Oblikovanje obogaćenoga društvenoga grafa na temelju koncepta udaljenosti

Humski, Luka

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

LUKA HUMSKI

OBLIKOVANJE OBOGAĆENOGA DRUŠTVENOGA GRAFA NA TEMELJU KONCEPTA UDALJENOSTI

DOKTORSKI RAD

Mentor: doc. dr. sc. Damir Pintar

Zagreb, 2019.

University of Zagreb
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

LUKA HUMSKI

MODELLING OF ENRICHED SOCIAL GRAPH BASED ON CONCEPT OF DISTANCE

DOCTORAL THESIS

Supervisor: Assistant Professor Damir Pintar, PhD

Zagreb, 2019

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za osnove elektrotehnike i električka mjerenja.

Mentor: doc. dr. sc. Damir Pintar

Doktorski rad ima 143 stranice

Doktorski rad br. _____

O MENTORU

Damir Pintar rođen je u Osijeku 1978., gdje je završio III. gimnaziju. 1996. godine upisuje Fakultet elektrotehnike i računarstva na Sveučilištu u Zagrebu, gdje 2001. diplomira u polju Elektrotehnike te brani magistarski i doktorski rad 2005., odnosno 2009. godine. Tijekom studija primio je četiri priznanja "Josip Lončar" za svaku godinu studija te brončanu plaketu "Josip Lončar" kao jedan od najboljih studenata generacije.

Od 2001. godine radi na Zavodu za osnove elektrotehnike i električka mjerenja na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu (FER), gdje 2012. prelazi u zvanje docenta. Radi kao istraživač na više uzastopnih znanstvenih projekata financiranih od strane Ministarstva znanosti, obrazovanja i sporta te na međunarodnom projektu financiranom iz strukturnih fondova EU. Također radi kao voditelj na znanstveno-istraživačkom projektu financiranom od strane HRZZ-a. Uz znanstvene projekte, radi na nizu stručnih projekata koje je FER ostvarivao s poduzećima kao što su: koncern Agrokor, Equidem d.o.o., Multicom d.o.o., Neos d. o. o. i Mercury Processing Services International.

Njegovi interesi uključuju internetske tehnologije, programiranje i analizu podataka. Znanstvena istraživanja uključuju razvoj rješenja za razne aspekte elektroničkog poslovanja, integraciju, skladištenje podataka te teoretsku i praktičnu primjenu metoda i tehnologija vezanih uz podatkovnu znanost. Autor je većeg broja znanstvenih članaka indeksiranih u vodećim bazama podataka kao što su Current Contents, SCI Expanded, SCOPUS, EBSCO i INSPEC. Recenzirao je članke za međunarodne konferencije i vodeće znanstvene časopise. Član je IEEE od 2001. Aktivno koristi engleski jezik u govoru i pismu. Oženjen je i ima dvoje djece.

ABOUT SUPERVISOR

Damir Pintar was born in Osijek in 1978, where he finished the III. gimnazija high school. In 1996 he enrolled in the Faculty of Electrical Engineering and Computing at the University of Zagreb, where he graduated in electrical engineering in 2001 and defended his master's and doctoral thesis in 2005 and 2009 respectively. During his studies he has received four "Josip Lončar" prizes for each year of study, and the bronze plaque "Josip Lončar" as one of the best students in his generation.

Since 2001 he has been working at the Department of Electrical Engineering Fundamentals and Measurement, Faculty of Electrical Engineering and Computing, University of Zagreb (FER), where he holds a position of Assistant Professor from 2012. He works as a researcher on several successive scientific projects funded by the Ministry of Science, Education and Sports, as well as on an international project funded from the EU Structural Funds. He also worked as a head of a scientific research project funded by the Croatian Science Foundation. In addition to scientific projects he has been working on a number of projects with the industry which involved cooperation of FER and companies such as Agrokor, Equidem d.o.o, Multicom d.o.o., Neos d. O. And Mercury Processing Services International.

His interests include Internet technologies, programming and data analysis. Scientific research includes the development of solutions for various aspects of electronic business, integration, data storage and theoretical and practical application of data science methods and technologies. He authored a number of scientific articles indexed in leading scientific databases such as Current Contents, SCI Expanded, SCOPUS, EBSCO and INSPEC. He has reviewed articles for international conferences and leading scientific journals. He is a member of IEEE since 2001. He actively uses English in speech and script. He is married and has two children.

ZAHVALE

Hvala cijeloj mojoj obitelji koja me podržavala kroz cijelo školovanje, a čija je kruna ovaj završeni doktorski studij.

Hvala svim mojim studentima na suradnji koja je rezultirala izvrsnim znanstvenim i stručnim rezultatima. Posebno hvala Juraju Iliću, mag. ing. na izvrsnoj suradnji u organizaciji i provođenju za ovaj rad ključnog društvenog istraživanja *NajFrend*.

Hvala svim mojim prijateljima i kolegama na podršci. Hvala prof. dr. sc. Zoranu Skočiru koji mi je predložio i omogućio rad na fakultetu, upis doktorskog studija te izgradnju karijere kroz akademsku zajednicu. Hvala mu što je uvijek bio uz mene kada je to bilo potrebno. Hvala mom mentoru doc. dr. sc. Damiru Pintaru na izvrsnoj suradnji, stručnom vodstvu i velikoj pomoći pri izradi doktorskoga rada.

Hvala i svima ostalima koji su bili uz mene.

Doktorski rad posvećujem svome ocu koji je najnestrpljivije iščekivao kraj moga doktorskoga studija, ali obranu nažalost nije dočekao.

SAŽETAK

Sustavi za društveno umrežavanje (Facebook, Twitter, Instagram i slični) široko su rasprostranjena digitalna platforma za komunikaciju i interakciju milijardi ljudi. Podatci o interakciji ljudi pohranjuju se na strukturiran – računalno čitljiv način, što (uz odgovarajuću suglasnost ljudi o čijim se podatcima radi) omogućuje njihovu obradu, analizu i izlučivanje različitih zaključaka. Dostupnost podataka o interakciji temelj je svih istraživanja društvenih mreža – kako onih sociologijskih i psihologijskih, tako i onih iz područja računalne znanosti.

U ovome doktorskome radu provedena su društvena istraživanja kroz koja su (uz suglasnost ispitanika) prikupljeni podatci o interakciji korisnika na sustavu za društveno umrežavanje Facebook te usporedno korisničke procjene njihova odnosa s ljudima s kojima na Facebooku interaktiraju. Time je oformljen referentni podatkovni skup za istraživanja društvenih mreža. Oformljen referentni podatkovni skup u radu je detaljno analiziran. Korištenjem referentnog podatkovnog skupa, osmišljen je, implementiran, ugođen i verificiran model za računanje društvene udaljenosti na temelju podataka o interakciji ljudi na Facebooku. Za razliku od prethodnih istraživanja koja su intenzitet odnosa među ljudima prikazivala primjenom koncepta težine ili snage veze, društvena udaljenost inovativan je način iskazivanja intenziteta odnosa među ljudima kroz primjenu koncepta udaljenosti. Osnovna značajka koncepta udaljenosti jest da se veći intenzitet prikazuje manjim brojem, što je vrlo pogodno za daljnje primjene kao što su određivanje vremena propagacije informacije ili određivanje približne geografske lokacije korisnika.

U radu je definiran model obogaćenoga društvenoga grafa. Obogaćeni društveni graf formalni je prikaz društvene mreže, tj. odnosa među ljudima. Obogaćenim društvenim grafom moguće je zapisati različite vrste i intenzitete odnosa među ljudima. Intenzitet odnosa među ljudima bilježi se kao njihova društvena udaljenost, a ona se računa na temelju modela koji se u ovom radu uvode.

Ključne riječi: analiza društvenih mreža, društvena udaljenost, društvene mreže, Facebook, obogaćeni društveni graf, sustavi za društveno umrežavanje

EXTENDED SUMMARY

Modelling of enriched social graph based on concept of distance

Introduction

Social networking sites (SNS), sometimes called *online social networks (OSNs)*, are widely used for communication and interaction by billions of people. With over two billion active users, Facebook is currently the most popular social networking site today.

Social networking sites are most often used by people to publish information about themselves, their attitudes and thoughts, as well as to digitally interact with other people. User interaction is performed in different ways: by liking and commenting other people's posts, photos, videos or shares, by exchanging private messages, appearance at mutual photos, etc. SNS users are mostly connected in an explicit fashion, i.e. they are required to make a direct decision regarding who will be their friend or who they are going to "follow". This explicit connection however does not differentiate between various natures and intensities of people's interpersonal relations.

The aim of this thesis is to develop a model that will be able to numerically represent the intensity of relations between people based on analysis of their interaction. The thesis offers several novelties:

1. It introduces and formally describes an **enriched social graph** as a means for representing nature and intensity of interpersonal relations
2. It describes a way of representing relationship intensity by using the concept of distance, i.e. it proposes and verifies a model for calculating social distance

During the research for this thesis three large-scale social surveys were held using the Facebook social networking site with the aim of collecting representative datasets for developing and verifying the proposed model for calculating social distance.

The thesis is structured in the following manner:

First section analyzes the emergence and evolution of social networking sites. The concept of distance is introduced and described in the second section. Third section formally defines an enriched social graph as a formal representation of relations between people. The fourth section is devoted to the description of the conducted social surveys

which are very important part of this thesis. Collected dataset is used for verifying the model for calculating social distance and the model of the enriched social graph. The fifth section presents the results of an exploratory analysis of reference dataset collected in the described surveys. Sixth section is devoted to social distance. It explains the advantages of expressing the intensity of interpersonal relations through the application of the concept of distance. In the sixth section various models for calculating social distance are proposed. These models are fine-tuned and verified on the reference dataset. Proposed models for calculating social distance are then used for verification of the enriched social graph model. The seventh section gives an overview of the trend of increased focus on personal data protection, which is particularly present after European General Data Protection Regulation (GDPR) came into effect. In that section conceptual model for the generation of an extended social graph (synthetic data about interaction of people on SNS) is proposed.

First section – Emergence and evolution of social networking sites

Social networking sites are a widespread platform for communication and interaction between billions of people. Users of social networking sites use them to interact with other people. Social networking sites are websites or mobile applications which are implemented in accordance with Web 2.0 principles, i.e. users can actively create content instead of being only passive consumers of previously prepared content. Although it is difficult to unambiguously define social networking sites and separate them from other similar systems, social networking sites should have the following 4 features:

1. Users are able to create their own profile
2. Users can explicitly create their connectivity (friendships, followings, etc.)
3. Users can publish content which will be visible to other users
4. User interaction mechanism needs to be the backbone of system functionality

First SNS was *SixDegrees.com* which was realized in 1997 and was active to 2000. Nowadays SNS with the largest number of users is *Facebook*. Other important SNSs are: Twitter, Instagram, QZone and VKontakte.

Second section – Concept of distance

Distance is a measure which shows how two close (or far) observed objects are from one another. The concept of distance most often relates to the physical or geographical distance – for example the distance from Zagreb to Rijeka. However, the concept of distance can also be applied to measure the dissimilarity between different objects.

Distance (or metrics) in the mathematical sense should be a positive-definite function and should have properties of symmetry and triangle inequality. Larger number means that the objects being compared are more different. In the field of machine learning sometimes it is enough to choose just a positive-definite function as a measure of distance. That kind of distance measures are called pseudometrics.

Distance can be calculated between various types of variables such as numeric, binary, ordered/unordered categorical, strings etc. Some of the most commonly used measures for calculating distance between objects described by numerical variables are the Euclid, Manhattan and Chebyshev distances, which are all special cases of the Minkowski distance, then the chord distance, geodesic distance and Mahalanobis distance.

Third section – Formal representation of relations between people – enriched social graph

Social network is a social structure composed of people connected by specific relations (family, friends, common interests, etc.). With analysis in mind, social network is often represented as a **social graph**. Social graph is composed of nodes and edges (or ties), where nodes represent people and edges represent relations between them. A **binary social graph** is a special case of social graph where only the existence of an edge is known, without type or weight assigned to it. Examples of binary graphs are a Facebook friendship graph or a Twitter followers graph. By analyzing interaction of people on SNSs, it is possible to enrich the semantically poor binary relations, i.e., instead of having only information if some relation exists or not, we can have the possibility to represent the intensity of some relation using a continuous numerical value. Enriched social graph enables storing of information about intensities of different types of relations between of pairs of people.

Fourth section – Conducted social surveys

In scope of this thesis three social surveys were conducted. All of them were based on the API provided by the social networking site Facebook. The goal of these surveys was to collect representative datasets which contain data about users' interactions on Facebook and users' evaluations of intensity of their relations between them and their Facebook friends, to be used as ground truth. That kind of dataset would be used in further researches as a basis for developing models which will be able to estimate intensity of users' relations by considering their interaction on Facebook.

In the first survey called *BestFriends* users were asked to identify up to 10 of their best friends and the system (after users' approval) then collected aggregated data about users' interaction with their friends. The resulting model was ultimately able to leverage the collected data and to approximately correctly pick 7 of (up to) 10 users' best friends.

Second survey called *NajFrend* is the most significant research held in the scope of this thesis. This survey included over 3000 of examinees. The survey had two basic goals: to verify the model developed as part of *BestFriend* research (1), to collect a much larger sample considering the number of examinees and in the sense of wealth of ground truth information gained by users (2). At scale from 1 to 5, where 5 is the best grade, users evaluated the model from research *BestFriends* by average grade 3.48.

Third survey was focused on users' experience of their interaction with their friends on Facebook. In previous researches the most important interaction parameters were found by using predictive techniques based on collected dataset. The goal of this survey was to examine users' impression about their interaction habits.

Fifth section – Exploratory analysis of users' interaction on Facebook

As data about users' interaction on SNS Facebook is a foundation for the model which calculates relations' intensities between users, the first step is to become acquainted with that data. Exploratory analysis is conducted for the interaction data summarized on ego-user and pairwise level. Through both analyses, correlations between different interaction parameters as well as distributions of each interaction parameter were extracted. Also, a relationship between total amount of Facebook friends and total amount of interaction was analyzed. For most of interaction parameters it was shown that the total amount of interaction does not linearly grow with the growth of number of friends. It

confirms the Dunbar's number hypothesis which says that total number of friends with whom ego-user can stay in contact is limited by neocortex size, i.e. although someone can have a lot of Facebook friends, only a small subset of that friends are his real life friends – others are only his acquaintances or random followers. In addition, users were clustered by considering their interaction habits.

Sixth section – Social distance

Social distance, in the scope of this thesis, represents a measure of interpersonal relations. Basic goal of this thesis is to develop model for measuring social distance, i.e. model which will be able to find a numerical representation of intensity of interpersonal relations based on people's interaction on Facebook and, based on it, build an enriched social graph.

This section presents various models for calculating social distance based on one or more interaction parameters. Some models use Minkowski distance, some average distance and some Mahalanobis distance. For all tested distance measures optimal coefficients of significance of particular interaction parameter are determined by using genetic algorithms. The worst results were gained for Mahalanobis distance, and the best for Minkowski distances. It turned out that changing the parameters of Minkowski distances does not significantly change the model's performance. It has also been shown that models which consider more than one interaction parameter perform better with experimentally obtained significance coefficients of each interaction parameter as opposed to the case when all significance coefficients are set to the same value. It was additionally checked whether models exhibit better performance if users are first clustered into groups considering their interaction habits (ratio of using each interaction parameter) and if significance coefficients would be found out for each cluster separately. It has been shown that it does not improve performance of models. However, it has been shown that the models provide different average performance for different clusters of users, so, based on user's cluster, it is possible to predict estimated accuracy of the model for that user. Also, it is verified whether the total amount of user's interaction affects the performance of the model for him. It turned out that model performs better for users who have more total interaction.

The model for calculation social distance can be used to build an enriched social graph.

Seventh section – The role of synthetic data in future research of social networks

This section comments recent trends of increased focus on personal data protection which is present after European General Data Protection Regulation (GDPR) came into effect. Conceptual model for synthetic expanded social graph generation is presented. Such artificially generated data for future research purposes could replace empirical data and thus enable research without the need for direct contact with (sensitive) personal data.

Conclusion

This thesis introduces, formally describes and verifies the concept of an enriched social graph. Intensity of interpersonal relations is numerically presented by using the concept of distance.

In the thesis a special attention is given to experimental research. Three social surveys were conducted and very valuable reference datasets were collected.

Thesis proposes and compares various models for calculating social distance and uses them to develop an enriched social graph.

As future research based on personal data are very sensitive, conceptual model for generating synthetic expanded social graph is proposed.

Keywords: social networks analysis, social distance, social networks, Facebook, enriched social graph, social networking sites

Sadržaj

UVOD

Sustavi za društveno umrežavanje (engl. *social networking sites – SNS* ili *online social networks - OSN*) široko su rasprostranjena platforma za komunikaciju i interakciju (međudjelovanje) milijardi ljudi. S više od dvije milijarde aktivnih korisnika, Facebook je uvjerljivo najpopularniji sustav za društveno umrežavanje današnjice [1].

Korisnici sustave za društveno umrežavanje najčešće koriste za objavljivanje informacija o sebi, svojih stavova i razmišljanja te za interakciju s drugim ljudima. Interakcija korisnika obavlja se na različite načine: iskazivanje odobravanja za objave drugih ljudi, komentiranje objava, razmjena privatnih poruka, zajedničko pojavljivanje na fotografijama i slično. Korisnici se na sustavima za društveno umrežavanje najčešće povezuju eksplicitno, tj. barem jedan od povezanih ljudi mora inicirati povezivanje, a ponekad i onaj drugi taj zahtjev mora odobriti. Povezani korisnici na Facebooku se nazivaju *prijateljima*, dok su na Twitteru i Instagramu to *sljedbenici*.

Informacija o tome jesu li ljudi eksplicitno povezani na sustavu za društveno umrežavanje ne otkriva puno o prirodi ili intenzitetu njihova odnosa. Primjerice, renomirana antropološka istraživanja pokazuju da prosječan čovjek može održavati do okvirno 150 stabilnih društvenih veza (imati do okvirno 150 prijatelja) [2], dok je u istraživanju koje je provedeno u sklopu izrade ovog doktorskog rada prosječan broj prijatelja na Facebooku bio 429. Jasno je da većina tih prijatelja nisu stvarni prijatelji ispitanika nego da se radi o mješavini pravih prijatelja, poznanika pa možda i ljudi koji se nikada uživo s promatranim ispitanikom nisu vidjeli.

Iako samo poznavanje povezanosti korisnika na sustavima za društveno umrežavanje nije dovoljno za određivanje prirode i intenziteta odnosa povezanih ljudi, na temelju analize interakcije korisnika na sustavima za društveno umrežavanje može se, manje ili više precizno, odrediti intenzitet njihova odnosa [3]–[19]. Dio istraživanja pokazuje da odnosi korisnika na sustavima za društveno umrežavanja u velikoj mjeri emuliraju njihove stvarnoživotne odnose [3], [4], [20], dok drugi dio tu povezanost detaljnije analizira te zaključuje da razina povezanosti ta dva *svijeta* ovisi o vrsti osobnosti promatranih korisnika [21]. Ipak, i jedni i drugi se slažu da postoji korelacija između odnosa ljudi u virtualnom i stvarnom svijetu.

Međutim, zašto je uopće bitno imati uvid u intenzitet odnosa različitih ljudi? Mogućih je primjena mnogo, a u nastavku će biti pobrojane neke od zanimljivijih. Sociolozi i socijalni psiholozi nastoje objasniti kako drugi ljudi utječu na naše osjećaje, razmišljanja i ponašanja

[20], [22]. Poznavanje povezanosti ljudi nužan je preduvjet za to. U telekomunikacijskom sektoru davatelji usluge često pokušavaju otkriti korisnike sklone promjeni mreže (engl. *churners*) [23]–[26]. To je moguće na temelju analize ponašanja pojedinaca, ali i kroz analizu njihova društvenog okruženja. Društveno okruženje može se analizirati kroz analizu društvenih grafova. Davatelji usluge trebaju izraditi društvene grafove u kojima veza ima značenje *utjecajnosti* jednog korisnika na drugog kako bi mogli predviđati hoće li odlazak jednog korisnika potaknuti i druge s njime povezane korisnike na isto. Uz telekome, odnose među ljudima žele analizirati i različite druge organizacije. Primjerice, radi povećavanja radne učinkovitosti svojih zaposlenika. Vrsta povezanosti koja je u tom slučaju zanimljiva jest razina komunikacije i suradnje među zaposlenicima [27]–[32]. Organizacije koje se bave prodajom ili najmom različitih proizvoda zainteresirane su poznavati interese svojih korisnika i njihovih prijatelja. Takve organizacije mogu izgraditi društveni graf u kojem povezanost ima značenje sličnosti interesa (povezani korisnici vole slične vrste filmova, knjiga, glazbe, kazališnih predstava i slično) ili razine povjerenja [12], [15], [33]. Sličnost interesa ili razina povjerenja mogu biti iskorišteni i za unapređenje rada preporučiteljskih sustava [34]–[37] kroz primjenu načela *homofilije*[1]. Uz poznavanje društvenog grafa povjerenja ili sličnosti interesa, preporuke preporučiteljskih sustava više nisu nužno zasnovane samo na prethodno iskazanim interesima korisnika već i na interesima koje su iskazali prijatelji korisnika. Na taj se način značajno smanjuje problem monotonije u listama preporuke[2]. Poznavanje odnosa među ljudima može imati primjenu i u obrazovnom sektoru. Analiza društvenoga grafa studenata može olakšati razumijevanje njihova ponašanja [38]. Na temelju analize društvenoga grafa studenata nastavnik može odrediti utjecajne studente u grupi, tj. studente koji su središnji čvorovi društvenoga grafa [39]. Ako ti studenti pravovremeno usvoje programom propisana znanja i vještine, vjerojatno je da će i propagacija tih znanja prema ostalim studentima biti uspješnija.

Za sve opisane primjene bitno je imati pristup (osobnim) podatcima ljudi koji su predmet interesa. Kako se radi o osjetljivim podatcima, za njihovu obradu i svako moguće korištenje prethodno je potrebno imati privolu ili neku drugu osnovu za obradu koju kao prihvatljivu propisuje *Opća uredba o zaštiti podataka*[3] [40].

[1] Načelo *homofilije* podrazumijeva da se ljudi druže s drugim ljudima s kojima dijele stavove, razmišljanja i interese.

[2] Monotonija u listama preporuke znači da se korisnicima učestalo preporučuju isti ili slični proizvodi zasnovani isključivo na njihovim prethodno iskazanim interesima. Slijedeći isključivo prethodno iskazane interese korisnika, teško je otkriti nešto novo, različito od prethodno konzumiranog, što bi korisniku bilo zanimljivo.

[3] Skraćeni naziv za *Uredba (EU) 2016/679 Europskog parlamenta i vijeća o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka* koja je donesena 27. travnja 2016., a u potpunosti se primjenjuje od 25. svibnja 2018. Koristi se još i engleska kratica GDPR (engl. *general data protection regulation*).

Ovaj doktorski rad nastavak je mnogih prethodnih istraživanja koja su za cilj imala odrediti intenzitet odnosa među ljudima na temelju njihove interakcije na sustavima za društveno umrežavanje. Međutim, u ovom se radu uvode dvije važne novosti:

1. Formalno se opisuje **obogaćeni društveni graf** kao zapis vrste i intenziteta različitih međuljudskih odnosa
2. Intenzitet odnosa iskazuje se primjenom **koncepta udaljenosti**, tj. uvodi se model za računanje **društvene udaljenosti**.

U radu se opisuje postupak izgradnje obogaćenoga društvenoga grafa primjenom koncepta udaljenosti. Društvene udaljenosti računaju se na temelju interakcije korisnika na sustavima za društveno umrežavanje. Razvoj modela za računanje društvene udaljenosti zasnovan je na primjeni principa nadziranog strojnog učenja. Za provođenje procesa učenja modela, potrebno je imati na raspolaganju reprezentativni podatkovni skup. S ciljem prikupljanja takvog skupa, u sklopu izrade rada provedeno je nekoliko društvenih istraživanja kroz koja su prikupljeni vrlo vrijedni podatkovni skupovi koji su iskorišteni za izgradnju i verifikaciju modela za računanje društvene udaljenosti. Kroz verifikaciju modela za računanje društvene udaljenosti verificiran je i model obogaćenog društvenog grafa, tj. pokazano je da je na temelju izračunatih društvenih udaljenosti među ljudima moguće izgraditi obogaćeni društveni graf kao model formalnog prikaza međuljudskih odnosa.

Doktorski rad strukturiran je kako slijedi. U prvom poglavlju analizira se pojava i razvoj sustava za društveno umrežavanje. U drugom poglavlju uvodi se koncept udaljenosti. Razmatra se pojam udaljenosti ili metrike u matematičkom smislu te se opisuju različiti načini računanja udaljenosti. U trećem poglavlju formalno se definira prikaz odnosa među ljudima kroz model obogaćenog društvenog grafa. Četvrto poglavlje posvećeno je opisu provedenih društvenih istraživanja koja čine vrlo važan dio ovog doktorskog rada i služe kao temelj za verifikaciju modela za računanje društvene udaljenosti te modela obogaćenog društvenog grafa. U petom su poglavlju izneseni rezultati eksploratorne analize istraživanjima prikupljenog referentnog podatkovnog skupa. Šesto poglavlje posvećeno je društvenoj udaljenosti. U njemu se obrazlažu prednosti iskazivanja intenziteta odnosa među ljudima kroz primjenu koncepta udaljenosti te se predlažu, uz opis postupka izgradnje, i verificiraju različiti modeli za računanje društvene udaljenosti. Na kraju poglavlja prethodno izneseni rezultati iskorišteni su za verifikaciju predloženog modela obogaćenog društvenog grafa zasnovanog na konceptu (društvene) udaljenosti. Sedmo, a ujedno i posljednje, poglavlje daje osvrt na trend povećanog fokusa na zaštitu osobnih podataka koji je posebno prisutan nakon stupanja na snagu *Opće uredbe o zaštiti*

podataka. S ciljem davanja novog poticaja istraživanjima za koja je pristup osobnim podatcima preduvjet i s ciljem davanja doprinosa sprječavanju mogućih zlouporaba osobnih podataka, izlaže se idejno rješenje za generiranje sintetičkog proširenog društvenog grafa, tj. umjetnih podataka o interakciji korisnika sustava za društveno umrežavanje. Takvi umjetno generirani podatci za potrebe istraživanja u budućnosti bi mogli zamijeniti stvarne podatke i na taj način omogućiti istraživački rad bez potrebe za izravnim kontaktom s (osjetljivim) osobnim podatcima.

