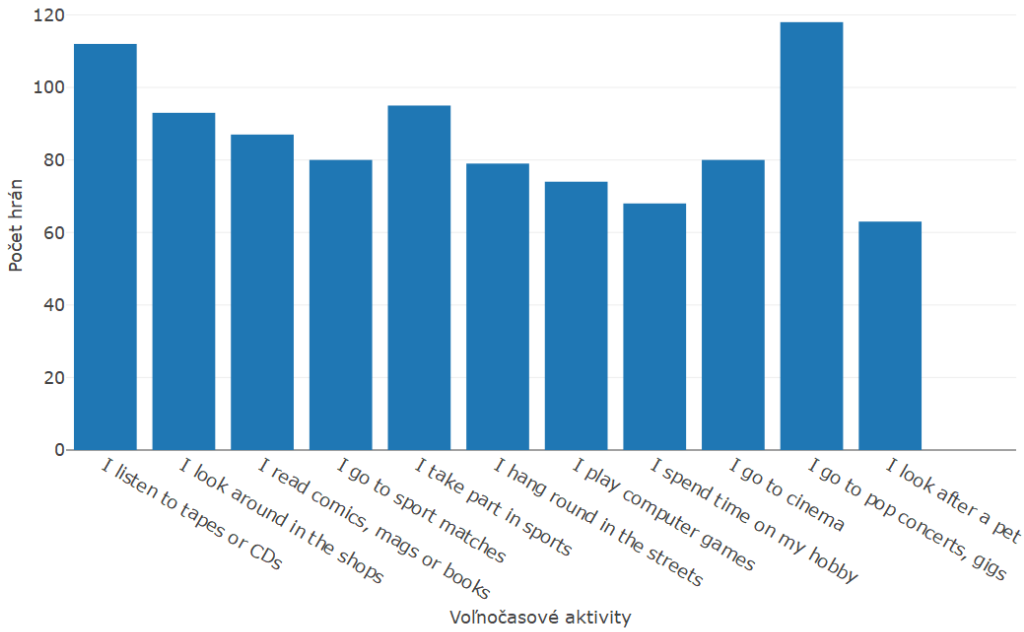
### 4.2.5 Modely sledujúce voľnočasové aktivity

Možností a dôvodov, ako a prečo sa ľudia kamarátia, je nespočetné množstvo. Veľa ľudí sa zoznámi pri voľnočasových aktivitách, prípadne sa aj pravidelne stretáva. Môžu sa taktiež skamarátiť na základe rovnakých záľub a záujmov, preto je vhodné sa pozrieť na informácie o voľnočasových aktivitách, ktoré máme v dátach. Niektoré aktivity sme museli z pozorovania vynechať, lebo obsahovali chýbajúce hodnoty. Odpovede v dátach pri voľnočasových aktivitách nadobúdali hodnoty 1 až 4, pričom tieto hodnoty reprezentovali interval, v akom žiak túto činnosť vykonáva. Sledoval sa interval vykonávania aktivity väčšinu dní, raz za týždeň, raz za mesiac a menej, respektíve nikdy. Tieto odpovede sme pre jednoduchosť a lepšie zavedenie do modelu upravili tak, že interval vykonávania väčšinu dní a raz za týždeň sme spojili do odpovede, že daný žiak túto aktivitu vykonáva a zvyšné dva intervaly sme označili, že aktivitu žiak nevykonáva.

Pozrime sa, aké sledované aktivity budeme zahŕňať do modelu a zároveň aj na početnosti odpovedí. Hodnoty vidíme zobrazené pomocou stĺpcového grafu na obrázku 12. Môžeme si všimnúť, že veľmi malý počet žiakov počúva kazety alebo CD. Na druhú stranu väčšina žiakov chodí na koncerty alebo do divadla. Taktiež viac ako polovica žiakov chodí na športové zápasy. Ostatné aktivity väčšina žiakov nevykonáva. Početnosti hrán medzi žiakmi, ktorí odpovedali rovnako vidíme v grafe na obrázku 13. Takmer pri každej voľnočasovej aktivite je viac ako polovica hrán medzi žiakmi s rovnakou odpoveďou.




**Obr. 12:** Pomer odpovedí pri voľnočasových aktivitách.




**Obr. 13:** Počty hrán medzi žiakmi, ktorí odpovedali rovnako pri voľnočasových aktivitách.

Po predchádzajúcich grafoch už máme predstavu o tom, aké máme informácie o voľnočasových aktivitách žiakov a ich vplyve na sieť. Pozrime sa, ako tieto premenné jednotlivo ovplyvňujú model. Odhadnuté modely aj s parametrami sú zahrnuté v prílohe A.1. Informačné kritériá vidíme v nasledujúcej tabuľke 6.

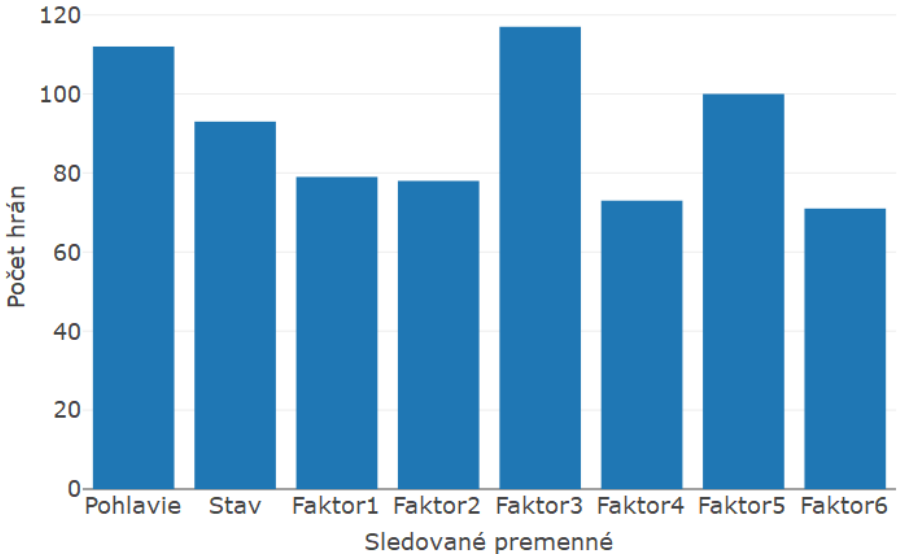
| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 1289 | 1296 |
| Počúvam kazety alebo CD | 1291 | 1305 |
| Chodím po obchodoch | 1291 | 1305 |
| Čítam knihy, komiksy | 1290 | 1304 |
| Chodím na športové zápasy | 1285 | 1299 |
| Športujem | 1291 | 1305 |
| Chodím po vonku | 1285 | 1299 |
| Hrám počítačové hry | 1291 | 1305 |
| Využívam čas na hobby ako umenie, hudobné nástroje | 1291 | 1305 |
| Chodím do kina | 1290 | 1304 |
| Chodím na popové koncerty | 1290 | 1304 |
| Starám sa o domáce zvieratko | 1291 | 1305 |

**Tabuľka 6:** Porovnanie AIC a BIC pre modely sledujúce voľnočasové aktivity.

Podobne, ako v časti sledujúcej zlozvyky, sa Bayesovo kritérium zhoršilo pridaním premenných do modelu. Efekt homofílie vzhľadom na voľnočasové aktivity nám nepomôže lepšie vysvetľovať kamarátstva v sieti.

### 4.2.6 Spojenie modelov

V tejto časti spojíme modely, ktoré sme vybrali v jednotlivých podkapitolách na základe Akaikeho a Bayesovho informačného kritéria v porovnaní s Null modelom, prípadne ich zoskupovaním a porovnávaním medzi sebou. Zo všetkých parametrov so základnými údajmi nám vyšlo najlepšie sledovať parametre pohlavie a stav. Následne sme na základe faktorovej analýzy vytvorili šesť faktorov, ktoré nám budú predstavovať skupiny hudobných štýlov, ktoré vložíme do modelu. Prekvapením bol výsledok podkapitoly, v ktorej sme sledovali parametre súvisiace s fajčením. V tejto časti nám nevyšiel žiadny parameter, ktorý by zlepšoval model. Podobný záver nás čakal aj v podkapitole sledujúcej voľnočasové aktivity. Vytvoríme teda spojený model sledujúci pohlavie, stav a faktory, vytvorené zoskupením hudobných štýlov. Počty hrán medzi žiakmi, ktorí odpovedali rovnako na jednotlivé parametre vidíme na grafe na obrázku 14. Najviac hrán medzi rovnakými odpoveďami je pri faktore 3, ktorý zoskupuje počúvanie hudobných štýlov *folk, jazz* a *classical*. Pripomeňme si, z akých hudobných štýlov sa skladali zvyšné faktory. Skupinu hudobných štýlov *rock, heavy, indie* sme zoskupili do faktora 1. Štýly *techno, rave* sme zoskupili do faktora 2. Pre tieto faktory nám vyšli podobné počty hrán medzi rovnakými odpoveďami. Do faktoru 4 sme zaradili štýly *reggea, rap, hiphop*. Faktor 5 predstavuje štýl *sixty* a faktor 6, pri ktorom máme najmenší počet hrán, predstavuje štýl *chart*.




**Obr. 14:** Počty hrán medzi žiakmi, ktorí odpovedali rovnako na sledované premenné.

Vytvorili sme model zahrňujúci tieto parametre. Po odhadnutí modelu dostávame výstup:

```
==========================
Summary of model fit
==========================

Formula:   net.sym ~ edges + nodematch("sex") + nodematch("romantic") +
    nodematch("faktor1") + nodematch("faktor2") +
    nodematch("faktor3") + nodematch("faktor4") +
    nodematch("faktor5") + nodematch("faktor6")

Iterations:  6 out of 20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges              -7.48887    0.55327      0 -13.536  < 1e-04 ***
nodematch.sex       2.23800    0.30483      0   7.342  < 1e-04 ***
nodematch.romantic  0.72573    0.21006      0   3.455 0.000551 ***
nodematch.faktor1   0.50651    0.18999      0   2.666 0.007677 **
nodematch.faktor2   0.30883    0.18931      0   1.631 0.102827
nodematch.faktor3   0.78178    0.40447      0   1.933 0.053254 .
nodematch.faktor4   0.21165    0.18593      0   1.138 0.254980
nodematch.faktor5  -0.16595    0.23872      0  -0.695 0.486964
nodematch.faktor6  -0.01702    0.18451      0  -0.092 0.926488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null Deviance: 11445  on 8256  degrees of freedom
Residual Deviance:  1166  on 8247  degrees of freedom

AIC: 1184    BIC: 1248    (Smaller is better.)
```

Vo výstupe máme zahrnuté rôzne informácie o odhadnutom modeli, nie len odhadnuté hodnoty parametrov. Pri MCML odhade parametrov dostávame aj informáciu o p-hodnote a následnej signifikantnosti parametra. V knihe [16] z roku 2014 uvádzajú, že tvrdenia o asymptotických rozdeleniach, z ktorých sa tieto hodnoty vypočítavajú neboli formálne dokázané. Odporúčajú tieto hodnoty interpretovať informatívne. Na druhej strane, v knihe [18] z roku 2015, už tieto hodnoty využívajú na určovanie signifikantnosti a na základe týchto hodnôt vyraďujú parametre z modelov. V dokumentácii k ergm balíku [28] v *R* určujú pri jednom parametri signifikantnosť na základe výstupu. Týmito informáciami sa inšpirujeme a pozrieme sa na model, ktorý bude obsahovať len parametre, ktoré sú v predchádzajúcom modeli signifikantné. Zvolíme si hraničnú hodnotu p-hodnoty na 0,05,

s tým, že faktor 3 ešte do modelu zahrnieme. Dostávame odhadnutý model s výstupom v tvare:

```
==========================
Summary of model fit
==========================

Formula:   net.sym ~ edges + nodematch("sex") + nodematch("romantic") +
    nodematch("faktor1") + nodematch("faktor3")

Iterations:  6 out of 20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges               -7.2985     0.5141      0 -14.197  < 1e-04 ***
nodematch.sex        2.2368     0.3047      0   7.341  < 1e-04 ***
nodematch.romantic   0.7099     0.2097      0   3.385 0.000713 ***
nodematch.faktor1    0.5200     0.1892      0   2.748 0.006005 **
nodematch.faktor3    0.7512     0.3925      0   1.914 0.055608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 11445  on 8256  degrees of freedom
 Residual Deviance:  1171  on 8251  degrees of freedom

AIC: 1181    BIC: 1216    (Smaller is better.)
```

Môžeme si všimnúť, že všetky parametre sledujúce charakteristiky kladne ovplyvňujú pravdepodobnosť existencie hrany. Jediným záporným odhadom je parameter, ktorý je považovaný za intercept. Hovorí o tendencii existencie hrán v sieti. Z predchádzajúcich podkapitol vieme, že tento parameter je záporný dôsledkom toho, že máme malý počet hrán oproti počtu vrcholov v sieti. Keď sa pozrieme ako dané parametre ovplyvňujú šance, dostaneme:

- pohlavie: $\Delta_1 odds(y_{ij} = 1) = \exp(\theta_2) = \exp(2{,}24) = 9{,}39$,
- stav: $\Delta_2 odds(y_{ij} = 1) = \exp(\theta_3) = \exp(0{,}71) = 2{,}03$,
- faktor 1: $\Delta_3 odds(y_{ij} = 1) = \exp(\theta_4) = \exp(0{,}52) = 1{,}68$,
- faktor 3: $\Delta_4 odds(y_{ij} = 1) = \exp(\theta_5) = \exp(0{,}75) = 2{,}12$.

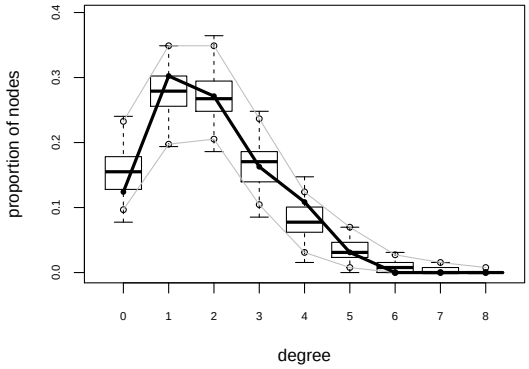
Vidíme, že najviac ovplyvňuje šance kamarátstva medzi rovnakým pohlavím, ktorým zodpovedá parameter $\theta_2$. Rovnaký stav a počúvanie rovnakých hudobných štýlov, ktoré sú zahrnuté vo faktore 3, znásobujú šance viac než dvojnásobne. Zo sledovaných parametrov najmenej ovplyvňuje šance faktor 1, ktorý zväčšuje šance 1,68-krát.

## 4.3 Testovanie modelov

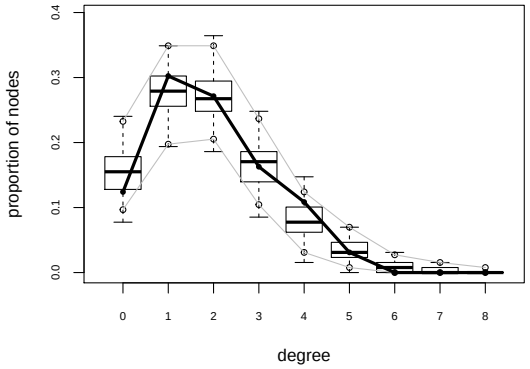
V predchádzajúcej podkapitole sme pre konkrétny model odhadli parametre. Predtým, ako model vyhlásime za finálny, mali by sme ho otestovať. Testovať model budeme na základe vlastností sietí, ktoré sme doteraz do modelu nezahrnuli. Tieto vlastnosti sme definovali v kapitole 1, a to rozdelenie stupňov vrcholov, rozdelenie počtu spoločných susedov a rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi. Zároveň budeme náš model porovnávať s Null modelom, aby sme mali lepšiu predstavu o tom, ako sa správajú jednotlivé kritériá aj pre jednoduchý model. Testovanie modelu spočíva vo vytvorení 100 náhodných sietí na základe odhadnutého modelu. Priblíženie a matematické pozadie, ako sa simulujú siete na základe odhadnutého modelu, sa nachádzajú v [4]. Následne sa pri testovaní pozrie na dané vlastnosti simulovaných sietí a porovnajú sa so zadanou sieťou, na ktorej sa model odhadol.

Ako prvé sa pozrieme na rozdelenie stupňov vrcholov v sieti. Ak si za $n$ označíme počet všetkých vrcholov $|V|$, hodnoty na horizontálnej osi dostaneme ako $\frac{d_1}{n}, \frac{d_2}{n}, \ldots, \frac{d_k}{n}$, pričom $d_k$, ako sme už spomenuli, predstavuje počet vrcholov, ktoré majú stupeň $k$. Na vertikálnej osi dostaneme hodnoty rozdelenia v tejto sieti. Boxploty na obrázku 15 nám predstavujú hodnoty rozdelení pre simulované siete a čiara, ktorá cez nich prechádza predstavuje spájajúcu čiaru pre hodnoty rozdelenia v našej sieti. Minimálne, maximálne a stredné hodnoty pre toto rozdelenie sú uvedené v prílohe A.2. Vidíme, že oba modely túto charakteristiku dobre vysvetľujú a odhadnuté siete majú podobné rozdelenie stupňov.

Ďalšou charakteristikou siete je rozdelenie počtu spoločných susedov. Na obrázku 16, podobne ako v predchádzajúcej charakteristike, na horizontálnej osi dostávame znormované hodnoty rozdelenia v tvare: $\frac{ep_0}{s_1}, \frac{ep_1}{s_1}, \ldots, \frac{ep_{n-2}}{s_1}$. Vidíme, že modely vôbec nezachytávajú túto vlastnosť, nakoľko hodnoty pre modelom simulované hodnoty sa vôbec nezhodujú s hodnotami v našej sieti. Simulované siete obsahovali väčšie množstvo hrán so žiadnym spoločným susedom. V našej sieti sa takýchto hrán nachádza 69 a model ich



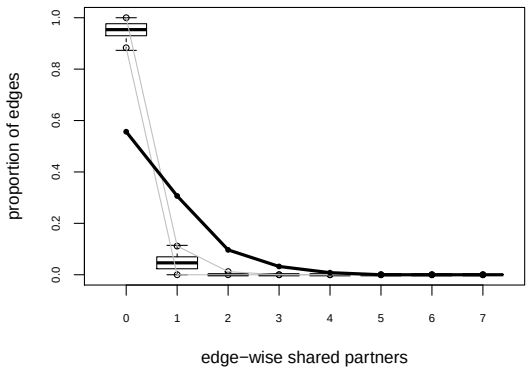

**(a)** Null model



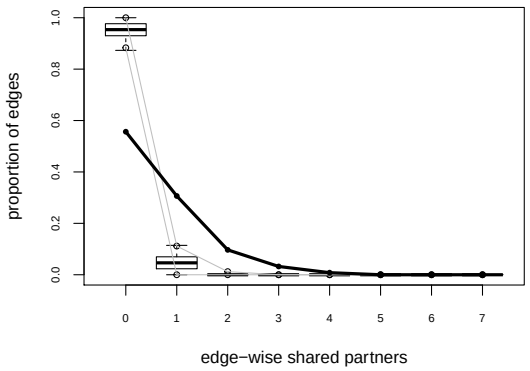

**(b)** Náš model

**Obr. 15:** Rozdelenie stupňov vrcholov v sieťach.

počet v priemere odhadol na 118. Vyššie počty spoločných susedov sa v simulovaných sieťach nachádzali už v menšom počte. Napríklad hrán s jedným spoločným susedom sa priemerne nachádzalo v simulovaných sieťach 5, pričom v pôvodnej sieti ich je 38. Zvyšné stupne a ich priemerné, ale aj minimálne a maximálne hodnoty pre simulované siete sa nachádzajú v prílohe A.2. Vlastnosti vrcholov nám očividne nedávajú dostatočne veľkú informáciu o pomyselných trojuholníkoch v sieti.