I. POJAVA I RAZVOJ SUSTAVA ZA DRUŠTVENO UMREŽAVANJE

Sustavi za društveno umrežavanje (engl. *social networking sites – SNS* ili *online social networks - OSN*) ili aplikacije za upravljanjem društvenim mrežama često se u razgovornom jeziku (pogrešno) nazivaju i samo *društvenim mrežama*. Društvene mreže sociološki su pojam puno duže izučavan od postojanja sustava za društveno umrežavanje. Pojam društvenih mreža podrazumijeva ljude i odnose među njima. Primjerice, svi ljudi na svijetu čine jednu veliku društvenu mrežu. Ta se velika društvena mreža može proučavati kao cjelina ili se mogu proučavati neki njezini izdvojeni dijelovi. Primjerice, mogu se proučavati ego-mreže pojedinih ljudi. Ego-mreža jest mreža koja sadrži odabranog korisnika (ego-korisnika) i njegove prijatelje, tj. ljude koji su izravno s njime povezani. Ego-mreža dalje se može proširivati dodavanjem ego-mreža prijatelja početnog ego-korisnika. Takvim dodavanjem ego-mreža korisnika zastupljenih u mreži u svega nekoliko koraka doći ćemo do potpune mreže. Više istraživanja pokazalo je da se u prosječno 6 koraka može povezati bilo koja dva čovjeka na svijetu [41], [42]. Općenito, pokazuje se da prosječan stupanj udaljenosti bilo koja dva čovjeka u društvenoj mreži nije veći od logaritma broja ljudi u mreži. Taj se fenomen zove *šest stupnjeva odvojenosti* ili *fenomen malog svijeta*.

Istraživanja društvenih mreža dugi su niz godina bila teško provediva zbog nedostatka podataka o društvenoj povezanosti. Pojava sustava za društveno umrežavanje omogućila je pohranu velike količine podataka o povezanosti ljudi na jednom mjestu u računalno čitljivom formatu. Na taj su se način otvorile neslućene mogućnosti za istraživanja društvenih mreža.

Sustavi za društveno umrežavanje vrlo su raširena platforma za komunikaciju i interakciju milijardi ljudi. Korisnici sustave za društveno umrežavanje koriste za objavljivanje informacija o sebi, objavljivanje svojih razmišljanja i komentiranje svijeta oko sebe te za interakciju s drugim ljudima. Sustavi za društveno umrežavanje u svojoj su osnovi *web*-stranice ili mobilne aplikacije izvedene u skladu s načelima *weba 2.0*, tj. omogućuju korisnicima da, umjesto da budu samo pasivni čitatelji prethodno pripremljenog im sadržaja, sami kreiraju sadržaj. Iako je teško nedvosmisleno definirati sustave za društveno umrežavanje i razdvojiti ih od drugih sličnih sustava, recimo da bi neki sustav, kako bi ga se moglo nazvati sustavom za društveno umrežavanje, morao omogućavati barem sljedeće:

1. Izradu profila korisnika
2. Eksplicitno povezivanje korisnika (sklapanje *prijateljstava*, *sljedbeništvo* i slično)
3. Objavu sadržaja koji će biti vidljiv drugim korisnicima sustava
4. Mehanizam za interakciju korisnika kao okosnicu funkcionalnosti sustava

S obzirom na to da ne postoji općeprihvaćen popis svojstava sustava za društveno umrežavanje, često se pod sustavima za društveno umrežavanje vode i sustavi koji to nisu. Primjerice, u sustave za društveno umrežavanje nerijetko se ubrajaju i sustavi koji služe isključivo kao porukatori (primjerice WhatsApp ili Viber) ili sustavi za dijeljenje video-sadržaja (primjerice YouTube[4]). Upravo zbog ovog zadnjeg bitno je naglasiti da sustavi za društveno umrežavanje moraju imati mehanizam za interakciju korisnika kao okosnicu funkcionalnosti sustava. Sustavi za društveno umrežavanje trebaju biti orijentirani prema interakciji korisnika, a ne biti mjesto na kojem će različiti korisnici samo ostavljati svoj sadržaj koji će tek sporadično biti komentiran – svi bi korisnici trebali i stvarati sadržaj i pregledavati sadržaj koji su drugi stvorili.

Sustavi za društveno umrežavanje ponekad se opisuju i kao sustavi koji olakšavaju pojedincima upravljanje njihovim ego-mrežama.

Različiti sustavi za društveno umrežavanje na različite načine upravljaju povezivanjem korisnika. Facebook svoje povezane korisnike naziva *prijateljima*. Da bi do povezivanja došlo, jedan od korisnika drugome prvo mora poslati *zahtjev za prijateljstvom* nakon čega ovaj drugi taj zahtjev treba odobriti. Korisnici postaju prijatelji tek nakon što oba uključena korisnika na to pristanu. S druge strane, Twitter omogućuje povezivanje korisnika kroz odnos *praćenja* ili *sljedbeništva*. Umjesto dvostrane potvrde prijateljstva, na Twitteru je dovoljno odlučiti pratiti objave nekog korisnika. Ako pak i taj drugi korisnik želi pratiti objave prvoga, i on će njega početi pratiti, ali obostrana povezanost nije nužna. Instagram je pak na neki način hibrid Twittera i Facebooka. Iako se veze praćenja (sljedbeništva) uspostavljaju u osnovi jednostrano, neki korisnici mogu imati zaključane profile koji traže eksplicitno odobrenje prije uspostavljanja sljedbeničke povezanosti[5].

[4] Činjenica jest da na YouTubeu korisnici imaju svoje profile, da mogu objavljivati sadržaj i da taj sadržaj drugi korisnici mogu komentirati, ali naglasak nije na interakciji korisnika nego na pregledavanju video-materijala, neovisno o tome tko ga je objavio. Većina korisnika ne objavljuje ništa nego samo pregledava sadržaj koji su drugi objavili.

[5] S vremenom je i Twitter uveo mogućnost ograničavanja popisa ljudi kojima će njihove objave biti vidljive, ali kako je srž Twittera objava do 140 znakova dugih tekstualnih objava, a Instagrama objava osobnih multimedijalnih sadržaja, razumljivo je da korisnici Instagrama imaju veću potrebu za *zatvaranjem* svojih profila.

Međutim, kako su se sustavi za društveno umrežavanje pojavili i kako su postali jedna od središnjih stvari u životima milijardi ljudi? Preduvjet za postojanje sustava za društveno umrežavanje jest postojanje električnih računala i računalnih mreža. Prva računalna mreža izrađena je 1969. godine u Sjedinjenim Američkim Državama pod nazivom ARPANET. Oformilo ju je američko Ministarstvo obrane. 1971. kroz mrežu ARPANET bilo je moguće po prvi puta poslati poruku s jednog na drugo računalo pa se ta godina smatra godinom začetka elektroničke pošte. 1979. pojavili su se BBS-ovi (engl. *Bulletin Board System* – elektronička oglasna ploča). Na BBS-ove korisnici su se mogli spajati korištenjem klasične telefonske linije. Po prijavi u sustav, korisnici su mogli razmjenjivati poruke s drugim korisnicima, čitati vijesti, preuzimati programe i slično. 1979. pokrenut je Usenet. Usenet je omogućavao formiranje javnih grupa čiji je sadržaj mogao čitati i u njih pisati bilo tko. Na taj su način u različitim grupama različiti ljudi međusobno mogli diskutirati o različitim temama. 1989. Tim Berners-Lee predlaže *World Wide Web* zasnovan na hipervezama, a 1991. godine izrađuje i prvu *web*-stranicu. 1993. napravljen je prvi *web*-preglednik s grafičkim sučeljem – Mosaic. U narednim godinama *web* se razvija i postaje temeljna platforma za mnoge druge sustave. U prvim je godinama *web* bio organiziran na način da je mali broj korisnika generirao sadržaj koji je većina korisnika tek pratila (čitala). *Web 2.0* trend je koji se počeo javljati na prijelazu tisućljeća, a koji znači transformaciju iz pasivnog u aktivni *web*, tj. transformaciju iz *weba* u kojem tek mali broj korisnika generira sadržaj u *web* na kojem svi korisnici imaju mogućnost jednostavnog generiranja sadržaja.

1996. u rad je pušten ICQ – prvi program za stvarnovremensko dopisivanje koji se mogao preuzeti i instalirati na klijentskom računalu. Godinu dana kasnije, 1997. godine, u rad je pušten prvi sustav koji bi se mogao nazvati sustavom za društveno umrežavanje – SixDegrees.com. Taj je sustav omogućavao stvaranje profila korisnika, međusobno povezivanje korisnika, stvaranje popisa prijatelja, rodbine i poznanika, pretraživanje drugih korisnika po iskazanim interesima te komunikaciju korisnika kroz javne objave na *oglasnoj ploči* ili kroz razmjenu privatnih poruka. Na svom vrhuncu sustav je imao oko 3,5 milijuna registriranih korisnika, ali potrajao je svega nešto više od 3 godine. Ugašen je 2000. godine uslijed nedovoljnog broja korisnika i posljedično nedovoljnih prihoda. Osnovni razlog za to je u to vrijeme još nedovoljno razvijena internetska infrastruktura što je za posljedicu imalo nedovoljno velik broj ljudi koji su imali stalni pristup Internetu. U to se vrijeme na Internet još uvijek većinski spajalo *dail-up* modemima, naplaćivalo se vrijeme provedeno na Internetu, a brzine su bile vrlo male.

1998. godine osniva se Google i s radom počinje istoimena tražilica koja će u narednim godinama postati najznačajnija tražilica *weba*. 1999. godine pušten je u rad Napster, sustav za P2P dijeljenje glazbe. 2001. godine s radom počinje Wikipedija – slobodna internetska enciklopedija čiji sadržaj pišu korisnici. 2003. godine s radom počinje LinkedIn – prvi sustav za poslovno društveno umrežavanje. LinkedIn služi korisnicima za prezentaciju svojih kompetencijama potencijalnim budućim poslodavcima. 2003. godine s radom počinje i Skype – aplikacija koja omogućuje stvarnovremensku audio i video komunikaciju korisnika putem Interneta. Pojava te aplikacije na neki se način može smatrati i revolucionarnom jer je postalo moguće razgovarati s ljudima s drugog kraja svijeta po cijeni nižoj od cijene lokalnog telefonskog poziva u vrijeme kada su međunarodni, a pogotovo međukontinentalni, pozivi bili izrazito skupi. 2003. godine s radom kreće i MySpace – prvi veliki sustav za društveno umrežavanje. U razdoblju od 2005. do 2008. MySpace je bio najveći sustav za društveno umrežavanje, a u lipnju 2006. pretekao je Google i postao najposjećenijom *web*-stranicom u SAD-u.

2004. godine među studentima Harvardova sveučilišta s radom počinje Facebook, 2005. godine Youtube, a 2006. Twitter. S pojavom Facebooka počeo je i streloviti uspon sustava za društveno umrežavanje. U svibnju 2008. Facebook je pretekao MySpace na mjestu najraširenijeg sustava za društveno umrežavanje, a čelnu poziciju uvjerljivo drži do današnjih dana. Prema statističkim podatcima iz listopada 2018. [1], Facebook je s preko 2,2 milijarde aktivnih korisnika uvjerljivo najpopularniji sustav za društveno umrežavanje. Slijede ga Youtube s 1,9 milijardi aktivnih korisnika, WhatsApp s 1,5 te Facebook Messenger s 1,3 milijarde aktivnih korisnika, ali ni jedan od tih sustava ne odgovara ranije danom opisu sustava za društveno umrežavanje. Instagram, kao realno drugi po broju korisnika sustav za društveno umrežavanje, ima milijardu aktivnih korisnika, a slijedi ga kineski sustav za društveno umrežavanje QZone s 548 milijuna aktivnih korisnika te Twitter s 335 milijuna aktivnih korisnika. Svoju neprikosnovenost na prvom mjestu Facebook je prije svega zaslužio svojim brzim promjenama i prilagodbama potrebama korisnika. 2012. godine Facebook je kupio, tada još relativno novu, aplikaciju Instagram. Instagram je danas po broju korisnika drugi najveći sustav za društveno umrežavanje s tendencijom da prestigne Facebook. Nudi puno manje mogućnosti od Facebooka, puno je jednostavniji i samim time posebno popularan među mlađom populacijom. Istraživanje koje je provodio Microsoft pokazalo je da su korisnici 2000. godine prosječno mogli pažnju držati 12 sekundi, a 2015. 8 sekundi [43]. Skraćeno vrijeme držanja pažnje za posljedicu ima sklonost korisnika prema jednostavnijim, vizualnijim oblicima komunikacije. S obzirom na to, u doglednom bi vremenu Facebook, kao platforma vrlo širokih

mogućnosti, mogao biti zamijenjen Instagramom – aplikacijom koja nudi dijeljenje fotografija i videa uz minimalne tekstualne opise.

Osim međunarodnih sustava za društveno umrežavanje, bilo je i domaćih pokušaja izrade takvih sustava. Među domaćim sustavima za društveno umrežavanje opće namjene najpoznatiji je Trosjed. Sustav je počeo s radom krajem 2007., a danas više nije aktivan. Na svjetskoj razini postoji nekoliko lokalno vrlo jakih sustava za društveno umrežavanje: VKontakte vrlo je raširen u Rusiji, a već spomenuti QZone u Kini.

Sustavi za društveno umrežavanje, u današnje vrijeme pametnih telefona i gotovo neprekinute povezanosti na Internet, neizostavni su dio svakodnevnice milijardi ljudi diljem svijeta.

II. KONCEPT UDALJENOSTI

Udaljenost je općenito govoreći mjera koja pokazuje koliko su dva objekta blizu (ili daleko) jedan od drugoga. Uz pojam udaljenosti najčešće se veže značenje fizičke ili geografske udaljenosti – primjerice udaljenost od Zagreba do Rijeke. Međutim, koncept udaljenosti može se primijeniti i u nekim drugim područjima. Recimo, može se računati „udaljenost" između dva filma ili dvije knjige. Tada ta udaljenost, naravno, nema značenje koliko su ti filmovi ili knjige udaljeni na policama na kojima se nalazi jedna od njihovih kopija, nego koliko su s obzirom na svoja svojstva (tematika, žanr, glumci, karakteri likova…) različiti (ili slični).

Značenje udaljenosti intuitivno je jasno kada se radi o dvodimenzionalnom ili trodimenzionalnom prostoru, ali računanje udaljenosti može se poopćiti i na N-dimenzionalan prostor. Pretpostavimo da film možemo opisati sljedećim značajkama: žanr filma (numerička reprezentacija), vrijeme proteklo od snimanja, ocjena na IMDbu[6] i broj dobivenih nagrada na međunarodnim filmskim festivalima. Spomenute značajke možemo zapisati kao uređenu četvorku koju onda možemo prikazati kao točku u 4-dimenzionalnom prostoru. Udaljenost između dva filma jest udaljenost između točaka koje predstavljaju filmove.

Udaljenost možemo računati i između riječi ili kodnih riječi (tj. nizova bitova) na način da, primjerice, kažemo da su (kodne) riječi udaljene onoliko u koliko se slova ili bitova razlikuju. Recimo, kodne riječi *11001* i *11000* udaljene se za 1 jer im se razlikuje samo zadnji bit, kao i riječi *pas* i *gas* koje se također razlikuju tek u jednom slovu.

Udaljenost između dva objekta može biti simetrična, tj. jednaka neovisno o ishodištu njezina računanja – primjerice, udaljenost između dvije riječi. Međutim, udaljenost između dva objekta može se i razlikovati ovisno o odabiru ishodišta. Primjerice, prema Googleovim kartama najkraća cestovna udaljenosti od Zagreba do Rijeke iznosi 161 km, dok je najkraća cestovna udaljenost od Rijeke do Zagreba 166 km.

Općenito, sličnost ili različitost objekata može se iskazivati primjenom koncepta sličnosti (engl. *similarity*) ili različitosti (engl. *dissimilarity*), tj. udaljenosti (engl. *distance*). Osnovna i najvažnija razlika između ta dva koncepta jest u tome što kod sličnosti veći broj znači da su uspoređivani objekti sličniji jedno drugome, dok kod udaljenosti (ili različitosti) veći broj znači da su objekti međusobno manje slični, tj. da su različitiji ili udaljeniji. U

[6] Internetska baza s informacijama o filmovima, televizijskom programu, video-igricama i sličnim sadržajima kroz koju korisnici imaju mogućnost ocjenjivati različite sadržaje. Uz filmove, u svrhu opisa zadovoljstva publike filmom, objavljuje se ocjena ostvarena na IMDbu (imdb.com).

nastavku će fokus biti na mjerenju udaljenosti, a u poglavlju VI bit će obrazloženo i zašto je to tako.

Osnovno svojstvo udaljenosti jest da udaljenost ne može poprimiti negativnu vrijednost te da manja vrijednost udaljenosti sugerira da su promatrani objekti bliže u odnosu na one za koje je utvrđena veća udaljenost.

Udaljenost između različitih vrsta objekata moguće je računati na različite načine. U nastavku će biti definirana udaljenost u matematičkom smislu, a nakon toga prezentirani različiti načini računanja udaljenosti među objektima opisanim varijablama različitih tipova.

II.1 UDALJENOST U MATEMATIČKOM SMISLU ILI METRIKA

Funkciju $d: R^n \times R^n \rightarrow [0, \infty\rangle$, sa svojstvom da za svaki $x, y \in R^n$ vrijedi:

1. $d(x, y) \geq 0, d(x, y) = 0 \Leftrightarrow x = y$ (pozitivna definitnost)
2. $d(x, y) = d(y, x)$ (simetričnost)
3. $d(x, y) \leq d(x, z) + d(z, y), \forall x, y \in R^n$ (nejednakost trokuta)

zovemo **metrikom na R^n**.

Vrijednost funkcije *d(x, y)* za neke $x, y \in R^n$ zovemo udaljenošću točaka *x* i *y*.

Ponekad je za potrebe strojnog učenja ovakva definicija metrike suviše stroga pa se uvode kvazimetrike. Kvazimetričke funkcije trebaju zadovoljiti samo svojstvo pozitivne definitnosti, dok ostala metrička svojstva ne moraju biti ispunjana. Za potrebe ovog rada kvaizmetričke funkcije dovoljno su dobre pa ćemo u nastavku pod pojmom udaljenosti podrazumijevati i metrike i kvazimetrike.

II.2 RAZLIČITI NAČINI RAČUNANJA UDALJENOSTI ZA RAZLIČITE TIPOVE PODATAKA

Udaljenosti se mogu računati između različitih objekata – artikala u prodavaonicama, filmova, knjiga, ljudi i slično. Objekti su opisani varijablama. Varijable mogu biti različitih tipova: nizovi znakova, kategorijske (ili nominalne) varijable, binarne varijable, poredane (ili ordinalne) varijable i brojčane (ili numeričke ili intervalne) varijable. Objekti mogu biti opisani varijablama istog tipa ili varijablama različitih tipova. U nastavku slijedi pregled različitih načina računanja udaljenosti za različite tipove varijabli. Pregled načina računanja udaljenosti najvećim se dijelom zasniva na [44]–[47].

II.2.1 Udaljenosti između nizova znakova

U području teorije informacije, teorije kodiranja, lingvistici ili računalnoj znanosti često je potrebno mjeriti sličnost (ili različitost) riječi, u smislu niza znakova, i kodnih riječi, u smislu niza bitova. U nastavku će biti dan pregled nekoliko najčešće korištenih mjera za računanje udaljenost riječi ili kodnih riječi.

II.2.1.1 Hammingova udaljenost

Hammingova udaljenost zasigurno je najpoznatija mjera udaljenosti među ponajprije kodnim riječima, a onda i riječima općenito. Hammingovom udaljenošću moguće je mjeriti udaljenost između nizova znakova jednake duljine. Iznos Hammingove udaljenosti jednak je broju različitih znakova na odgovarajućima mjestima u promatranim nizovima znakova. U nastavku je dano nekoliko primjera računanja Hammingove udaljenosti:

1. *10010110* i *11010010* – Hammingova udaljenost iznosi 2 jer su različiti tek 2. i 6. bit
2. *plutati* i *gledati* – Hammingova udaljenost iznosi 3 jer su različiti 1., 3. i 4. znak
3. *vrata* i *ratar* – Hammingova udaljenost iznosi 5 jer su odgovarajući znakovi na svih 5 mjesta različiti.

Kod binarnih kodnih riječi Hammingova udaljenost jednaka je broju *jedinica* u rezultatu operacije *isključivo ili* (XOR) nad promatranim binarnim kodnim riječima.

Hammingova udaljenost ponajviše se koristi u području otkrivanja ili ispravljanja pogrešaka[7] pri prijenosu kodnih riječi (ili niza bitova). U kodnom sustavu K može se detektirati do g grešaka ako je najmanja Hammingova udaljenosti između bilo koje dvije kodne riječi u kodnom sustavu $g+1$ (u slučaju da se dogodi $g+1$ greška, početna kodna riječ pretvorila bi se u drugu kodnu riječ pa bi bilo nemoguće zaključiti da to nije izvorno poslana kodna riječ nego da se dogodila greška pri prijenosu). Kako bi se g grešaka moglo ispraviti, najmanja Hammingova udaljenost između bilo koje dvije kodne riječi u kodnom sustavu mora biti $2g+1$ (pretpostavlja se da je izvorno poslana kodna riječ ona kojoj je primljena kodna riječ najsličnija – zato za otkrivanje g grešaka najmanja udaljenost između dvije kodne riječi mora biti $2g+1$).

II.2.1.2 Leejeva udaljenost

Leejevu udaljenost može se promatrati kao proširenje Hammingove udaljenosti. Hammingovom udaljenošću provjeravalo se jesu li znakovi na odgovarajućim mjestima jednaki

[7] Otkriti pogrešku znači biti svjestan da nije primljena ispravna kodna riječ, ali bez mogućnosti otkrivanja koja je kodna riječ poslana. Ispraviti pogrešku znači biti svjestan da nije primljena ispravna kodna riječ i biti u mogućnosti grešku ispraviti, tj. odrediti izvorno poslanu kodnu riječ.

ili različiti, a konačan rezultat bio je jednak broju mjesta na kojima su odgovarajući znakovi različiti, pri čemu je svaka različitost znakova bila jednako vrednovana. Leejeva udaljenost mjeri različitost između znakova na način da svakom znaku u *abecedi* (skupu mogućih znakova) pridijeli redni broj pa se udaljenošću smatra udaljenost njihovih rednih brojeva u *abecedi*. Pretpostavlja se da su redni brojevi znakova *abecede* nalaze na prstenu, a kod izračuna udaljenosti uzima se manja od dvije moguće udaljenosti između dva znaka. Tako su, primjerice, prvo i zadnje slovo abecede jedno do drugoga, tj. udaljeni tek za 1.

Primjerice, u abecedi engleskog jezika Leejeva udaljenost između znakova *a* i *d* iznosi 3, a u hrvatskoj abecedi 5. Naime, redni broj slova *a* u obje je abecede 1, ali je redni broj slova *d* u engleskoj abecedi 4, a u hrvatskoj 6. Udaljenost slova *a* i *z* u hrvatskoj abecedi iznosi 2, a u engleskoj 1.

Leejeva udaljenost dviju kodnih riječi jednaka je zbroju udaljenosti po odgovarajućim mjestima.

Leejevu udaljenost može se izraziti sljedećim izrazom:

$$Leejeva\ udaljenost = \sum_{i=1}^{n} \min(|x_i - y_i|, q - |x_i - y_i|), \qquad \text{(II.1)}$$

pri čemu je n dužina kodne riječi promatranog kodnog sustava (broj znakova), q broj znakova *abecede* promatranog kodnog sustava, a i je indeks mjesta u kodnoj riječi – primjerice x_3 je redni broj u *abecedi* znaka na trećem mjestu u prvoj kodnoj riječi, a y_3 redni broj u *abecedi* znaka na trećem mjestu u drugoj kodnoj riječi.

Ako abeceda kodnog sustava ima 2 ili 3 različita znaka, Leejeva udaljenost jednaka je Hammingovoj. U nastavku je dano nekoliko primjera računanja Leejeve udaljenosti:

1. *1011* i *1101* – ako pretpostavimo da je q=2, tj. u abecedi postoje znakovi samo *0* i *1*, Leejeva udaljenost je: 0+1+1+0 = 2, tj. jednaka Hammingovoj udaljenosti ta dva niza.
2. *3521* i *8245* – ako pretpostavimo da je q=*9*, tj. u abecedi postoje znakovi *1*, *2*, *3*, *4*, *5*, *6*, *7*, *8* i *9*, Leejeva udaljenost je: (9-(8-3)) + (5-2) + (4-2) + (5-1) = 13
3. *vrata* i *ratar* – ako pretpostavimo da je q=*30*, tj. da se radi o abecedi hrvatskog jezika, Leejeva udaljenost je: (28-23) + (30-(23-1)) + (30-(26-1)) + (30-(26-1)) + (30-(23-1)) = 31

II.2.1.3 Levenštejnova udaljenost

Za razliku od Hammingove i Leejeve udaljenosti koje su uspoređivale znakove na odgovarajućim mjestima u promatranim (kodnim) riječima, Levenštejnova udaljenost mjeri najmanji broj promjena znakova (dodavanja, brisanja ili zamjene) u jednoj od promatranih riječi kako bi ona postala jednaka kao ona druga promatrana riječ. S obzirom na to da su prihvatljive i operacije brisanja i dodavanja znakova, nije nužno da promatrane riječi budu jednake duljine. Levenštejnovu udaljenost može se promatrati i kroz perspektivu gornje i donje granice vrijednosti koje ona može poprimiti:

- Najmanja vrijednost Levenštejnove udaljenosti razlika je između duljina uspoređivanih riječi
- Najveća vrijednost Levenštejnove udaljenosti duljina je veće u paru uspoređivanih riječi
- Levenštejnova udaljenost iznosi 0 samo kada su uspoređivane riječi jednake
- Ako su uspoređivane riječi jednake duljine, Levenštejnova udaljenost bit će manja ili jednaka Hammingovoj udaljenosti.

Levenštejnova udaljenost najčešće se koristi za pronalaženje odabranih riječi u tekstovima. Pri tome je bitno moći otkriti riječi i u slučaju kada se dogodio zatipak ili se na temelju riječi u osnovnom obliku (nominativ, jednina, prvo lice, prezent i slično) pokušava otkriti ista ta riječ u nekom drugom obliku. Pronalaženje riječi s malom Levenštejnovom udaljenošću u odnosu na unesenu riječ koristi se i za rad pravopisnih provjerenika, koji za cilj imaju otkriti zatipke i manje pravopisne greške te predložiti ispravan oblik riječi. Sve se to radi na način da se, za riječi u tekstu koje ne postoje u rječniku, predlažu iz rječnika riječi s malom Levenštejnovom udaljenošću u odnosu na pojavnicu riječi u tekstu. Levenštejnova udaljenost koristi se i za računanje lingvističke udaljenosti, tj. određivanja koliko su dva jezika različita.

U nastavku je dano nekoliko primjera računanja Levenštejnove udaljenosti:

1. *10010110* i *11010010* – potrebno je promijeniti 2. i 6. znak druge kodne riječi kako bi ona postala jednaka prvoj pa Levenštejnova udaljenost između te dvije riječi iznosi 2 (jednako kao i Hammingova)
2. *plutati* i *gledati* – potrebno je promijeniti 1., 3. i 4. znak druge riječi kako bi ona postala jednaka prvoj pa Levenštejnova udaljenost između te dvije riječi iznosi 3 (jednako kao i Hammingova)
3. *vrata* i *ratar* – potrebno je na početak druge riječi dodati znak *v* te obrisati znak *r* s kraja te riječi kako bi ona postala jednaka prvoj riječi. Dakle, Levenštejnova

udaljenost ovih dviju riječi iznosi tek 2, iako se riječi izvorno razlikuju na svim odgovarajućim mjestima pa je Hammingova udaljenost među njima 5.

II.2.2 Udaljenosti između kategorijskih varijabli

Kategorijske ili nominalne varijable mogu poprimiti vrijednost iz konačnog skupa od dvije ili više kategorija. Među kategorijskim varijablama nije moguće definirati redoslijed. Kategorijska varijabla jest, primjerice, *boja kose* ili *spol*.

Udaljenost između dvije kategorijske varijable *x* i *y* – *d(x, y)* najjednostavnije se može definirati kao:

$$d(x,y) = \begin{cases} 0, ako\ je\ x = y \\ 1, ako\ je\ x \neq y \end{cases} \tag{II.2}$$

Udaljenost između dva objekta *A* i *B* – *D(A, B)*, od kojih je svaki opisan s po *n* kategorijskih varijabli a_i i b_i, može se izračunati prema izrazu:

$$D(A,B) = \sum_{i=1}^{n} d(a_i, b_i) \tag{II.3}$$

Želimo li uzeti u obzir i frekventnost pojavljivanja različitih kategorija u podatkovnom skupu, tada udaljenost između objekata *a* i *b* možemo izraziti na sljedeći način:

$$D(A,B) = \sum_{i=1}^{n} \frac{\left(N_{a_i} + N_{b_i}\right)}{N_{a_i} N_{b_i}} d(a_i, b_i), \tag{II.4}$$

pri čemu je N_{a_i} broj objekata u podatkovnom skupu koji vrijednost *i*-te varijable imaju jednaku kao objekt *A*.

II.2.3 Udaljenosti između binarnih varijabli

Binarne varijable posebna su vrsta kategorijskih varijabli kod kojih postoje samo dvije moguće vrijednosti. Najčešće su to *0* ili *1*, *postoji* ili *ne postoji*, *istina* ili *laž*, ali mogu biti i drugačije.

Objekti opisani binarnim varijablama mogu se opisati binarnim vektorima – vektorima kod kojih bilo koji element vektora može poprimiti vrijednost iz skupa od dvije moguće vrijednosti.

Binarne varijable mogu biti simetrične i asimetrične. Kod simetričnih varijabli oba moguća stanja jednako su vrijedna, dok se kod asimetričnih preferira jedno stanje. S obzirom na to, mogu se računati simetrične i asimetrične binarne udaljenosti. Primjerice, ako dva vektora s asimetričnim binarnim varijablama na odgovarajućim mjestima imaju vrijednost *1* (pozitivno podudaranje), to je za računanje udaljenosti značajnije nego kada na odgovarajućim mjestima imaju vrijednost *0* (negativno podudaranje). Štoviše, poklapanje vrijednosti *0* ponekad se zanemaruje u izračunu udaljenosti. Asimetrične binarne varijable ponekad se nazivaju i varijablama sa samo jednim stanjem (jer drugo moguće stanje nema značaj).

Prije prezentiranja različitih načina računanja udaljenosti između binarnih vektora, nužno je definirati varijable za opis nekih vrsta odnosa između dva binarna vektora koje će pojednostavniti definiranje različitih vrsta mjera udaljenosti između binarnih vektora:

$$A = S_{11}(a,b) = ab^T = \sum_i^n a_i\, b_i \tag{II.5a}$$

$$B = S_{01}(a,b) = \bar{a}b^T = \sum_i^n (1-a_i)\, b_i \tag{II.5b}$$

$$C = S_{10}(a,b) = a\bar{b}^T = \sum_i^n a_i\, (1-b_i) \tag{II.5c}$$

$$D = S_{00}(a,b) = \bar{a}\bar{b}^T = \sum_i^n (1-a_i)(1-b_i) \tag{II.5d}$$

$$\sigma = \sqrt{(A+B)(A+C)(B+D)(C+D)} \tag{II.5e}$$

$$n = A + B + C + D \tag{II.5f}$$

U gornjim izrazima korištene oznake imaju sljedeće značenje:

- *a* i *b* – promatrani binarni vektori
- *n* – duljina (broj elemenata) vektora *a* i *b*
- *A* – broj mjesta na kojima vektori *a* i *b* imaju vrijednost *1*
- *B* – broj mjesta na kojima vektor *a* ima vrijednost *0*, a vektor *b 1*
- *C* – broj mjesta na kojima vektor *a* ima vrijednost *1*, a vektor *b 0*
- *D* – broj mjesta na kojima vektori *a* i *b* imaju vrijednost *0*
- ^{T} – oznaka za transponiranje matrice (vektora)
- $\bar{x}$ – inverz vektora *x* (0 zamijene s 1, a 1 s 0)

U nastavku će biti prikazane različite simetrične i asimetrične mjere za računanje udaljenosti između binarnih vektora te primjer njihova korištenja.

II.2.3.1 Simetrične mjere za računanje udaljenosti između binarnih vektora

Simetrične mjere za računanje udaljenosti jednako vrednuju preklapanja vrijednosti *1* i vrijednosti *0* u promatranim vektorima. U nastavku će biti prezentirana tri primjera takvih mjera: Yuleova udaljenost, Pearsonova udaljenost te Rogers-Tanimotova udaljenost.

Yuleova udaljenost

Yuleova udaljenost mjera je simetrične udaljenosti koja uspoređuje umnožak slučajeva kada se odgovarajući elementi promatranih vektora ne podudaraju sa zbrojem umnožaka slučajeva kada se odgovarajući elementi podudaraju i kada se ne podudaraju. Definirana je sljedećim izrazom:

$$Yuleova\ udaljenost = \frac{BC}{AD + BC} \tag{II.6}$$

Pearsonova udaljenost

Pearsonova udaljenost uspoređuje umnožak slučajeva u kojima se odgovarajući elementi promatranih vektora podudaraju i umnožak slučajeva u kojima se odgovarajući elementi ne podudaraju. Za slučaj kada je umnožak slučajeva podudaranja veći od umnoška slučajeva nepodudaranja, vrijednost udaljenosti bit će manja od 0,5, a u slučaju kada je umnožak nepodudaranja veći, vrijednost udaljenosti bit će veća od 0,5. Pearsonova udaljenost definirana je sljedećim izrazom:

$$Pearsonova\ udaljenost = \frac{1}{2} - \frac{AD - BC}{2\sigma} \tag{II.7}$$

Rogers-Tanimotova udaljenost

Rogers-Tanimotova udaljenost mjera je simetrične udaljenosti koja značajnije vrednuje nepodudaranje vrijednosti odgovarajućih elemenata uspoređivanih binarnih vektora. Definirana je sljedećim izrazom:

$$Rogers - Tanimotova\ udaljenost = \frac{2(B + C)}{A + 2(B + C) + D} \tag{II.8}$$

II.2.3.2 Asimetrične mjere za računanje udaljenosti između binarnih vektora

Asimetrične mjere za računanje udaljenosti zanemaruju mjesta na kojima oba promatrana vektora imaju vrijednost *0*. Udaljenost između dva objekta to je manja na što više mjesta oba vektora poprimaju vrijednost *1*.

Jaccardova udaljenost

Jaccardova udaljenost mjera je asimetrične udaljenosti koja jednako vrednuje podudaranja i nepodudaranja, tj. računa udio nepodudaranja u odnosu na sve promatrane slučajeve (zanemaruju se slučaju podudaranja vrijednosti *0*). Jaccardova udaljenost definirana je sljedećim izrazom:

$$Jaccardova\ udaljenost = \frac{B + C}{A + B + C} \tag{II.9}$$

Diceova udaljenost

Diceova udaljenost ponekad se naziva Czekanowskijevom ili Sørensenovom udaljenošću. Za razliku od Jaccardove, Diceova udaljenost podudaranje vrednuje dvostruko u odnosu na nepodudaranje.

$$Diceova\ udaljenost = \frac{B + C}{2A + B + C} \tag{II.10}$$

Russel-Raova udaljenost

Russel-Raova udaljenost mjera je koja promatra udio pozitivnih podudaranja u odnosu na ukupan broj podudaranja, tj. broj elemenata vektora. Definirana je sljedećim izrazom:

$$Russell - Raova\ udaljenost = 1 - \frac{A}{n} \tag{II.11}$$

Kulzinskyjeva udaljenost

Kulzinskyjeva udaljenost promatra razliku između podudaranja vrijednosti *1* i nepodudaranja. Definirana je sljedećim izrazom:

$$Kulzinskyjeva\ udaljenost = \frac{B + C - A + n}{B + C + n} \tag{II.12}$$

II.2.3.3 Primjeri izračuna udaljenosti između binarnih vektora

Izračune različitih mjera udaljenosti između binarnih vektora prikazat ćemo računajući udaljenosti između sljedeća dva binarna vektora – *a* i *b*:

a	1	1	0	1	1	0	1	0	1	1	0	0	0	0	0	0
b	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0

Prvo ćemo prema izrazima (II-5a-f) izračunati varijable koje opisuju odnose između promatranih vektora:

$$\boldsymbol{A} = S_{11}(a,b) = ab^T = \sum_i^n a_i\, b_i = \mathbf{5}$$

$$\boldsymbol{B} = S_{01}(a,b) = \bar{a}b^T = \sum_i^n (1-a_i)\, b_i = \mathbf{2}$$

$$\boldsymbol{C} = S_{10}(a,b) = a\bar{b}^T = \sum_i^n a_i\, (1-b_i) = \mathbf{2}$$

$$\boldsymbol{D} = S_{00}(a,b) = \bar{a}\bar{b}^T = \sum_i^n (1-a_i)(1-b_i) = \mathbf{7}$$

$$\boldsymbol{\sigma} = \sqrt{(A+B)(A+C)(B+D)(C+D)} = \sqrt{7\times 7\times 9\times 9} = \mathbf{63}$$

$$\boldsymbol{n} = A+B+C+D = \mathbf{16}$$

Na temelju izračunatih ovih podataka može pristupiti računanju udaljenosti. Prvo ćemo izračunati udaljenosti simetričnim mjerama (II.6-8):

$$\textbf{\textit{Yuleova udaljenost}} = \frac{BC}{AD+BC} = \frac{4}{35+4} = \frac{4}{39} = \mathbf{0,103}$$

$$\textbf{\textit{Pearsonova udaljenost}} = \frac{1}{2} - \frac{AD-BC}{2\sigma} = \frac{1}{2} - \frac{35-4}{2\times 63} = \frac{1}{2} - \frac{31}{126} = \mathbf{0,254}$$

$$\textbf{\textit{Rogers}} - \textbf{\textit{Tanimotova udaljenost}} = \frac{2(B+C)}{A+2(B+C)+D} = \frac{2\times 4}{5+2\times 4+7} = \frac{8}{20} = \mathbf{0,4}$$

A zatim i asimetričnim mjerama (II.9-12):

$$\textbf{\textit{Jaccardova udaljenost}} = \frac{B+C}{A+B+C} = \frac{2+2}{5+2+2} = \frac{4}{9} = \mathbf{0,444}$$

$$\textbf{Diceova udaljenost} = \frac{B+C}{2A+B+C} = \frac{2+2}{10+2+2} = \frac{4}{14} = \mathbf{0,286}$$

$$\textbf{Russell} - \textbf{Raova udaljenost} = 1 - \frac{A}{n} = 1 - \frac{5}{16} = \mathbf{0,688}$$

$$\textbf{Kulzinskyjeva udaljenost} = \frac{B+C-A+n}{B+C+n} = \frac{2+2-5+16}{2+2+16} = \frac{15}{20} = \mathbf{0,75}$$

Iako postoje velike razlike u izračunatoj udaljenosti među različitim načinima računanja udaljenosti, ipak se može uočiti da u ovom primjeru mjere simetrične udaljenosti u prosjeku daju manju udaljenost u odnosu na asimetrične mjere. Razlog za to leži u činjenici da gotovo 50% slučajeva u primjeru čini podudaranje vrijednosti *0* što kod simetričnih mjera rezultira smanjenjem udaljenosti, a asimetrične mjere takve elemente ili zanemaruju (primjerice: Jaccardova ili Diceova udaljenost) ili čak na njih gledaju kao na primjere nepodudaranja (primjerice: Russel-Raova udaljenost).

II.2.4 Udaljenosti između poredanih varijabli

Poredane varijable slične su kategorijskim varijablama. Svaka varijabla može poprimiti ograničen broj vrijednosti (kategorija), ali je kategorije moguće poredati. Primjerice, pretpostavimo da je promatrana varijabla umijeće igranja odbojke i da su moguće vrijednosti: *početnik*, *napredni početnik*, *niželigaški igrač*, *prvoligaški igrač*, *reprezentativac* i *svjetski priznat igrač*. Kategorije umijeća igranja odbojke moguće je poredati gornjim redoslijedom navođenja. Primjerice, *prvoligaški igrač* bolji je od *naprednog početnika*. Međutim, razlike među kategorijama nisu ujednačene. Primjerice, nije nužno da je *napredni početnik* bolji od *početnika* onoliko koliko je *niželigaški igrač* bolji od *naprednog početnika*.

Kako bismo odredili udaljenosti među objektima opisanima poredanim varijablama, potrebno je kategorije koje poprimaju varijable pretvoriti u brojeve. Pretpostavimo da varijabla v može poprimiti K_v kategorija. Svaka kategorija u poredanom skupu svih kategorija ima svoje mjesto označeno rednim brojem R_v. Vrijednost svake varijable, tj. njezinu kategoriju možemo pretvoriti u njezinu numeričku reprezentaciju ($N(R_v)$) primjenom sljedećeg izraza:

$$N(R_v) = \frac{R_v - 1}{K_v - 1} \tag{II.13}$$

Opisani postupak prikazat će se na gore spomenutom primjeru. Pretpostavimo da umijeće igranja odbojke može poprimiti sljedeće kategorije ($K_{umijeće\ igranja\ odbojke}$): *početnik*, *napredni početnik*, *niželigaški igrač*, *prvoligaški igrač*, *reprezentativac* i *svjetski priznat igrač*.

Kategorije su poredane prema njihovu rednom broju ($R_{umijeće\ igranja\ odbojke}$) pri čemu vrijednosti *početnik* dodjeljujemo redni broj 1, a vrijednost *svjetski priznati igrač* redni broj 6. Prikazat će se računanje numeričke reprezentacije kategorija za kategorije *početnik* (R=1), *napredni početnik*, *niželigaški igrač* (R=3), *prvoligaški igrač*, *reprezentativac* i *svjetski priznat igrač* (R=6).

$$\boldsymbol{N(početnik-1)=\frac{1-1}{6-1}=0}$$

$$\boldsymbol{N(niželigaški\ igrač-3)=\frac{3-1}{6-1}=0,4}$$

$$\boldsymbol{N(svjetski\ priznati\ igrač-6)=\frac{6-1}{6-1}=1}$$

Nakon što se kategorije svih poredanih varijabli pretvore u njihove numeričke reprezentacije, udaljenosti između objekata opisanih poredanim varijablama mogu se odrediti nekom od mjera računanja udaljenosti između brojčanih varijabli. Te će mjere biti predstavljene u narednom odjeljku.

II.2.5 Udaljenosti između brojčanih varijabli

Brojčane (ili numeričke ili intervalne) varijable vrlo su slične poredanim varijablama. Razlika je utoliko što vrijednosti koje varijable poprimaju već same po sebi jesu brojčane pa im ne treba tražiti brojčanu reprezentaciju te što su intervali između različitih vrijednosti jednoliko raspoređeni. Primjerice, razlika između plaće od 5.000 i 6.000 kuna jednaka je kao i razlika između plaća od 12.000 i 13.000 kuna, a iznosi 1.000 kuna, dok, primjerice, razlika u umijeću igranje odbojke, spomenuta u odjeljku II.2.4 kao primjer poredane varijable, između *početnika* i *naprednog početnika*, nije jednaka kao i između *reprezentativca* i *svjetski priznatog igrača*, iako je u oba slučaja razlika numeričkih reprezentacija tih kategorija 1.

Objekti opisani brojčanim varijablama često se zamišljaju kao točke *n*-dimenzionalnog prostora, pri čemu je *n* broj varijabli. Temelj takvih prikaza jesu 2-dimenzionalni i 3-dimenzionalni prostori koje si jednostavno možemo vizualizirati, a onda se odnosi koji vrijede u takvim prostorima nastoje poopćiti na višedimenzionalne prostore.

Kada govorimo o prostoru, prvo što imamo na umu jest euklidski prostor – najstarija matematička reprezentacija prostora i reprezentacija prostora kakvim ga intuitivno zamišljamo. Međutim, euklidski prostor nije jedina vrsta prostora. Razvoj tzv. neeuklidske geometrije značajnije počinje u XIX. stoljeću uvođenjem eliptičke i hiperboličke geometrije. Te geometrije i pripadajući prostori zasnivaju se na promjeni pretpostavke euklidske geometrija da je, ako

zamislimo točku A i pravac p, moguće kroz točku A provući samo jedan pravac p' koji je paralelan s pravcem p. U hiperboličkom prostoru moguće je kroz točku A provući beskonačno mnogo pravaca paralelnih s pravcem p, dok u eliptičkom prostoru paralelni pravci ne postoje.

Mjere udaljenosti mogu biti euklidske i neeuklidske te metričke i nemetričke. U nastavku će biti prikazane različite vrste mjera udaljenosti za brojčane podatke. Nakon prikaza različitih mjera udaljenosti, bit će objašnjeni pojmovi standardizacije i normalizacije varijabli, koji su važni u postupku računanja udaljenosti među varijablama brojčanog tipa.

II.2.5.1 Mjere za računanje udaljenosti između objekata opisanih brojčanim varijablama

Euklidska udaljenost

Kada govorimo o udaljenostima i o prostoru, prvo što imamo na umu jest euklidski prostor i zračna (najmanja) udaljenost između promatranih točaka u prostoru. Ta najmanja udaljenost zove se euklidskom udaljenošću, a predstavlja udaljenost koju bismo izmjerili kada bismo stavili ravnalo između promatranih točaka u prostoru. Računanje euklidske udaljenosti u dvodimenzionalnom i trodimenzionalnom prostoru zasniva se na primjeni Pitagorina poučka. euklidska (ponekad zvana i Pitagorina) udaljenost točaka $A(x_a, y_a)$ i $B(x_b, y_b)$ u dvodimenzionalnom euklidskom prostoru računa se prema sljedećem izrazu:

$$euklidska\ udaljenost(A,B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{II.14}$$

Ako gornji izraz poopćimo na N-dimenzionalni prostor, euklidska udaljenost između točaka $A(x_a, y_a, \ldots, n_a)$ i $B(x_b, y_b, \ldots, n_b)$ može se izraziti sljedećim izrazom:

$$euklidska\ udaljenost(A,B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + \cdots + (n_a - n_b)^2} \tag{II.15}$$

što za točke $A(a_1, a_2, \ldots, a_N)$ i $B(b_1, b_2, \ldots, b_N)$ možemo kraće zapisati:

$$euklidska\ udaljenost(A,B) = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2} \tag{II.16}$$

Minkowskijeve ili L_p udaljenost

Euklidska udaljenost tek je poseban slučaj puno općenitije definirane Minkowskijeve ili L_p udaljenosti. Minkowskijeva udaljenost reda *p* između točaka $A(a_1, a_2, \ldots, a_N)$ i $B(b_1, b_2, \ldots, b_N)$ definira se sljedećim izrazom:

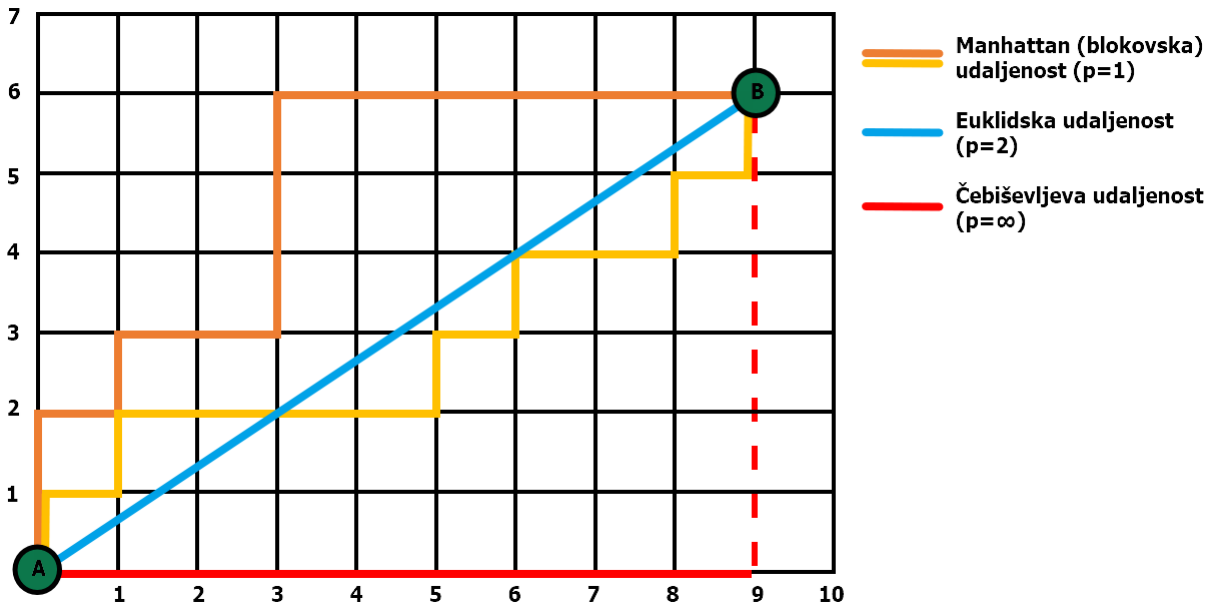
$$minkowskijeva\ udaljenost(A,B) = \left(\sum_{i=1}^{N} |a_i - b_i|^p\right)^{\frac{1}{p}} \qquad \text{(II.17)}$$

Za slučaj kada je $p=2$, dobiva se euklidska udaljenost, tj. L_2 udaljenost. Slučaj u kojem je $p=1$ poznat je i kao Manhattan, blokovska, L_1 ili *taxicab* udaljenost. Manhattan udaljenost predstavlja udaljenost koju vozilo (ili pješak) treba proći između dviju točaka krećući se *rešetkom* gradskih okomito i vodoravno postavljenih ulica. Zapravo, računa se broj blokova koje treba proći između dviju točaka.

Kada pak *p* teži u ∞, dobiva se Čebiševljeva (ili najveća) udaljenost. Čebiševljeva udaljenost u dvodimenzionalnom prostoru često se naziva i šahovskom udaljenošću jer predstavlja najmanju udaljenost koju kralj u šahu treba prijeći kako bi prošao put između dva promatrana polja, a predstavljana je razlikom koordinata centara polja. Čebiševljeva udaljenost dvaju točaka jednaka je udaljenosti po onoj osi po kojoj je udaljenost najveća. Primjerice, u dvodimenzionalnom prostoru točke *A(2,2)* i *B (4,5)* udaljene su za 2 po osi *x* i za 3 po osi *y*. Čebiševljeva udaljenost između te dvije točke iznosi 3. Čebiševljeva udaljenost ponekad se koristi u skladištima kako bi se odredilo vrijeme potrebno kranu da pomakne neki objekt. Naime, kran istodobno može istom brzinom pomicati objekt i po osi *x* i po osi *y* pa je trajanje premještanja jednako vremenu potrebnom za savladavanje osi na kojoj je potrebno prevaliti veći put.