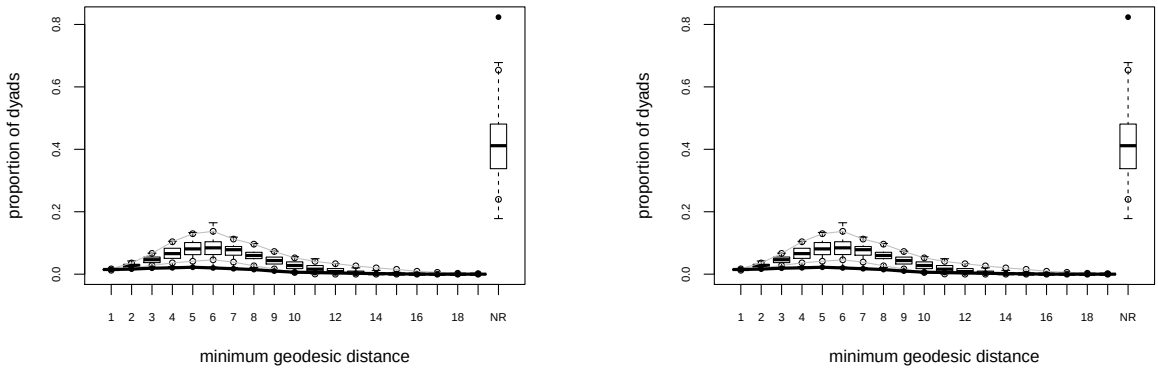

**(a)** Null model




**(b)** Náš model

**Obr. 16:** Rozdelenie počtu spoločných susedov v sieťach.

Poslednou sledovanou charakteristikou je rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi. Na obrázku 17 máme zobrazené grafy pre toto rozdelenie. Na vertikálnej osi máme počty vrcholov, ktoré majú príslušnú vzdialenosť znormalizované súčtom všetkých možných vzdialeností. Pre vzdialenosť dĺžky 1, ktorá predstavuje spojenie vrcholov hranou, si simulované siete viedli dobre. V priemere počet takýchto vzdialeností bol v simulovaných sieťach 124 zo skutočných 124. Tento efekt by sme mohli očakávať, keďže v modeli máme zahrnutý parameter, ktorý sleduje počet hrán v sieti. Počet vrcholov, ktoré majú medzi sebou vzdialenosť dĺžky 2, mali však už simulované siete v priemere 230 a zo skutočných 134. Pre vzdialenosti dĺžky 3 až 8 sa hodnoty rozdelenia pre simulované siete viac odlišovali od hodnôt v zadanej sieti. Dvojice vrcholov, ktoré nemajú spoločnú cestu, predstavuje hodnota *NR* v grafoch. Takýchto dvojíc máme v sieti 6798 a v simulovaných sieťach sme mali ich počet v priemere 3514. Všetky hodnoty pre toto rozdelenie sa taktiež nachádzajú v prílohe A.2. Môžeme zhodnotiť, že modely si ani v tejto charakteristike neviedli veľmi dobre.




**(a)** Null model **(b)** Náš model

**Obr. 17:** Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi v sieťach.

Na základe týchto zistení sa v nasledujúcej podkapitole pozrieme bližšie na charakteristiky siete, ich modifikácie a spôsob, ako ich zahrnúť do modelu.

## 4.4 Modely s charakteristikami siete

Pripomeňme si, čo sme zatiaľ spravili. Ako sme už videli, našli sme parametre, ktoré zahŕňajú vplyv homofílie. Zároveň sme vybrali len tie parametre, ktoré nám zlepšovali model vzhľadom na informačné kritéria. Pri testovaní vytvoreného modelu sme však zistili, že aj keď parametre nám zlepšujú model, nevysvetľujú dostatočne sieť, respektíve charakteristiky siete. Na základe tohto zistenia sa v tejto časti práce zameriame na zahrnutie charakteristík siete do modelu.

Jednou z vlastností siete, ktoré sme nezahrnuli v modeli, je sledovanie stupňov v sieti. Jednotlivo sme zahŕňali stupne od 1 po 5, nakoľko žiaden vrchol nemá väčší stupeň. Zahrnutie jednotlivého stupňa do siete predstavuje dodatočnú informáciu pre model, koľko vrcholov má daný stupeň. Výstupy pre modely sú zahrnuté v prílohe A.3. Poďme sa pozrieť na informačné kritériá pre modely v porovnaní s modelom, ktorý sme vybrali za najlepší spomedzi modelov sledujúcich vlastnosti vrcholov. Tento model sme uviedli v podkapitole 4.2.6 a v tejto časti ho nazveme *model s vlastnosťami*. Ako môžeme vidieť v tabuľke 7, všetkým modelom sa zhoršili informačné kritériá po pridaní parametra sledujúceho stupeň, tým pádom tieto stupne nebudeme zahŕňať do finálneho modelu.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model s vlastnosťami | 1181 | 1216 |
| Model sledujúci stupeň 0 | 1181 | 1223 |
| Model sledujúci stupeň 1 | 1182 | 1224 |
| Model sledujúci stupeň 2 | 1183 | 1225 |
| Model sledujúci stupeň 3 | 1183 | 1225 |
| Model sledujúci stupeň 4 | 1182 | 1224 |
| Model sledujúci stupeň 5 | 1183 | 1225 |

**Tabuľka 7:** Porovnanie AIC a BIC pre modely sledujúce stupne.

Pozrime sa na alternatívnu možnosť ku sledovaniu stupňov, ktorú vieme zahrnúť do modelov, a to sledovanie počtu k-star. Do modelov sme nezahrnuli 1-star, nakoľko, ako sme už spomenuli v kapitole 1, ich počet je rovný dvojnásobku počtu hrán a parameter sledujúci počet hrán, už v modeli máme. Zahrnuli sme postupne parametre sledujúce počet k-star do veľkosti 5. Dospeli sme k záveru, že ani pridanie týchto parametrov nám vo vylepšení modelu nepomohlo, ako si môžeme všimnúť na nasledujúcej tabuľke 8.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model s vlastnosťami | 1181 | 1216 |
| Model sledujúci 2-star | 1181 | 1223 |
| Model sledujúci 3-star | 1180 | 1222 |
| Model sledujúci 4-star | 1179 | 1221 |
| Model sledujúci 5-star | 1178 | 1220 |

**Tabuľka 8:** Porovnanie AIC a BIC pre modely sledujúce k-star.

Poslednou alternatívou k rozdeleniu stupňov, na ktorú sa pozrieme v tejto práci, je geometrické rozdelenie stupňov v sieti. Rozdelenie sme definovali v kapitole 2. Pripomeňme si, že ho vieme zahrnúť do modelu dvomi spôsobmi, nakoľko má v sebe parameter [29]. Ak parameter necháme zafixovaný, dostaneme vlastnosť pre ERG modely. Ak by sme parameter nezafixovali, tak by ho model musel tiež odhadnúť a dostali by sme CERG modely (z angl. *curved exponential random graph models*). Pri zvolení vhodnej hodnoty parametra $\tau$ pre geometrické rozdelenie stupňov, ako aj pre geometrické rozdelenie počtu spoločných susedov, v literatúre odporúčajú testovať rôzne hodnoty parametra a sledovať, ako sa model mení [5]. V článku [13] po preskúmaní rôznych zafixovaných hodnôt $\tau$ zvolili hodnotu 1, pričom porovnávali hodnoty z intervalu 0,5 až 2. Možnosť nezafixovaného parametra neuvažovali, nakoľko by to skomplikovalo model. Taktiež v poznámkach k prednáškam na *ECPR Winter School Inferential Network Analysis* [2] autori odporúčajú pri voľbe parametra $\tau$ jeho hodnotu zvyšovať a sledovať, ako sa správa Bayesovo informačné kritérium. V kapitole 3 z knihy [10] sa pri geometrických rozdeleniach pozreli najprv na rôzne hodnoty zafixovaného parametra $\tau$ od hodnoty 0,1 po 1, pričom najlepší výsledok dostali pre hodnotu 1. Následne sa práve pre túto hodnotu rozhodli pozrieť na možnosť nezafixovať parameter, čím dostali najlepšie hodnoty informačných kritérií.

Po preštudovaní rôznych prístupov z literatúry sme sa rozhodli vyskúšať najprv zafixované hodnoty parametra $\tau$ od 0,1 po 2. Hodnoty informačných kritérií sme pre každý z týchto modelov dostali rovnaké, a to 1181, 1223. Na ilustráciu výstupu sme zvolili model, preferovaný aj v literatúre, so zafixovanou hodnotou parametra $\tau = 1$, ktorý uvádzame v prílohe A.3. Poďme sa pozrieť, čo sa stane, ak modelu parameter nezafixujeme a nastavíme začiatočnú hodnotu 1. V tomto prípade sme dostali vysoké hodnoty informačných kritérií, a to 1366, 1415. Zároveň vo výstupe, ktorý sa nachádza v prílohe A.3, nám odhadnuté parametre pre premenné výrazne stúpli, čo môže signalizovať degenerovanosť. Dostávame, že v oboch prípadoch sa Bayesovo informačné kritérium zhoršilo pridaním parametra. V tomto prípade to pre nás nie je veľký problém, nakoľko v testovaní v predchádzajúcej podkapitole si model v tejto charakteristike viedol dobre.

Druhou vlastnosťou, ktorú sme nezahŕňali v modeli, je rozdelenie počtu spoločných susedov. Pozreli sme sa na toto rozdelenie pre hrany s 1 a 2 spoločnými susedmi. Pre vyšší počet nám model vykazoval degenerovanosť. Môžeme si všimnúť v tabuľke 9, že ich zahrnutie nám model výrazne zlepšilo. Kompletné výstupy sú priložené v prílohe A.3.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model s vlastnosťami | 1181 | 1216 |
| Model sledujúci 1 spol. suseda | 1147 | 1189 |
| Model sledujúci 2 spol. susedov | 1155 | 1197 |
| Model sledujúci 1 a 2 spol. susedov | 1123 | 1172 |

**Tabuľka 9:** Porovnanie AIC a BIC pre modely sledujúce rozdelenie počtu spoločných susedov.

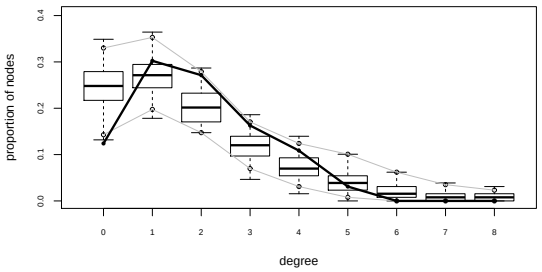
Pozrime sa ešte na alternatívu k rozdeleniu počtu spoločných susedov, a to geometricky vážené rozdelenie, ktoré sme taktiež definovali v kapitole 2. Toto rozdelenie nám pomôže zahrnúť všetky možnosti počtu spoločných susedov. Podobne, ako pre geometrické rozdelenie stupňov v sieti, aj toto rozdelenie má parameter. Taktiež tento parameter môžeme zafixovať na konkrétnu hodnotu alebo ho môžeme nechať nezafixovaný, len s počiatočnou hodnotou. Ako sme už spomínali, pri geometrickom rozdelení stupňov, v literatúrach volia parameter tak, aby sa model čo najviac zlepšil. Napríklad aj v článku [6] postupovali tak, že volili hodnoty parametra $\tau$ od 0, pokým nedostali dobrý odhad modelu.

V našej práci sme si zvolili zafixovanú hodnotu parametra $\tau$ pre toto rozdelenie 0,1 a postupne sme ho zvyšovali až po hodnotu 0,7. Pre hodnoty parametra vyššie ako 0,8 nám model nedokonvergoval. V prílohe A.3 uvádzame jednotlivé hodnoty informačných kritérií pre modely. Najlepšie hodnoty informačných kritérií sme dostali pre modely s hodnotou parametra 0,6 a 0,7, a to 1108, 1150. Rovnako, ako v prípade geometrického rozdelenia stupňov, sa pozrieme aj na možnosť nezafixovania hodnoty parametra. V prípade hodnoty nezafixovaného parametra s počiatočnou hodnotou $\tau = 0{,}6$ sme dostali hodnoty informačných kritérií 1110 a 1159, pričom model odhadol parameter na hodnotu 0,7140. Bližšie informácie k modelu sa nachádzajú v prílohe A.3. Na druhú stranu model s počiatočnou hodnotu nezafixovaného parametra 0,7 nám už nedokonvergoval.

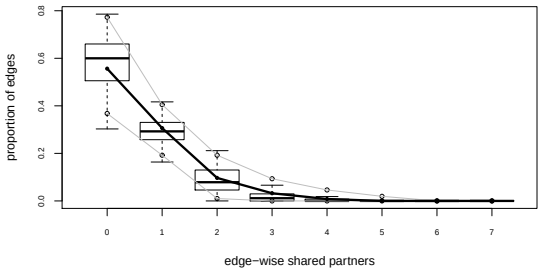
Po pozorovaní rôznych charakteristík siete, ktoré sme zahrnuli do modelov, sme dostali, že model najviac zlepšuje pozorovanie geometricky váženého rozdelenia počtu spoločných susedov pre hodnoty zafixovaného parametra 0,6 a 0,7. Pri nezafixovaní parametra nám model odhadol jeho hodnotu na približne 0,7, preto sme sa rozhodli zvoliť túto hodnotu pre zafixovaný parameter vo finálnom modeli.

Poďme model znovu otestovať, podobne ako v podkapitole 4.3. Na obrázku 18 vidíme, že v časti (*b*) sa model výrazne zlepšil. Tento efekt by sme mohli očakávať, nakoľko presne parameter sledujúci túto závislosť sme pridali do modelu. Stále nám model zaostáva v tretej testovacej premennej, ktorú do modelu už nevieme pridať. Hodnoty rozdelení pre našu sieť a aj simulované siete sú priložené v prílohe A.4. Môžeme však usúdiť, že sa model pri testovaní celkovo zlepšil.

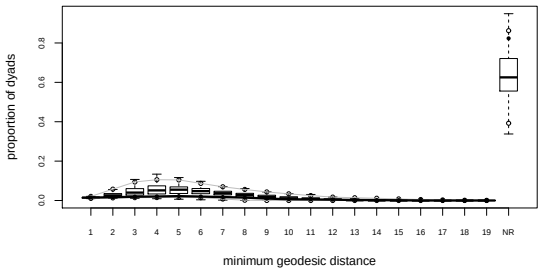
Poďme sa pozrieť aké hodnoty parametrov nám model odhadol. Pripomeňme si, že výsledný model sa skladal z premenných sledujúcich počet hrán, geometricky vážené rozdelenie počtu spoločných susedov, počty hrán medzi rovnakými údajmi o pohlaví, stave a vzniknutých faktoroch 1 a 3. Faktor 1 zahŕňal zoskupenie hudobných štýlov *rock, heavy* a *indie* a faktor 3 zahŕňal počúvanie hudobných štýlov *folk, jazz* a *classical*.




**(a)** Rozdelenie stupňov vrcholov




**(b)** Rozdelenie počtu spoločných susedov




**(c)** Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi

**Obr. 18:** Testovanie modelu s charakteristikou.

Na základe výstupu sme dostali vektor odhadnutých parametrov:

$$\theta = \begin{pmatrix} -7{,}024 \\ 0{,}999 \\ 1{,}784 \\ 0{,}486 \\ 0{,}466 \\ 0{,}514 \end{pmatrix}. \tag{43}$$

Ako sme už videli v predchádzajúcich častiach, odhadnutý parameter pre počet sietí sme dostali záporný. Ostatné odhadnuté parametre nám kladne pôsobia na logaritmus šancí hrán. Na interpretáciu parametrov sa pozrieme, už priblíženým spôsobom, a to ich vplyvom na zmenu šancí. Vieme, že logaritmus šancí pre hranu je závislý od všetkých parametrov v modeli, preto nasledovné porovnanie vplyvu daného parametra uvažujeme pri ostatných hodnotách parametrov nezmenených [16]. Pozrieme sa na hodnoty parametrov,

ktoré sledujú homofíliu. Dostávame:

- pohlavie: $\Delta_3 odds(y_{ij} = 1) = \exp(\theta_3) = \exp(1{,}78) = 5{,}95$,
- stav: $\Delta_4 odds(y_{ij} = 1) = \exp(\theta_4) = \exp(0{,}49) = 1{,}63$,
- faktor 1: $\Delta_5 odds(y_{ij} = 1) = \exp(\theta_5) = \exp(0{,}47) = 1{,}59$,
- faktor 3: $\Delta_6 odds(y_{ij} = 1) = \exp(\theta_6) = \exp(0{,}51) = 1{,}67$.

Najväčší vplyv na výslednú hodnotu má kamarátstvo medzi rovnakým pohlavím, a to faktorom 5,95. Porovnateľný vplyv má rovnaký stav, ktorému prislúcha parameter $\theta_4$ a počúvanie hudobných štýlov *rock, heavy* a *indie*, ktoré sú zahrnuté v faktore 1 a príslúcha im parameter $\theta_5$. Šance sa zvýšia približne faktorom 1,60. Posledným parametrom zahrnutým v modeli je počúvanie hudobných štýlov *folk, jazz* a *classical*. Tieto hudobné štýly sú zahrnuté vo faktore 3 a šance zvýšia faktorom 1,67.

Interpretácia parametra zahrňujúceho geometricky vážene rozdelenie počtu spoločných susedov je komplikovanejšia [12]. Zvolili sme si zafixovanú hodnotu parametra $\tau = 0{,}7$. Odhadnutý parameter vieme interpretovať na príklade, ak by sme si zobrali dve dvojice vrcholov, medzi ktorými existuje hrana. Budeme uvažovať, že tieto dvojice sa líšia iba v ich počte spoločných susedov. Pre nejaké $k$ bude mať jedna dvojica vrcholov hodnoty rozdelenia $(ep_k, ep_{k+1})$ a druhá dvojica $(ep_k - 1, ep_{k+1} + 1)$. Pozrieme sa, ako sa zmenia šance druhej dvojice voči prvej dvojici. Túto zmenu si pre potreby zapísania označme $\Delta ep_k$. Logaritmus šancí dostaneme v tvare:

$$log(odds(\Delta ep_k)) = 0{,}999 \cdot (1 - \exp(-0{,}7))^k = 0{,}999 \cdot 0{,}5^k. \tag{44}$$

Ak by počet spoločných susedov druhej dvojice narástol na 2, pričom prvá dvojica by mala jedného spoločného suseda, tak logaritmus šancí by sa zmenil faktorom $0{,}99 \cdot 0{,}5$, keďže v tomto prípade sa $k$=1. Ďalšou alternatívou by bolo, ak by počet spoločných susedov druhej dvojice bol 3, pričom prvá dvojica by mala dvoch spoločných susedov tak dostaneme $0{,}99 \cdot 0{,}5^2$ a podobne. Vidíme, ako rýchlo sa bude faktor ovplyvňujúci šance zmenšovať. Dostali sme zároveň, že všetky počty spoločných susedov budú znižovať výsledné šance.

# 5 Aplikácia modelov na dáta od priateľov na Facebooku

V tejto kapitole aplikujeme ERG modely na dáta, ktoré sme zozbierali na Facebooku. Priateľom sme dali vyplniť dotazník, ktorý sa skladal z rôznych otázok. Po vyplnení dotazníku sme si zaznačili našich spoločných priateľov, ktorých mali uverejnených. Z týchto priateľstiev nám vznikla sieť. V prvej podkapitole bližšie priblížime dáta a zároveň aj dotazník, ktorý nám priatelia vypĺňali. Následne vytvoríme modely s informáciami, ktoré sme získali z dotazníkov. Z modelov, podobne ako v kapitole 4, vyberieme na základe informačných kritérií tie, ktoré najviac zlepšovali Null model. Podľa najväčšieho zlepšenia budeme vlastnosti zahŕňať do jedného modelu, ktorý otestujeme charakteristikami siete. Z predchádzajúcej skúsenosti môžeme očakávať, že si model nebude viesť dobre, a preto sa pokúsime do modelov zahrnúť aj charakteristiky. Ku koncu kapitoly výsledný model otestujeme a pozrieme sa na odhadnuté parametre a ich interpretáciu.

## 5.1 Dáta

Na zozbieranie informácií o priateľoch sme vytvorili dotazník. Rozposielali sme ho jednotlivo kamarátom na Facebooku a vyplnilo ho z nich 202 osôb. V dotazníku sme sa pýtali na ich základné informácie ako vek, pohlavie, stav, národnosť. Venovali sme otázku aj viere. Ohľadne zamestnania nám zodpovedali informáciu, či pracujú na trvalý pracovný pomer, prípadne na čiastočný úväzok. Ďalšiu časť dotazníku sme venovali záľubám ako napríklad šport, počúvanie hudby. Dotazník sa skladal z otázky, v ktorej vpisovali odpovede:

- vek.