Slika II.1 prikazuje usporedbu različitih vrsta Minkowskijevih udaljenosti u dvodimenzionalnom prostoru kroz prikaz udaljenosti između točaka *A* i *B*. Žutom i narančastom bojom prikazani su primjeri Manhattan ili blokovske udaljenosti. Blokovska udaljenost jednaka je duljini puta koji treba proći *rešetkom* okomito postavljenih gradskih ulica kako bi se došlo od polazišta do odredišta. Moguće je više takvih putova, a udaljenost je jednaka neovisno o odabiru puta. U primjeru sa slike blokovska udaljenost iznosi 15, tj. jednaka je zbroju udaljenosti po obje promatrane osi. Euklidska udaljenost predstavlja najkraći put između dvije točke, tj. njihovu *zračnu udaljenosti*. Na slici je prikazana plavnom bojom i iznosi $\sqrt{9^2 + 6^2} = 10{,}82$. Čebiševljeva udaljenost, na slici prikazana crvenom bojom, jednaka je većoj od

promatranih udaljenosti po koordinatnim osima. U primjeru su te udaljenosti 9 i 6 pa Čebiševljeva udaljenost iznosi 9.



Slika II.1 Usporedba različitih vrsta Minkowskijevih udaljenosti

Prosječna udaljenost

Prosječna udaljenost modifikacija je euklidske udaljenosti. Euklidska udaljenost može dati zavaravajuće rezultate kada se uspoređuju objekti koji imaju malo zajedničkih atributa. Kako bi se to spriječilo, osmišljena je prosječna udaljenost koja je definirana sljedećim izrazom:

$$prosječna\ udaljenost(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - b_i)^2} \qquad \text{(II.18)}$$

pri čemu je N broj zajedničkih (usporedivih) atributa, a $A(a_1, a_2, a_3, \ldots, a_N)$ i $B(b_1, b_2, b_3, \ldots, b_N)$ vektori koji sadrže zajedničke atribute promatranih objekata na odgovarajućim mjestima. Prosječnu udaljenost moguće je poopćiti i na prosječnu Minkowskijevu udaljenost.

Udaljenost tetive

Udaljenost tetive također je modifikacija euklidske udaljenosti. Osim što može biti problematična pri uspoređivanju heterogenih objekata, euklidska udaljenost daje vrlo loše rezultate ako se provodi nad nenormaliziranim ili nestandardiziranim podatcima (problem je detaljnije opisan u odjeljku II.2.5.2). Udaljenost tetive može se računati i iz nenormaliziranih podataka. Udaljenost tetive definirana je kao dužina tetive koja povezuje dvije normalizirane

točke unutar hiperkugle radijusa 1. Za točke $A(a_1, a_2, a_3, \ldots, a_N)$ i $B(b_1, b_2, b_3, \ldots, b_N)$ N-dimenzionalnog prostora može je se izračunati sljedećim izrazom:

$$udaljenost\ tetive(A,B) = \sqrt{2 - 2\frac{\sum_{i=1}^{N} a_i b_i}{\|a\|_2 \|b\|_2}} \tag{II.19}$$

pri čemu je $\|.\|_2$ L_2-norma, tj. euklidska udaljenost točke od ishodišta:

$$\|a\|_2 = \sqrt{\sum_{i=1}^{N} a_i^2} \tag{II.20}$$

Geodetska udaljenost

Geodetska udaljenost transformacija je udaljenosti tetive. Definirana je kao duljina kraćeg luka koji povezuje dvije normalizirane točke na površini jedinične hiperkugle. Za točke $A(a_1, a_2, a_3, \ldots, a_N)$ i $B(b_1, b_2, b_3, \ldots, b_N)$ N-dimenzionalnog prostora računa je se sljedećim izrazom:

$$geodetska\ udaljenost(A,B) = \arccos\left(1 - \frac{udaljenost\ tetive(A,B)}{2}\right) \tag{II.21}$$

Mahalanobisova udaljenost

Mahalanobisova udaljenost ponešto je drugačija od do sada prezentiranih mjera udaljenosti. Mahalanobisova udaljenost može neutralizirati probleme pri računanju udaljenosti uzrokovane linearnom korelacijom varijabli. Računanje Mahalanobisove udaljenosti uključuje rad s matricama. Mahalanobisova udaljenost točaka $A(a_1, a_2, a_3, \ldots, a_N)$ i $B(b_1, b_2, b_3, \ldots, b_N)$ N-dimenzionalnog prostora definirana je sljedećim izrazom:

$$mahalanobisova\ udaljenost(A,B) = \sqrt{(A-B)^T S^{-1} (A-B)} \tag{II.22}$$

S je matrica kovarijance, a S^{-1} njezin inverz. T u eksponentu znači da treba transponirati matricu.

Za razumijevanje računanja Mahalanobisove udaljenosti bitno je razumjeti što je to varijanca, kovarijanca i matrica kovarijance pa će ti pojmovi u nastavku biti objašnjeni.

Pretpostavljamo postojanje varijabli[8] A i B, pri čemu a_i i b_i predstavljaju i-te vrijednosti promatrane varijable, a $\bar{A}$ srednju vrijednost varijable A.

Varijanca opisuje promjene vrijednosti jedne promatrane varijable, tj. njezino osciliranje oko srednje vrijednosti. Varijanca se računa primjenom sljedećeg izraza:

$$var(A) = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{A})^2 \tag{II.23}$$

Kovarijanca pokazuje kako se dvije varijable mijenjaju zajedno, a definirana je sljedećim izrazom:

$$cov(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B}) \tag{II.24}$$

Matrica kovarijance ima broj redaka i broj stupaca jednak broju varijabli za koje se matrica kovarijance definira. Svaki stupac i svaki redak predstavlja jednu od varijabli, a na mjesto sjecišta određenog stupca i retka u matrici kovarijance stavlja se vrijednost kovarijance varijabli predstavljenih stupcem i retkom koji se sijeku. Ako i stupac i redak predstavljaju istu varijablu, umjesto vrijednosti kovarijance, unosi se vrijednost varijance. Primjer matrice kovarijance za 5 varijabli (A, B, C, D i X):

$$S(A,B,C,D,X) = \begin{bmatrix} var(A) & cov(A,B) & cov(A,C) & cov(A,D) & cov(A,X) \\ cov(A,B) & var(B) & cov(B,C) & cov(B,D) & cov(B,X) \\ cov(A,C) & cov(B,C) & var(C) & cov(C,D) & cov(C,X) \\ cov(A,D) & cov(B,D) & cov(C,D) & var(D) & cov(D,X) \\ cov(A,X) & cov(B,X) & cov(C,X) & cov(D,X) & var(X) \end{bmatrix} \tag{II.25}$$

Najveći nedostatak Mahalanobisove udaljenosti jest složenost njezina izračuna. Dok je složenost euklidske udaljenost *O(n)*, složenost Mahalanobisove udaljenosti jest *O(3n)*.

[8] Za razumijevanje bitno je naglasiti da varijable i točke nisu isto. Točka u sebi sadrži vrijednosti odsječaka na različitim osima, tj. vrijednosti za različite dimenzije. Varijabla je povezana s dimenzijom i sadrži vrijednosti svih točaka za promatranu dimenziju.

II.2.5.2 Standardizacija i normalizacija varijabli

Kod objekata opisanih brojčanim varijablama čest je slučaj da su različite varijable koje opisuju isti objekt mjerene različitim mjernim jedinicama, tj. poprimaju vrijednosti na različitim skalama (različitim intervalima). Primjerice, pretpostavimo da osobu muškog spola opisujemo njegovom masom u kilogramima, visinom u centimetrima i prosječnom duljinom vlasi kose u centimetrima. Možemo pretpostavite da će većina muškaraca imati masu između 70 i 100 kilograma, visinu između 170 i 190 centimetara, a duljinu vlasi kose između 1 i 5 centimetara. Svaka od tri varijable poprima vrijednosti iz različitih intervala. Ako pretpostavimo da primjenom euklidske udaljenosti (opisana u II.2.5.1) trebamo odrediti udaljenost između osoba:

A (visina: 185 cm, masa: 90 kg, prosječna duljina vlasi kose: 2 cm) i

B (visina: 165 cm, masa: 70 kg, prosječna duljina vlasi kose: 4 cm),

dobivamo sljedeće:

$$udaljenost(A,B) = \sqrt{(185-165)^2 + (90-70)^2 + (4-2)^2} = \sqrt{400+400+4} = 28{,}35$$

Može se primijetiti da prosječna duljina vlasi kose gotovo uopće ne doprinosi ukupnoj udaljenosti, tj. da ovakav način izračunavanja udaljenosti u potpunosti zanemaruje tu varijablu. Kako bi se ovom problemu doskočilo, prije računanja udaljenosti preporučljivo je sve varijable svesti na istu skalu (tj. isti interval), iako neke mjere udaljenosti mogu raditi i s vrijednostima različitih skala. Uobičajeni postupak svođenja na istu skalu jest standardizacija i normalizacija. Standardizacijom se varijable transformira u oblik u kojemu im je srednja vrijednost *0*, a varijanca *1*. Zadaća normalizacije jest sve vrijednosti varijabli transformirati na interval [0, 1].

Postupak standardizacije opisan je sljedećim izrazom:

$$standardizirana\ vrijednost = \frac{originalna\ vrijednost - srednja\ vrijednost}{standardna\ devijacija} \qquad \text{(II.26)}$$

Postupak normalizacije opisan je sljedećim izrazom:

$$normalizirana\ vrijednost = \frac{originalna\ vrijednost - min\ vrijednost\ varijable}{max\ vrijednost\ varijable - min\ vrijednost\ varijable} \qquad \text{(II.27)}$$

I postupak normalizacije i postupak standardizacije imaju svoje nedostatke. Ako u skupu postoje iskočnice (engl. *outliers*), normalizacija će rezultirati time da će najveći dio vrijednosti biti smješten u vrlo malom intervalu. S druge strane, standardizacija nema taj problem, ali ne ograničava vrijednosti na zatvoreni interval.

Učinak normalizacije i standardizacije bit će prikazan na prethodno iskazanom primjeru na način da će se vrijednosti korištenih varijabli standardizirati i normalizirati te će se zatim izračunati udaljenosti na temelju takvih vrijednosti.

Pretpostavimo sljedeće vrijednosti standardne devijacije, srednje vrijednosti te najveće i najmanje vrijednosti za promatrane varijable:

srednja vrijednost (visina) = 175 cm
standardna devijacija (visina) = 10 cm
najmanja vrijednost (visina) = 220 cm
najveća vrijednost (visina) = 130 cm

srednja vrijednost (masa) = 75 kg
standardna devijacija (masa) = 5 kg
najmanja vrijednost (masa) = 50 kg
najveća vrijednost (masa) = 220 kg

srednja vrijednost (prosječna duljina vlasi kose) = 3 cm
standardna devijacija (prosječna duljina vlasi kose) = 1 cm
najmanja vrijednost (prosječna duljina vlasi kose) = 50 cm
najveća vrijednost (prosječna duljina vlasi kose) = 0 cm

Na temelju tih pretpostavljenih vrijednosti provest će se standardizacija i normalizacija varijabli prema (II.26-27) pa će udaljenost između osoba *A* i *B* biti izračunata na temelju standardiziranih i normaliziranih vrijednosti.

Standardizirane su vrijednosti sljedeće:

$$\boldsymbol{standardizirana\ vrijednost}(\boldsymbol{A}, \boldsymbol{visina}) = \frac{185 - 175}{10} = \mathbf{1}$$

$$\boldsymbol{standardizirana\ vrijednost}(\boldsymbol{A}, \boldsymbol{masa}) = \frac{90 - 75}{5} = \mathbf{3}$$

$$\boldsymbol{standardizirana\ vrijednost}(\boldsymbol{A}, \boldsymbol{prosje\check{c}na\ duljina\ vlasi\ kose}) = \frac{2 - 3}{1} = \mathbf{-1}$$

$$\boldsymbol{standardizirana\ vrijednost(B, visina)} = \frac{165 - 175}{10} = \mathbf{-1}$$

$$\boldsymbol{standardizirana\ vrijednost(B, masa)} = \frac{70 - 75}{5} = \mathbf{-1}$$

$$\boldsymbol{standardizirana\ vrijednost(B, prosječna\ duljina\ vlasi\ kose)} = \frac{4 - 3}{1} = \mathbf{1}$$

Normalizirane su vrijednosti sljedeće:

$$\boldsymbol{normalizirana\ vrijednost(A, visina)} = \frac{185 - 130}{220 - 130} = \mathbf{0,61}$$

$$\boldsymbol{normalizirana\ vrijednost(A, masa)} = \frac{\mathbf{90 - 50}}{\mathbf{220 - 50}} = \mathbf{0,24}$$

$$\boldsymbol{normalizirana\ vrijednost(A, prosječna\ duljina\ vlasi\ kose)} = \frac{2 - 0}{50 - 0} = \mathbf{0,04}$$

$$\boldsymbol{normalizirana\ vrijednost(B, visina)} = \frac{165 - 130}{220 - 130} = \mathbf{0,39}$$

$$\boldsymbol{normalizirana\ vrijednost(B, masa)} = \frac{70 - 50}{220 - 50} = \mathbf{0,12}$$

$$\boldsymbol{normalizirana\ vrijednost(B, prosječna\ duljina\ vlasi\ kose)} = \frac{4 - 0}{50 - 0} = \mathbf{0,08}$$

Dakle, ako se primjenjuju standardizirane vrijednosti, osoba *A* i *B* opisane su sljedećim vrijednostima:

A (visina: 1, masa: 3, prosječna duljina vlasi kose: -1) i

B (visina: -1, masa: -1, prosječna duljina vlasi kose: 1).

Ako na temelju standardiziranih vrijednosti računamo euklidsku udaljenosti između osoba *A* i *B*, dobivamo sljedeće:

$$udaljenost(A, B) = \sqrt{\big(1 - (-1)\big)^2 + \big(3 - (-1)\big)^2 + (-1 - 1)^2} = \sqrt{4 + 16 + 4} = 4{,}9$$

Ako se pak primjenjuju normalizirane vrijednosti, osoba *A* i *B* opisane su sljedećim vrijednostima:

A (visina: 0,61, masa: 0,24, prosječna duljina vlasi kose: 0,04) i

B (visina: 0,39, masa: 0,12, prosječna duljina vlasi kose: 0,08).

Ako na temelju normaliziranih vrijednosti računamo euklidsku udaljenosti između osoba *A* i *B*, dobivamo sljedeće:

$$udaljenost(A,B) = \sqrt{(0{,}63-0{,}39)^2 + (0{,}24-0{,}12)^2 + (0{,}04-0{,}08)^2}$$
$$= \sqrt{0{,}0576 + 0{,}0144 + 0{,}0016} = 0{,}27$$

Iz primjera računanja udaljenosti na temelju standardiziranih i(li) normaliziranih vrijednosti vidljivo je da je značaj različitih varijabli kojima su opisani objekti uravnotežen. Osim toga, na primjeru prosječne duljine vlasi kose može se uočiti negativan utjecaj iskočnica na izračun normaliziranih vrijednosti. Naime, duljina kose uobičajeno je nekoliko centimetara, ali postoji i rijetki primjeri ćelavosti (0 cm) te jako duge kose (50 cm) koji onda smanjuju značajnost razlika u intervalu u kojem je najveći dio vrijednosti, tj. sve uobičajene vrijednosti transformiraju se u vrijednosti blizu 0.

II.2.6 Udaljenosti između mješovitih podataka

U prethodnim odjeljcima predstavljeni su načini računanja udaljenosti za različite tipove podataka. U praksi je čest slučaj da objekti nisu opisani samo varijablama jedne vrste. Postavlja se pitanje kako računati udaljenosti između objekata opisanih varijablama različitih tipova. Osnovna ideja jest izračunati udaljenost za svaku varijablu zasebno načinom odgovarajućim za taj tip varijable. Takve zasebne udaljenosti treba kasnije povezati u jedinstvenu udaljenost između dva objekta. To se može napraviti prema sljedećem izrazu:

$$d(A,B) = \frac{\sum_{i=1}^{N} d(a_i,b_i)\delta(a_i,b_i)}{\sum_{i=1}^{N} \delta(a_i,b_i)} \qquad \text{(II.28)}$$

pri čemu je $d(a_i, b_i)$ udaljenost između dva objekta na temelju *i*-te varijable. Bitno je da vrijednosti udaljenosti budu normalizirane. $\delta(a_i,b_i)$ poprima u pravilu vrijednost *1*. Vrijednost *0* poprima kada (a) jedan od promatranih objekata nije opisan *i*-tom varijablom ili (b) kada je *i*-ta varijabla asimetrična binarna varijabla, a njezina je vrijednost *0* kod oba promatrana objekta.

Na ovaj će način i ukupna udaljenost između promatranih objekata biti normalizirana, tj. njezina će vrijednost biti u intervalu [0, 1].

Pretpostavimo da želimo izračunati udaljenosti između objekata *A* i *B* koji su opisani varijablama *x*, *y*, *z*. Varijabla *x* numerička je varijabla, a udaljenost između normaliziranih

vrijednosti varijabli *x* koje opisuju objekte *A* i *B* iznosi 0,3. Varijable *y* i *z* binarne su varijable. Varijabla *y* poprima vrijednost 0 kod oba promatrana objekta, dok je vrijednost varijable *z* kod objekta *A* 1, a kod objekta *B* 0. S obzirom na navedeno, udaljenost između objekta *A* i *B* može se prema izrazu (II.28) izračunati na sljedeći način:

$$d(A,B) = \frac{\sum_{i=1}^{N} d(a_i, b_i)\delta(a_i, b_i)}{\sum_{i=1}^{N} \delta(a_i, b_i)} = \frac{0{,}3 \times 1 + 1 \times 0 + 0 \times 1}{1 + 0 + 1} = \frac{0{,}3}{2} = 0{,}15$$

II.2.7 Računanje udaljenosti uz dodavanje težinskih faktora

U odjeljku II.2.5.2 prikazano je da, kod nestandardiziranih (ili nenormaliziranih) vrijednosti brojčanih varijabli koje opisuju promatrane objekte, neke varijable nenamjerno mogu poprimiti veći značaj od nekih drugih. Tom se problemu može doskočiti standardizacijom ili normalizacijom. Međutim, ponekad pri računanju udaljenosti neke varijable zaista trebaju imati veći značaj od drugih. Tada tim varijablama treba pridijeliti odgovarajuće težinske faktore, tj. njihove vrijednosti, prije računanja udaljenosti, pomnožiti odgovarajućim brojem – njihovim težinskim faktorom.

Računanje težinske udaljenosti bit će demonstrirano na primjeru iz odvojka II.2.5.2. Tada su izvorne vrijednosti varijabli normalizirane kako bi se eliminirao veći značaj koji su pridobivale neke varijable samo zato što poprimaju po apsolutnom iznosu veće vrijednosti. Sada će se ponovo koristiti normalizirane vrijednosti, ali će varijablama biti pridijeljeni težinski faktori kako bi im se ciljano pridijelio različit značaj.

Prisjetimo se normaliziranih vrijednosti varijabli kojima su opisani objekti *A* i *B*:

A (visina: 0,61, masa: 0,24, prosječna duljina vlasi kose: 0,04 i

B (visina: 0,39, masa: 0,12, prosječna duljina vlasi kose: 0,08).

Pretpostavimo da su varijablama pridijeljeni sljedeći težinski faktori *W*:

$W_{visina} = 5$

$W_{masa} = 3$

$W_{prosječna\ duljina\ vlasi\ kose} = 1$

Na temelju tih podataka moguće je izračunati težinsku euklidsku udaljenost između objekata A i B prema sljedećem izrazu:

$$težinska\ euklidska\ udaljenost(A,B) = \sqrt{\sum_{i=1}^{N} W_i \times (a_i - b_i)^2}$$

$$= \sqrt{5 \times (0{,}63 - 0{,}39)^2 + 3 \times (0{,}24 - 0{,}12)^2 + 1 \times (0{,}04 - 0{,}08)^2} = 0{,}577$$

Odgovarajući težinski faktori u praksi se uobičajeno određuju na temelju eksperimentalno prikupljenih podataka. Eksperimentom se prikupe prediktivne i ciljne varijable, a potom se razvija model kojem je cilj na temelju prediktivnih varijabli odrediti ciljne. Ako je ciljna varijabla udaljenost, tada su parametri modela težinski faktori pridruženi pojedinoj prediktivnoj (ulaznoj) varijabli. Cilj je pronaći one težinske faktore uz koje će model na temelju prediktivnih varijabli najtočnije računati ciljne varijable. Grana strojnog učenja u kojoj se parametri modela uče na temelju poznatih ulaznih i izlaznih vrijednosti naziva se nadziranim učenjem.

Određivanje optimalnih težinskih faktora optimizacijski je problem. U svrhu rješavanja tog optimizacijskog problema kojem je cijeli skup potencijalnih rješenja često nepretraživ, uobičajeno se koriste heurističke metode.

Proces određivanja težinskih faktora primjenom principa nadziranog učenja i heurističkih optimizacijskih algoritama bit će detaljnije objašnjen u dijelu vezanom za računanje društvene udaljenosti na temelju podataka o interakciji korisnika na Facebooku u odjeljku VI.1.2.

III. FORMALNI PRIKAZ ODNOSA MEĐU LJUDIMA – OBOGAĆENI DRUŠTVENI GRAF

U ovom će poglavlju biti definirani osnovni pojmovi vezani za društvene mreže i njihov prikaz. Definirat će se model obogaćenog društvenog grafa kao formalni prikaz odnosa među ljudima, tj. kao formalni prikaz društvene mreže.

Društvena mreža jest društvena struktura koja se sastoji od ljudi povezanih različitim odnosima (prijateljstvo, zajednički interesi, obiteljski odnosi, pripadnost istoj grupi i slično). Kako bi je se moglo analizirati, društvena se mreža formalno prikazuje kao društveni graf. **Društveni** se **graf** sastoji od čvorova (tj. vrhova) i veza (tj. grana ili bridova). Čvorovi predstavljaju ljude, a veze odnose među njima.

Analiza društvenih mreža (ili analiza društvenih grafova - engl. *social network analysis* – SNA) nije zasnovana na proučavanju ljudi kao izdvojenih jedinki, već je naglasak na analizi njihovih odnosa, tj. njihovoj povezanosti. SNA može pružiti uvid u međuljudske odnose i olakšati njihovo razumijevanje. Sociolozi i društveni psiholozi koriste analizu društvenih mreža kako bi objasnili kako različiti ljudi utječu na naše osjećaje, razmišljanja i ponašanja [20], [22]. Međutim, znanstvenici koji su izučavali društvene mreže dugo su se suočavali s problemom prikupljanja reprezentativnog podatkovnog skupa. Pojavom sustava za društveno umrežavanje i digitalnim (i strukturiranim) pohranjivanjem podataka o interakciji milijardi ljudi, taj je problem značajno ublažen. Pokazalo se da interakcija ljudi na sustavima za društveno umrežavanje u velikoj mjeri opisuje njihove odnose u stvarnom životu [3], [4], [20] pa se upravo podatci o povezanosti ljudi preuzeti sa sustava za društveno umrežavanje mogu iskoristiti kao reprezentativan podatkovni skup za analizu društvenih mreža.

Podatci o povezanosti ljudi na sustavima za društveno umrežavanje u osnovi su vrlo šturi – zna se tek jesu li ljudi međusobno povezani. Poseban slučaj društvenog grafa u kojem je poznato samo jesu li neki čvorovi međusobno povezani, bez poznavanja intenziteta povezanosti, naziva se **binarnim društvenim grafom**. Primjeri binarnih društvenih grafova jesu fejsbukovski graf prijateljstva ili graf sljedbeništva na Twitteru. Takvi grafovi ne razlikuju najbolje prijatelje od poznanika ili potpunih stranaca s kojima su korisnici povezani na sustavima za društveno umrežavanje. Međutim, osim podataka o *prijateljstvu (ili sljedbeništvu)*, sa sustava za društveno umrežavanje korištenjem različitih API-ja moguće je dohvaćati podatke i o interakciji njihovih korisnika[9]. Interakcija se na sustavima za društveno

[9] Kako bi se mogli dohvaćati podatci o interakciji korisnika, korisnici prethodno moraju dati svoju suglasnost za to. Na Facebooku se to radi na način da se korisniku prikaže prozorčić u kojem piše koja aplikacije traži pristup

umrežavanje obavlja na različite načine: pisanje objava, komentiranje, označavanje oznakom *sviđa mi se*, označavanje na zajedničkim fotografijama, razmjena privatnih poruka i slično. Ti različiti načini interakcije zovu se **interakcijskim parametrima**.

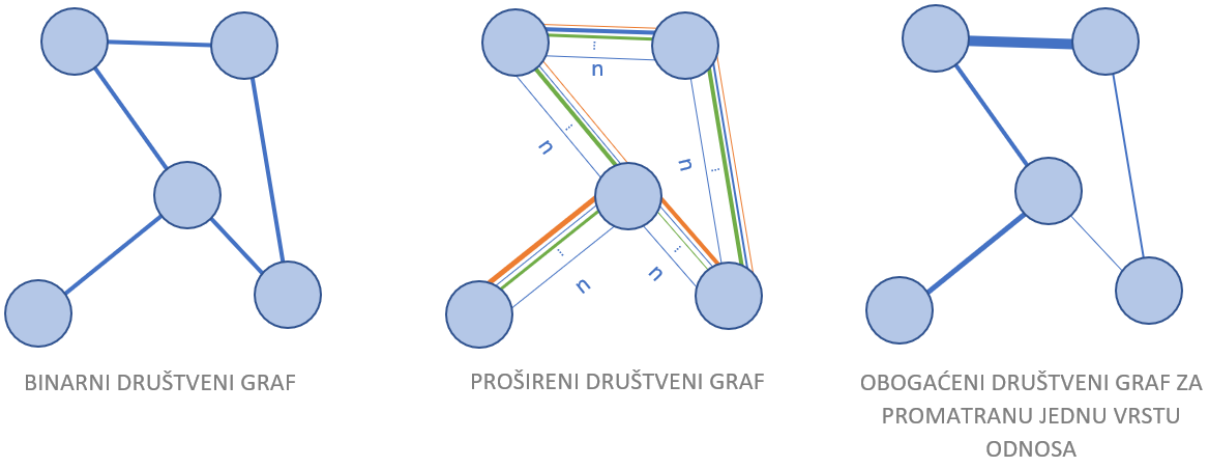
Ako binarnom društvenom grafu dodamo više veza između dvaju čvorova, od kojih svaka predstavlja (ukupnu) interakciju određenim interakcijskim parametrom, tada dobivamo **prošireni društveni graf**. Primjer takvog grafa može biti graf koji ima zasebnu vezu između promatranih čvorova za svaki parametar interakcije na Facebooku. Svaka takva veza bilježi intenzitet interakcije predmetnim interakcijskim parametrom između promatranog para korisnika. Primjerice, jedna veza predstavlja broj razmijenjenih poruka između promatranih korisnika, druga broj oznaka *sviđa mi se* na objave i slično.