Pri nasledujúcich údajoch si mohli vypĺňajúci vybrať jednu odpoveď z uvedených:

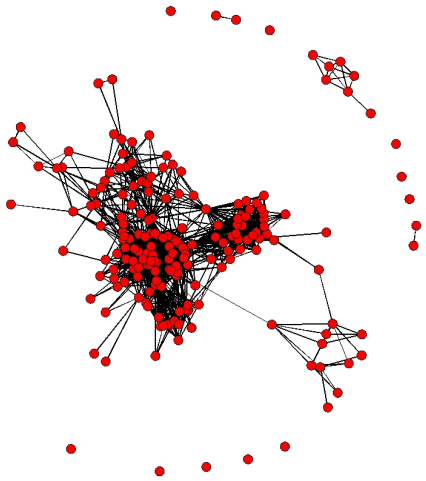
- pohlavie - muž, žena,
- stav - slobodný/á, mám priateľa/priateľku, ženatý/vydatá,
- som veriaci/a - áno, nie,
- národnosť - slovenská, česká a možnosť vpísania inej národnosti,

- fajčím - áno, nie,
- pracujem na trvalý pracovný pomer - áno, nie,
- pracujem na čiastočný úväzok - áno, nie.

V nasledujúcej skupine otázok vyberali konkrétne možnosti, ktoré priblížime neskôr v kapitolách:

- chodím na alebo zúčastňujem sa zadaných voľnočasových aktivít,
- venujem sa zadaným športom,
- počúvam zadané hudobné štýly.

Na základe zverejnených spoločných priateľov sme následne vytvorili sieť. Osoby nám budú predstavovať vrcholy a hrany ich kamarátstva medzi sebou. Túto sieť vykresľujeme na obrázku 19. Na rozdiel od kapitoly 4, v sieti máme veľa hrán, a to konkrétne 1815. Môžeme si všimnúť, že máme málo osôb, s ktorými sme nemali žiadnych spoločných kamarátov.



**Obr. 19:** Vytvorená sieť na základe dát z Facebooku.

## 5.2 Modely s charakteristikami vrcholov

Po predstavení siete a informácií, ktoré sme zozbierali od kamarátov sa pozrieme, ako tieto informácie ovplyvňujú hrany v sieti. Jednotlivo budeme zahŕňať efekt homofílie, a teda rovnakej odpovede na otázky z dotazníku, na hrany siete. Zo začiatku sa zameriame na Bernoulliho model, ktorý sleduje iba počet hrán. Následne jednotlivo zahrnieme základné údaje do modelov a pozrieme sa, ako zlepšujú prvý uvedený model, a to Bernoulliho model. Na záver časti sledujúcej základné údaje vyberieme model zoskupujúci tieto informácie. Pozrieme sa na odhadnuté parametre a ich interpretáciu. Následne priblížime, ako kamaráti odpovedali na počúvanie hudobných štýlov a pre lepšie zahrnutie do modelu využijeme faktorovú analýzu. Pri voľnočasových aktivitách sa taktiež pozrieme na jednotlivé aktivity a odpovede kamarátov. Postupne budeme zahŕňať jednotlivé aktivity do modelu. Podobne ako pri základných údajoch sa pozrieme, ako ovplyvňujú model. Na základe veľkého počtu športov, o ktorých máme informáciu sa po preskúmaní odpovedí, pozrieme na menší počet faktorov, ktoré budú reprezentovať zoskupenia športov. Na záver podkapitoly spojíme vybrané modely z každej časti a pozrieme sa na signifikantnosť premenných. Po prípadnom vyradení sledovaných premenných za účelom dosiahnutia lepšieho modelu sa pozrieme na interpretáciu koeficientov.

### 5.2.1 Bernoulliho model

Ako sme už spomenuli, ako prvé sa pozrieme na základný model. Bernoulliho model, nazývaný aj Null model, sleduje iba počet hrán. Model sme priblížili v predchádzajúcej kapitole 4 pri aplikovaní modelu na dáta od žiakov z Glasgowu. V tomto prípade máme v sieti 1815 hrán a 202 vrcholov. Po odhadnutí Bernoulliho modelu sme dostali odhadnutý parameter z rovnice (10):

$$\theta = -2{,}32. \tag{45}$$

Výstup pre celý model sa nachádza v prílohe B.1. Na pravdepodobnosti existencie hrany použijeme rovnicu (29) a dosadením dostaneme:

$$p(y_{ij} = 1) = \frac{\exp(-2{,}32)}{1 + \exp(-2{,}32)} = 0{,}0895. \tag{46}$$

Táto hodnota je už väčšia oproti hodnote pri dátach od žiakov z Glasgowu. Zväčšenie pravdepodobnosti sme mohli očakávať na základe zväčšenia pomeru medzi počtom hrán

a počtom vrcholov. Pravdepodobnosť existencie hrany v tejto sieti je 0,09. Stále to nie je veľká hodnota, nakoľko v sieti s 202 vrcholmi je $\binom{202}{2} = 20301$ možných hrán a našich 1815 hrán je z nich malá časť.

### 5.2.2 Modely sledujúce základné údaje

Rovnakým spôsobom, ako v predchádzajúcej kapitole sa pozrieme na základné údaje o kamarátoch, ktoré nám vyplnili v dotazníku. Údaje, ktoré budeme sledovať v tejto kapitole sú:

- pohlavie,
- vek,
- stav,
- viera v Boha,
- národnosť,
- práca na trvalý pracovný pomer,
- práca na čiastočný úväzok,
- fajčenie.

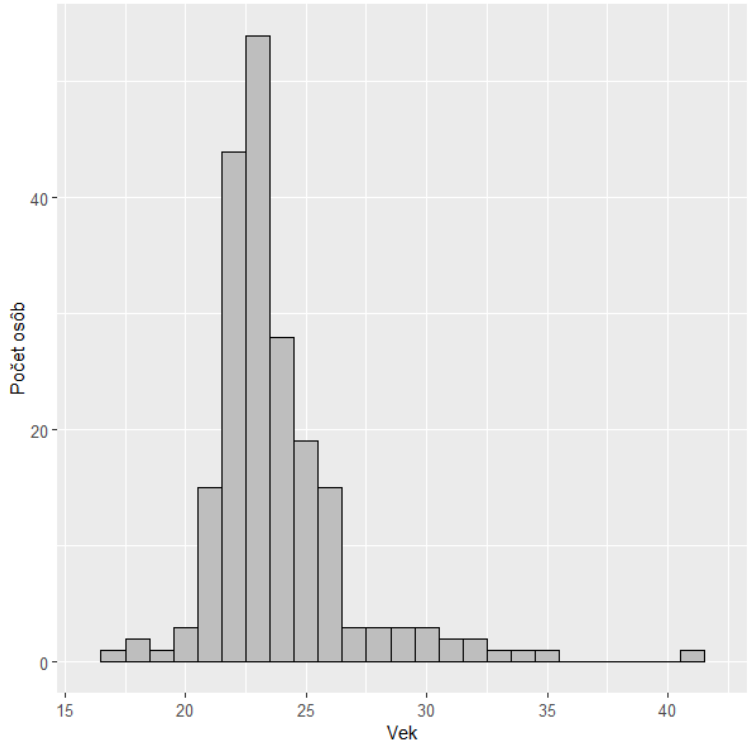
Vytvoríme postupne modely zahrňujúce údaje jednotlivo. Rovnako ako v kapitole 4, modely porovnáme pomocou informačných kritérií, aby sme vedeli vybrať vhodný model. Ak dostaneme viacero údajov, ktoré nám pomôžu zlepšiť model, budeme ich spájať a znovu sa pozerať na informačné kritériá.

Prvým parametrom, ktorý ideme sledovať, je pohlavie. V sieti máme 87 mužov a 115 žien. Zároveň máme 947 hrán z celkových 1815 medzi osobami s rovnakým pohlavím. Ak by sme rozlišovali medzi pohlaviami, tak by sme zistili, že 368 hrán je práve medzi dvomi mužmi a zvyšných 579 hrán je medzi dvomi ženami. Vytvorili sme dva modely, ktoré buď rozlišovali alebo nerozlišovali medzi pohlaviami. Výstupy z modelov sa nachádzajú v prílohe B.1. V tabuľke 10 vidíme, že pridaním parametra sa model v oboch prípadoch zhoršil, tým pádom nebudeme ďalej uvažovať sledovanie tohto parametra.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 12 230 | 12 238 |
| Model nerozlišujúci medzi pohlaviami | 12 230 | 12 246 |
| Model rozlišujúci medzi pohlaviami | 12 229 | 12 253 |

**Tabuľka 10:** Porovnanie AIC a BIC pre modely sledujúce pohlavie.

Druhou informáciou, pre ktorú otestujeme, ako vplýva na hrany v sieti, je vek. V tomto prípade máme informácie o veku v celých číslach, pričom najmladší respondent mal 17 rokov a naopak najstarší mal 41 rokov. Na obrázku 20 vidíme histogram počtu osôb vzhľadom na vek. Môžeme si všimnúť, že najviac priateľov máme rovesníkov. Ani v týchto dátach nemáme veľa hrán medzi osobami, ktoré zadali rovnaký vek, a to iba 331. Po odhadnutí modelu sme dostali hodnoty informačných kritérií 12 221 a 12 237, ktoré sú zahrnuté vo výstupe v prílohe B.1. Tieto hodnoty sú o niečo menšie v porovnaní s Bernoulliho modelom, preto ku koncu tejto podkapitoly zahrnieme vek do ďalšieho pozorovania.




**Obr. 20:** Histogram počtu osôb vzhľadom na vek.

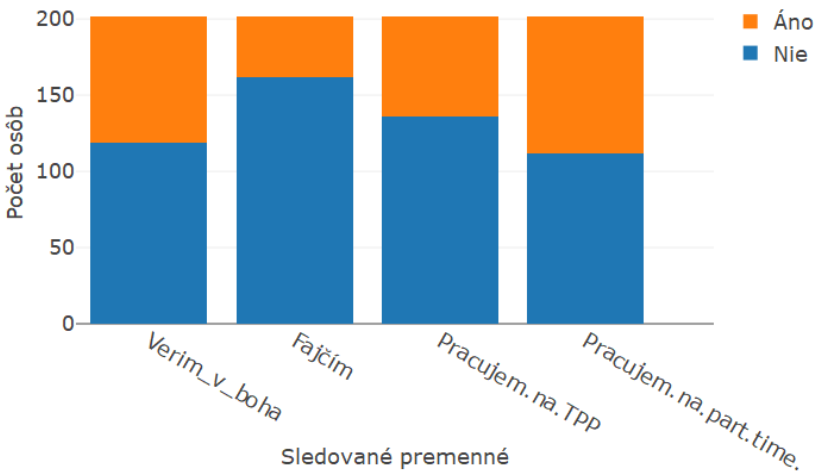
Ďalšou informáciou, ktorú nám kamaráti uviedli, je stav. V dotazníku sme rozlišovali, či je daná osoba slobodná, vo vzťahu alebo už vydatá, respektíve ženatý. Z kamarátov máme 96 osôb nezadaných, 95 osôb vo vzťahu. Vydaných, respektíve ženatých osôb máme v našom pozorovaní iba 11. Najprv vytvoríme model, ktorý nebude rozlišovať medzi stavmi. Z 1815 hrán máme 812 medzi osobami s rovnakým stavom. Zároveň, ak sa pozrieme na túto hodnotu bližšie, tak 344 kamarátstiev bolo medzi slobodnými osobami, 463 kamarátstiev medzi osobami vo vzťahu a len 5 medzi vydatými, ženatými osobami. Po vytvorení modelov si môžeme všimnúť, že hodnoty Akaikeho informačného kritéria sa zlepšili, ak sme rozlišovali medzi stavmi. Bayesovo kritérium sa ale v oboch prípadoch zhoršilo, preto údaje nebudeme ďalej sledovať. Hodnoty informačných kritérií vidíme v tabuľke 11 a celé výstupy odhadnutých modelov sa nachádzajú v prílohe B.1.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 12 230 | 12 238 |
| Model nerozlišujúci medzi stavmi | 12 232 | 12 247 |
| Model rozlišujúci medzi stavmi | 12 219 | 12 245 |

**Tabuľka 11:** Porovnanie AIC a BIC pre modely sledujúce stav.

Medzi základné údaje v tejto časti považujeme aj národnosť. Z respondentov je 187 osôb slovenskej národnosti a zvyšok ďalších štyroch rôznych. Keď sa pozrieme na sieť, dostaneme, že 1615 hrán je medzi osobami s rovnakou národnosťou, a to práve medzi osobami so slovenskou národnosťou. Počet je veľmi veľký, nakoľko 92,5% osôb je slovenskej národnosti. Z výstupu, ktorý sa nachádza v prílohe B.1 vidíme, že sme dostali hodnoty informačných kritérií 12 214 a 12 230. Dostali sme zlepšenie oproti Bernoulliho modelu, preto sa neskôr pozrieme na zahrnutie aj tohto vplyvu do modelu s viacerými parametrami.

V informáciách ohľadne viery, fajčenia, trvalého pracovného pomeru a čiastočného úväzku nám respondenti odpovedali áno, respektíve nie. Na obrázku 21 vidíme na grafe pomer jednotlivých odpovedí. Viac ako polovica osôb pri každom údaji odpovedala nie.




**Obr. 21:** Pomer odpovedí pri jednotlivých premenných.

Keď sa pozrieme bližšie na rozloženie hrán medzi rovnakými odpoveďami, dostaneme, že najviac hrán v sieti je medzi osobami, ktoré rovnako odpovedali na otázku ohľadne fajčenia, a to 1203. O niečo menej, a to 1099 hrán, je medzi osobami, ktoré sú buď zamestnané na trvalý pracovný pomer alebo obe osoby nie sú zamestnané na trvalý pracovný pomer. Taktiež viac ako polovica hrán je medzi osobami s rovnakou vierou, a to 982 hrán. Posledným údajom, na ktorý sme sa v tejto časti pozreli, je informácia o zamestnaní na čiastočný úväzok. V sieti máme 883 hrán medzi osobami, ktoré odpovedali rovnako na túto otázku. Pozrieme sa rovno na vytvorené modely a ich jednotlivé kritériá, ktoré uvádzame v nasledujúcej tabuľke 12. Do tabuľky sme pridali aj informácie o modeloch, o ktorých sme už zistili, že nám zlepšujú Null model. Celé výstupy sa nachádzajú v prílohe B.1. Z tabuľky vidíme, že model sledujúci, či osoba pracuje na trvalý pracovný pomer najviac zlepšil Null model. Neveľmi za ním zaostáva model sledujúci národnosť, preto ďalší model, ktorý vytvoríme, bude vytvorený s týmito vysvetľujúcimi premennými. Následne pridáme premennú, ktorá sleduje stav a zároveň aj rozlišuje medzi stavmi. Pozrieme sa, či pridanie premennej sledujúcej vek ešte zlepší model alebo už bude preťažený počtom premenných.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 12 230 | 12 238 |
| Model sledujúci vek | 12 221 | 12 237 |
| Model sledujúci národnosti | 12 214 | 12 230 |
| Model sledujúci vieru | 12 226 | 12 241 |
| Model sledujúci trvalý pracovný pomer | 12 213 | 12 229 |
| Model sledujúci čiastočný úväzok | 12 229 | 12 245 |
| Model sledujúci fajčenie | 12 229 | 12 245 |

**Tabuľka 12:** Porovnanie AIC a BIC pre modely sledujúce základné osobné údaje.

V tabuľke 13 vidíme, že hoci Akaikeho kritérium sa pridaním parametra sledujúceho vek zlepšilo, Bayesovo kritérium, ktoré je prísnejšie, sa už zhoršilo. Na základe Bayesovho kritéria môžeme vybrať model sledujúci informácie o trvalom pracovnom pomere a národnosti ako najlepší model.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 12 230 | 12 238 |
| Model sledujúci TPP | 12 213 | 12 229 |
| Model sledujúci TPP, národnosť | 12 200 | 12 224 |
| Model sledujúci TPP, národnosť, stav | 12 184 | 12 232 |
| Model sledujúci TPP, národnosť, stav, vek | 12 179 | 12 234 |

**Tabuľka 13:** Porovnanie AIC a BIC pre modely spájajúce základné osobné údaje.

Poďme sa pozrieť na odhadnuté parametre $\theta$ pre tento model, na základe výstupu uvedeného v prílohe B.1. Dostali sme vektor parametrov:

$$\theta = \begin{pmatrix} -2{,}70 \\ 0{,}20 \\ 0{,}29 \end{pmatrix}. \tag{47}$$

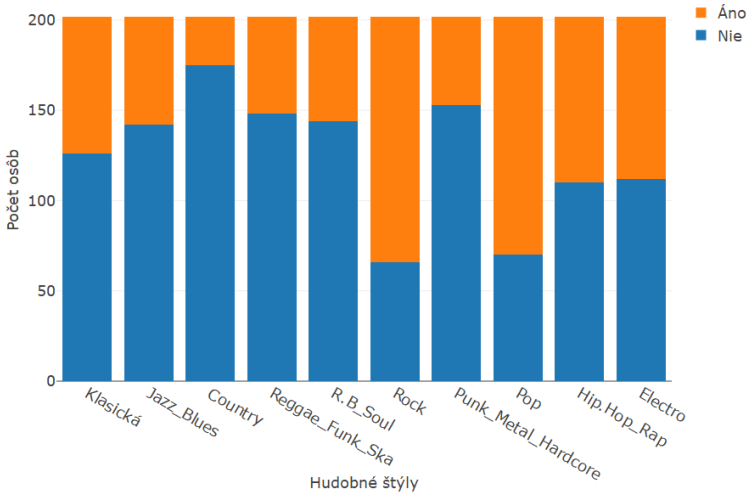
Pre parametre sledujúce homofíliu sme dostali kladné hodnoty vo vektore $\theta$. Ich hodnoty sú však malé a teda vplyv homofílie, vzhľadom na tieto vlastnosti nebude veľký. Pozrime sa, ako vplývajú tieto parametre na šance hrán. Ak obe osoby pracujú, respektíve obe osoby nepracujú na trvalý pracovný pomer, šance existencie hrany medzi nimi sa zvýšia faktorom $\exp(0{,}2) = 1{,}22$ oproti osobám s rozdielnou odpoveďou. O niečo viac ovplyvňujú šance, ak by osoby mali rovnakú národnosť, a to faktorom $\exp(0{,}29) = 1{,}34$.

### 5.2.3 Modely sledujúce hudobné preferencie

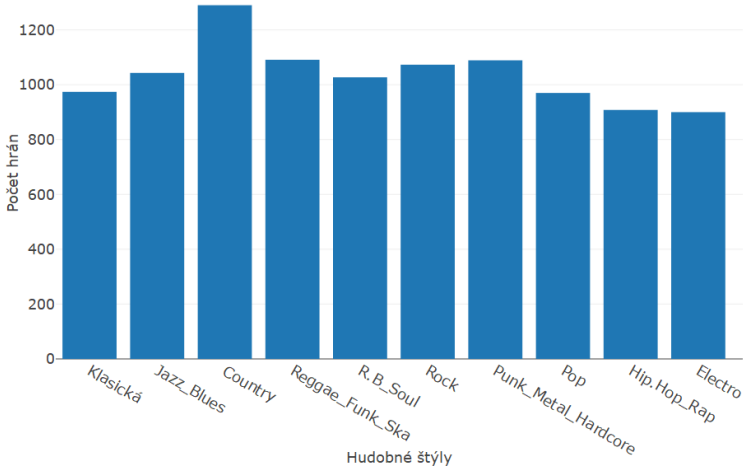
Podobne ako v predchádzajúcej kapitole aj pri týchto dátach sa pozrieme na hudobné preferencie. V dotazníku sme sa pýtali kamarátov na počúvanie viacerých hudobných štýlov. Osoby označovali, či daný štýl, respektíve skupinu štýlov počúvajú. Niektoré podobné štýly sme už zoskupili do jednej skupiny. Zoskupili sme napríklad hudobné štýly metal, punk a hardcore. Hudobné štýly ska, funk a reggea sme tiež zaradili pod jednu skupinu. V dotazníku mali kamaráti na výber z hudobných štýlov:

- klasická (vážna) hudba,
- jazz, blues,
- country,
- reggea, funk, ska,
- R&B, soul,
- rock,
- punk, metal, hardcore,
- pop,
- hip-hop, rap,
- electro - dubstep, house, techno.