Povezivanjem binarnog i proširenog društvenog grafa može se izgraditi obogaćeni društveni graf. Naime, društveni se grafovi u osnovi mogu kreirati implicitno i eksplicitno [48], [49]. Kod **eksplicitnih grafova** korisnik eksplicitno određuje s kime želi biti povezan, dok se **implicitni društveni grafovi** mogu izgraditi na temelju analize korisničkih interesa i njihove interakcije [27]. Primjer eksplicitnog društvenog grafa jest graf prijateljstva ili sljedbeništva sa sustava za društveno umrežavanje. Takvi su grafovi u svojoj osnovi binarni. Na temelju takvih binarnih grafova te analizom interakcije na sustavu za društveno umrežavanje, tj. uzimanjem u obzir proširenog društvenog grafa, moguće je inicijalno binarne veze obogatiti informacijom o intenzitetu odnosa promatranih korisnika. Takav se onda graf naziva **obogaćenim društvenim grafom**. Obogaćeni društveni graf nastaje hibridnim pristupom, tj. kombinacijom eksplicitne i implicitne izgradnje društvenog grafa.

Slika III.1 grafički prikazuje odnos binarnog, proširenog i obogaćenog društvenog grafa. Pretpostavimo da se radi o društvenim grafovima izgrađenim na temelju podataka sa sustava za društveno umrežavanje Facebook. Binarni graf izgrađuje se na temelju informacije o tome tko je kome *prijatelj*. Ako, osim informacije o tome jesu li neka dva korisnika prijatelji, želimo prikazati i intenzitet interakcije promatranih korisnika primjenom različitih interakcijskih parametara, tada ćemo to učiniti kroz prošireni društveni graf, u kojem između dva korisnika može biti više veza i svakoj od njih pridijeljena je određena težina. Svaka veza opisuje jedan interakcijski parametar i prikazuje njegovu ukupnu količinu. Interakcija korisnika na Facebooku opisuje odnos *prijateljstva*. Na temelju proširenog društvenog grafa možemo izgraditi obogaćeni društveni graf u kojem će za svaki odnos između promatranog para korisnika postojati točno jedna veza koja opisuje intenzitet toga odnosa. Obogaćeni društveni

kojim njegovim podatcima. Da bi aplikacija mogla dobiti pristup traženim podatcima, korisnik u prozorčiću mora kliknuti na gumb koji znači suglasnost. Korisniku se, naravno, nudi i mogućnost uskraćivanja suglasnosti.

graf sažima informacije dostupne u proširenom društvenom grafu, tj. skup vrijednosti koje opisuju odnos između promatranih čvorova zamjenjuje jednom vrijednošću, što omogućuje kasnije praktične primjene.



Slika III.1 Grafički prikaz odnosa binarnog, proširenog i obogaćenog društvenog grafa za promatranu jednu vrstu odnosa

Obogaćeni društveni graf formalni je način prikaza društvene mreže, tj. odnosa među ljudima. Sastoji se od čvorova, koji predstavljaju ljude, te veza, koje predstavljaju njihove odnose. Čvorovi međusobno mogu biti povezani jednom ili više veza. Svaka veza ima svoju vrstu i intenzitet. Vrsta veze govori o tome koji odnos ta veza predstavlja, a intenzitet koliko je spominjani odnos jak.

Formalno možemo reći da se obogaćeni društveni graf G sastoji od konačnog nepraznog skupa $Č$ čije elemente zovemo *čvorovima*, a koji predstavljaju ljude i konačnog skupa V čije elemente zovemo *vezama*, a koji predstavljaju odnose među ljudima. Možemo reći da je $G=(Č, V)$.

Skup $Č$ ima elemente *čvorove* pa ga možemo formalno zapisati kao $Č=\{č_1, č_2, \ldots, č_{nč}\}$, pri čemu je $n_č$ broj elemenata skupa $Č$. Skup V ima elemente *veze* pa ga možemo formalno zapisati kao $V=\{v_1, v_2, \ldots, v_{nv}\}$, pri čemu je n_v broj elemenata skupa V.

Svaki element skupa $Č$ može se prikazati kao skup svojstava koja ga opisuju: $č_i=\{s_{i1}, s_{i2}, \ldots, s_{ins}\}$, pri čemu je n_s broj svojstava koja opisuju pojedini element skupa $Č$.

Svaki element skupa V može se prikazati kao uređeni par $v_j=(č_{ishodišni}, č_{odredišni}, t, i)$. $č_{ishodišni}$ i $č_{odredišni}$ elementi su skupa $Č$, t je oznaka vrste (tipa) veze, a i je intenzitet veze. t je član nepraznog konačnog skupa T koji su sebi sadrži različite vrste veza – $T=\{t_1, t_2, \ldots t_{nt}\}$, pri čemu je n_t broj elemenata skupa T.

Veze mogu biti usmjerene ili neusmjerene. Ako su veze neusmjerene, tada su ($č_1$, $č_2$, t, i) i ($č_2$, $č_1$, t, i) posve jednake uređene četvorke. Usmjerene veze opisuju intenzitet i promatranog odnosa t između čvorova $č_{ishodišni}$ i $č_{odredišni}$ iz perspektive čvora $č_{ishodišni}$.

Intenzitet veze prikazuje se primjenom koncepta udaljenosti. S obzirom na to da se primjenom koncepta udaljenosti opisuje intenzitet odnosa među ljudima, takva će se udaljenost zvati **društvenom udaljenošću**. Društvena udaljenost detaljno će biti razmatrana u poglavlju VI. U tom će poglavlju biti uveden i model za računanje društvene udaljenosti.

Obogaćeni društveni graf prezentirat će se i primjerom. Pretpostavimo da obogaćenim društvenim grafom želimo formalno prikazati društvenu mrežu koja se sastoji od 5 ljudi: *A*, *B*, *C*, *D* i *E*. Promatrat ćemo odnose *prijateljstva*, *obiteljske povezanosti*, *sličnosti interesa* te *sličnosti u političkim stavovima*. Prikazat će se one veze koje su poznate, tj. ne će za svaki par čvorova biti prikazane baš sve vrste povezanosti uz pretpostavku da one nisu poznate.

Navedeno možemo formalno zapisati na sljedeći način:

Č={*A*, *B*, *C*, *D*, *E*}

T={*prijateljstvo*, *obiteljska povezanost*, *sličnost interesa*, *sličnost u političkim stavovima*}

Pretpostavimo da se skup veza *V* sastoji od sljedećih veza:

V={v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{11}}

koje možemo detaljnije prikazati na sljedeći način:

v_1=(A, B, obiteljska povezanost, 1)

v_2=(A, B, prijateljstvo, 1)

v_3=(A, B, sličnost interesa, 2)

v_4=(A, C, sličnost u političkim stavovima, 1)

v_5=(A, C, sličnost interesa, 1)

v_6=(B, E, prijateljstvo, 1)

v_7=(B, E, sličnost u političkim stavovima, 3)

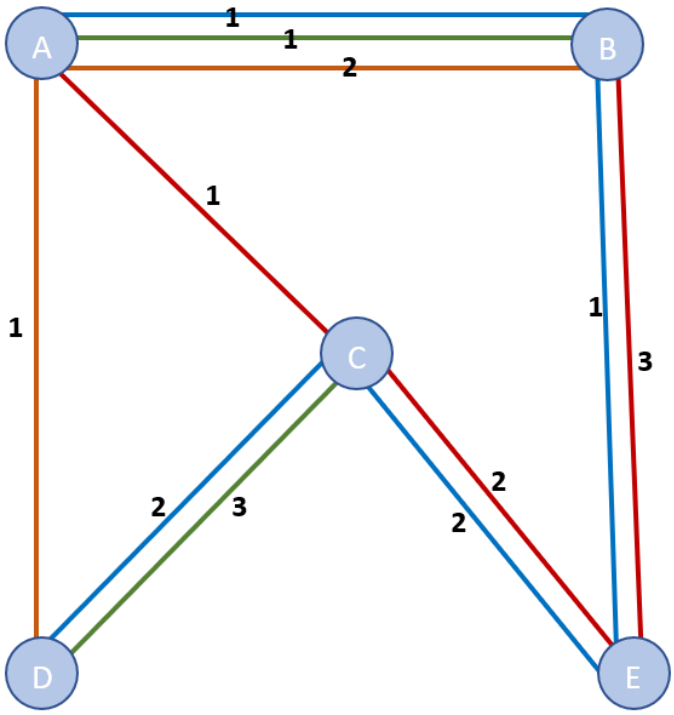
v_8=(C, D, obiteljska povezanost, 2)

v_9=(C, D, prijateljstvo, 3)

v_{10}=(C, E, prijateljstvo, 2)

v_{11}=(C, E, sličnost u političkim stavovima, 2)

Ovako formalno opisani društveni graf vizualizira Slika III.2. Vrste veza na slici prikazane su bojama kojima su određene vrste veza obojane i u formalnom zapisu.



Slika III.2 Vizualizacija primjera obogaćenog društvenog grafa

IV. PROVEDENA DRUŠTVENA ISTRAŽIVANJA

Vrlo važan dio ovog doktorskog rada čine provedena društvena istraživanja. Ovo poglavlje posvećeno je njima. Provedena društvena istraživanja bit će kronološki prezentirana. Svako provedeno istraživanje bit će detaljno opisano. Bit će naglašeno što se njime htjelo postići i koji su rezultati dobiveni. Istraživanjima dobiveni referentni podatkovni skupovi u narednim će poglavljima biti korišteni za definiranje modela za računanje društvene udaljenosti te za verifikaciju modela obogaćenog društvenog grafa.

IV.1 ISTRAŽIVANJE *BESTFRIENDS*

Istraživanja sustava za društveno umrežavanje s ciljem oblikovanja modela za računanje intenziteta odnosa među njihovim korisnicima, u sklopu izrade ovog doktorskog rada, započeta su 2013. godine. S obzirom na to da je u to vrijeme, kao i danas, Facebook bio uvjerljivo najraširenij sustav za društveno umrežavanje [1], [50], kako u Hrvatskoj tako i u svijetu, odlučeno je usredotočiti se upravo na taj sustav. Prvo provedeno istraživanje nazvano je *BestFriends*[10]. Polazišna točka toga istraživanja bili su prethodno objavljeni znanstveni i popularni radovi koji su bilo objašnjavali kako Facebook primjenom algoritma EdgeRank rangira objave korisnika [51] bilo provodili eksperimente i predlagali različite modele izračuna intenziteta odnosa među ljudima [15], [52]. Želju za određivanjem intenziteta prijateljstva na temelju interakcije na društvenim mrežama dodatno je potaknula rečenica iz rada talijanskog računalnog znanstvenika, ali i psihologa Andrea Biancinija [20]: „Facebook je vrlo moćan alat s mogućnošću emuliranja društvenog života svojih korisnika“. Dakle, na temelju interakcije ljudi na Facebooku trebalo bi biti moguće odrediti intenzitet njihova odnosa u stvarnom životu. Kako bi se ta tvrdnja provjerila, osmišljeno je društveno istraživanje na Facebooku u kojem se od korisnika tražilo da navedu svojih do 10 najboljih prijatelja u stvarnom životu. U svrhu provođenja istraživanja razvijena je *web*-aplikaciju *BestFriends* povezana s istoimenom fejsbukovskom aplikacijom. Ta je aplikacija, s jedne strane, korištenjem Facebookova API-ja dohvaćala sumarne podatke o interakciji ispitanika s njihovim fejsbukovskim prijateljima, a s druge strane kroz jednostavan upitnik (Slika IV.2) od korisnika tražila da unesu rangirani popis svojih do 10 najboljih prijatelja u stvarnom životu – neovisno o tome koriste li oni Facebook.

Interakcija korisnika na Facebooku obavlja se kroz različite načine interakcije koji se nazivaju interakcijskim parametrima, a to su, primjerice: komentari, oznake *sviđa mi se*,

[10] Istraživanje *BestFriend* proveo je autor ovoga rada u suradnji sa studentima koje je vodio na predmetu Diplomski projekt – Majom Majić (udana Skorin) i Juricom Skorinom

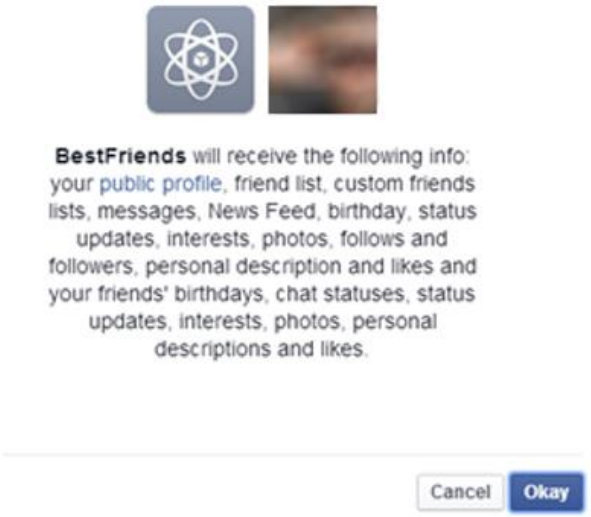
zajedničko pojavljivanje na fotografijama, razmijenjene poruke i slično. Popis parametara interakcije koji su dohvaćani u ovom istraživanju prikazuje Tablica IV.1.

Kako bi bilo moguće prikupiti podatke o interakciji korisnika na Facebooku, korisnici su aplikaciji prethodno trebali dati dopuštenje za pristup tim podatcima. Takav su pristup dali odabirom opcije slaganja u prozorčiću, generiranom od strane Facebooka, u kojem je bilo popisano kojim će sve njihovim podatcima aplikacija dobiti pristup (Slika IV.1). U istraživanju je sudjelovalo 200 ispitanika. Svi ispitanici govornici su hrvatskog jezika. Geografski je istraživanje smješteno u grad Zagreb i okolicu te u manjoj mjeri u ostatak Hrvatske i susjedne države.

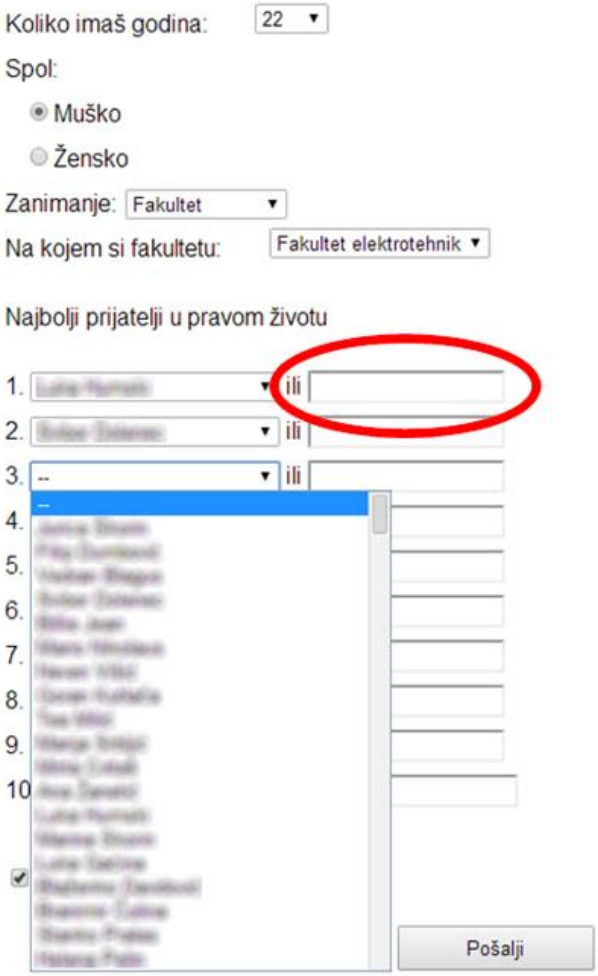
Prijatelje korisnike Facebooka trebalo je odabrati iz padajućeg izbornika, a imena prijatelja koji nisu korisnici Facebooka upisati u obližnje tekstualno polje. Kroz unos i onih prijatelja koji ne koriste Facebook, htjelo se utvrditi koliki je njihov udio, tj. je li on dovoljno mali da bi se analizirajući samo interakciju korisnika na Facebooku moglo donositi zaključke o njihovim odnosima u stvarnom životu i koliko bi mogla biti prosječna uspješnost pri tome. U nekim ranijim istraživanjima [15] korisnike se pitalo da odaberu korisnike s kojima najviše komuniciraju na društvenim mrežama. Međutim, smatrali smo to pogrešnim pristupom. Naime, analizom interakcije korisnika na Facebooku možemo objektivno pronaći one ljude s kojima naši ispitanici na Facebooku najviše komuniciraju, ali to nije poanta istraživanja. Poanta je utvrditi ima li povezanosti između intenziteta komunikacije na društvenoj mreži i odnosa prijateljstva u stvarnom životu te pronaći način kako povezati ta dva *svijeta*, tj. kako na temelju poznatog intenziteta komunikacije (ili općenitije – interakcije) na Facebooku čim preciznije odrediti intenzitet odnosa prijateljstva u stvarnom životu. Pri tome smo bili potpuno svjesni činjenice da je, s obzirom na to da nemaju svi prijatelji naših ispitanika profil na Facebooku, najveća teorijska uspješnost našeg modela manja od 100%. Primjerice, ako od 10 najboljih prijatelja naših ispitanika prosječno 1 ne koristi Facebook, uspješnost našeg modela u otkrivanju najboljih stvarnoživotnih prijatelja na temelju interakcije na Facebooku može najviše biti $\frac{10-1}{10} \times 100\% = 90\%$.

U sklopu ovog prvog istraživanja, na temelju prikupljenih podataka, osmišljen je model za određivanje intenziteta odnosa prijateljstva na temelju analize interakcije na Facebooku. Model je uzimao u obzir činjenicu da prijateljstvo nije nužno refleksivna veza, tj. da je moguće da *korisnik A* smatra *korisnika B* boljim prijateljem no što *korisnik B* smatra *korisnika A*, pa je intenzitetu povezanosti pridjeljivao i svojstvo usmjerenosti. U model je unesena pretpostavka da svaki parametar interakcije ima pridruženi koeficijent općenite značajnosti (W_g) i koeficijent

značajnosti specifične za pojedinog korisnika (W_s). Koeficijent općenite značajnosti opisuje univerzalnu značajnost pojedinog parametra interakcije za sve korisnike, dok specifični koeficijent nastoji opisati koliko je pojedini parametar značajan za određenog korisnika.



Slika IV.1 Prozorčić kroz koji korisnici dodjeljuju aplikaciji pravo pristupa njihovim podatcima s Facebooka



Slika IV.2 Upitnik u istraživanju BestFriends

Model za računanje intenziteta odnosa prijateljstva od *korisnika A* prema *korisniku B*, odnosno koliko *korisnik A* smatra *korisnika B* dobrim prijateljem, definirali smo na sljedeći način:

$$intenzitet\ prijateljstva(A,B) = \sum_{i=0}^{n} W_{si}(A) * W_{gi} * N_i(A,B)$$

pri čemu je $W_{si}(A)$ koeficijent specifične značajnosti parametra interakcije *i* za *korisnika A*, W_{gi} jest koeficijent globalne značajnosti parametra interakcije *i*, a $N_i(A,B)$ jest ukupna količina interakcije između korisnika *A* i *B* po parametru interakcije *i*, iz perspektive korisnika A.

Koeficijent specifične značajnosti parametra interakcije *i* za *ego-korisnika A* – $W_{si}(A)$ obrnuto je proporcionalan udjelu tog parametra komunikacije kod *ego-korisnika A* u odnosu ukupnu interakciju *ego-korisnika A*. Primjerice, što korisnik češće (u odnosu na ostale oblike interakcije) komentira objave, svaki njegov komentar zasebno manje vrijedi. Koeficijent specifične značajnosti računa se prema sljedećem izrazu:

$$W_{si}(A) = 1 - \frac{N_i(A)}{N(A)}$$

pri čemu je $N_i(A)$ količina parametra interakcije *i* između ego-korisnika *A* i svih njegovih prijatelja (primjerice, ukupan broj poruka *ego-korisnika A*), dok je $N(A)$ ukupna interakcija ego-korisnika, tj. zbroj svih parametara interakcije *ego-korisnika A* prema svim njegovim prijateljima (ukupan broj poruka, komentara, oznaka *sviđa mi se* i slično).

Koeficijenti općenite značajnosti određeni su eksperimentalno kroz proces treniranja (kalibriranja) modela. Kako bismo mogli objasniti kako je model treniran, prvo ćemo objasniti kako ćemo računati uspješnost rada modela.

Zadaća modela jest odrediti numeričku reprezentaciju intenziteta odnosa prijateljstva između svih ispitanika i svih njihovih fejsbukovskih prijatelja. Model je tim uspješniji što su vrijednosti koje izračunava bliže stvarnim vrijednostima. Međutim, prijateljstvo je suviše apstraktna kategorija da bismo ispitanike mogli tražiti da numerički opišu intenzitet svoga prijateljstva pa onda nemamo niti referentne numeričke vrijednosti intenziteta prijateljstva koje bismo koristili pri određivanju uspješnosti modela. S obzirom na to, odlučili smo uspješnost rada modela mjeriti posredno. Zadaća ispitanika bila je da unesu svojih do 10 najboljih prijatelja. Isti smo zadatak dali i modelu – treba izračunati intenzitet prijateljstva između

ispitanika i svih njegovih prijatelja te vratiti 10 prijatelja s izračunatim najvećim intenzitetom prijateljstva. Uspješnost rada modela za promatranog ispitanika jednaka je postotku u kojem se dva spomenuta skupa prijatelja preklapaju (točan redoslijed prijatelja unutar ovih skupova ne će se uzimati u obzir). Ukupna uspješnost modela jednaka je prosječnoj uspješnosti rada modela nad svim ispitanicima u skupu.

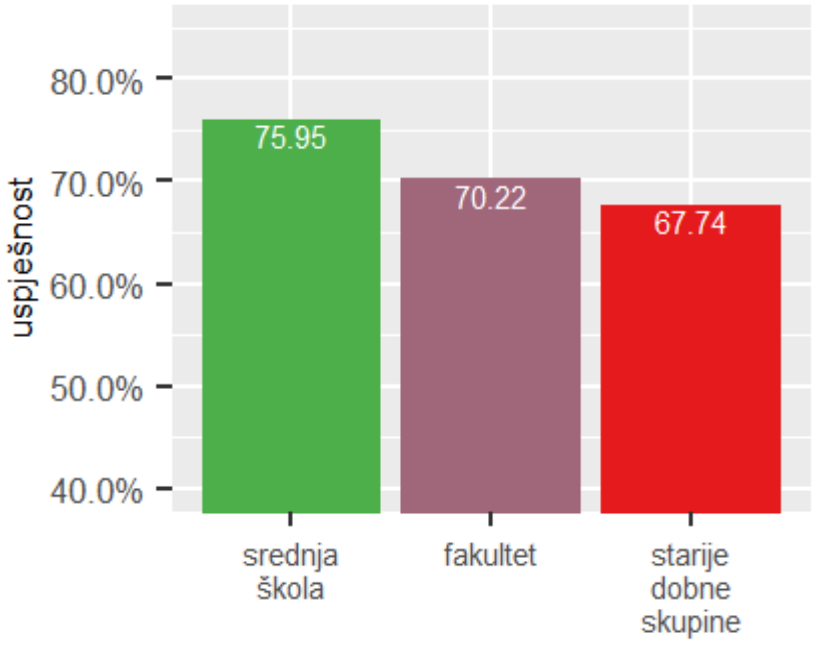
Istraživanjem prikupljen skup podataka podijeljen je na dva dijela – podskup za treniranje modela i podskup za testiranje. Podskup za treniranje modela iskorišten je za traženje najboljih mogućih vrijednosti koeficijenata općenite značajnosti interakcijskih parametara. Cilj je pronaći takve vrijednosti koeficijenata općenite značajnosti koje će dati najveću uspješnost rada modela. Optimalne vrijednosti (ako postoje) pronašli bismo kada bismo pretražili cijeli prostor mogućih rješenja, međutim taj pristup ima u sebi dva nerješiva problema. Prvi je problem to što prostor rješenja nije konačan (naime, ne postoji najveća moguća vrijednost koeficijenata općenite značajnosti parametara komunikacije), a drugi je taj što je postupak ispitivanja svih mogućih kombinacija koeficijenata općenite značajnosti parametara komunikacije – čak i kada bismo se ograničili na neki konačan prostor rješenja – vrlo računski složen. Odlučili smo se za pristup koji koristi znanje analitičara za postavljanje početnih vrijednosti koeficijenata te smo primjenom metode Monte Carlo došli do vrijednosti koeficijenata općenite značajnosti. Metodu Monte Carlo ostvarili smo tako da smo po vlastitom nahođenju postavili početne vrijednosti koeficijenata nakon čega smo programski provjeravali uspješnost rada modela za nasumično odabrane različite kombinacije koeficijenata različitih parametara interakcije. Prostor mogućih rješenja ograničili smo na vrijednosti ±10 u odnosu na početne vrijednosti s granulacijom od 0,1. Dobivene koeficijente skalirali smo tako da u zbroju daju 1. Dobivene vrijednosti koeficijenata općenite značajnosti različitih parametara interakcije prikazuje Tablica IV.1.