Poďme sa pozrieť ako osoby odpovedali. Na obrázku 22 vidíme pomer odpovedí na jednotlivé hudobné štýly. Najviac ľudí počúva pop a rock. Na druhú stranu si môžeme všimnúť, že country už v dnešnej dobe počúva menej ľudí. Zároveň na obrázku 23 vidíme, aké máme počty hrán medzi rovnakými odpoveďami. Môžeme si všimnúť, že pri každom hudobnom štýle máme väčšinu hrán medzi osobami s rovnakou odpoveďou, tým pádom by hudobné štýly mohli ovplyvňovať kamarátstva.



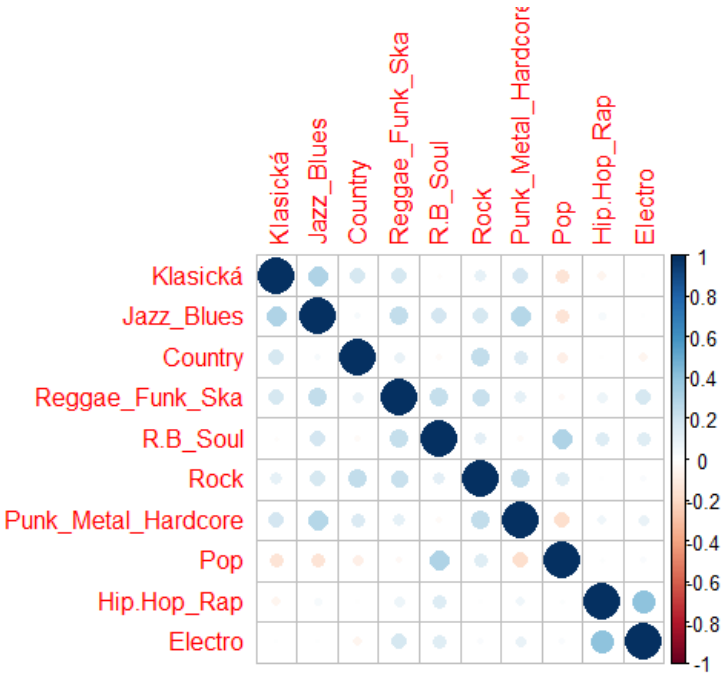

**Obr. 22:** Pomer odpovedí na počúvanie hudobných štýlov.



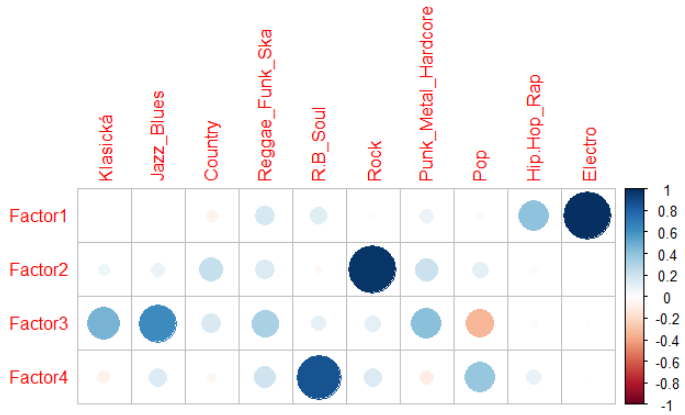

**Obr. 23:** Počty hrán medzi osobami, ktoré odpovedali rovnako pri počúvaní hudobných štýlov.

Korelácie jednotlivých hudobných štýlov medzi sebou vidíme na obrázku 24. Môžeme si všimnúť, že hip-hop spojený s rapom kladne koreluje s elektronickou hudbou. Taktiež, ako by sme mohli očakávať, klasická hudba kladne koreluje s jazzom spojeným so žánrom blues. Prekvapivá korelácia je medzi jazzom a skupinou hudobných štýlov punk, metal, hardcore. Vzhľadom na stále veľký počet hudobných štýlov, aj keď sme už niektoré štýly pozoskupovali, pozrieme sa na faktorovú analýzu nad týmito skupinami. Pomocou varimaxovej metódy sme vytvorili rôzne počty faktorov. Dostali sme, že tri faktory vysvetľujú iba 30% kumulovanej variancie v dátach. Na druhú stranu štyri faktory nám vysvetľujú už 44%. Pri zvýšení počtu faktorov na päť, sme už nedostali veľké zlepšenie, a to 47%.

Keď sa pozrieme na p-hodnoty, ktoré nám hovoria o dostačujúcnosti počtu faktorov, tak pre tri faktory sme dostali hodnotu 0,01. Pri štyroch faktoroch nám p-hodnota stúpla na 0,49. Ak by sme zvýšili počet faktorov na päť, dostali by rovnakú hodnotu, preto sme sa rozhodli vybrať štyri faktory, ktoré budú reprezentovať hudobné štýly. Na obrázku 25 vidíme korelácie faktorov s jednotlivými hudobnými štýlmi.




**Obr. 24:** Korelácia hudobných štýlov medzi sebou.




**Obr. 25:** Korelácia hudobných štýlov s faktormi.

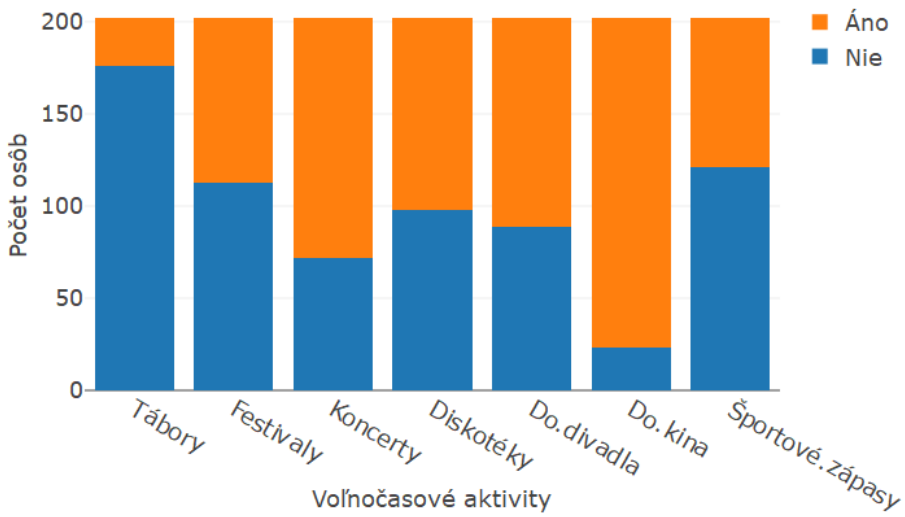
Ak zvolíme hranicu 0,4 ako najmenšiu hodnotu korelácie faktora s hudobnými štýlmi, vieme vytvoriť nasledovné zoskupenia, ktoré budú predstavovať naše faktory:

- 1. faktor: hip-hop, rap, electro,
- 2. faktor: rock,
- 3. faktor: klasická, jazz, blues, punk, metal, hardcore,
- 4. faktor: R&B, soul, pop.

Rovnako, ako v predchádzajúcej kapitole 4, faktory pridáme do výsledného modelu pred jeho testovaním.

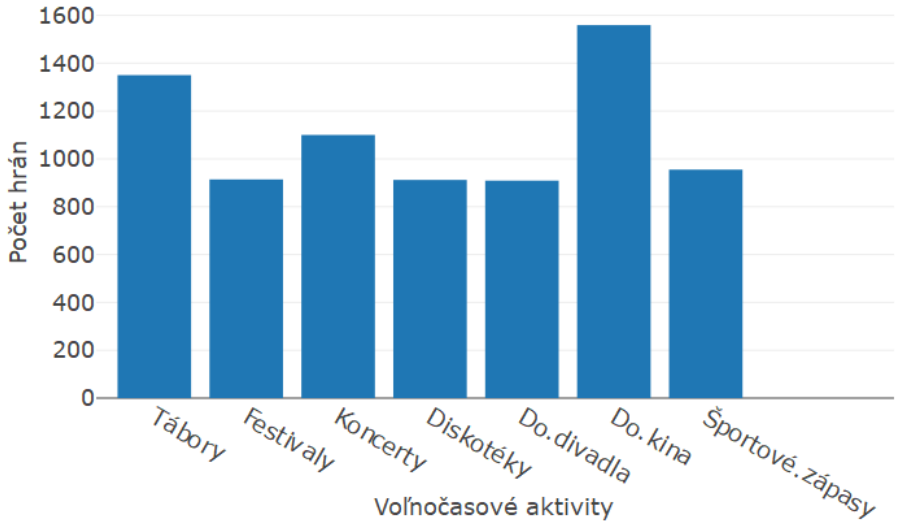
### 5.2.4 Modely sledujúce voľnočasové aktivity

Ľudia sa môžu zoznamovať aj na spoločenských a voľnočasových aktivitách. Následne pri pravidelnom stretávaní sa môžu skamarátiť a na základe toho môžu byť aj priatelia na Facebooku. V dotazníku sme sa pýtali, akých aktivít sa daná osoba zúčastňuje. Zahrnuli sme tábory, festivaly, koncerty, diskotéky, divadlo, kino a športové zápasy. Na nasledujúcom grafe vidíme pomer odpovedí pri jednotlivých aktivitách. Môžeme si všimnúť, že väčšina ľudí chodí do kina a naopak na tábory už menšia časť ľudí.




**Obr. 26:** Pomer odpovedí pri voľnočasových aktivitách.

Zároveň keď sa pozrieme na obrázok 27 vidíme v grafe počty hrán, ktoré sú medzi osobami s rovnakou odpoveďou. Môžeme si všimnúť, že takmer všetky hrany máme medzi osobami, ktoré odpovedali rovnako na otázku ohľadne chodenia do kina. O niečo menej hrán máme medzi osobami, ktoré sa rovnako zúčastňujú alebo nezúčastňujú táborov. Tieto väčšie počty hrán môžu byť následkom toho, že pri týchto aktivitách väčšina osôb odpovedala rovnako, ako sme videli na obrázku 26. Pri ostatných voľnočasových aktivitách sa približne polovica hrán nachádza medzi osobami s rovnakou odpoveďou.




**Obr. 27:** Počty hrán medzi osobami, ktoré odpovedali rovnako pri voľnočasových aktivitách.

Po predchádzajúcich grafoch už máme dobrú predstavu o tom, aké informácie o voľnočasových aktivitách máme v našej sieti. S každou aktivitou sme jednotlivo vytvorili modely, ktorých výstupy sú priložené v prílohe B.1. Poďme sa pozrieť na informačné kritériá a ako sa modely zlepšujú, prípadne zhoršujú pridaním parametra. Pri pozretí sa na tabuľku 14, v ktorej uvádzame informačné kritériá pre jednotlivé modely, si môžeme všimnúť, že pridanie parametra, ktorý sleduje kamarátstva osôb, ktoré buď chodia alebo nechodia do kina, najviac zlepšilo Null model. Za zmienku stojí aj vplyv chodenia na koncerty, ktorý taktiež výraznejšie zlepšil Null model. Pri spájaní premenných sa pozrieme aj na vplyv chodenia na tábory, nakoľko tento parameter zlepšil kritériá, aj keď nie veľmi výrazne. Ostatné premenné nám zhoršili Bayesovo informačné kritérium, tým pádom ich do modelov nezahrnieme. Poďme jednotlivé voľnočasové aktivity zoskupovať podľa najväčšieho zlepšenia Null modelu.

| | Akaikeho | Bayesovo |
|---|---|---|
| Null model | 12 230 | 12 238 |
| Model zahrňujúci účasť na táboroch | 12 221 | 12 237 |
| Model zahrňujúci účasť na festivaloch | 12 232 | 12 247 |
| Model zahrňujúci účasť na koncertoch | 12 195 | 12 211 |
| Model zahrňujúci účasť na diskotékach | 12 231 | 12 247 |
| Model zahrňujúci účasť na divadelných predstaveniach | 12 231 | 12 247 |
| Model zahrňujúci chodenie do kina | 12181 | 12 197 |
| Model zahrňujúci účasť na športových zápasoch | 12 231 | 12 247 |

**Tabuľka 14:** Porovnanie AIC a BIC pre modely sledujúce voľnočasové aktivity.

| | Akaikeho | Bayesovo |
|---|---|---|
| Model zahrňujúci chodenie do kina | 12181 | 12 197 |
| Model zahrňujúci chodenie do kina, na koncerty | 12 155 | 12 178 |
| Model zahrňujúci chodenie do kina, na koncerty, tábory | 12 148 | 12 179 |

**Tabuľka 15:** Porovnanie AIC a BIC pre modely spájajúce voľnočasové aktivity.

Na základe tabuľky 15 vidíme, že podľa kritérií si najlepšie vedie model sledujúci chodenie do kina a na koncerty. V kapitole 5.2.6 preto pridáme práve tieto aktivity do modelu. Celý výstup je uvedený v prílohe B.1.

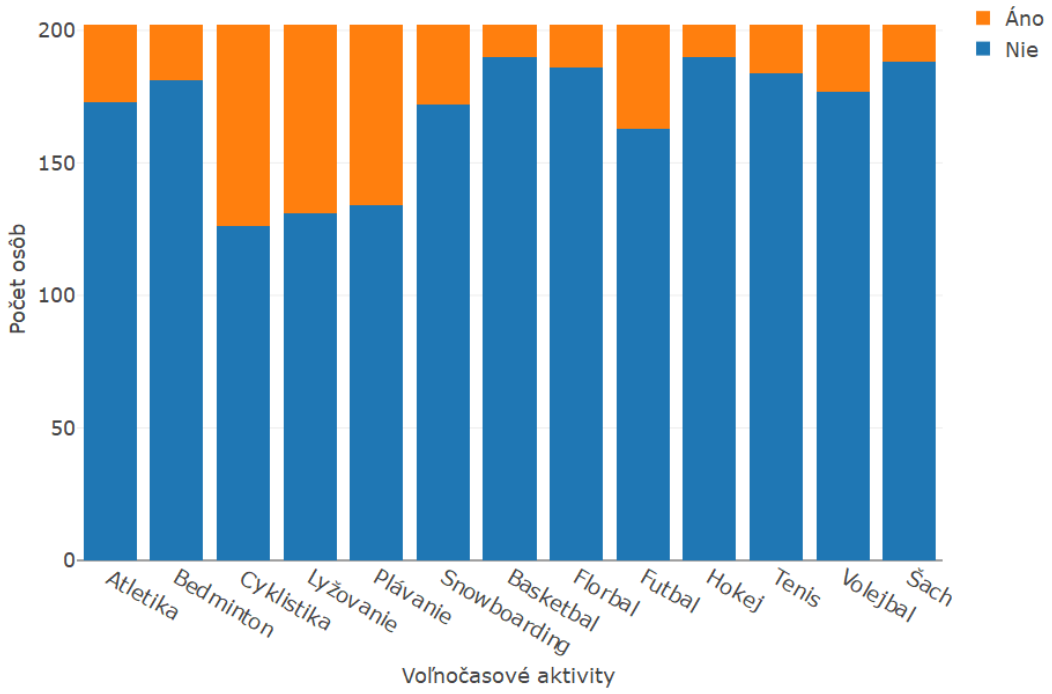
### 5.2.5 Modely sledujúce rôzne druhy športu

Podobne ako pri voľnočasových aktivitách, aj pri vykonávaní jednotlivých športov sa vedia ľudia zoznámiť, prípadne aj pravidelne stretávať. Tým pádom sa môžu skamarátiť a následnej aj spriateliť na sociálnej sieti Facebook. V dotazníku nám respondenti odpovedali, či vykonávajú dané športy, a to:

- atletika,
- bedminton,
- cyklistika,
- lyžovanie,
- plávanie,
- snowboarding,

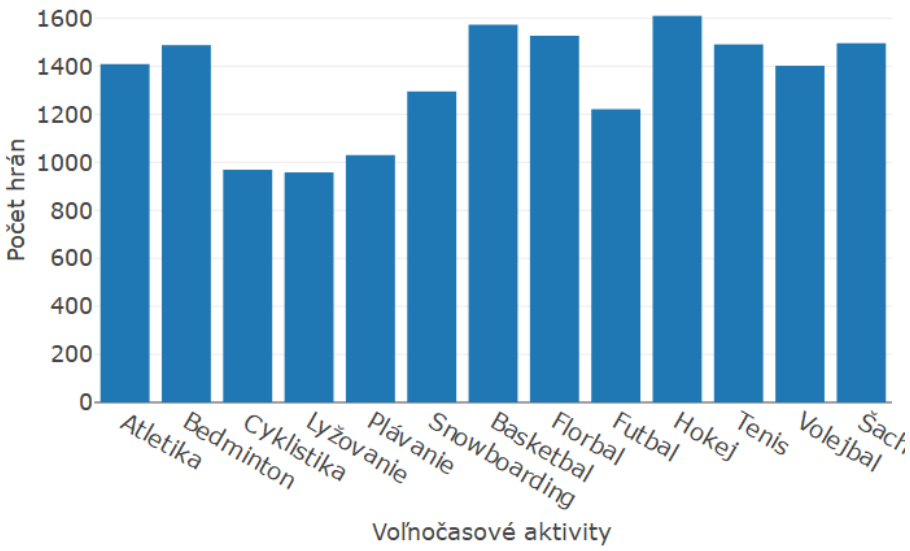
- basketbal,
- florbal,
- futbal,
- hokej,
- tenis,
- volejbal,
- šach.

Pozrime sa ako kamaráti odpovedali. Pri každom športe viac ako polovica ľudí označila, že daný šport nevykonáva. Taktiež väčšinu športov vykonáva v priemere len 20 osôb. Športy ako cyklistika, lyžovanie a plávanie najviac prevládajú medzi kamarátmi. Najviac osôb, a to konkrétne 76, sa chodí bicyklovať. Lyžovať chodí 71 osôb. Trojicu obľúbených športov nám uzatvára plávanie, ktoré vykonáva 68 osôb. Všetky počty vidíme na grafe na obrázku 28.



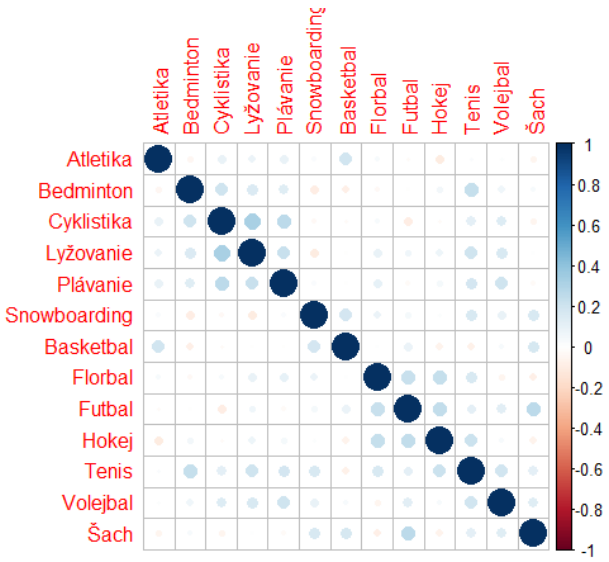

**Obr. 28:** Pomer odpovedí pri športoch.

Na obrázku 29 vidíme počty hrán medzi osobami, ktoré odpovedali rovnako. Celkom prekvapivé počty dostávame pre cyklistiku, lyžovanie a plávanie, pri ktorých najviac osôb označilo, že ich vykonáva. Práve pri týchto odpovediach sme dostali najmenšie počty hrán.




**Obr. 29:** Počty hrán medzi osobami, ktoré odpovedali rovnako pri športoch.

Vo všeobecnosti vieme, že niektoré športy sú podobné, preto sa pozrieme na korelácie medzi jednotlivými športami v našich dátach. Na obrázku 30 nevidíme extrémne korelácie, ale môžeme si všimnúť, že kladná korelácia je už medzi spomínanými športami, a to cyklistikou, lyžovaním a plávaním. Druhá skupina športov, ktoré medzi sebou kladne korelujú, sa skladá z florbalu, futbalu a hokeja.




**Obr. 30:** Korelácie medzi jednotlivými športami.

Pozrime sa na varimaxovú metódu faktorovej analýzy. Pozorovanie správania sa počtu faktorov a p-hodnoty vidíme na obrázku 31 v časti $(a)$. Táto hodnota nám hovorí, ako dostačujúci je počet faktorov. Najväčšiu p-hodnotu dostávame pri piatich faktoroch. Zároveň, ak sa pozrieme na kumulovanú varianciu dostaneme, že dva faktory zachytávajú 17%. Pre tri faktory dostávame 23%, pre štyri faktory už 28%. Pre päť faktorov nám hodnota stúpne na 33% a pri najväčšom uvažovanom počte faktorov, a to šiestich faktoroch dostaneme 39% kumulovanej variancie. Aj keď sa nám kumulovaná variancia zväčšovala, model podľa p-hodnoty usúdil, že päť faktorov je dostačujúcich, preto si zvolíme päť faktorov, ktoré budeme vytvárať zoskupovaním hudobných štýlov. Na obrázku 31 v časti $(b)$ vidíme už korelácie jednotlivých športov s vytvorenými faktormi.




**(a)** Graf závislosti počtu faktorov a p-hodnoty

**(b)** Korelácia medzi faktormi a jednotlivými športami

**Obr. 31:** Faktory a ich korelácie na základe športov.

Na základe korelácií vieme vytvoriť nasledovné faktory:

- 1. faktor: cyklistika, lyžovanie, plávanie,
- 2. faktor: florbal, futbal, hokej,
- 3. faktor: futbal, šach,
- 4. faktor: basketbal,
- 5. faktor: snowboarding, tenis.

Môžeme si všimnúť zaujímavý faktor 3, ktorý zahŕňa futbal a šach. Naopak, faktor 2 nás neprekvapuje, nakoľko zahŕňa podobné športy. Niektoré športy nám týmto štýlom vypadnú a nebudeme ich nezahŕňať do sledovania. V nasledujúcej kapitole využijeme tieto faktory a zahrnieme ich spolu s ďalšími parametrami do modelu.

### 5.2.6 Spojenie modelov

Ako sme už spomenuli, v tejto časti vytvoríme model na základe predchádzajúcich podkapitol a zistení, ktoré parametre máme zahrnúť do modelu. Zároveň sa podobne ako kapitole 4 pozrieme na signifikantnosť parametrov. Uvidíme, či vynechanie niektorých parametrov, ktoré by mohli komplikovať model, nám naopak model zlepší.

Na základe podkapitoly 5.2.2, v ktorej sme sledovali základné údaje, do modelu zahrnieme parameter sledujúci informáciu o trvalom pracovnom pomere a národnosti. Z podkapitoly 5.2.4, v ktorej sme sledovali vplyv voľnočasových aktivít, sme zistili, že nám zlepšuje model sledovanie, či chodí človek do kina a na koncerty. Z faktorovej analýzy na základe hudobných štýlov z podkapitoly 5.2.3 sme dostali štyri faktory. Následne z faktorovej analýzy v podkapitole 5.2.5 sme dostali päť faktorov, vytvorených zo športov, ktoré taktiež zahrnieme do modelu. Po odhadnutí modelu dostávame výstup:

```
==========================
Summary of model fit
==========================

Formula:   net ~ edges + nodematch("Pracujem.na.TPP.") + nodematch("Narodnost") +
    nodematch("music.faktor1") + nodematch("music.faktor2") +
    nodematch("music.faktor3") + nodematch("music.faktor4") +
    nodematch("Do.kina") + nodematch("Koncerty") +
    nodematch("sport.faktor1") + nodematch("sport.faktor2") +
    nodematch("sport.faktor3") + nodematch("sport.faktor4") +
    nodematch("sport.faktor5")

Iterations:  4 out of 20

Monte Carlo MLE Results:
                           Estimate Std. Error MCMC % z value Pr(>|z|)
edges                     -3.038250   0.141694      0 -21.442  < 1e-04 ***
nodematch.Pracujem.na.TPP.  0.186444   0.050690      0   3.678 0.000235 ***
nodematch.Narodnost         0.341220   0.078620      0   4.340  < 1e-04 ***
nodematch.music.faktor1     0.054419   0.049466      0   1.100 0.271276
nodematch.music.faktor2     0.140246   0.050464      0   2.779 0.005451 **
nodematch.music.faktor3     0.035402   0.049463      0   0.716 0.474167
```

```
nodematch.music.faktor4    -0.075890   0.050011      0  -1.517 0.129151
nodematch.Do.kina           0.458677   0.070808      0   6.478  < 1e-04 ***
nodematch.Koncerty          0.263388   0.050812      0   5.184  < 1e-04 ***
nodematch.sport.faktor1    -0.002838   0.049586      0  -0.057 0.954367
nodematch.sport.faktor2    -0.165014   0.065270      0  -2.528 0.011466 *
nodematch.sport.faktor3     0.065925   0.067392      0   0.978 0.327963
nodematch.sport.faktor4    -0.177524   0.074621      0  -2.379 0.017360 *
nodematch.sport.faktor5    -0.133727   0.051865      0  -2.578 0.009928 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 28143  on 20301  degrees of freedom
 Residual Deviance: 12079  on 20287  degrees of freedom

AIC: 12107    BIC: 12217    (Smaller is better.)
```

Vidíme, že iba faktor 2 označil model za signifikatný voči ostatným faktorom vytvorených z hudobných štýlov. Z faktorov, ktoré sú vytvorené na základe rôznych športov nám model označil za nesignifikantné faktor 1 a 3. Poďme sa pozrieť, čo sa stane ak vynecháme parametre, ktoré nám model neoznačil za signifikantné. Ponecháme si sledovanie homofílie vzhľadom na pracovný pomer, národnosť, faktor 2 vytvorený na základe hudobných štýlov, chodenie do kina, na koncerty a faktory 2, 4, 5, vytvorené na základe športov. Pripomeňme si, že faktor 2 sa skladal iba z hudobného štýlu rock. Pri faktoroch vytvorených zoskupovaním športov, sa faktor 2 skladal z florbalu, futbalu a hokeja. Faktor 4 zahŕňal basketbal a faktor 5 snowboarding a tenis. Po odhadnutí modelu dostávame výstup, na ktorom si môžeme všimnúť, že vynechaním parametrov sa nám model o niečo zlepšil.

```
==========================
Summary of model fit
==========================

Formula:   net ~ edges + nodematch("Pracujem.na.TPP.") + nodematch("Narodnost") +
    nodematch("music.faktor2") + nodematch("Do.kina") +
    nodematch("Koncerty") + nodematch("sport.faktor2") +
    nodematch("sport.faktor4") + nodematch("sport.faktor5")

Iterations:  5 out of 20

Monte Carlo MLE Results:
                          Estimate Std. Error MCMC % z value Pr(>|z|)
edges                     -3.02038    0.13270      0 -22.761  < 1e-04 ***
nodematch.Pracujem.na.TPP.  0.18537    0.05059      0   3.664 0.000248 ***
nodematch.Narodnost         0.33793    0.07859      0   4.300  < 1e-04 ***
nodematch.music.faktor2     0.13815    0.05036      0   2.743 0.006088 **
```

```
nodematch.Do.kina            0.45245    0.07061       0   6.408  < 1e-04 ***
nodematch.Koncerty           0.26602    0.05070       0   5.247  < 1e-04 ***
nodematch.sport.faktor2     -0.12199    0.05036       0  -2.422 0.015420 *
nodematch.sport.faktor4     -0.16767    0.07394       0  -2.268 0.023341 *
nodematch.sport.faktor5     -0.13305    0.05183       0  -2.567 0.010253 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 28143  on 20301  degrees of freedom
 Residual Deviance: 12083  on 20292  degrees of freedom

AIC: 12101    BIC: 12173    (Smaller is better.)
```

Pozrime sa na odhadnuté koeficienty $\theta$. Taktiež ako v predchádzajúcej kapitole sme dostali záporný odhadnutý koeficient $\theta_1$, pre parameter sledujúci hrany. Mohlo by to byť prekvapujúce, nakoľko už máme v sieti väčší počet hrán, avšak stále je to menej ako 9% všetkých možných hrán. Záporné koeficienty $\theta_7, \theta_8, \theta_9$ nám vyšli aj pri faktoroch vytvorených zo športov. Hodnoty týchto koeficientov sú však celkom malé, preto sa poďme pozrieť na ich vplyv na zmenu šancí. Dostávame:

- faktor 2: $\Delta_7 odds(y_{ij} = 1) = \exp(\theta_7) = \exp(-0{,}12) = 0{,}89$,
- faktor 4: $\Delta_8 odds(y_{ij} = 1) = \exp(\theta_8) = \exp(-0{,}17) = 0{,}84$,
- faktor 5: $\Delta_9 odds(y_{ij} = 1) = \exp(\theta_9) = \exp(-0{,}13) = 0{,}88$.

Šance na kamarátstvo dvoch osôb, ktoré vykonávajú rovnaké športy mierne klesajú. Teda dostávame opačný vplyv od homofílie. Model však tieto parametre neoznačil sa výrazne signifikantné, tým pádom neovplyvňujú model vo veľkej miere. Pozrime sa na parametre pri ktorých nám vyšli odhady $\theta_i$ kladné. Najväčší koeficient sme dostali pre parameter sledujúci, či osoby chodia do kina, a to $\theta_5 = 0{,}45$. Druhým najväčším koeficientom je $\theta_3 = 0{,}34$, ktorý prislúcha parametru sledujúceho národnosť. O niečo menšiu hodnotu má koeficient $\theta_6 = 0{,}27$, ktorý patrí k parametru sledujúceho efekt homofílie vzhľadom na účasť na koncertoch. Následne nám už zostali najnižšie koeficienty $\theta_1 = 0{,}19$ a $\theta_4 = 0{,}14$, ktoré veľmi nebudú ovplyvňovať šance. Pozrime sa na jednotlivé hodnoty zmeny šancí:

- práca na TPP: $\Delta_2 odds(y_{ij} = 1) = \exp(\theta_2) = \exp(0{,}19) = 1{,}21$,
- národnosť: $\Delta_3 odds(y_{ij} = 1) = \exp(\theta_3) = \exp(0{,}34) = 1{,}40$,

- hudobný faktor 2: $\Delta_4 odds(y_{ij} = 1) = \exp(\theta_4) = \exp(0{,}14) = 1{,}15$,
- chodenie do kina: $\Delta_5 odds(y_{ij} = 1) = \exp(\theta_5) = \exp(0{,}45) = 1{,}57$,
- chodenie do divadla: $\Delta_6 odds(y_{ij} = 1) = \exp(\theta_6) = \exp(0{,}27) = 1{,}31$.

Vidíme, že najväčší vplyv, a teda najviac narastú šance osôb, ktoré chodievajú, respektíve nechodievajú do kina, avšak narastú iba faktorom približne 1,6. Žiaden z parametrov viac nezvyšuje šance hrán. V nasledujúcej kapitole tento model otestujeme vlastnosťami, ktoré sme do modelu nezahrnuli a pozrieme sa, ako dobre interpretuje našu sieť.

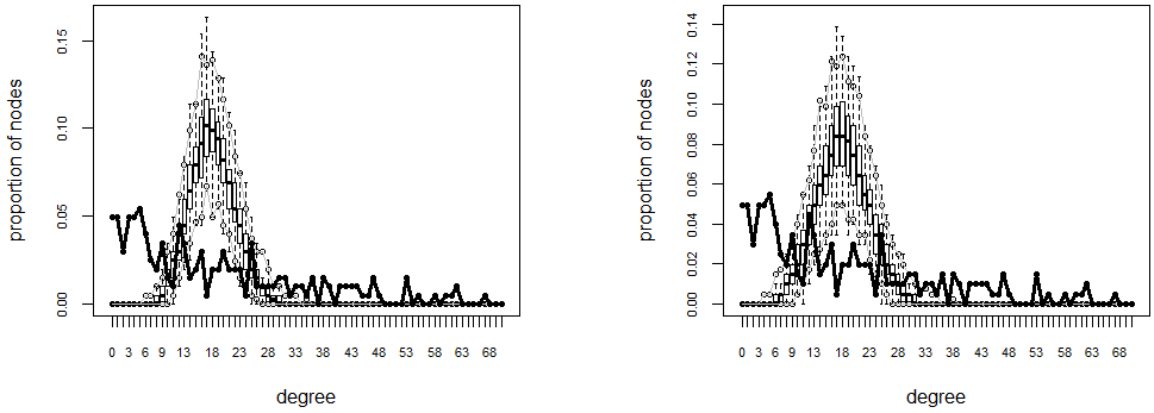
## 5.3 Testovanie modelu

Podobne ako v dátach od žiakov z Glasgowu, model otestujeme tromi vlastnosťami siete, ktoré sme doteraz nezahrnuli do modelu. Taktiež porovnáme hodnoty rozdelení pre Null model a model, ktorý sme vybrali v podkapitole 5.2.6. V tejto a nasledujúcej podkapitole ho budeme nazývať *náš model*.

Ako prvé sa pozrieme na rozdelenie stupňov vrcholov v sieti. Na obrázku 32 vidíme, že máme veľkú škálu stupňov v sieti a žiaden z nich výrazne neprevláda. Najväčší stupeň máme 67 a tento stupeň má v sieti práve jeden vrchol. Zároveň vidíme, že si model s touto vlastnosťou veľmi neporadil. Pripomeňme si, že boxploty v grafoch predstavujú hodnoty rozdelenia pre 100 simulovaných sietí. Krivka spájajúca body pre každý stupeň rozdelenia hovorí o hodnotách rozdelenia v našej sieti. Nakoľko sa grafy pre náš model a Null model veľmi nelíšia, v prílohe B.2 uvádzame hodnoty tohto rozdelenia a aj simulovaných sietí, ktoré využívame na testovanie, iba pre náš model.

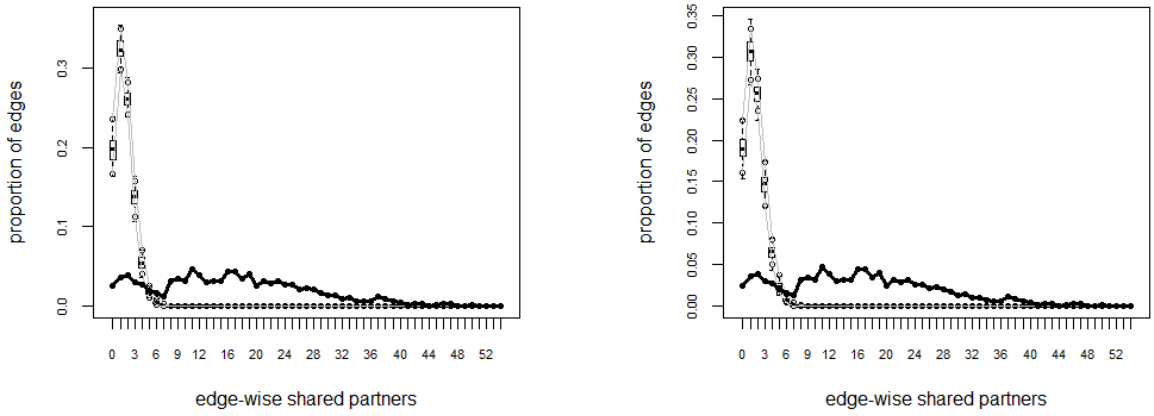
Druhou vlastnosťou, ktorou otestujeme model, je rozdelenie počtu spoločných susedov. Na obrázku 33 vidíme, že aj toto rozdelenie máme v našej sieti veľmi rôznorodé. V sieti máme zastúpené hrany, ktoré majú od 0 až po 51 spoločných susedov. Taktiež si môžeme všimnúť, že žiadne počty spoločných susedov výrazne neprevyšujú. V simulovaných sieťach sa však najviac vyskytovali počty spoločných susedov od 0 po 4. Keď sa pozrieme na konkrétne hodnoty, ktoré sú priložené v prílohe B.2, v našej sieti máme 45 hrán, ktoré nemajú spoločného suseda a v simulovaných sieťach sa v priemere takýchto hrán nachádzalo 347. Ešte väčšiu odchýlku dostávame pre hrany, ktoré majú jedného spoločného suseda. V simulovaných sieťach sa v priemere nachádzalo 557 takýchto hrán z

reálnych 66. Taktiež môžeme konštatovať, že model túto vlastnosť nezachytil na základe parametrov, ktoré zahŕňajú sledovanie homofílie.




**(a)** Null model **(b)** Náš model

**Obr. 32:** Rozdelenie stupňov vrcholov v sieťach.

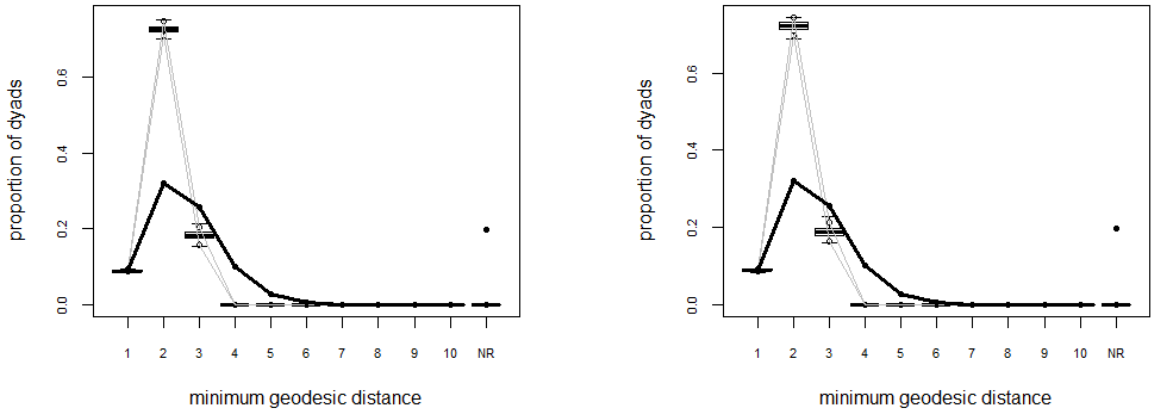



**(a)** Null model **(b)** Náš model

**Obr. 33:** Rozdelenie počtu spoločných susedov v sieťach.

Poslednou vlastnosťou je rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi. Správanie sa sietí pre toto rozdelenie vidíme na obrázku 34. Oproti predchádzajúcim vlastnostiam si náš model vedie lepšie, avšak v porovnaní s Null modelom nevidíme výraznú zmenu. Pripomeňme si, že hodnota $NR$ v grafe predstavuje počet dvojíc vrcholov, ktoré

nemajú spoločnú cestu. Zaujímavé pozorovanie je, že v simulovaných sieťach sa žiadna takáto dvojica vrcholov nenachádzala, pričom v našej sieti máme 3988 takýchto dvojíc.