Tablica IV.1 Eksperimentalno dobiveni koeficijenti općenite značajnosti parametara interakcije za sve promatrane parametre interakcije

Parametar interakcije	**Koeficijent W_g**
zajednička fotografija korisnika A i B koju je objavio korisnik A	0.539
komentar korisnika B na **fotografiju** koju je objavio korisnik A	0.183
objava korisnika B na zid korisnika A	0.086
komentar korisnika B na **objavu** koju je korisnik A objavio na svom zidu	0.054
privatne poruke između korisnika A i B	0.042
zajednička fotografija korisnika A i B koju je objavio korisnik B	0.036
oznaka *sviđa mi se* korisnika B na **objavu** koju je korisnik A objavio na svom zidu	0.027
oznaka *sviđa mi se* korisnika B na **fotografiju** koju je objavio korisnik A	0.027
zajednički prijatelji korisnika A i B	0.006

Uspješnost rada modela ispitana je korištenjem eksperimentalno dobivenih koeficijenata općenite značajnosti nad podskupom za testiranje i iznosi 70,9%, tj. prosječno model može *pogoditi* ~7 od 10 najboljih prijatelja ispitanika. Najbolji ispitanikov prijatelj s liste prijatelja koju je ispitanik sam unio na prvom je mjestu i na listi generiranoj na temelju izračuna modela u 43% slučajeva. U 80% slučajeva preklapaju se ili prvi ili drugi prijatelj s ispitanikove liste najboljih prijatelja i liste dobivene modelom.

Uspješnost rada modela analizirana je i za različite dobne skupine. Pretpostavka istraživanja bila je da mlađi korisnici sa svojim najboljim prijateljima iz stvarnog života više komuniciraju preko Facebooka nego stariji korisnici pa je za mlađe korisnike lakše predvidjeti najbolje prijatelje iz stvarnog života. Pretpostavka je da će s porastom godina korisnika padati uspješnost rada modela. S obzirom na dob, korisnike smo podijeliti na one koji pohađaju srednju školu, korisnike koji su na fakultetu i one koji su zaposleni, tj. stariji. Provedeno istraživanje potvrdilo je naša očekivanja. Slika IV.3 prikazuje rezultate uspješnosti rada modela za različite dobne skupine.



Slika IV.3 Uspješnost rada modela za različite dobne skupine

Do sada opisani rezultati objavljeni su u radu [4].

Istraživanja su nastavljena 2014. godine kada je odlučeno provesti još jednom isto istraživanje, ali uz povećanje broja parametara interakcije čije su sumarne vrijednosti dohvaćane[11]. Odlučeno je analizirati i prediktivnu snagu svakog interakcijskog parametra

[11] Istraživanja je provodio autor ovog rada u suradnji sa studentom Juricom Skorinom kojem je to bio diplomski zadatak

zasebno, tj. odrediti koliko bi bili uspješni modeli koji bi radili samo na temelju poznavanja jednog interakcijskog parametara. Također, unaprijeđen je proces treniranja modela te povećan broj različitih kombinacija parametara općenite značajnosti modela za koje se ispitivala uspješnost rada modela s ciljem pronalaženja one najbolje. U novom istraživanju sudjelovala su 282 ispitanika najvećim dijelom iz Zagreba.

Interakcijskim parametrima koji su u prethodnom istraživanju dohvaćani, dodano je sljedećih 6 interakcijskih parametara:

1. korisnik A označio je korisnika B kao **bliskog prijatelja**
2. **oznaka *sviđa mi se*** korisnika A na **poveznicu** koju je objavio korisnik B
3. **oznaka *sviđa mi se*** korisnika A na **status** koju je objavio korisnik B
4. **oznaka *sviđa mi se*** korisnika A na **fotografiju** koju je objavio korisnik B
5. **zajednička fotografija** korisnika A i B koju je objavio netko treći
6. korisnici A i B **zajedno označeni u objavi**

Uspješnosti rada modela koji rade samo na temelju poznavanja jednog parametra interakcije, tj. prediktivnu snagu tih interakcijskih parametara prikazuje Tablica IV.2 u drugom stupcu. Pokazuje se da privatne poruke imaju najveću prediktivnu snagu od 55,83%, dok najmanju prediktivnu snagu imaju oznake *sviđa mi se* ispitanika na statuse prijatelja. Odlučili smo provjeriti i kolika bi bila uspješnost modela kada bismo uzeli u obzir sve parametre interakcije, ali im dodijelili jednake vrijednosti koeficijenata općenite značajnosti. Uspješnost rada takvog modela iznosi 56,94%. Činjenica da je taj broj veći od prediktivne snage bilo kojeg parametra interakcije zasebno, pokazuje da različiti parametri interakcije zajedno postižu sinergijski učinak čak i kada su im pridijeljene jednake vrijednosti koeficijenata općenite značajnosti.

Za određivanje koeficijenata općenite značajnosti pojedinih parametara interakcije, podatkovni skup prikupljen provedenim društvenim istraživanjem podijeljen je na dva dijela – podskup za treniranje i podskup za testiranje modela. Koeficijenti općenite značajnosti parametara interakcije i ovaj su puta određivani primjenom metode Monte Carlo na podskupu za treniranje modela. Cilj treniranja modela bio je, kao i prethodni put, pronaći one kombinacije koeficijenata općenite značajnosti koje daju najveće preklapanje liste najboljih prijatelja dobivene modelom i liste koju je ispitanik unio kroz obrazac ankete.

Metoda Monte Carlo ovaj je puta provedena u dvije faze kako bi se uklonila mogućnost pristranosti istraživača koji su u prvom istraživanju postavili inicijalne vrijednosti koeficijenata općenite značajnosti. Ovoga su puta „inicijalne“ vrijednosti pojedinih koeficijenata također

određene eksperimentalno. Naime, metodom Monte Carlo u prvoj su fazi nasumično pogađani koeficijente općenite značajnosti parametara interakcije na skupu vrijednosti [1,60] s granulacijom 1. Parametar interakcije *oznaka prijatelja kao bliski prijatelj* odskače od ovog pravila i njegova je vrijednost u prvoj fazi postavljena na 4.000, jer u sebi krije izravan odgovor dobiven od ispitanika da se radi o njegovu vrlo bliskom prijatelju.

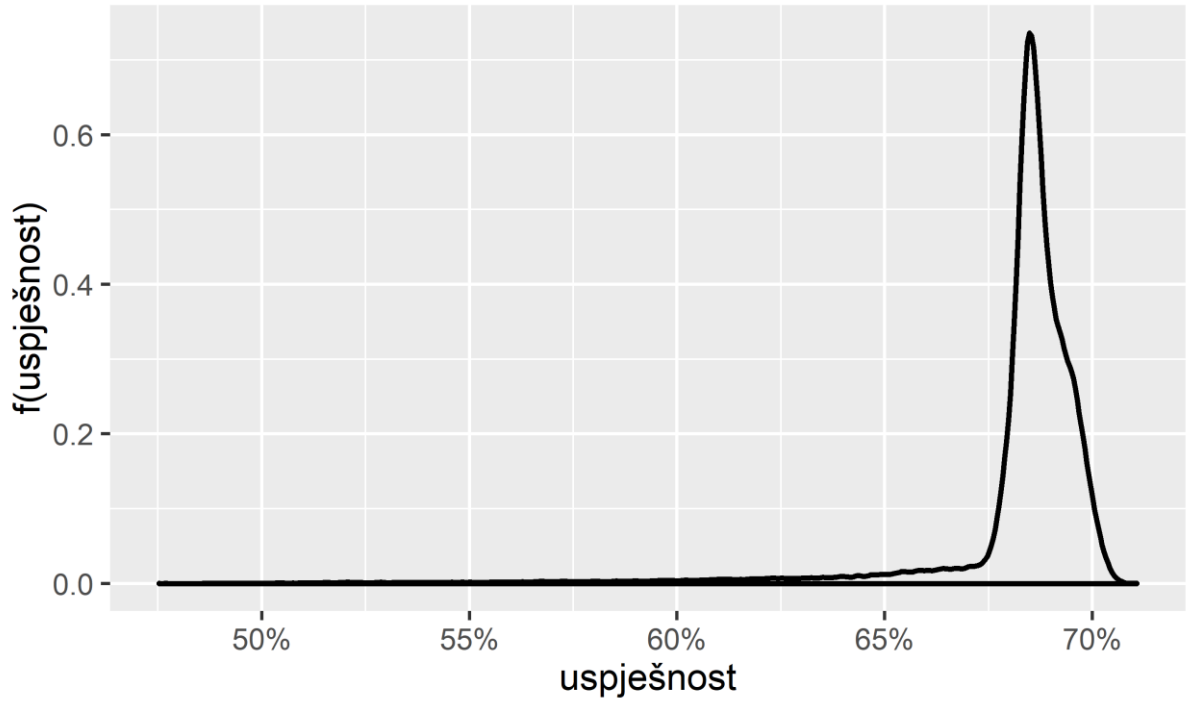
U prvoj fazi isprobano je 181.279 kombinacija, od kojih je odabrano 10 najuspješnijih „inicijalnih" kombinacija koeficijenata. Za svaku od tih 10 kombinacija, u drugoj fazi (fazi finog ugađanja koeficijenata) testirano je još oko 15.000 kombinacija tako da se za svaki koeficijent pogađalo iz skupa ±10% od vrijednosti dobivene u prvoj fazi.

Iz skupa od 332.919 kombinacija odabrana je ona kombinacija koeficijenata globalne značajnosti parametara interakcije uz koju je model postizao najbolje rezultate na skupu za treniranje. Te vrijednosti prikazuje Tablica IV.2 u trećem stupcu.

Tablica IV.2 Popis promatranih parametara interakcije, njihove pojedinačne prediktivne snage i koeficijenata njihove općenite značajnosti

Parametar interakcije	**Prediktivna snaga**	W_{gi}
korisnik A označio je korisnika B kao **bliskog prijatelja**	19,24%	3735
oznaka *sviđa mi se* korisnika A na **poveznicu** koju je objavio korisnik B	22,24%	47
oznaka *sviđa mi se* korisnika A na **status** koju je objavio korisnik B	17,05%	9
oznaka *sviđa mi se* korisnika A na **fotografiju** koju je objavio korisnik B	22,65%	15
privatne poruke između korisnika A i B	55,83%	11
zajednički prijatelji korisnika A i B	21,33%	13
komentar korisnika B na **fotografiju** koju je objavio korisnik A	25,02%	8
oznaka *sviđa mi se* korisnika B na **fotografiju** koju je objavio korisnik A	25,19%	1
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio netko treći	30,52%	19
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio korisnik B	19,26%	54
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio korisnik A	26,24%	47
korisnici A i B **zajedno označeni u objavi**	19,95%	43
objava korisnika B na zid korisnika A	24,56%	3
komentar korisnika B na **objavu** koju je korisnik A objavio na svom zidu	32,71%	37
oznaka *sviđa mi se* korisnika B na **objavu** koju je korisnik A objavio na svom zidu	33,45%	16

Slika IV.4 prikazuje funkciju gustoće razdiobe uspješnosti različitih kombinacija koeficijenata općenite značajnosti pojedinih parametara interakcije za isprobanih 181.279 kombinacija koeficijenata iz prvog koraka na podskupu za treniranje modela.



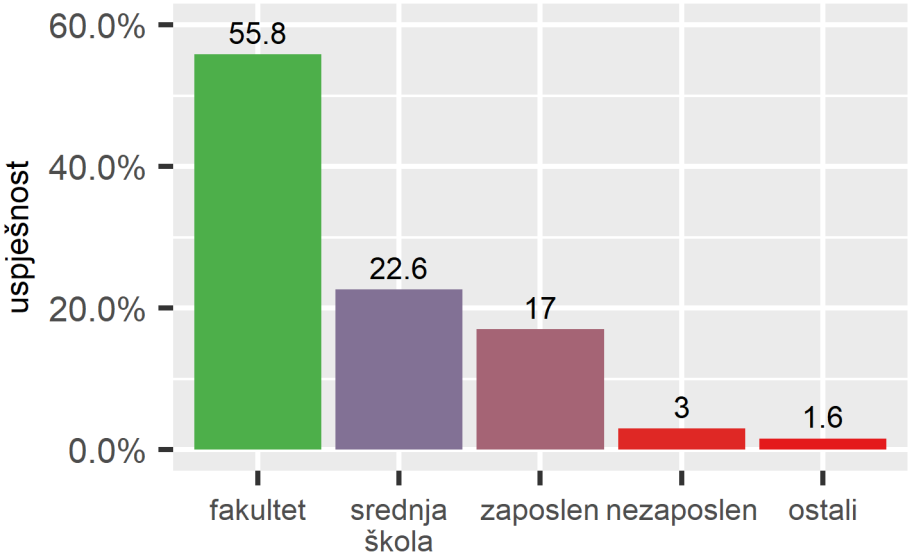
Slika IV.4 Prikaz uspješnosti različitih kombinacija koeficijenata općenite značajnosti parametara interakcije – funkcija gustoće razdiobe

Usporedbom odnosa prediktivnih snaga pojedinih parametara interakcije i njihovih koeficijenata općenite značajnosti može se uočiti da to dvoje nije korelirano u mjeri u kojoj bi se možda očekivalo. Razlog tome jest taj da prediktivna snaga govori koliko je neki parametar interakcije značajan sam za sebe, a koeficijent općenite značajnosti govori koliko vrijedi jedna pojavnost pojedinog parametra interakcije. Tako, primjerice, parametar interakcije *oznaka prijatelja kao „bliski prijatelj“* ima općeniti koeficijent 3735, a razmijenjena privatna poruka 11, iako je prediktivna snaga drugog višestruko veća od prvog. Međutim, korisnici samo jednom nekog svog prijatelja mogu označiti za *bliskog prijatelja*, a poruke razmjenjuju u tisućama ili desecima tisuća. Tada umnožak količine parametra interakcije i njegova koeficijenta općenite značajnosti ponovo značajno veću važnost daje razmijenjenim privatnim porukama (primjerice, $11 \times 5.000 > 3.735 \times 1$).

Model s koeficijentima općenite značajnosti parametara interakcije određenim kroz prethodno opisane dvije faze na podskupu za testiranje postiže uspješnost od 71,26%. Taj je rezultat značajno bolji o rezultata modela koji je koeficijente općenite značajnosti svih interakcijskih parametara postavio na istu vrijednosti (56,94%), a bolji je i od uspješnosti dobivene modelom iz prvog istraživanja koji je polučio uspješnost od 70,9%.

IV.2 ISTRAŽIVANJE *NAJFREND*

Istraživanje *NajFrend*[12] središnje je i najznačajnije istraživanje provedeno u sklopu izrade ovog doktorskog rada. Istraživanje je provedeno u travnju i prvim danima svibnja 2015. godine. Prikupljen je podatkovni skup od 3.277 ispitanika i više od 1.400.000 njihovih fejsbukovskih prijatelja, od čega je 2.626 ispitanika uspješno odgovorilo na sva pitanja iz upitnika. U istraživanju je sudjelovalo najviše studenata i srednjoškolaca (Slika IV.5), što je ponajprije posljedica organiziranog poticanja studenata i srednjoškolaca na sudjelovanje u istraživanju. Među studentima najzastupljeniji su oni sa zagrebačkog FER-a s 48,43%, ali u istraživanju je sudjelovalo i 79 studenata Filozofskog fakulteta u Zagrebu, 78 studenata s osječkog FERIT-a (tadašnjeg ETF-a), 43 studenta ETF-a iz Beograda te studenti mnogih drugih fakulteta iz Hrvatske i susjedstva. Kada se pogleda dobna struktura ispitanika, najviše ispitanika je u dobi od 18 do 30 godina. Žene čine 42,3% ispitanika, a muškarci 57,7%. Većina od 94,9% sudionika istraživanja stanovnici su Hrvatske, dok je 5,1% sudionika iz drugih država. Pitanja u upitniku postavljena su na hrvatskom jeziku pa je za ispunjavanje upitnika bilo nužno njegovo razumijevanje.



Slika IV.5 Raspodjela ispitanika po zanimanju

U usporedbi s podatkovnim skupovima prikupljenim u nekim drugim srodnim istraživanjima, podatkovni skup prikupljen u istraživanju *NajFrend* po broju ispitanika za red je ili gotovo dva reda veličine veći. Primjerice, u [12] podatkovni skup uključuje 59 ispitanika,

[12] Istraživanje je, uz autora ovog rada, u sklopu svog završnog rada pod vodstvom autora ovog rada, provodio i Juraj Ilić. U kasnoj fazi istraživanja, priključili su se i Vanja Smailović i Darko Štriga, koji su istraživanju doprinijeli kroz programsku izvedbu sustava za dohvaćanje detaljnih (*sirovih*) podataka o interakciji korisnika Facebooka.

u [15] 150 ispitanika, u [18] 35 ispitanika, a u [19] 177 ispitanika. U istraživanju *NajFrend* podatci o interakciji korisnika Facebooka dohvaćani su Facebookovim *Graph API-jem 1.0*. *Graph API 1.0* s početkom svibnja 2015. povučen iz uporabe, a novije inačice API-ja značajno su restriktivnije u razini pristupa podatcima. S obzirom na to i s obzirom na literaturu koju smo pregledali, prikupljeni podatkovni skup u sklopu ovog istraživanja najveći je i najnoviji podatkovni skup koji sadrži podatke o interakciji korisnika Facebooka.

Istraživanje *NajFrend* logičan je nastavak prethodno provedenog istraživanja *BestFriends*. Istraživanjem *NajFrend* željeli smo ostvariti sljedeće ciljeve:

1. provjeriti zadovoljstvo korisnika radom modela za određivanje intenziteta prijateljstva razvijenog u sklopu istraživanja *BestFriends*
2. prikupiti od ispitanika informacije o njihovu poimanju prijateljstva na cijelom intervalu prijateljstva, a koji samo na malom dijelu uključuje samo najbolje prijatelje – što je bio slučaj u istraživanju *BestFriends*
3. provjeriti može li se prethodno razvijeni model za određivanje intenziteta prijateljstva primijeniti za pronalazak po stavovima sličnih ljudi ispitanicima te za izračun povjerenja između ispitanika i njegovih prijatelja
4. prikupiti reprezentativan podatkovni skup za buduća istraživanja, a prije svega za razvoj budućih modela kojima je cilj odrediti intenzitet povezanosti ljudi na temelju analize interakcije na Facebooku. Prikupljeni podatkovni skup s jedne strane treba sadržavati podatke o interakciji svih ispitanika sa svim njihovim fejsbukovskim prijateljima, a s druge strane od ispitanika izlučiti njihovo (subjektivno) poimanje prijateljstva sa svojim fejsbukovskim prijateljima.

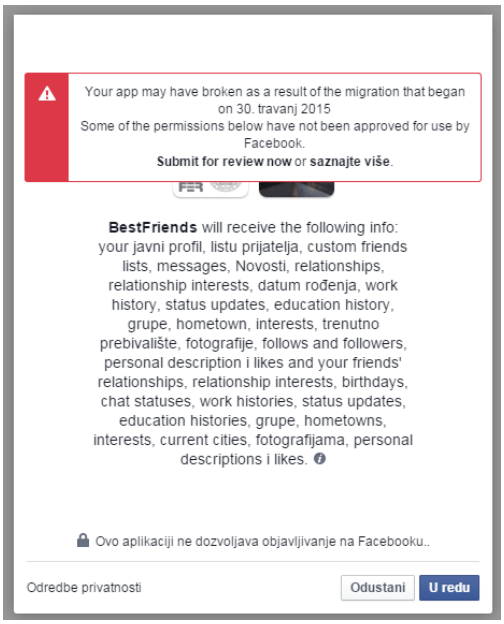
U nastavku će detaljno biti opisan anketni upitnik koji je svaki ispitanik istraživanja *NajFrend* trebao popuniti te rezultati istraživanja. Ispitanici su upitnik ispunjavali kroz posebnu, u svrhu provođenja istraživanja izrađenu, *web*-aplikaciju.

IV.2.2 Prikaz upitnika i aplikacije za provođenje istraživanja

Slika IV.6 prikazuje početni zaslon *web*-aplikacije *NajFrend* izrađene u svrhu provođenja istraživanja.

Prije otvaranja anketnog upitnika, svaki ispitanik trebao je pridruženoj fejsbukovskoj aplikaciji dozvoliti pristup podatcima o interakciji njega i njegovih fejsbukovskih prijatelja (Slika IV.7). Prikupljeni sumarni podatci o interakciji ispitanika s njegovim prijateljima na Facebooku (isti kao u drugom dijelu istraživanja *BestFriends* – Tablica IV.2) pohranjuju se u bazi podataka i koriste pri generiranju personaliziranog anketnog upitnika za svakog ispitanika.

Slika IV.6 Početni zaslon web-*aplikacije* NajFrend

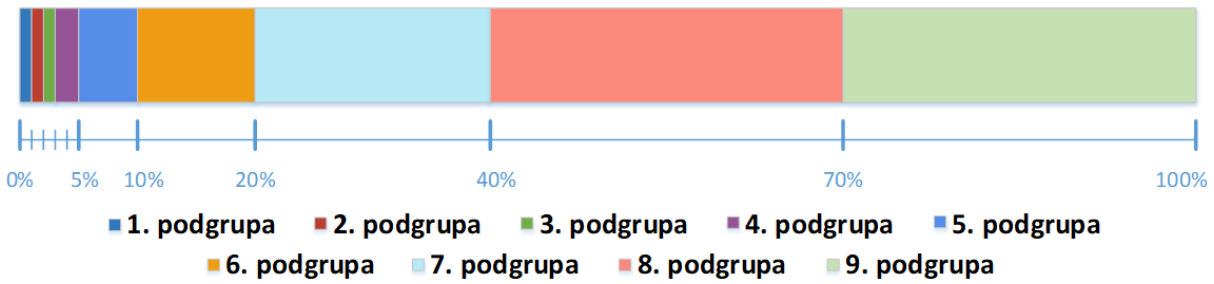


Slika IV.7 Prozorčić kroz koji korisnici dodjeljuju aplikaciji pravo pristupa njihovim podatcima s Facebooka

Primjenom modela za računanje intenziteta prijateljstva razvijenog u sklopu istraživanja *BestFriends* i primjenom koeficijenata općenite značajnosti interakcijskih parametara određenih u drugom dijelu istraživanja *BestFriends* (Tablica IV.2), za svakog se ispitanika kreira po izračunatom intenzitetu prijateljstva silazno poredana lista njegovih prijatelja. Na

prvom je mjestu po izračunu modela *najbolji* prijatelj ispitanika, a na zadnjem onaj *najmanje dobar*. S obzirom na to da ispitanici prosječno imaju nekoliko stotina fejsbukovskih prijatelja (429 u ovom istraživanju), nemoguće je od ispitanika tražiti da zasebno odgovaraju za svako prijateljstvo. Zbog toga smo prijatelje ispitanika podijelili u podgrupe s obzirom na izračunati intenzitet njihova prijateljstva s ispitanikom. U upitniku ispitanicima postavljamo pitanja vezana za konkretne njihove prijatelje, ali svaki prijatelj koji se nalazi u anketi zapravo je predstavnik podgrupe prijateljstva u koju je svrstan.

Ispitanikove prijatelj podijeljeni su u ukupno 9 podgrupa. U 1. podgrupi nalaze prijatelji s izračunatim najvećim intenzitetom prijateljstva, dok su u 9. prijatelji s najmanjim. Prvu grupu čini 1% najboljih prijatelja prema izračunu modela, drugu grupu narednih 1%, treću narednih 1%, četvrtu narednih 2%, petu narednih 5%, šestu narednih 10%, sedmu narednih 20%, osmu narednih 30% i devetu zadnjih 30%. Slika IV.8 vizualizira podjelu u podgrupe. Osigurano je da se u svakoj podgrupi nalazi barem jedan prijatelj, a isti prijatelj ne može se istovremeno nalaziti u dvije podgrupe. Tako je ukupna lista prijatelja podijeljena na 9 disjunktnih podlisti. Podgrupe su različite veličine zbog pretpostavke istraživanja da se intenzitet prijateljstva najviše razlikuje između ispitanika i njegovih najbližih prijatelja. Što je intenzitet prijateljstva slabiji, to su i odnosi međusobno sličniji – podjednako slabi, tj. i sam se ispitanik teško odlučuje s kojim je prijateljem jače (tj. slabije) povezan.



Slika IV.8 Podjela prijatelja ispitanika u podgrupe

Anketni upitnik sastoji se od 4 cjeline u kojima ispitanik odgovara na pitanja vezana za djelomično slučajno odabrane prijatelje. Za svakog je prijatelja koji se pojavljuje u upitniku unaprijed definirana podgrupa iz koje će biti odabran, ali je unutar te podgrupe sam prijatelj nasumično odabran. U prvoj cjelini pitanja od ispitanika se traži da odgovori tko mu je bolji prijatelj u ponuđenim parovima prijatelja, u drugoj cjelini ispitanik ponuđene prijatelje treba rasporediti u različite ranije definirane kategorije prijateljstva, u trećoj cjelini odgovara na niz pitanja vezanih za povjerenje prema svojim prijateljima i sličnost s njima u stavovima i razmišljanjima, a u zadnjoj cjelini ispitaniku se prikazuje lista njegovih 10 najboljih prijatelja prema izračunu modela, a on je treba ocijeniti. Cilj je bio doznati što više o odnosu ispitanika i

njegovih prijatelja, a opet zadržati upitnik dovoljno kratkim da ispitanik ne izgubi interes za ispunjavanjem. U nastavku će biti prezentirana svaka cjelina upitnika, zajedno s dobivenim rezultatima i diskusijom.

Osim odgovora na 4 cjeline pitanja u upitniku, ispitanike smo zamolili da odgovore i na nekoliko pitanja o sebi: koliko imaju godina, kojeg su spola, iz koje su države te koje je njihovo zanimanje. Studente i školarce zamolili smo da navedu naziv svoje škole ili fakulteta. Svakom ispitaniku po završetku ispunjavanja upitnika prikazano je nekoliko zanimljivih podataka o interakciji njega i njegovih prijatelja: s kim su razmijenili najviše poruka, s kim imaju najviše zajedničkih fotografija, a s kim najviše zajedničkih prijatelja, tko je najviše puta njihove fotografije označio sa *sviđa mi se* te tko je najviše puta njihove objave označio sa *sviđa mi se*. Svi ispitanici imali su priliku ostaviti komentar na istraživanje. Prikupljeno je čak 218 vrlo korisnih komentara.

IV.2.2.1 Usporedba parova prijatelja

Usporedba parova prijatelja prva je cjelina pitanja na koju je ispitanik u upitniku trebao odgovoriti. Ponuđena su mu 24 para prijatelja, a ispitanik u svakom paru treba odabrati boljeg prijatelja ili odgovoriti da se ne može odlučiti. Svaki ponuđeni prijatelj slučajno je odabran iz jedne od 9 podgrupa prijateljstva. Kroz ponuđena 24 para prijatelja ispitanicima se nudi da usporede prijatelje iz podgrupa prema prikazu u Tablica IV.3.

Tablica IV.3 Usporedbe podgrupa prijateljstva

1. – 9.	2. – 8.	3. – 7.	4. – 6.	4. – 6.	5. – 6.
5. – 6.	1. – 2.	1. – 3.	1. – 4.	2. – 3.	2. – 4.
3. – 4.	1. – 5.	2. – 5.	3. – 5.	4. – 5.	7. – 9.
8. – 9.	7. – 8.	5. – 7.	5. – 8.	5. – 9.	4. – 7.

Cilj ove cjeline pitanja jest utvrditi koliko model koji je treniran nad podatcima o najboljim prijateljima može uspješno uspoređivati prijatelje iz cijelog intervala prijateljstva. Model bi savršeno radio kada bi uvijek izračunao veći intenzitet prijateljstva za onog prijatelja kojeg je i ispitanik u upitniku označio boljim. Slika IV.9 prikazuje dio anketnog upitnika za usporedbu parova prijatelja (imena i fotografije zbog zaštite privatnosti zamućeni su za potrebe doktorskog rada).

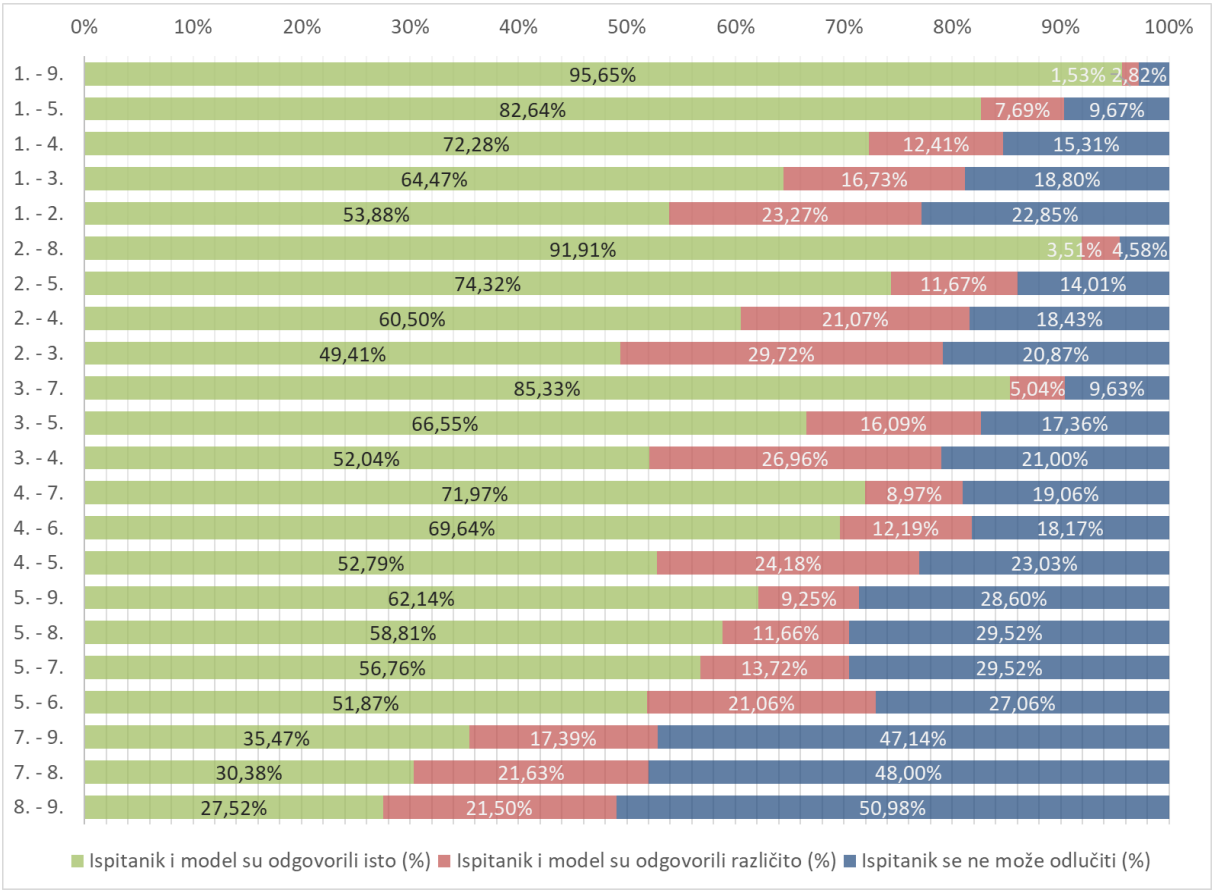
Slika IV.9 Upitnik za usporedbu parova prijatelja

Rezultati

Slika IV.10 prikazuje uspješnost rada modela za različite parove podgrupa. Slučajevi u kojima je ispitanik kao boljega odabrao prijatelja iz više podgrupe označeni su zelenom bojom (model je ispravno boljeg prijatelja svrstao u višu podgrupu), slučajevi u kojima je ispitanik kao boljega odabrao prijatelja iz niže podgrupe označeni su crvenom bojom (model je za lošijeg prijatelja po mišljenju ispitanika pogrešno izračunao veći intenzitet prijateljstva), dok su slučajevi u kojima se ispitanik nije mogao odlučiti za boljeg prijatelja označeni plavom bojom.

Rasprava

Za sve kombinacije podgrupa ispitanici su češće kao boljeg prijatelja odabirali onog prijatelja za kojeg je i model izračunao veći intenzitet prijateljstva no što su za boljeg odabrali onoga koji je po izračunu modela lošiji. Međutim, može se uočiti da se odgovori ispitanika i modela u većem postotku podudaraju kada su uspoređivane podgrupe udaljenije. Primjerice, pri usporedbi 1. i 9. podgrupe model i ispitanik istog prijatelja smatraju boljim u čak 95,65% slučajeva. Kada se pak radi o usporedbi 1. i 2. podgrupe, odgovori ispitanika i modela podudaraju se u 53,88% slučajeva, dok se u 23,27% slučajeva razlikuju. U 22,85% slučajeva ispitanik nije mogao odlučiti koji mu je prijatelj bolji, što pokazuje da se radi o toliko malim razlikama da i sam ispitanik često ne može reći koji mu je prijatelj bolji. Ta neodlučnost ispitanika najviše dolazi do izražaja pri usporedbi 8. i 9. podgrupe. Tada u preko 50% slučajeva ispitanik nije u mogućnosti procijeniti tko mu je bolji prijatelj. To potvrđuje i jednu od pretpostavki istraživanja koja kaže da se oni *najlošiji* prijatelji međusobno najmanje razlikuju, tj. da to zapravo i nisu prijatelji nego samo poznanici među kojima je teško odrediti tko je od koga bolji (ili lošiji). Međutim, i za usporedbu 8. i 9. podgrupe za slučaj kada ispitanik ipak uspije odlučiti, model češće daje odgovor jednak odgovoru ispitanika no što pogriješi u procjeni.



Slika IV.10 Usporedba parova prijatelja – rezultati

Zaključno se može reći da model vrlo dobro razlikuje jaka od slabih prijateljstava te da prilično dobro uspijeva rangirati najbolje prijatelje. Kod rangiranja slabijih prijatelja model je nešto lošiji, ali to je i posve očekivano jer se pokazuje da u tom intervalu prijateljstva čak ni ispitanik vrlo često ne može odlučiti koji mu je prijatelj (ili u tom slučaju poznanik) bolji. Općenito, uspješnost modela manja je što su podgrupe bliže.

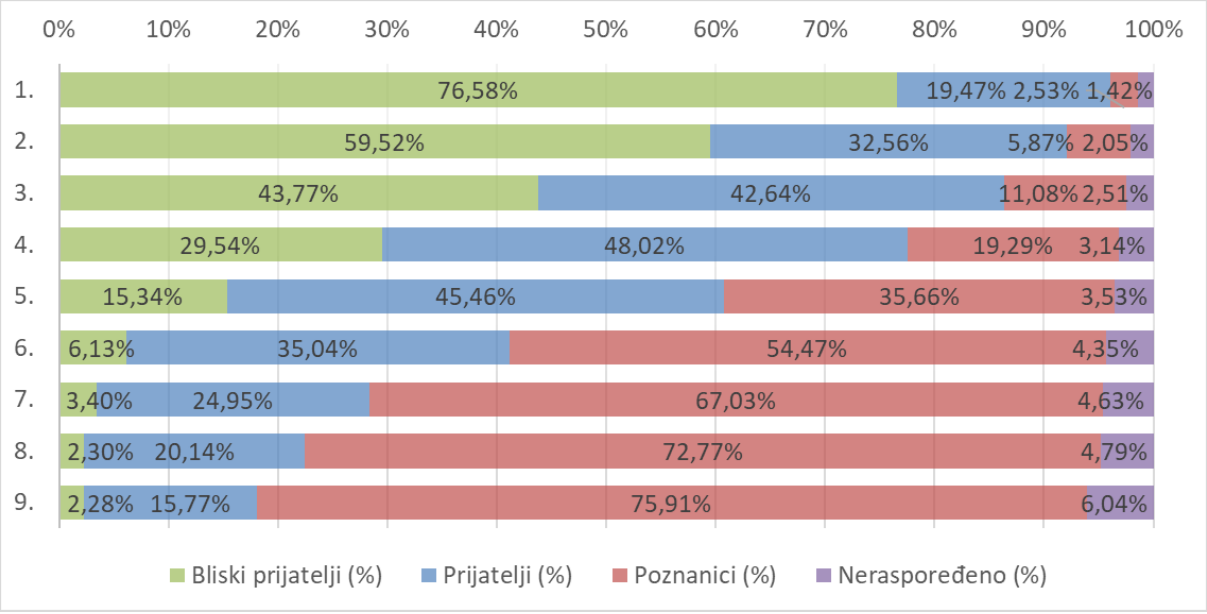
IV.2.2.2 Raspoređivanje prijatelja u različite kategorije prijateljstva

U drugoj cjelini pitanja u upitniku od ispitanika se tražilo da ponuđena 34 prijatelja rasporede u 3 kategorije prijateljstva: *bliski prijatelji*, *prijatelji* i *poznanici*. 34 ponuđena prijatelja nasumično su odabrana iz različitih podgrupa prijateljstva na sljedeći način: iz prvih 7 podgrupa slučajnim su odabirom izabrana po 4 prijatelja, a iz 8. i 9. podgrupe po 3 prijatelja. „Prijatelje“ koji nisu fizičke osobe ispitanici su trebali ostaviti neraspoređenima (iako to nije u skladu s pravilima Facebooka, nije rijetkost da profile otvaraju i različite organizacije – poduzeća, sportski klubovi i slično). Slika IV.11 prikazuje upitnik u kojem su ispitanici korištenjem mehanizma *povuci i ispusti* trebali rasporediti ponuđene prijatelje u različite kategorije prijateljstva.

Slika IV.11 Upitnik za raspoređivanje prijatelja u različite kategorije prijateljstva

Rezultati

Slika IV.12 prikazuje u koje su kategorije prijateljstva ispitanici raspoređivali prijatelje predstavnike određenih podgrupa prijateljstva. Za svaku podgrupu prijateljstva prikazano je u kojem su postotku prijatelji koji su predstavljali tu podgrupu raspoređeni u svaku od ponuđenih kategorija te u kojem su postotku ostali neraspoređeni.



Slika IV.12 Raspoređivanje prijatelja u različite kategorije prijateljstva - rezultati

Rasprava

Ova cjelina upitnika zasnovana je na pretpostavci da se prijatelji općenito mogu rasporediti u 3 osnovne kategorije prijateljstva: *bliski prijatelji* (prijatelji najvišeg intenziteta prijateljstva), *prijatelji* (prijatelji srednjeg intenziteta prijateljstva) i *poznanici* (prijatelji slabog ili gotovo nikakvog intenziteta prijateljstva). S obzirom na to, očekivalo se da će *bliskih prijatelja* biti najviše u visokim podgrupama prijateljstva i da će njihov broj padati što bude

promatrana niža podgrupa prijateljstva. Za *poznanike* se očekivalo da će ih najviše biti u niskim podgrupama prijateljstva i da će njihov broj padati kako se bude išlo prema višim podgrupama. *Prijatelji* su zamišljeni kao kategorija između *bliskih prijatelja* i *poznanika* pa se za njih pretpostavljalo da će ih najviše biti u srednjim skupinama prijateljstva.

Rezultati se upravo savršeno poklapaju s inicijalnim pretpostavkama. Doduše, idealno bi bilo kada bi postojale fiksne granice, u smislu izračunatog intenziteta prijateljstva, između kategorija prijateljstva. Naravno, u realnom slučaju to nije moguće pa se, recimo, *bliske prijatelje*, iako u vrlo malom postotku, može pronaći i u 9. podgrupi prijateljstva, kao i poznanike u 1. podgrupi prijateljstva.

Može se primijetiti i da je postotak neraspoređenih prijatelja (prijatelja koji nisu fizičke osobe) najmanji u visokim podgrupama, a najveći u najnižim. To sugerira da ispitanici imaju malu interakciju s „prijateljima“ koji su zapravo lažni profili nekih organizacija, tj. da su oni većinom svrstani u niske skupine prijateljstva.

IV.2.2.3 Određivanje razine povjerenje i sličnosti

U trećoj cjelini pitanja kroz 5 pitanja nastojalo se provjeriti kako se odnos povjerenja i sličnosti ispitanika s njegovim prijateljima mijenja s promjenom podgrupe prijateljstva. Ispitanicima je postavljeno sljedećih 5 pitanja koja opisuju realne životne situacije:

1. *Biste li vjerovali traču koji ova osoba podijeli s Vama?*
2. *Biste li otvorili povjerljivu poštu pred ovom osobom?*
3. *Biste li bili spremni ovoj osobi posuditi značajnu količinu novaca?*
4. *Mislite li da biste se mogli osloniti na ovu osobu u teškoj situaciji?*
5. *Koliko ste Vi i ova osoba slični u vjerskim, političkim i društvenim stavovima?*

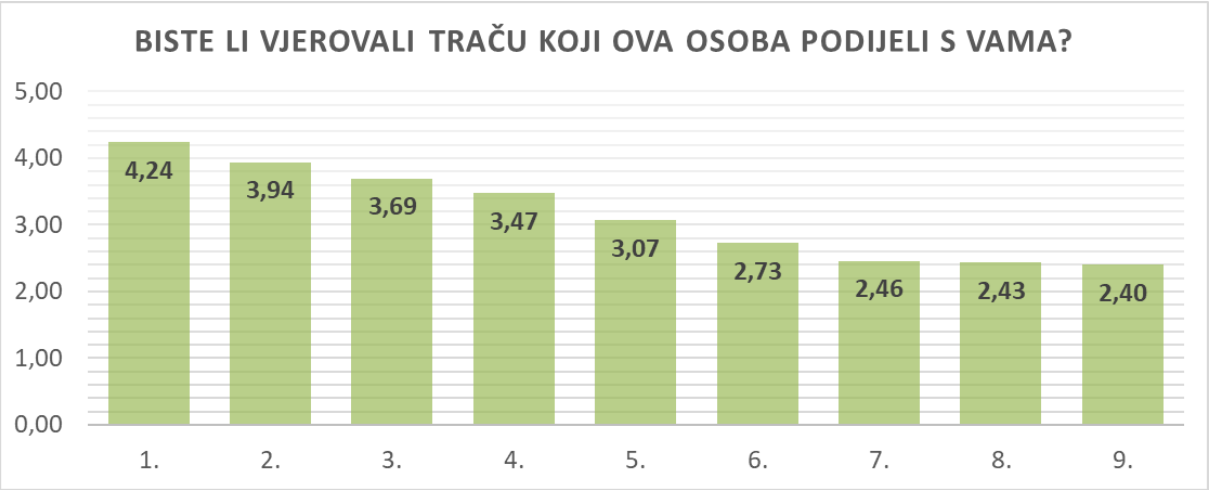
Za svako od 5 ponuđenih pitanja, ispitanik je trebao odgovoriti zasebno za svojih 9 djelomično slučajno odabranih prijatelja. U svih 5 pitanja pojavljuje se istih 9 prijatelja, a oni su ujedno i predstavnici svake od 9 podgrupa prijateljstva. Za svakog prijatelja, ispitanik može odgovoriti na skali od 1 do 5 (1 – potpuno neslaganje (ne), 5 – potpuno slaganje (da)), a može odlučiti i ne odgovoriti.

Slika IV.13 prikazuje primjer jednog pitanja u iz treće cjeline upitnika.

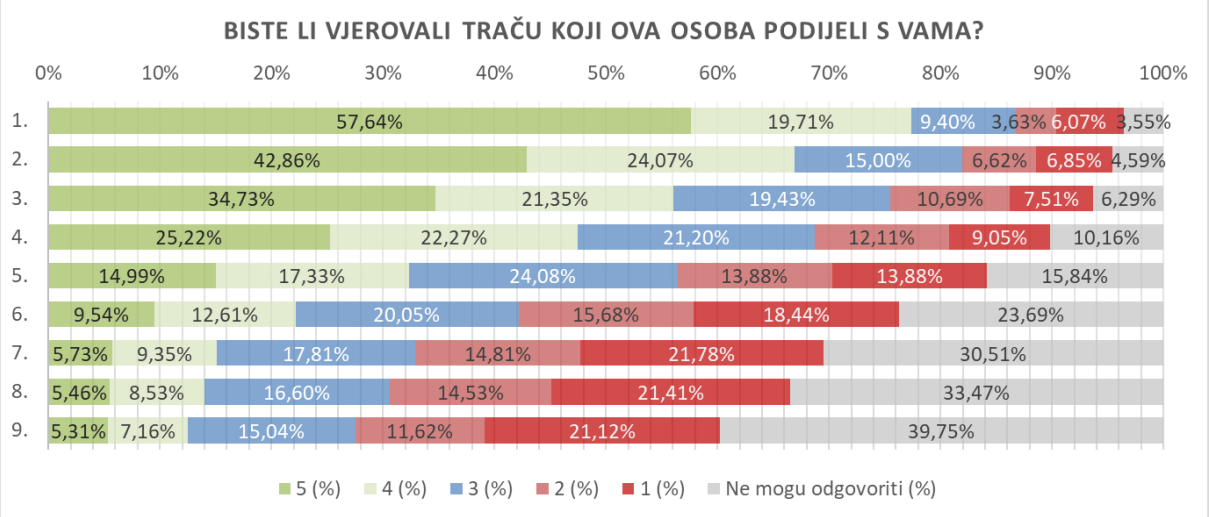
Slika IV.13 Ocjenjivanje razine povjerenja i sličnosti – primjer pitanja u upitniku

Rezultati

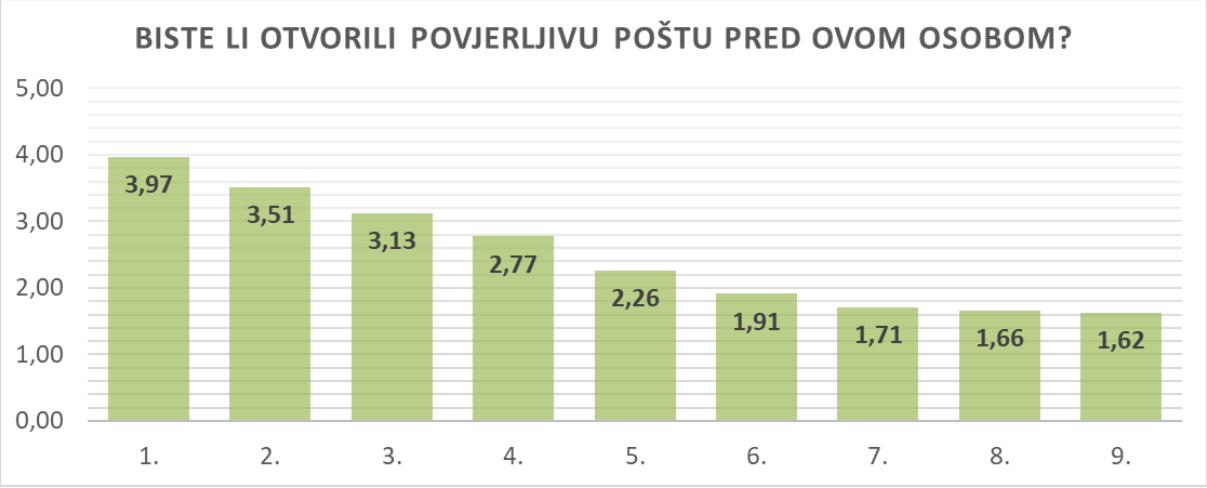
Slika IV.14 prikazuje prosjek ocjena koje su ispitanici dodijelili prijateljima predstavnicima različitih podgrupa na pitanje: „*Biste li vjerovali traču koji ova osoba podijeli s Vama?*“. Slika IV.15 prikazuje u kojem su postotku ispitanici odgovarali s kojom ocjenom za prijatelje predstavnike određene podgrupe na spomenuto pitanje. Za pitanje: „*Biste li otvorili svoju povjerljivu poštu pred ovom osobom?*“, Slika IV.16 prikazuje prosječne ocjene po grupama, a Slika IV.17 raspodjelu ocjena. Slika IV.18 i Slika IV.19 odnose se na pitanje: „*Biste li bili spremni ovoj osobi posuditi značajniju količinu novaca?*“, Slika IV.20 i Slika IV.21 na pitanje: „*Mislite li da biste se mogli osloniti na ovu osobu u teškoj situaciji?*“, a Slika IV.22 i Slika IV.23 na pitanje: „*Koliko ste Vi i ova osoba slični u vjerskim, političkim i društvenim stavovima*“.