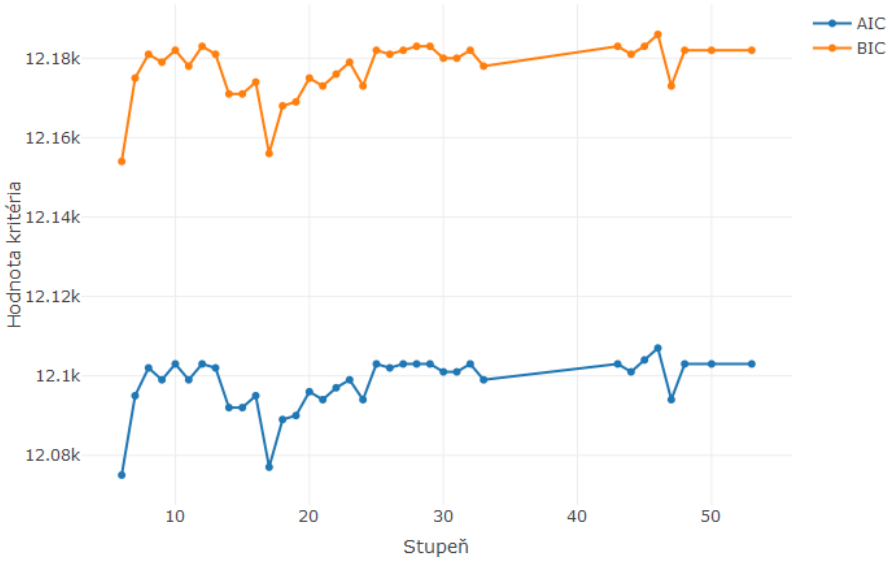

**(a)** Null model

**(b)** Náš model

**Obr. 34:** Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi v sieťach.

## 5.4 Modely s charakteristikami siete

V predchádzajúcej podkapitole sme videli, aké máme rozdelenie stupňov v našej sieti. Taktiež sme videli, že si náš model v tejto charakteristiky neviedol veľmi dobre. Postupne sme sa snažili zahrnúť tieto stupne do modelu. Pri prvých 5 stupňoch nám model vykazoval degenerovanosť, a to tak, že model nedokonvergoval alebo vykazoval možnú koreláciu s inými parametrami. Pre rôznorodosť rozdelenia sme sa postupne pozreli na rôzne stupne, pri ktorých nám model dokonvergoval. Na obrázku 35 vidíme hodnoty kritérií pre modely s jednotlivými stupňami zobrazené v grafe. Môžeme si všimnúť, že pridanie stupňov 6 a 17 nám najviac zlepšilo model, a to na hodnoty Akaikeho kritéria 12 075, 12 077 a Bayesovho kritéria 12 154, 12 156. Modely sme si následne zoradili podľa toho, ako zlepšovali model a postupne zahŕňali stupne do spoločného modelu. Ako sa správali informačné kritériá pri zoskupovaní stupňov do modelu vidíme v nasledujúcej tabuľke 16.



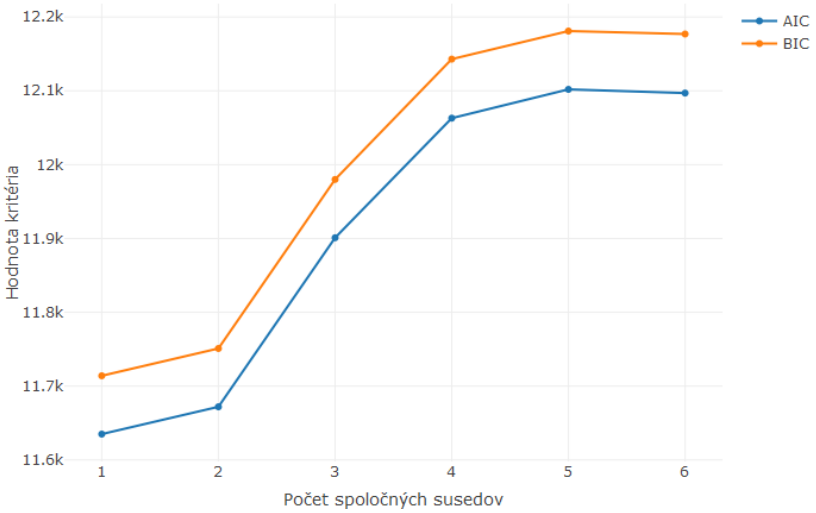

**Obr. 35:** Hodnoty informačných kritérií pre modely s jednotlivými stupňami zobrazené v grafe.

| | Akaikeho | Bayesovo |
|---|---|---|
| Náš model | 12 101 | 12 173 |
| Model sledujúci stupeň 6 | 12 075 | 12 154 |
| Model sledujúci stupne 6, 17 | 12 051 | 12 138 |
| Model sledujúci stupeň 6, 17, 18 | 12 036 | 12 131 |
| Model sledujúci stupeň 6, 17, 18, 19 | 12 019 | 12 122 |
| Model sledujúci stupeň 6, 17, 18, 19, 14, 15 | 11 984 | 12 103 |
| Model sledujúci stupeň 6, 17, 18, 19, 14, 15, 21, 24 | 11 953 | 12 087 |
| Model sledujúci stupeň 6, 17, 18, 19, 14, 15, 21, 24, 16 | 11 924 | 12 067 |
| Model sledujúci stupeň 6, 17, 18, 19, 14, 15, 21, 24, 16, 20 | 11 886 | 12 037 |

**Tabuľka 16:** Porovnanie AIC a BIC pre modely spájajúce stupne.

Pri zahŕňaní ďalších stupňov do modelu sme dostali buď koreláciu alebo nám model nedokonvergoval. Zároveň, aj keď jednotlivé možnosti zlepšovali model, takéto zoskupovanie nemá interpretáciu. Rozdelenie stupňov vrcholov v našej sieti je tak špecifické, že aj pri alternatívnych rozdeleniach sledujúcich stupne v sieti sme dospeli k problémom.

Poďme sa pozrieť na druhú vlastnosť, a to rozdelenie počtu spoločných susedov. Ako sme videli na obrázku 33 v časti ($b$) pri testovaní modelu, v našej sieti máme toto rozdelenie veľmi rôznorodé. Podobne ako pri rozdelení stupňov v sieti sme do modelu zahŕňali jednotlivo rôzne počty spoločných susedov. Pre počty spoločných susedov vyšších ako 6 nám model vykazoval degenerovanosť. Môžeme si avšak všimnúť, že práve pre tieto počty spoločných susedov, pri ktorých model nemal problém s degeneráciou, sme pri testovaní modelu dostali veľmi odlišné hodnoty. Na obrázku 36 vidíme ako zahrnutie týchto premenných zlepšilo model.




**Obr. 36:** Hodnoty informačných kritérií pre modely s počtami spoločných susedov.

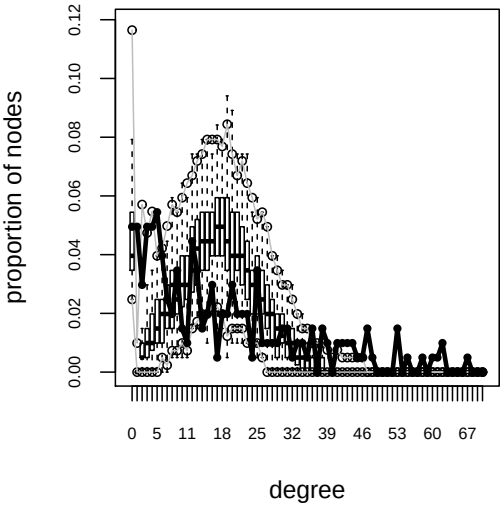
Najviac model zlepšilo pridanie sledovania počtu hrán, ktoré majú jedného spoločného suseda. Pri sledovaní vyšších stupňov sa už hodnoty informačných kritérií zvyšovali. Keď sme sa pokúsili jednotlivé počty spoločných susedov zahŕňať do spoločného modelu, model mal problém dokonvergovať. Rozhodli sme sa preto pozrieť na geometrické rozdelenie počtu spoločných susedov, ktoré zahŕňa celé rozdelenie a všetky možné hodnoty počty spoločných susedov. Toto rozdelenie taktiež obsahuje parameter $\tau$, ktorý sme podobne ako v predchádzajúcej kapitole 4 najprv zvolili zafixovaný parameter s hodnotou 0.1 a postupne zvyšovali. V tabuľke 17 vidíme ako sa hodnoty informačných kritérií zmenšovali zvyšovaním parametra $\tau$. Vidíme, že zvyšovaním parametra sa model zlepšuje, avšak pre ešte väčšie hodnoty zafixovaného parametra $\tau$ nám model nedokonvergoval. Modely s nezafixovaným parametrom ani neuvádzame, nakoľko nám každý z nich vykazoval známky degenerácie. Pri pokuse kombinovať rôzne charakteristiky sietí do jedného modelu sme

dostali buď problém s koreláciou už po prvých iteráciách alebo model nedokonvergoval. Ak sa pozrieme na informačné kritériá všetkých modelov, ktoré obsahovali pridanú charakteristiku siete, najviac zlepšilo model pridanie parametra sledujúceho geometrické rozdelenie počtu spoločných susedov s parametrom $\tau = 0{,}3$. Tým pádom, môžeme tento model vybrať ako najlepší zo sledovaných modelov.

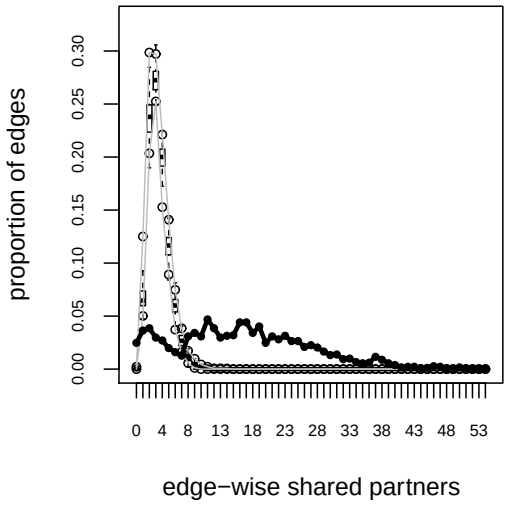
| | Akaikeho | Bayesovo |
|---|---|---|
| Náš model | 12 101 | 12 173 |
| Model s hodnotou $\tau = 0{,}1$ | 11 553 | 11 633 |
| Model s hodnotou $\tau = 0{,}2$ | 11 328 | 11 407 |
| Model s hodnotou $\tau = 0{,}3$ | 11 106 | 11 185 |

**Tabuľka 17:** Porovnanie AIC a BIC pre modely sledujúce gwesp s rôznymi hodnotami parametra $\tau$.

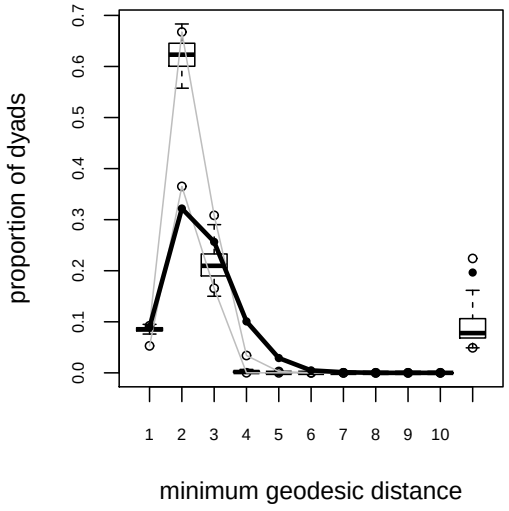
Poďme však model otestovať a pozrieť sa, či sa zlepší aj vzhľadom na tri charakteristiky siete oproti *nášmu modelu*. Na obrázku 37 vidíme jednotlivé charakteristiky a hodnoty pre nový model a simulované siete. Pri prvom pohľade na grafy sa nám môže zdať, že sa model nezlepšil. Ak sa pozrieme bližšie na hodnoty jednotlivých charakteristík uvidíme rozdiel. Pozrime sa na rozdelenie stupňov vrcholov v sieti. V sieťach vytvorených na základe modelu bez zahrnutej charakteristiky, ktorej graf sa nachádza na obrázku 32 v časti $(b)$, sa prvé stupne rozdelenia takmer nenachádzali. V týchto simulovaných sieťach na základe nového modelu sa už nachádzajú vo väčšom množstve aj keď nie dostačujúcom. Všetky hodnoty pre nový model sa nachádzajú v prílohe B.4 a pre model bez charakteristiky siete v prílohe B.2. Pri rozdelení počtu spoločných susedov, zobrazenom na obrázku 37 v časti $(b)$, sa model nezlepšil o veľa, čo je celkom prekvapivé, nakoľko sme pridali parameter sledujúci práve túto charakteristiku. Napríklad počet hrán, ktoré majú jedného spoločného suseda sa nachádzalo v simulovaných sieťach pri testovaní modelu bez charakteristiky v priemere 557. V tomto prípade sa počet hrán zmenšil na 113, avšak v sieti máme len 66 takýchto hrán. Naopak pri počte hrán, ktoré majú troch spoločných susedov sa model zhoršil a nachádzalo sa v sieti v priemere 458 takýchto hrán a v sieťach, vytvorených na základe modelu bez charakteristiky siete sa ich v priemere nachádzalo 268.




**(a)** Rozdelenie stupňov vrcholov




**(b)** Rozdelenie počtu spoločných susedov




**(c)** Rozdelenie pre najmenšiu vzdialenosť medzi vrcholmi

**Obr. 37:** Testovanie modelu s charakteristikou.

Pri poslednej charakteristike, a to rozdelení najmenšej vzdialenosti medzi vrcholmi, sme dostali zaujímavé výsledky. Model bez charakteristiky si viedol pre vzdialenosť 1 dobre a nový model sa zhoršil. Vo väčších vzdialenostiach sa naopak nový model zlepšil

proti modelu bez charakteristiky. Na základe týchto pozorovaní môžeme usúdiť, že model sa v niektorých stránkach zlepšil, ale aj zhoršil.

Poďme sa pozrieť na odhadnuté parametre a jednotlivé premenné, ktoré model sledoval. Vo výstupe z programu *R* dostávame:

```
==========================
Summary of model fit
==========================

Formula:   net ~ edges + gwesp(0.3, fixed = T) + nodematch("Pracujem.na.TPP.") +
    nodematch("Narodnost") + nodematch("music.faktor2") +
    nodematch("Do.kina") + nodematch("Koncerty") +
    nodematch("sport.faktor2") + nodematch("sport.faktor4") +
    nodematch("sport.faktor5")

Iterations:  20 out of 60

Monte Carlo MLE Results:
                          Estimate Std. Error MCMC % z value Pr(>|z|)
edges                     -9.17781    0.30398      0 -30.192  < 1e-04 ***
gwesp.fixed.0.3            4.51305    0.22478      0  20.078  < 1e-04 ***
nodematch.Pracujem.na.TPP.  0.10929    0.03960      0   2.760  0.00578 **
nodematch.Narodnost         0.09886    0.04292      1   2.304  0.02125 *
nodematch.music.faktor2     0.08414    0.04213      0   1.997  0.04580 *
nodematch.Do.kina           0.12762    0.04220      0   3.024  0.00249 **
nodematch.Koncerty          0.16634    0.04020      0   4.138  < 1e-04 ***
nodematch.sport.faktor2    -0.06833    0.03727      0  -1.834  0.06671 .
nodematch.sport.faktor4    -0.05790    0.04292      1  -1.349  0.17734
nodematch.sport.faktor5    -0.05835    0.03499      0  -1.667  0.09542 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null Deviance: 28143  on 20301  degrees of freedom
Residual Deviance: 11086  on 20291  degrees of freedom

AIC: 11106    BIC: 11185    (Smaller is better.)
```

Vidíme, že pri parametroch sledujúcich homofíliu sme dostali veľmi nízke odhady parametrov. Keď sa pozrieme ako ovplyvňujú šance dostaneme:

- práca na TPP: $\Delta_3 odds(y_{ij} = 1) = \exp(\theta_3) = \exp(0{,}11) = 1{,}12$,
- národnosť: $\Delta_4 odds(y_{ij} = 1) = \exp(\theta_4) = \exp(0{,}10) = 1{,}11$,
- hudobný faktor 2: $\Delta_5 odds(y_{ij} = 1) = \exp(\theta_5) = \exp(0{,}08) = 1{,}08$,
- chodenie do kina: $\Delta_6 odds(y_{ij} = 1) = \exp(\theta_6) = \exp(0{,}13) = 1{,}14$,
- chodenie na koncerty: $\Delta_7 odds(y_{ij} = 1) = \exp(\theta_7) = \exp(0{,}17) = 1{,}19$,
- športový faktor 2: $\Delta_8 odds(y_{ij} = 1) = \exp(\theta_8) = \exp(-0{,}07) = 0{,}93$,
- športový faktor 4: $\Delta_9 odds(y_{ij} = 1) = \exp(\theta_9) = \exp(-0{,}06) = 0{,}94$,
- športový faktor 5: $\Delta_{10} odds(y_{ij} = 1) = \exp(\theta_{10}) = \exp(-0{,}06) = 0{,}94$.

Hoci nám parametre sledujúce homofíliu zlepšovali model, ich vplyv na šance nie sú veľké. Dostali sme taktiež aj záporné hodnoty odhadnutých parametrov. Model však premenné, pri ktorých vyšli kladné hodnoty, označil za signifikantné na rozdiel od záporných odhadov. Zo všetkých sledovaných parametrov v modeli, najviac ovplyvňuje šance, ak osoby rovnako odpovedali na chodenie na koncerty.

Interpretácia parametra sledujúceho geometricky vážené rozdelenie počtu spoločných susedov je zložitejšia, ako sme už videli v predchádzajúcej kapitole 4. Logaritmus šancí pri zmene tohto parametra dostaneme v tvare:

$$log(odds(\Delta ep_k)) = 4{,}51 \cdot (1 - \exp(-0{,}3))^k = 4{,}51 \cdot 0{,}26^k. \tag{48}$$

Z toho dostávame, že napríklad logaritmus šancí existencie hrany medzi dvojicou vrcholov, ktoré majú dvoch spoločných susedov by narástol faktorom $4{,}51 \cdot 0{,}26 = 1{,}17$ oproti dvojici vrcholov, ktoré majú jedného spoločného suseda. Na druhú stranu, ak by počet spoločných susedov jednej dvojice bol 3, tak ich šance by už klesli oproti dvojici s dvomi spoločnými susedmi faktorom $4{,}51 \cdot 0{,}26^2 = 0{,}30$. Vidíme práve efekt klesajúcej váhy pri zvyšujúcom sa počte spoločných susedov.

# Záver

V našej diplomovej práci sme sa venovali analýze sociálnych sietí, špeciálne detegovaní prítomnosti homofílie. Predstavili sme ERG modely, ktoré pracujú so sieťami. Taktiež sme uviedli charakteristiky sietí, ich alternatívy a ich zahrnutie do modelu. Aplikovali sme rôzne modely na dáta, ktoré sme našli na internete. Tieto dáta a aj aplikáciu sme predstavili v kapitole 4. Následne sme si zaobstarali vlastné dáta od priateľov. Zvolili sme si Facebook ako sociálnu sieť, na ktorej sme testovali prítomnosť efektu homofílie.

V jednotlivých kapitolách sme pre sady dát predstavili základný Bernoulliho model a následne nami zvolené modely na základe informácií, ktoré sme mali o daných osobách. V pokračovaní jednotlivých kapitol sme otestovali naše modely a porovnali ich medzi sebou na základe Akaikeho a Bayesovho informačného kritéria. Modely sme otestovali tromi charakteristikami sietí, ktoré nie sú zahrnuté v modeli. Zistili sme, že v našom prípade pri oboch skupinách dát, ich modely ťažko opisovali, preto sme ich následne pozahŕňali do modelov. Ako sme sa dočítali už v spomínanej literatúre, ERG modely sú náchylné na degeneráciu. Inak nebolo ani v našom prípade a na degeneráciu sme narážali pomerne často.

Pri sade dát o žiakoch z Glasgowu sme dospeli, že najväčší vplyv na ich kamarátstva bol v homofílii vzhľadom na pohlavie. Na výsledné šance vplývali faktorom 5,95. Nakoľko žiaci boli veľmi mladí, mohol by byť tento efekt očakávaný. Dostali sme však, že napriek tomu, že sú žiaci mladí, ich stav už taktiež ovplyvňoval kamarátstva. Vplyv na výsledné šance sa už zmenšil, avšak rovnaký stav vplýval faktorom 1,63. Láska očividne kvitne v každom veku. Kladný efekt na kamarátstva sme dostali aj pre počúvanie hudobných štýlov *rock, heavy* a *indie*, ktoré sme pomocou faktorovej analýzy zoskupili do jedného faktora. Do výsledného modelu sa nám dostalo aj pozorovanie vplyvu hudobných štýlov *folk, jazz* a *classical*. Ostatné spomenuté štýly neovplyvňovali hrany v sieti a zároveň nezlepšovali model.

Sledovaním priateľov z Facebooku sme zistili, že vlastnosti, na ktoré sme sa ich pýtali pomocou dotazníka, neovplyvňovali veľmi hrany v sieti. Ako sa ľudia skamarátia môže byť niekedy dielom náhody. Dospeli sme, že aj keď nie výrazne ale kladne ovplyvňujú kamarátstva spoločné aktivity. Najvýraznejšie ovplyvňovalo hrany, ak osoby rovnako odpovedali pri chodení na koncert a do kina. Efekt homofílie sme dostali aj vzhľadom na to,

či daná osoba pracuje na trvalý pracovný pomer. Pri národnosti sme dostali tiež kladný vplyv homofílie, avšak treba pripomenúť, že sme mali v dátach veľkú prevahu ľudí slovenskej národnosti. Vplyv sledovaných premenných na výsledné šance neboli väčšie ako 1,2, i keď ich model označil za signifikantné.

V práci sme sa snažili zlepšiť modely, aby čo najlepšie opisovali dáta. V dnešnej dobe už existuje veľa alternatív, respektíve rozšírení ERG modelov. Jednu z nich, ktorú sme využili z časti aj my, sú CERG modely, s ktorými pracovali autori Hunter a Handcock v článku [14]. Stále sme však predstavili len jednu z možností testovania homofílie. V článku [19] uvádzajú prehľad literatúry a štúdií zaoberajúcou sa homofíliou. Jedna z prvých štúdií sledujúce homofíliu vzhľadom na pohlavie, vek, národnosť a vzdelanie, vznikla už v roku 1929. Počet štúdií sa časom stále zväčšuje a môžeme usúdiť, že analýza sociálnych sietí a správania sa ľudí bola a aj stále je populárna.

# Zoznam použitej literatúry


[1] Akaike H. Information theory and an extension of the maximum likelihood principle. In *B N Petrov and F Csaki (Eds.), Second International Symposium on Information Theory*. 1973, p 267–281. Budapest: Akademiai Kaid o.

[2] Brandenberger L. - Martínez S. Intro to ERGMs, marginal effects and goodness-of-fit. ECPR Winter School Inferential Network Analysis. 2019. dostupné na internete (05.04.2019): *https://rpubs.com/sbmrtnz/ina_day2*.

[3] Frank O. - Strauss D. Markov graphs. *Journal of the American Statistical Association*. 1986, vol 81, no. 395, p. 832–842. DOI:10.1080/01621459.1986.10478342.

[4] Fronczak A.: Exponential random graph models.*Chapter in Encyclopedia of Social Network Analysis and Mining*. R. Alhajj, J. Rokne (Eds.), Springer-Verlag, 2014.

[5] Goodreau S. M. et. al. A statnet Tutorial. In *Journal of statistical software*. 2008, vol. 24, no.9, p. 1–27.

[6] Goodreau S. M. - Kitts A. J. - Morris M. Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks. In *Demography*. 2009, vol. 46, no. 1, p. 103-125. DOI:10.1353/dem.0.0045.

[7] Handcock S. M. Assessing degeneracy in statistical models of social networks. In *Journal of the American Statistical Association*. 2003, vol. 76, p. 33-50.

[8] Handcock S. M. et. al. ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks. 2018. The Statnet Project. R package version 3.9.4, *https://CRAN.R-project.org/package=ergm*.

[9] Harman R. Multivariate Statistical Analysis. Selected lecture notes for the students of Faculty of Mathematics, Physics and Informatics. Comenius University in Bratislava. 2018.

[10] Harris J. K. 2014. *An Introduction to Exponentional Random Graph Modeling*. SAGE Publications. ISBN: 9781483322056. 136 s.

[11] Hunter R. D. et. al. ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. In *Journal of Statistical Software*. 2018, vol. 24, no. 3, p. 1-29.

[12] Hunter D. R. Curved Exponential Family Models for Social Network. In *Socail Networks* 2007, vol.29, no. 2, p. 216–230. DOI:10.1016/j.socnet.2006.08.005.

[13] Hunter D. R.- Goodreau M. S. - Handcock S. M. 2005. Goodness of Fit of Social Network Models. In *Journal of the American Statistical Association*. 2008, vol. 103, no. 481, p. 248-258. DOI: 10.1198/016214507000000446.

[14] Hunter D. R. - Handcock M. S. Inference in Curved Exponential Family Models for Networks. In *Journal of Computational and Graphical Statistics*. 2012, vol. 15, no. 3, p. 565–583. DOI: 10.1198/106186006X133069.

[15] Jackson O. M.2008. *Social and economic networks*. New Jersey: Princeton University Press, 2008.504 s. ISBN 978-0-691-14820-5.

[16] Kolaczyk E.D. - Csárdi G. 2014. *Statistical Analysis of Network Data with R*. New York: Springer, 2014. 207 s. ISBN 978-1-4939-0982-7.

[17] Kuha J. AIC and BIC: Comparisons of Assumptions and Performance. In *Sociological Methods & Research*. 2004, vol. 33, no. 2, p. 188–229.

[18] Luke A. D. 2015. *A User's Guide to Network Analysis in R*. New York: Springer, 2015. 238 s. ISBN 978-3-319-23882-1.

[19] Mcpherson M.- Smith-Lovin L. - Cook M J. Birds of a Feather: Homophily in Social Networks. In*Annual Review of Sociology*. ISSN 0360-0572, 2001, vol. 27, no. 1, p. 415-444.

[20] Michell L. - Amos A. Girls, pecking order and smoking. In *Social Science & Medicine*.1997, vol. 44, no. 12, p. 1861-1869.

[21] Michell L. - West P. Peer pressure to smoke: the meaning depends on the method. In *Health Education Research*. 1996, vol. 11, no. 1, p. 39–49. DOI: 10.1353/dem.0.0045.

[22] SIENA: Description 'Teenage Friends and Lifestyle Study' data, dostupné na internete: (15.10.2018) *https://www.stats.ox.ac.uk/ snijders/siena/Glasgow_data.htm.*

[23] Snijders T. A. - Pattison P. E. - Robins G. L. - Handcock M. S. NEW SPECIFICATIONS FOR EXPONENTIAL RANDOM GRAPH MODELS. In *Sociological Methodology.* 2006, vol. 36, no. 1, p. 99-153. doi:10.1111/j.1467-9531.2006.00176.x.

[24] Steglich C. - Snijders T. A. - Pearson, M. DYNAMIC NETWORKS AND BEHAVIOR: SEPARATING SELECTION FROM INFLUENCE. In *Sociological Methodology.* 2010, vol. 40, p. 329-393. DOI:10.1111/j.1467-9531.2010.01225.x.

[25] Steglich Ch. - Snijders T. A. B. - West P. An illustrative Analysis of the Coevolution of Adolescents' Friendship Networks, Taste in Music, and Alcohol Consumption. In *Methodology.* 2006, vol. 2, p. 48-56.

[26] Steiger H. J. - Exploratory Factor Analysis with R. Great Lakes Institute Of Management. PGPBA SMDM. dostupné na internte (19.4.2019): *http://www.statpower.net/Content/312/R%20Stuff/Exploratory%20Factor%20Analysis%20with%20R.pdf.*

[27] Term H. - Reinert G. 2009. *Statistical Interence for Networks.* Boston: 2009. dostupné na internete (20.11.2018): *http://www.stats.ox.ac.uk/wp-content/uploads/2018/08/Statistical_Inference_for_Networks.pdf.*

[28] The Statnet Development Team. 2018. Introduction to Exponential-family Random Graph (ERG or p*) modeling with ergm. version 3.9.4. dostupné na internete (13.3.2019): *https://cran.r-project.org/web/packages/ergm/vignettes/ergm.pdf.*

[29] Van Der Pol J. (2017): *Introduction to network modeling using Exponential Random Graph models (ERGM)*, Gretha UMR-CNRS 5113, university of Bordeaux, dostupné na internete (07.01.2019): *https://hal.archives-ouvertes.fr/hal-01284994/document.*

# Príloha A - dáta z Glasgowu

Kvôli veľkému počtu výstupov, pri jednotlivých modeloch uvádzame názov súboru, v ktorom je tento výstup uložený. Súbory sa nachádzajú na priloženom CD.

## A.1 Výstupy pre modely s vlastnosťami vrcholov

**Model sledujúci pohlavie**

- Priloha_A_model_pohlavie1.txt

**Model sledujúci a rozlišujúci medzi pohlaviami**

- Priloha_A_model_pohlavie2.txt

**Model sledujúci vek**

- Priloha_A_model_vek.txt

**Model sledujúci stav**

- Priloha_A_model_stav1.txt

**Model sledujúci a rozlišujúci medzi stavmi**

- Priloha_A_model_stav2.txt

**Model sledujúci pohlavie a stav**

- Priloha_A_model_pohlavie_stav.txt

**Model s 6 faktormi**

- Priloha_A_model_hudba6.txt

**Model s 7 faktormi**

- Priloha_A_model_hudba7.txt

**Model sledujúci fajčenie tabaku**

- Priloha_A_model_tabak.txt

**Model sledujúci fajčenie marihuany**

- Priloha_A_model_marihuana.txt

**Model sledujúci fajčenie v mieste bydliska**

- Priloha_A_model_fajcenie_bydlisko.txt

**Model sledujúci fajčenie rodičov**

- Priloha_A_model_fajcenie_rodicia.txt

**Model sledujúci fajčenie súrodencov**

- Priloha_A_model_fajcenie_surodenci.txt

**Model sledujúci počúvanie kaziet alebo CD**

- Priloha_A_model_pocuvanie_CD.txt

**Model sledujúci efekt chodenia po obchodoch**

- Priloha_A_model_chodenie_po_obchodoch.txt

**Model sledujúci efekt čítania kníh a komiksov**

- Priloha_A_model_citanie_knih.txt

**Model sledujúci efekt chodenia na športové zápasy**

- Priloha_A_model_chodenie_na_zapasy.txt

**Model sledujúci efekt športovania**

- Priloha_A_model_sportovania.txt

**Model sledujúci efekt chodenia po vonku**

- Priloha_A_model_chodenie_po_vonku.txt

**Model sledujúci efekt hrania počítačových hier**

- Priloha_A_model_hranie_PC_hier.txt

**Model sledujúci efekt využívania času na záujmy ako umenie, hudobné nástroje**

- Priloha_A_model_umenie.txt

**Model sledujúci efekt chodenia do kina**

- Priloha_A_model_kino.txt

**Model sledujúci efekt chodenia na popové koncerty**

- Priloha_A_model_pop_koncerty.txt

**Model sledujúci efekt starania sa o domáce zvieratko**

- Priloha_A_model_domace_zvieratko.txt

## A.2 Testovanie modelu s vlastnosťami vrcholov

**Hodnoty rozdelenia stupňov pre simulované siete**

```
Goodness-of-fit for degree

  obs min  mean max MC p-value
0  16   9 19.92  35        0.52
1  39  20 36.11  49        0.64
2  35  21 33.95  45        0.88
3  21  12 21.59  36        1.00
4  14   5 11.04  19        0.44
5   4   0  4.69  10        1.00
6   0   0  1.28   5        0.58
7   0   0  0.33   2        1.00
8   0   0  0.09   2        1.00
```

**Hodnoty rozdelenia počtu spoločných susedov pre simulované siete**

```
Goodness-of-fit for edgewise shared partner

     obs min   mean max MC p-value
esp0  69  91 118.35 148          0
esp1  38   0   5.07  21          0
esp2  12   0   0.12   2          0
esp3   4   0   0.01   1          0
esp4   1   0   0.00   0          0
```

**Hodnoty rozdelenia pre najmenšiu vzdialenosť medzi vrcholmi pre simulované siete**

Goodness-of-fit for minimum geodesic distance

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 1 | 124 | 97 | 123.55 | 151 | 0.94 |
| 2 | 134 | 149 | 229.85 | 340 | 0.00 |
| 3 | 158 | 194 | 390.17 | 708 | 0.00 |
| 4 | 168 | 244 | 566.48 | 1153 | 0.00 |
| 5 | 183 | 298 | 686.96 | 1380 | 0.00 |
| 6 | 165 | 338 | 706.77 | 1240 | 0.00 |
| 7 | 144 | 278 | 622.23 | 1033 | 0.00 |
| 8 | 122 | 215 | 486.93 | 781 | 0.00 |
| 9 | 81 | 102 | 348.43 | 722 | 0.00 |
| 10 | 52 | 34 | 231.47 | 586 | 0.06 |
| 11 | 44 | 3 | 146.57 | 447 | 0.26 |
| 12 | 34 | 0 | 89.64 | 321 | 0.64 |
| 13 | 22 | 0 | 52.62 | 237 | 0.86 |
| 14 | 16 | 0 | 29.97 | 178 | 0.92 |
| 15 | 9 | 0 | 16.18 | 135 | 0.74 |
| 16 | 2 | 0 | 8.11 | 108 | 0.90 |
| 17 | 0 | 0 | 3.47 | 77 | 1.00 |
| 18 | 0 | 0 | 1.49 | 58 | 1.00 |
| 19 | 0 | 0 | 0.80 | 43 | 1.00 |
| 20 | 0 | 0 | 0.33 | 22 | 1.00 |
| 21 | 0 | 0 | 0.10 | 7 | 1.00 |
| 22 | 0 | 0 | 0.01 | 1 | 1.00 |
| Inf | 6798 | 1586 | 3513.87 | 5831 | 0.00 |

## A.3 Výstupy pre modely s charakteristikami sietí

**Model s pridaním sledovania stupňa veľkosti 0**

- Priloha_A_model_stupen_0.txt

**Model s pridaním sledovania stupňa veľkosti 1**

- Priloha_A_model_stupen_1.txt

**Model s pridaním sledovania stupňa veľkosti 2**

- Priloha_A_model_stupen_2.txt

**Model s pridaním sledovania stupňa veľkosti 3**

- Priloha_A_model_stupen_3.txt

**Model s pridaním sledovania stupňa veľkosti 4**

- Priloha_A_model_stupen_4.txt

**Model s pridaním sledovania stupňa veľkosti 5**

- Priloha_A_model_stupen_5.txt

**Model s pridaním sledovania 2-star**

- Priloha_A_model_2star.txt

**Model s pridaním sledovania 3-star**

- Priloha_A_model_3star.txt

**Model s pridaním sledovania 4-star**

- Priloha_A_model_4star.txt

**Model s pridaním sledovania 5-star**

- Priloha_A_model_5star.txt

**Model s pridaním sledovania geometricky váženého rozdelenia stupňov v sieti so zafixovaným $\tau = 1$**

- Priloha_A_model_gwdeg1_fixed.txt

**Model s pridaním sledovania geometricky váženého rozdelenia stupňov v sieti s nezafixovaným $\tau = 1$**

- Priloha_A_model_gwdeg1_unfixed.txt

**Model s pridaním parametra sledujúceho vrcholy s 1 spoločným susedom**

- Priloha_A_model_esp1.txt

**Model s pridaním parametra sledujúceho vrcholy s 1 spoločným susedom**

- Priloha_A_model_esp2.txt

**Model s pridaním parametra sledujúceho vrcholy s 1 a 2 spoločnými susedmi**

- Priloha_A_model_esp1-2.txt

**Hodnoty informačných kritérií pre rôzne hodnoty parametra $\tau$ pre geometrické rozdelenie počtu spoločných susedov**

| | Akaikeho | Bayesovo |
|---|---|---|
| Model s hodnotou parametra $\tau = 0.1$ | 1114 | 1156 |
| Model s hodnotou parametra $\tau = 0.2$ | 1113 | 1155 |
| Model s hodnotou parametra $\tau = 0.3$ | 1111 | 1153 |
| Model s hodnotou parametra $\tau = 0.4$ | 1109 | 1151 |
| Model s hodnotou parametra $\tau = 0.5$ | 1109 | 1151 |
| Model s hodnotou parametra $\tau = 0.6$ | 1108 | 1150 |
| Model s hodnotou parametra $\tau = 0.7$ | 1108 | 1150 |

**Model so zafixovanou hodnotou parametra $\tau = 0.6$ pre geometrické rozdelenie počtu spoločných susedov**

- Priloha_A_model_gwesp0.6_fixed.txt

**Model so zafixovanou hodnotou parametra $\tau = 0.7$ pre geometrické rozdelenie počtu spoločných susedov**

- Priloha_A_model_gwesp0.7_fixed.txt

**Model so nezafixovaným parametrom s počiatočnou hodnotou $\tau = 0.6$ pre geometrické rozdelenie počtu spoločných susedov**

- Priloha_A_model_gwesp0.6_unfixed.txt

## A.4 Testovanie modelu so zahrnutou charakteristikou siete

**Hodnoty rozdelenia stupňov pre simulované siete**

```
Goodness-of-fit for degree                    0   16  19 32.38  51       0.00
                                              1   39  22 34.82  50       0.52
   obs min  mean max MC p-value               2   35  16 25.69  41       0.06
```

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 21 | 8 | 15.85 | 26 | 0.26 | 9 | 0 | 0 | 0.47 | 3 | 1.00 |
| 4 | 14 | 1 | 8.81 | 18 | 0.14 | 10 | 0 | 0 | 0.21 | 3 | 1.00 |
| 5 | 4 | 0 | 5.00 | 10 | 0.86 | 11 | 0 | 0 | 0.24 | 2 | 1.00 |
| 6 | 0 | 0 | 3.20 | 9 | 0.22 | 12 | 0 | 0 | 0.08 | 2 | 1.00 |
| 7 | 0 | 0 | 1.34 | 5 | 0.60 | 13 | 0 | 0 | 0.04 | 1 | 1.00 |
| 8 | 0 | 0 | 0.86 | 3 | 1.00 | 14 | 0 | 0 | 0.01 | 1 | 1.00 |

### Hodnoty rozdelenia počtu spoločných susedov pre simulované siete

Goodness-of-fit for edgewise shared partner

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| esp0 | 69 | 51 | 68.29 | 90 | 0.90 |
| esp1 | 38 | 9 | 36.60 | 57 | 0.98 |
| esp2 | 12 | 0 | 11.66 | 41 | 0.88 |
| esp3 | 4 | 0 | 2.78 | 15 | 0.60 |
| esp4 | 1 | 0 | 0.55 | 7 | 0.52 |
| esp5 | 0 | 0 | 0.11 | 2 | 1.00 |
| esp6 | 0 | 0 | 0.03 | 1 | 1.00 |

### Hodnoty rozdelenia pre najmenšiu vzdialenosť medzi vrcholmi pre simulované siete

Goodness-of-fit for minimum geodesic distance

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 1 | 124 | 66 | 120.02 | 164 | 0.86 |
| 2 | 134 | 67 | 231.08 | 422 | 0.16 |
| 3 | 158 | 64 | 359.19 | 811 | 0.10 |
| 4 | 168 | 65 | 440.04 | 1093 | 0.10 |
| 5 | 183 | 56 | 448.14 | 985 | 0.18 |
| 6 | 165 | 49 | 396.08 | 729 | 0.18 |
| 7 | 144 | 22 | 316.07 | 576 | 0.16 |
| 8 | 122 | 8 | 233.87 | 518 | 0.40 |
| 9 | 81 | 2 | 161.33 | 387 | 0.46 |
| 10 | 52 | 0 | 103.94 | 319 | 0.62 |
| 11 | 44 | 0 | 64.04 | 262 | 1.00 |
| 12 | 34 | 0 | 39.24 | 226 | 0.74 |
| 13 | 22 | 0 | 23.38 | 195 | 0.64 |
| 14 | 16 | 0 | 14.15 | 191 | 0.48 |
| 15 | 9 | 0 | 8.97 | 185 | 0.44 |
| 16 | 2 | 0 | 5.87 | 153 | 0.46 |
| 17 | 0 | 0 | 3.96 | 110 | 1.00 |
| 18 | 0 | 0 | 2.37 | 92 | 1.00 |
| 19 | 0 | 0 | 1.64 | 75 | 1.00 |
| 20 | 0 | 0 | 1.07 | 53 | 1.00 |
| 21 | 0 | 0 | 0.55 | 34 | 1.00 |
| 22 | 0 | 0 | 0.20 | 18 | 1.00 |
| 23 | 0 | 0 | 0.07 | 7 | 1.00 |
| 24 | 0 | 0 | 0.02 | 2 | 1.00 |
| Inf | 6798 | 3498 | 5280.71 | 7808 | 0.18 |

# Príloha B - dáta z Facebooku

Kvôli veľkému počtu výstupov, pri jednotlivých modeloch uvádzame názov súboru, v ktorom je tento výstup uložený. Súbory sa nachádzajú na priloženom CD.

## B.1 Výstupy pre modely s vlastnosťami vrcholov

**Bernoulliho model**

- Priloha_B_model_Bernoulliho.txt

**Model sledujúci pohlavie**

- Priloha_B_model_pohlavie1.txt

**Model sledujúci a rozlišujúci medzi pohlaviami**

- Priloha_B_model_pohlavie2.txt

**Model sledujúci vek**

- Priloha_B_model_vek.txt

**Model sledujúci stav**

- Priloha_B_model_stav1.txt

**Model sledujúci a rozlišujúci medzi stavmi**

- Priloha_B_model_stav2.txt

**Model sledujúci národnosť**

- Priloha_B_model_narodnost.txt

**Model sledujúci vieru**

- Priloha_B_model_viera.txt

**Model sledujúci fajčenie**

- Priloha_B_model_fajcenie.txt

**Model sledujúci prácu na TPP**

- Priloha_B_model_TPP.txt

**Model sledujúci prácu na čiastočný uväzok**

- Priloha_B_model_part_time.txt

**Model sledujúci prácu na TPP a národnosť**

- Priloha_B_model_TPP_narodnost.txt

**Model sledujúci prácu na TPP, národnosť a stav**

- Priloha_B_model_TPP_narodnost_stav.txt

**Model sledujúci prácu na TPP, národnosť, stav a vek**

- Priloha_B_model_TPP_narodnost_stav_vek.txt

**Model sledujúci účasť na táboroch**

- Priloha_B_model_tabory.txt

**Model sledujúci účasť na festivaloch**

- Priloha_B_model_festivaly.txt

**Model sledujúci účasť na koncertoch**

- Priloha_B_model_koncerty.txt

**Model sledujúci účasť na diskotékach**

- Priloha_B_model_diskoteky.txt

**Model sledujúci chodenie do divadla**

- Priloha_B_model_divadlo.txt

**Model sledujúci chodenie do kina**

- Priloha_B_model_kino.txt

**Model sledujúci chodenie na športové zápasy**

- Priloha_B_model_sportove_zapasy.txt

**Model sledujúci chodenie do kina a na koncerty**

- Priloha_B_model_kino_koncerty.txt

**Model sledujúci chodenie do kina, na koncerty a tábory**

- Priloha_B_model_kino_koncerty_tabory.txt

## B.2 Testovanie modelu s vlastnosťami vrcholov

**Hodnoty rozdelenia stupňov pre simulované siete**

Goodness-of-fit for degree

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 0 | 10 | 0 | 0.00 | 0 | 0.00 |
| 1 | 10 | 0 | 0.00 | 0 | 0.00 |
| 2 | 6 | 0 | 0.00 | 0 | 0.00 |
| 3 | 10 | 0 | 0.00 | 0 | 0.00 |
| 4 | 10 | 0 | 0.09 | 2 | 0.00 |
| 5 | 11 | 0 | 0.20 | 2 | 0.00 |
| 6 | 8 | 0 | 0.55 | 3 | 0.00 |
| 7 | 5 | 0 | 0.98 | 5 | 0.02 |
| 8 | 4 | 0 | 1.64 | 6 | 0.22 |
| 9 | 7 | 0 | 2.76 | 6 | 0.00 |
| 10 | 3 | 1 | 4.69 | 11 | 0.66 |
| 11 | 2 | 1 | 5.72 | 14 | 0.10 |
| 12 | 9 | 2 | 8.18 | 16 | 0.90 |
| 13 | 7 | 5 | 10.62 | 20 | 0.26 |
| 14 | 3 | 3 | 12.43 | 22 | 0.02 |
| 15 | 4 | 7 | 14.02 | 23 | 0.00 |
| 16 | 6 | 6 | 15.68 | 27 | 0.02 |
| 17 | 1 | 9 | 16.79 | 26 | 0.00 |
| 18 | 4 | 8 | 17.33 | 27 | 0.00 |
| 19 | 4 | 8 | 16.32 | 31 | 0.00 |
| 20 | 6 | 7 | 14.77 | 25 | 0.00 |
| 21 | 4 | 6 | 13.32 | 20 | 0.00 |
| 22 | 4 | 5 | 11.22 | 19 | 0.00 |
| 23 | 4 | 3 | 9.92 | 20 | 0.04 |
| 24 | 1 | 2 | 7.53 | 18 | 0.00 |
| 25 | 7 | 0 | 5.55 | 11 | 0.70 |
| 26 | 2 | 0 | 3.93 | 10 | 0.56 |
| 27 | 2 | 0 | 2.68 | 7 | 0.94 |
| 28 | 2 | 0 | 1.97 | 7 | 1.00 |
| 29 | 2 | 0 | 1.33 | 5 | 0.76 |
| 30 | 3 | 0 | 0.83 | 3 | 0.08 |
| 31 | 3 | 0 | 0.41 | 3 | 0.02 |
| 32 | 1 | 0 | 0.25 | 2 | 0.44 |
| 33 | 2 | 0 | 0.17 | 1 | 0.00 |
| 34 | 2 | 0 | 0.05 | 1 | 0.00 |
| 35 | 1 | 0 | 0.02 | 1 | 0.04 |
| 36 | 3 | 0 | 0.04 | 1 | 0.00 |
| 37 | 0 | 0 | 0.01 | 1 | 1.00 |
| 38 | 3 | 0 | 0.00 | 0 | 0.00 |
| 39 | 2 | 0 | 0.00 | 0 | 0.00 |
| 41 | 2 | 0 | 0.00 | 0 | 0.00 |
| 42 | 2 | 0 | 0.00 | 0 | 0.00 |
| 43 | 2 | 0 | 0.00 | 0 | 0.00 |
| 44 | 2 | 0 | 0.00 | 0 | 0.00 |
| 45 | 1 | 0 | 0.00 | 0 | 0.00 |
| 46 | 1 | 0 | 0.00 | 0 | 0.00 |
| 47 | 3 | 0 | 0.00 | 0 | 0.00 |
| 48 | 1 | 0 | 0.00 | 0 | 0.00 |
| 53 | 3 | 0 | 0.00 | 0 | 0.00 |
| 55 | 1 | 0 | 0.00 | 0 | 0.00 |
| 58 | 1 | 0 | 0.00 | 0 | 0.00 |
| 60 | 1 | 0 | 0.00 | 0 | 0.00 |
| 61 | 1 | 0 | 0.00 | 0 | 0.00 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 62 | 2 | 0 | 0.00 | 0 | 0.00 | 67 | 1 | 0 | 0.00 | 0 | 0.00 |

### Hodnoty rozdelenia počtu spoločných susedov pre simulované siete

Goodness-of-fit for edgewise shared partner

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| esp0 | 45 | 281 | 346.67 | 415 | 0.00 |
| esp1 | 66 | 510 | 556.54 | 610 | 0.00 |
| esp2 | 70 | 405 | 465.13 | 543 | 0.00 |
| esp3 | 54 | 205 | 267.89 | 327 | 0.00 |
| esp4 | 49 | 70 | 119.01 | 166 | 0.00 |
| esp5 | 36 | 19 | 43.15 | 72 | 0.46 |
| esp6 | 29 | 4 | 13.19 | 29 | 0.02 |
| esp7 | 23 | 0 | 3.15 | 9 | 0.00 |
| esp8 | 56 | 0 | 0.84 | 4 | 0.00 |
| esp9 | 62 | 0 | 0.13 | 3 | 0.00 |
| esp10 | 56 | 0 | 0.02 | 1 | 0.00 |
| esp11 | 85 | 0 | 0.01 | 1 | 0.00 |
| esp12 | 70 | 0 | 0.00 | 0 | 0.00 |
| esp13 | 54 | 0 | 0.00 | 0 | 0.00 |
| esp14 | 57 | 0 | 0.00 | 0 | 0.00 |
| esp15 | 58 | 0 | 0.00 | 0 | 0.00 |
| esp16 | 80 | 0 | 0.00 | 0 | 0.00 |
| esp17 | 80 | 0 | 0.00 | 0 | 0.00 |
| esp18 | 62 | 0 | 0.00 | 0 | 0.00 |
| esp19 | 73 | 0 | 0.00 | 0 | 0.00 |
| esp20 | 45 | 0 | 0.00 | 0 | 0.00 |
| esp21 | 56 | 0 | 0.00 | 0 | 0.00 |
| esp22 | 51 | 0 | 0.00 | 0 | 0.00 |
| esp23 | 57 | 0 | 0.00 | 0 | 0.00 |
| esp24 | 48 | 0 | 0.00 | 0 | 0.00 |
| esp25 | 48 | 0 | 0.00 | 0 | 0.00 |
| esp26 | 38 | 0 | 0.00 | 0 | 0.00 |
| esp27 | 41 | 0 | 0.00 | 0 | 0.00 |
| esp28 | 37 | 0 | 0.00 | 0 | 0.00 |
| esp29 | 30 | 0 | 0.00 | 0 | 0.00 |
| esp30 | 24 | 0 | 0.00 | 0 | 0.00 |
| esp31 | 25 | 0 | 0.00 | 0 | 0.00 |
| esp32 | 17 | 0 | 0.00 | 0 | 0.00 |
| esp33 | 18 | 0 | 0.00 | 0 | 0.00 |
| esp34 | 12 | 0 | 0.00 | 0 | 0.00 |
| esp35 | 10 | 0 | 0.00 | 0 | 0.00 |
| esp36 | 11 | 0 | 0.00 | 0 | 0.00 |
| esp37 | 21 | 0 | 0.00 | 0 | 0.00 |
| esp38 | 16 | 0 | 0.00 | 0 | 0.00 |
| esp39 | 10 | 0 | 0.00 | 0 | 0.00 |
| esp40 | 7 | 0 | 0.00 | 0 | 0.00 |
| esp41 | 3 | 0 | 0.00 | 0 | 0.00 |
| esp42 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp43 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp44 | 1 | 0 | 0.00 | 0 | 0.00 |
| esp45 | 2 | 0 | 0.00 | 0 | 0.00 |
| esp46 | 5 | 0 | 0.00 | 0 | 0.00 |
| esp47 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp48 | 1 | 0 | 0.00 | 0 | 0.00 |
| esp50 | 3 | 0 | 0.00 | 0 | 0.00 |
| esp51 | 1 | 0 | 0.00 | 0 | 0.00 |

### Hodnoty rozdelenia pre najmenšiu vzdialenosť medzi vrcholmi pre simulované siete

Goodness-of-fit for minimum geodesic distance

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 1 | 1815 | 1707 | 1815.73 | 1927 | 1 |
| 2 | 6534 | 13971 | 14646.61 | 15328 | 0 |
| 3 | 5208 | 3046 | 3838.45 | 4622 | 0 |
| 4 | 2044 | 0 | 0.21 | 2 | 0 |
| 5 | 589 | 0 | 0.00 | 0 | 0 |
| 6 | 98 | 0 | 0.00 | 0 | 0 |
| 7 | 25 | 0 | 0.00 | 0 | 0 |
| Inf | 3988 | 0 | 0.00 | 0 | 0 |

## B.3 Výstupy pre modely s charakteristikami sietí

**Model s parametrom $\tau = 0.3$**

- Priloha_B_model_gwesp0.3_fixed.txt

## B.4 Testovanie modelu so zahrnutou charakteriskou siete

**Hodnoty rozdelenia stupňov pre simulované siete**

Goodness-of-fit for degree

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 0 | 10 | 5 | 9.84 | 24 | 0.66 |
| 1 | 10 | 0 | 0.30 | 3 | 0.00 |
| 2 | 6 | 0 | 2.09 | 16 | 0.16 |
| 3 | 10 | 0 | 2.63 | 12 | 0.06 |
| 4 | 10 | 0 | 3.16 | 14 | 0.08 |
| 5 | 11 | 0 | 3.42 | 15 | 0.04 |
| 6 | 8 | 0 | 3.73 | 11 | 0.10 |
| 7 | 5 | 0 | 4.59 | 13 | 0.86 |
| 8 | 4 | 1 | 5.66 | 14 | 0.76 |
| 9 | 7 | 1 | 5.56 | 13 | 0.62 |
| 10 | 3 | 1 | 6.56 | 13 | 0.30 |
| 11 | 2 | 1 | 6.67 | 15 | 0.10 |
| 12 | 9 | 3 | 7.73 | 18 | 0.74 |
| 13 | 7 | 3 | 8.64 | 18 | 0.74 |
| 14 | 3 | 4 | 9.24 | 16 | 0.00 |
| 15 | 4 | 2 | 9.34 | 23 | 0.08 |
| 16 | 6 | 3 | 9.63 | 16 | 0.30 |
| 17 | 1 | 3 | 10.30 | 21 | 0.00 |
| 18 | 4 | 2 | 9.83 | 19 | 0.08 |
| 19 | 4 | 1 | 9.46 | 19 | 0.16 |
| 20 | 6 | 2 | 8.57 | 18 | 0.60 |
| 21 | 4 | 2 | 8.59 | 15 | 0.14 |
| 22 | 4 | 2 | 7.82 | 17 | 0.34 |
| 23 | 4 | 2 | 7.37 | 15 | 0.34 |
| 24 | 1 | 1 | 6.15 | 14 | 0.02 |
| 25 | 7 | 1 | 5.78 | 12 | 0.90 |
| 26 | 2 | 1 | 5.42 | 13 | 0.26 |
| 27 | 2 | 0 | 4.44 | 11 | 0.48 |
| 28 | 2 | 0 | 3.56 | 8 | 0.66 |
| 29 | 2 | 0 | 3.05 | 9 | 0.88 |
| 30 | 3 | 0 | 2.71 | 7 | 1.00 |
| 31 | 3 | 0 | 2.16 | 7 | 0.66 |
| 32 | 1 | 0 | 1.73 | 6 | 0.94 |
| 33 | 2 | 0 | 1.33 | 5 | 0.72 |
| 34 | 2 | 0 | 1.23 | 4 | 0.70 |
| 35 | 1 | 0 | 0.90 | 4 | 1.00 |
| 36 | 3 | 0 | 0.65 | 5 | 0.04 |
| 37 | 0 | 0 | 0.49 | 3 | 1.00 |
| 38 | 3 | 0 | 0.47 | 4 | 0.04 |
| 39 | 2 | 0 | 0.27 | 2 | 0.06 |
| 40 | 0 | 0 | 0.20 | 2 | 1.00 |
| 41 | 2 | 0 | 0.17 | 2 | 0.06 |
| 42 | 2 | 0 | 0.15 | 3 | 0.02 |
| 43 | 2 | 0 | 0.13 | 1 | 0.00 |
| 44 | 2 | 0 | 0.05 | 1 | 0.00 |
| 45 | 1 | 0 | 0.08 | 1 | 0.16 |
| 46 | 1 | 0 | 0.06 | 1 | 0.12 |
| 47 | 3 | 0 | 0.02 | 1 | 0.00 |
| 48 | 1 | 0 | 0.00 | 0 | 0.00 |
| 51 | 0 | 0 | 0.01 | 1 | 1.00 |
| 52 | 0 | 0 | 0.02 | 1 | 1.00 |
| 53 | 3 | 0 | 0.00 | 0 | 0.00 |
| 54 | 0 | 0 | 0.01 | 1 | 1.00 |
| 55 | 1 | 0 | 0.00 | 0 | 0.00 |
| 56 | 0 | 0 | 0.02 | 1 | 1.00 |
| 58 | 1 | 0 | 0.01 | 1 | 0.02 |
| 60 | 1 | 0 | 0.00 | 0 | 0.00 |
| 61 | 1 | 0 | 0.00 | 0 | 0.00 |
| 62 | 2 | 0 | 0.00 | 0 | 0.00 |
| 67 | 1 | 0 | 0.00 | 0 | 0.00 |

### Hodnoty rozdelenia počtu spoločných susedov pre simulované siete

Goodness-of-fit for edgewise shared partner

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| esp0 | 45 | 0 | 0.75 | 5 | 0.00 |
| esp1 | 66 | 71 | 113.24 | 144 | 0.00 |
| esp2 | 70 | 291 | 399.12 | 457 | 0.00 |
| esp3 | 54 | 251 | 458.00 | 560 | 0.00 |
| esp4 | 49 | 141 | 333.67 | 416 | 0.00 |
| esp5 | 36 | 62 | 196.35 | 292 | 0.00 |
| esp6 | 29 | 29 | 100.56 | 150 | 0.02 |
| esp7 | 23 | 15 | 45.85 | 74 | 0.08 |
| esp8 | 56 | 2 | 18.56 | 32 | 0.00 |
| esp9 | 62 | 0 | 7.89 | 21 | 0.00 |
| esp10 | 56 | 0 | 2.93 | 10 | 0.00 |
| esp11 | 85 | 0 | 0.95 | 4 | 0.00 |
| esp12 | 70 | 0 | 0.16 | 2 | 0.00 |
| esp13 | 54 | 0 | 0.18 | 2 | 0.00 |
| esp14 | 57 | 0 | 0.06 | 2 | 0.00 |
| esp15 | 58 | 0 | 0.01 | 1 | 0.00 |
| esp16 | 80 | 0 | 0.01 | 1 | 0.00 |
| esp17 | 80 | 0 | 0.00 | 0 | 0.00 |
| esp18 | 62 | 0 | 0.00 | 0 | 0.00 |
| esp19 | 73 | 0 | 0.00 | 0 | 0.00 |
| esp20 | 45 | 0 | 0.00 | 0 | 0.00 |
| esp21 | 56 | 0 | 0.00 | 0 | 0.00 |
| esp22 | 51 | 0 | 0.00 | 0 | 0.00 |
| esp23 | 57 | 0 | 0.00 | 0 | 0.00 |
| esp24 | 48 | 0 | 0.00 | 0 | 0.00 |
| esp25 | 48 | 0 | 0.00 | 0 | 0.00 |
| esp26 | 38 | 0 | 0.00 | 0 | 0.00 |
| esp27 | 41 | 0 | 0.00 | 0 | 0.00 |
| esp28 | 37 | 0 | 0.00 | 0 | 0.00 |
| esp29 | 30 | 0 | 0.00 | 0 | 0.00 |
| esp30 | 24 | 0 | 0.00 | 0 | 0.00 |
| esp31 | 25 | 0 | 0.00 | 0 | 0.00 |
| esp32 | 17 | 0 | 0.00 | 0 | 0.00 |
| esp33 | 18 | 0 | 0.00 | 0 | 0.00 |
| esp34 | 12 | 0 | 0.00 | 0 | 0.00 |
| esp35 | 10 | 0 | 0.00 | 0 | 0.00 |
| esp36 | 11 | 0 | 0.00 | 0 | 0.00 |
| esp37 | 21 | 0 | 0.00 | 0 | 0.00 |
| esp38 | 16 | 0 | 0.00 | 0 | 0.00 |
| esp39 | 10 | 0 | 0.00 | 0 | 0.00 |
| esp40 | 7 | 0 | 0.00 | 0 | 0.00 |
| esp41 | 3 | 0 | 0.00 | 0 | 0.00 |
| esp42 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp43 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp44 | 1 | 0 | 0.00 | 0 | 0.00 |
| esp45 | 2 | 0 | 0.00 | 0 | 0.00 |
| esp46 | 5 | 0 | 0.00 | 0 | 0.00 |
| esp47 | 4 | 0 | 0.00 | 0 | 0.00 |
| esp48 | 1 | 0 | 0.00 | 0 | 0.00 |
| esp50 | 3 | 0 | 0.00 | 0 | 0.00 |
| esp51 | 1 | 0 | 0.00 | 0 | 0.00 |

### Hodnoty rozdelenia pre najmenšiu vzdialenosť medzi vrcholmi pre simulované siete

Goodness-of-fit for minimum geodesic distance

| | obs | min | mean | max | MC p-value |
|---|---|---|---|---|---|
| 1 | 1815 | 935 | 1678.29 | 1928 | 0.38 |
| 2 | 6534 | 5939 | 12194.32 | 13870 | 0.04 |
| 3 | 5208 | 3046 | 4375.38 | 6538 | 0.22 |
| 4 | 2044 | 1 | 92.79 | 1341 | 0.00 |
| 5 | 589 | 0 | 6.48 | 425 | 0.00 |
| 6 | 98 | 0 | 0.92 | 90 | 0.00 |
| 7 | 25 | 0 | 0.00 | 0 | 0.00 |
| Inf | 3988 | 995 | 1952.82 | 5580 | 0.16 |