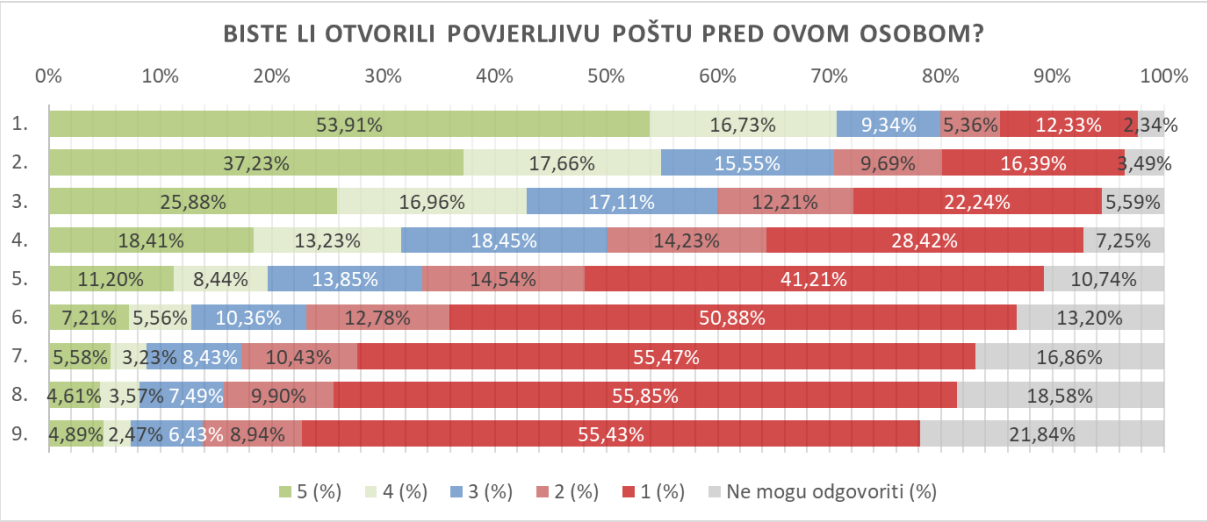
Slika IV.14 Prosječna ocjena po podgrupama za pitanje: „Biste li vjerovali traču koji ova osoba podijeli s Vama?“



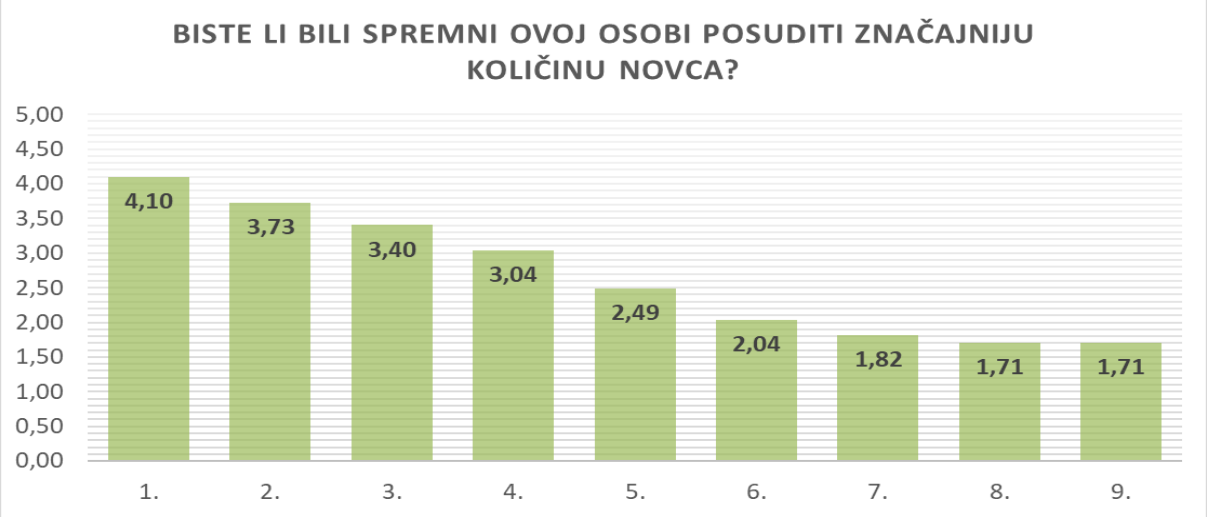
Slika IV.15 Raspodjela ocjena po podgrupama za pitanje: „Biste li vjerovali traču koji ova osoba podijeli s Vama?“



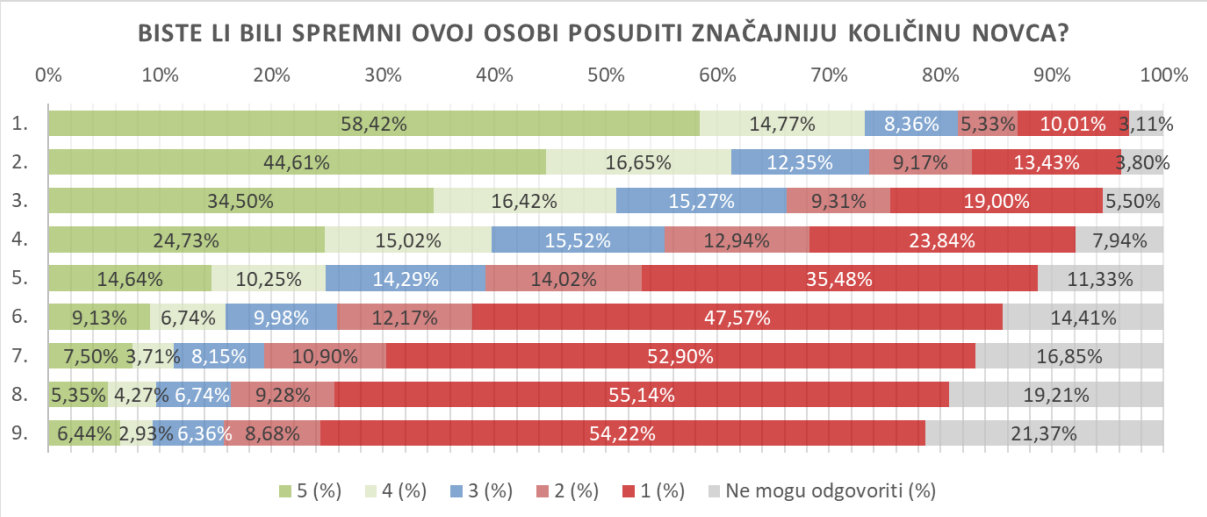
Slika IV.16 Prosječna ocjena po podgrupama za pitanje: „Biste li otvorili povjerljivu poštu pred ovom osobom?“



Slika IV.17 Raspodjela ocjena po podgrupama za pitanje: „Biste li otvorili povjerljivu poštu pred ovom osobom?“

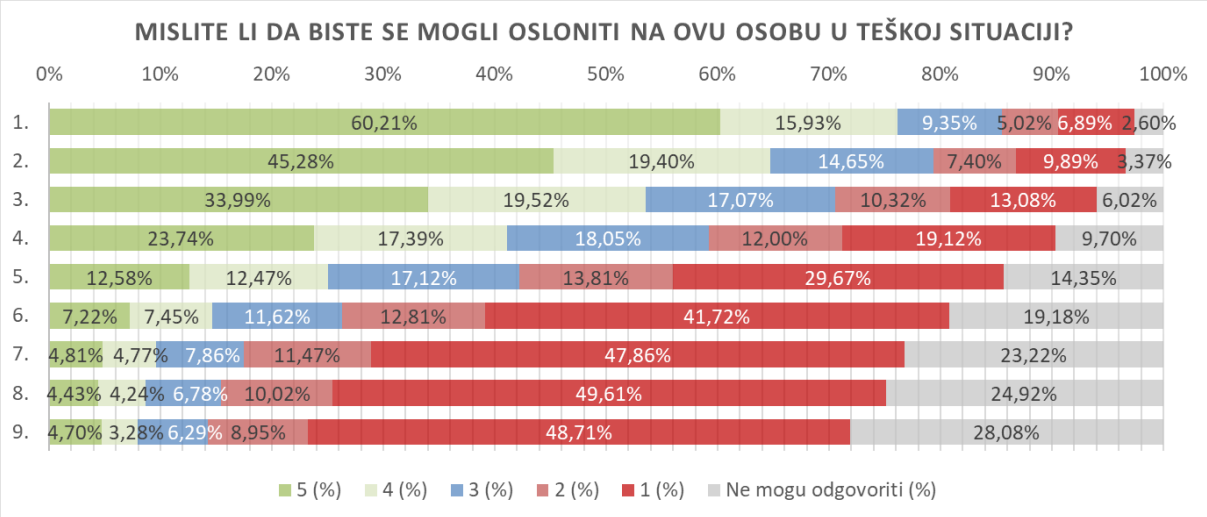


Slika IV.18 Prosječna ocjena po podgrupama za pitanje: „Biste li bili spremni ovoj osobi posuditi značajniju količinu novaca?“



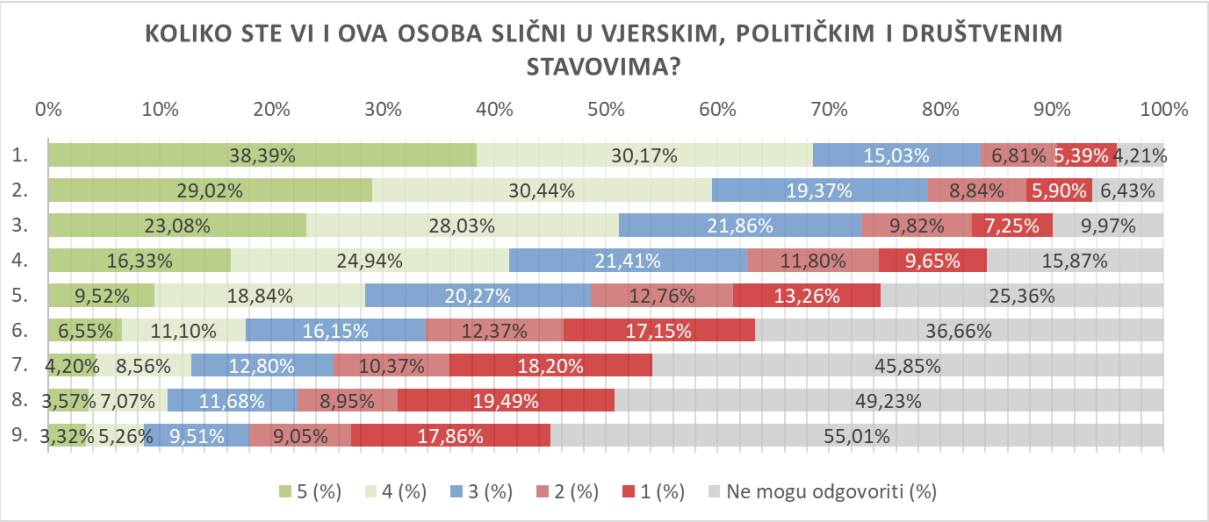
Slika IV.19 Raspodjela ocjena po podgrupama za pitanje: „Biste li bili spremni ovoj osobi posuditi značajniju količinu novaca?“

Slika IV.20 Prosječna ocjena po podgrupama za pitanje: „Mislite li da biste se na ovu osobu mogli osloniti u teškoj situaciji?“



Slika IV.21 Raspodjela ocjena po podgrupama za pitanje: „Mislite li da biste se na ovu osobu mogli osloniti u teškoj situaciji?“

Slika IV.22 Prosječna ocjena po podgrupama za pitanje: „Koliko ste Vi i ova osoba slični u vjerskim, političkim i društvenim stavovima?“



Slika IV.23 Raspodjela ocjena po podgrupama za pitanje: „Mislite li da biste se na ovu osobu mogli osloniti u teškoj situaciji?“

Rasprava

Od 5 pitanja postavljenih ispitanicima, 4 ispituju odnos povjerenja kroz primjere različitih životnih situacija, a jedno pitanje sličnost između prijatelja i ispitanika u stavovima i razmišljanjima. Kod rezultata za sva pitanja vidi se da prosječna ocjena pada s padom važnosti podgrupe koju prikazani prijatelj prema izračunu modela predstavlja. Također, može se uočiti i da s padom važnosti podgrupe, pada i broj visokih ocjena, a raste broj niskih. Kod svih pitanja s padom značaja podgrupe izraženo je povećanje postotka u kojem ispitanik ne može odgovoriti na pitanje. To je posebno izraženo u zadnjem pitanju koje ispituje sličnosti u stavovima i razmišljanjima, a može se objasniti pretpostavkom istraživanja da ispitanik prijatelje s kojima je slabo povezan vrlo slabo poznaje (tek su poznanici) pa vjerojatno ni ne zna kakvi su njegovi stavovi i razmišljanja, a što je preduvjet za odgovor na postavljeno pitanje.

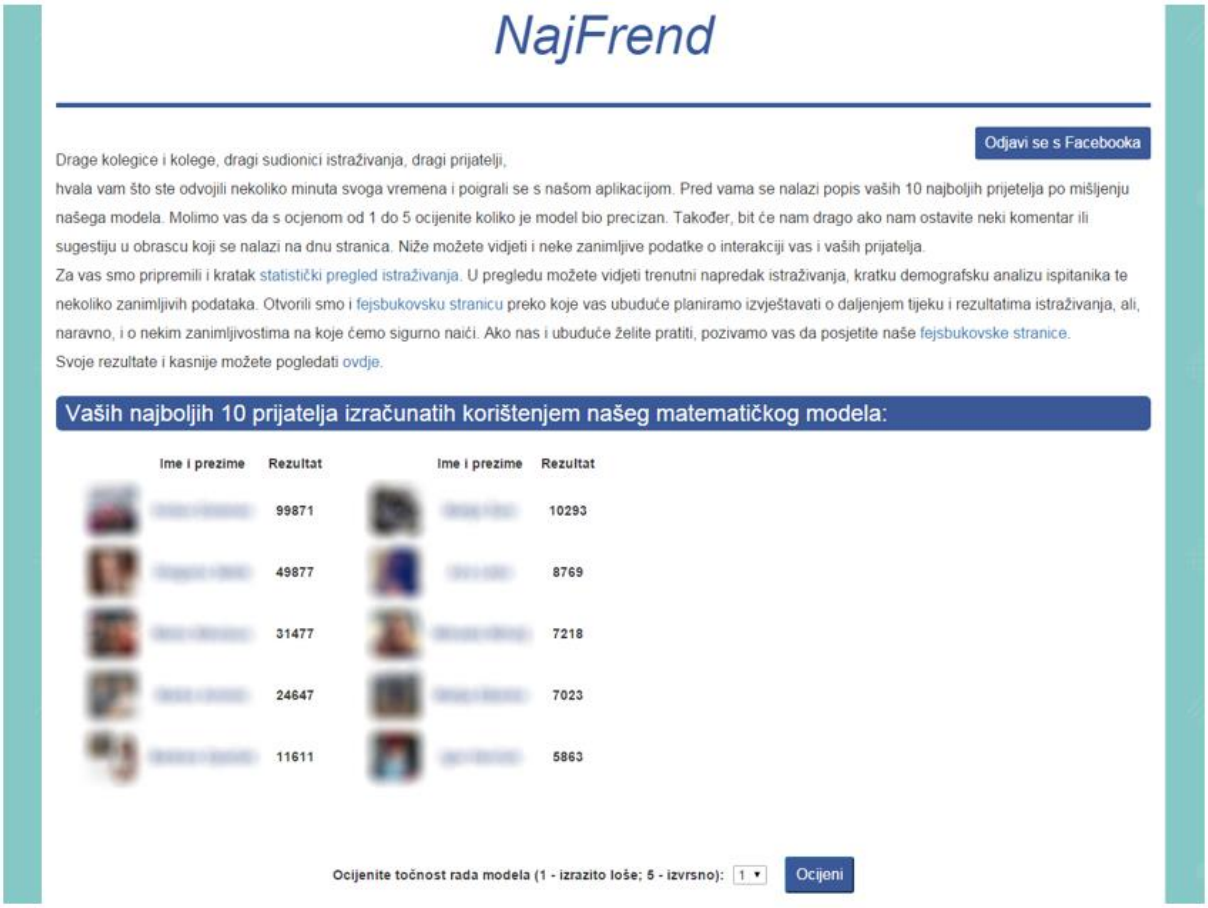
Iz odgovora na sva pitanja može se uočiti da se predloženi model, koji na temelju interakcije na Facebooku računa intenzitet prijateljstva, može iskoristiti za dobivanje uvida u različite odnose među ljudima te da su ti različiti odnosi međusobno vrlo korelirani. Izazov za daljnja istraživanja jest provjeriti je li moguće odrediti posebne općenite koeficijente značajnosti pojedinog parametra interakcije takve da model čim bolje detektira specifične društvene odnose.

IV.2.2.4 Evaluacija liste najboljih prijatelja

U zadnjoj cjelini pitanja ispitaniku je prikazano njegovih 10 najboljih prijatelja prema izračunu modela, a od ispitanika se tražilo da tu listu ocijeni ocjenom od 1 do 5, pri čemu 1 znači potpuno nezadovoljstvo, a 5 potpuno zadovoljstvo. Prosječna ocjena dobivena od ispitanika iznosi 3,48. Ako bi se taj rezultat prikazao postotno, tj. skalirao na interval o 0 do

100, dobili bismo 69,6%. Ta uspješnost gotovo je identična uspješnosti koju je model polučio u drugom dijelu istraživanja *BestFriends*. Tada je prosječno preklapanje ispitanikove liste od 10 najboljih prijatelja i liste od 10 najboljih prijatelja generirane na temelju izračuna modela bilo ~7, tj. ~70%. Na ovaj je način uspješnost rada modela verificirana nad posve novim ispitanicima i s promijenjenim zadatkom za ispitanike – umjesto da unose svojih do 10 najboljih prijatelja, a da model onda pokuša kroz izračun intenziteta prijateljstva doći do iste te liste, ovaj puta je model odmah generirao listu najboljih prijatelja po svom izračunu, a ispitanik je trebao procijeniti koliko je model pri tome bio uspješan.

Slika IV.24 prikazuje isječak iz upitnika u kojem se ispitaniku prezentira lista njegovih 10 najboljih prijatelja prema izračunu modela. Uz svakog prijatelja prikazan je i izračunati intenzitet njegova prijateljstva s ispitanikom.



Slika IV.24 Ocjenjivanje točnosti liste od 10 najboljih prijatelja koju je generirao model – isječak iz upitnika

Dio rezultata istraživanja *NajFrend* objavljen je u radu [3].

IV.3 ISTRAŽIVANJE KORISNIČKOG DOŽIVLJAJA INTERAKCIJE NA FACEBOOKU

Istraživanje korisničkog doživljaja interakcije na Facebooku provedeno je 2016. godine. U istraživanju su sudjelovala 144 ispitanika[13]. U prethodno provedenom istraživanju *BestFriends* nastojalo se u procesu treniranja modela odrediti koliko je koji parametar interakcije značajan. Određene su neovisno i prediktivne snage za svaki parametar. U istraživanju *NajFrend* kroz mnoštvo komentara ispitanika doznali smo kako ispitanici gledaju na interakciju na Facebooku. U ovom smo istraživanju odlučili kroz konkretna pitanja ispitati stavove i navike korisnika. Provjerili smo kako sami korisnici Facebooka doživljavaju pojedine parametre interakcije, koje parametre koriste primarno za interakciju s bliskim prijateljima i slično. Kako se već više puta pokazalo da broj razmijenjenih privatnih poruka ima vrlo veliku prediktivnu snagu pri određivanju intenziteta prijateljstva, odlučili smo provjeriti u kojem udjelu korisnici koriste najpoznatije porukatore. U nastavku ćemo ukratko prikazati rezultate, tj. spoznaje do kojim smo došli analizirajući odgovore korisnika u upitniku.

IV.3.1 Rezultati upitnika i rasprava

Na početku upitnika, ispitanike smo pitali koliko često koriste Facebook te koju vrstu interakcije najviše koriste. Gotovo 95% ispitanika odgovorilo je da Facebook koristi na dnevnoj bazi, a glavne aktivnosti na Facebooku im je razmjena privatnih poruka te pregledavanje objava prijatelja. Među različitim porukatorima, najviše ispitanika koristi *Whatsapp* (53%), dok je Facebookov *Messenger* drugi (39%). Doznali smo i da korisnici za interakciju s bliskim prijateljima na Facebooku najviše koriste oznake *sviđa mi se* na objave prijatelja, komentiraju njihove objave, označavaju se na zajedničkim fotografijama, ali i razmjenjuju poruke kroz Facebookov *Messenger*. U proteklim istraživanjima imali smo dosta nedoumica oko toga što znači prijatelj označen kao *blizak prijatelj*. Naime, događalo se da bliski prijatelji ispitanika ne budu na popisu korisnikovih 10 najboljih prijatelja, a budu prijatelji koji nisu označeni kao bliski. Većina od 52% ispitanika odgovorila nam je da ne koristi opciju *bliski prijatelj*, 42% ispitaniku tu opciju koristi zaista za bliske prijatelje iz stvarnog života, dok njih 6% prijatelje označava *bliskima* samo zato da bi primali obavijesti o njihovim novim objavama, iako to nisu njihovi bliski prijatelji u stvarnom životu. Ovi rezultati dosta dobro objašnjavaju zašto je uzimanje u obzir oznake *blizak prijatelj* pri izračunu intenziteta prijateljstva ujedno bilo vrlo korisno, ali i zašto je nekad posve pokvarilo rezultat rada modela.

[13] Istraživanje je u sklopu svoga diplomskog rada, pod vodstvom autora ovoga rada, provodila Sanja Krakan.

Pitali smo korisnike i misle li da imaju više zajedničkih prijatelja na Facebooku s bliskim prijateljima iz stvarnog života. Čak 84% ispitanika smatra da veći broj zajedničkih prijatelja s nekim fejsbukovskim prijateljem ne znači da im je to blizak prijatelj u stvarnom životu. To je, zapravo, vrlo u skladu s rezultatima istraživanja *BestFriends*. Tada se u oba dijela istraživanja pokazalo da je broj zajedničkih prijatelja na Facebooku jedan od najmanje značajnih parametara interakcije pri određivanju intenziteta prijateljstva.

Na koncu smo pitali što misle koji parametar interakcije najbolje oslikava jačinu prijateljstva u stvarnom životu. Čak 60% ispitanika smatra da su to razmijenjene privatne poruke, a 22% da su to zajedničke fotografije. Broj razmijenjenih privatnih poruka pokazao se i u istraživanju *BestFriends* uvjerljivo najprediktivnijim interakcijskim parametrom, a i zajedničke fotografije nalaze se među najznačajnijim interakcijskim parametrima.

Rezultati ovog istraživanja objavljeni su u [13]. U radu je predstavljen i novi model za računanje intenziteta prijateljstva koji uzima u obzir subjektivnu procjenu korisnika ispitanika Facebooka o značajnosti pojedinog parametra interakcije te značajnost pojedinog parametra interakcije dobivenu primjenom algoritma nadziranog učenja slučajne šume nad podatcima prikupljenim u istraživanju *NajFrend*.

V. EKSPLORATORNA ANALIZA PODATAKA O INTERAKCIJI KORISNIKA NA FACEBOOKU

U okviru ovog doktorskog rada predlaže se model za računanje društvene udaljenosti na temelju analize interakcije ljudi na sustavu za društveno umrežavanje Facebook. U sklopu prethodno opisanog istraživanja *NajFrend* (potpoglavlje IV.2) prikupljeni su sumarni podatci[14] o interakciji više od 1.400.000 parova prijatelja. Kako podatci o interakciji korisnika na Facebooku u predloženom modelu imaju ulogu prediktora, valja prije njihova korištenja detaljno analizirati njihova svojstva. U tu je svrhu provedena eksploratorna analiza podataka, a njezine se rezultati prezentiraju u ovom poglavlju. Rezultati provedene eksploratorne analize prezentirani su i u radovima [53], [54].

Eksploratorna analiza provedena je nad dva podatkovna skupa. Prvi podatkovni skup čine podatci o interakciji korisnika na Facebooku sumirani na razini ispitanika, tj. ukupan broj poruka koje je ispitanik razmijenio sa svim svojim prijateljima na Facebooku, ukupan broj oznaka *sviđa mi se* na objave prijatelja i slično. Drugi podatkovni skup čine podatci o interakciji korisnika na Facebooku sumirani na razini parova prijatelja (ukupan broj razmijenjenih poruka između svakog para ispitanik-prijatelj, ukupan broj oznaka *sviđa mi se* između parova prijatelja i slično). Prvi podatkovni skup opisuje ponašanje pojedinog ispitanika, a njegovom se analizom može utvrditi kako se ponašaju različiti ispitanici. Drugi podatkovni skup opisuje interakciju koju ispitanici imaju sa svojim prijateljima. Njegovom se analizom utvrđuju različitosti u interakciji između različitih parova prijatelja

Model za računanje društvene udaljenosti uzimat će u obzir 14 parametara interakcije koje prikazuje Tablica V.1, pa je za te parametre provedena i eksploratorna analiza. Dodatno, prvom je podatkovnom skupu, uz 14 interakcijskih parametara, dodan i atribut *ukupan broj fejsbukovskih prijatelja ispitanika*. Tablica V.1, uz opis parametra interakcije, sadrži i njegov kraći naziv koji ćemo koristiti u tekstu te kraticu koja će se koristiti na slikama na kojima nema mjesta za kraći naziv.

[14] Pod sumarnim podatcima misli se na ukupnu količinu pojedinog interakcijskog parametra – primjerice, koliko je promatrani par prijatelja ukupno razmijenio poruka, koliko je puta jedan od prijatelja drugom fotografiju označio oznakom *sviđa mi se* i slično.

Tablica V.1 Parametri interakcije koje koristi model za računanje društvene udaljenosti – opis, kratki naziv i kratica

Parametar interakcije	Kraći naziv	Kratica
korisnik A označio je korisnika B kao **bliskog prijatelja**	is_close_friend	icf
oznaka *sviđa mi se* korisnika A na **poveznicu** koju je objavio korisnik B	my_link_likes	mll
oznaka *sviđa mi se* korisnika A na **fotografiju** koju je objavio korisnik B	my_photo_likes	mpl
privatne poruke između korisnika A i B	inbox_chat	ic
zajednički prijatelji korisnika A i B	friend_mutual	fm
komentar korisnika B na **fotografiju** koju je objavio korisnik A	photo_comment	pc
oznaka *sviđa mi se* korisnika B na **fotografiju** koju je objavio korisnik A	photo_like	pl
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio netko treći	mutual_photo_published_by_others	mpo
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio korisnik B	mutual_photo_published_by_friend	mpf
zajednička fotografija korisnika A i B koju je objavio korisnik koju je objavio korisnik A	mutual_photo_published_by_user	mpu
korisnici A i B **zajedno označeni u objavi**	feed_together_in_post	ftp
objava korisnika B na zid korisnika A	feed_addressed	fa
komentar korisnika B na **objavu** koju je korisnik A objavio na svom zidu	feed_comment	fc
oznaka *sviđa mi se* korisnika B na **objavu** koju je korisnik A objavio na svom zidu	feed_like	fl
broj prijatelja korisnika A	friends_total	ft

U sklopu eksploratorne analize podataka izlučeni su *Pearsonovi koeficijenti korelacije*[15] [55] za sve parove atributa u promatranim podatkovnim skupovima. Također, za svaki atribut podatkovnog skupa iščitane su njegove empirijske distribucije te određene teorijske distribucije koje su najsličnije empirijskim distribucijama atributa. Analizirano je i kako se mijenja ukupna interakcija po ispitaniku s porastom broja fejsbukovskih prijatelja. Primjenom algoritma *k-means* ispitanici su podijeljeni u klastere s obzirom na njihove interakcijske navike – sklonost različitim interakcijskim parametrima, a što će se kasnije nastojati primijeniti za zasebno treniranje modela za svaki klaster ispitanika s ciljem povećanja uspješnosti modela za računanje društvene udaljenosti. Nad podatkovnim skupom s podatcima sumiranim na razini para prijatelja provedena je analiza glavnih komponenata (PCA) kako bi se utvrdilo može li se skup parametara koji ulaze u model reducirati uz zadržavanje glavnine informacija.

[15] Pearsonov koeficijent korelacije mjera je koja opisuje linearnu korelaciju između dvije varijable te kao rezultat daje vrijednost iz [-1,1], pri čemu -1 predstavlja potpunu negativnu koreliranost, 1 potpunu pozitivnu koreliranost, a 0 znači da ne postoji koreliranost između promatranih varijabli.

Sve analize rađene su u jeziku R[16].

U potpoglavlju VII.3 prezentira se idejno rješenje generatora sintetičkog proširenog društvenog grafa koji kao ulaz koristi rezultate analize koja se opisuje u ovom poglavlju. Zaštita privatnosti i osobnih podataka posebno je naglašena od 25. svibnja 2018. kada je stupila na snagu *Uredba (EU) 2016/679 Europskog parlamenta i vijeća od 27. travnja 2016. o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka te o stavljanju izvan snage Direktive 95/46/EZ (Opća uredba o zaštiti podataka – GDPR)* [40]. S obzirom na pojačavanje mjera za zaštitu privatnosti i mogućnosti deanonimizacije prethodno anonimiziranih podatkovnih skupova, svako dijeljenje empirijskih podatkovnih skupova, pa čak i kada su iz njih uklonjeni identifikatori stvarnih osoba, vrlo je rizično zbog mogućnosti deanonimizacije takvih podatkovnih skupova. Iz toga slijedi pretpostavka da će u budućem razdoblju biti značajno povećana potreba za sintetičkim podatkovnim skupovima. Kako u literaturi nije pronađen generator sintetičkog proširenog društvenog grafa, rezultati provedene eksploratorne analize bit će iskorišteni i za izradu idejnog rješenja generatora proširenog društvenog grafa.

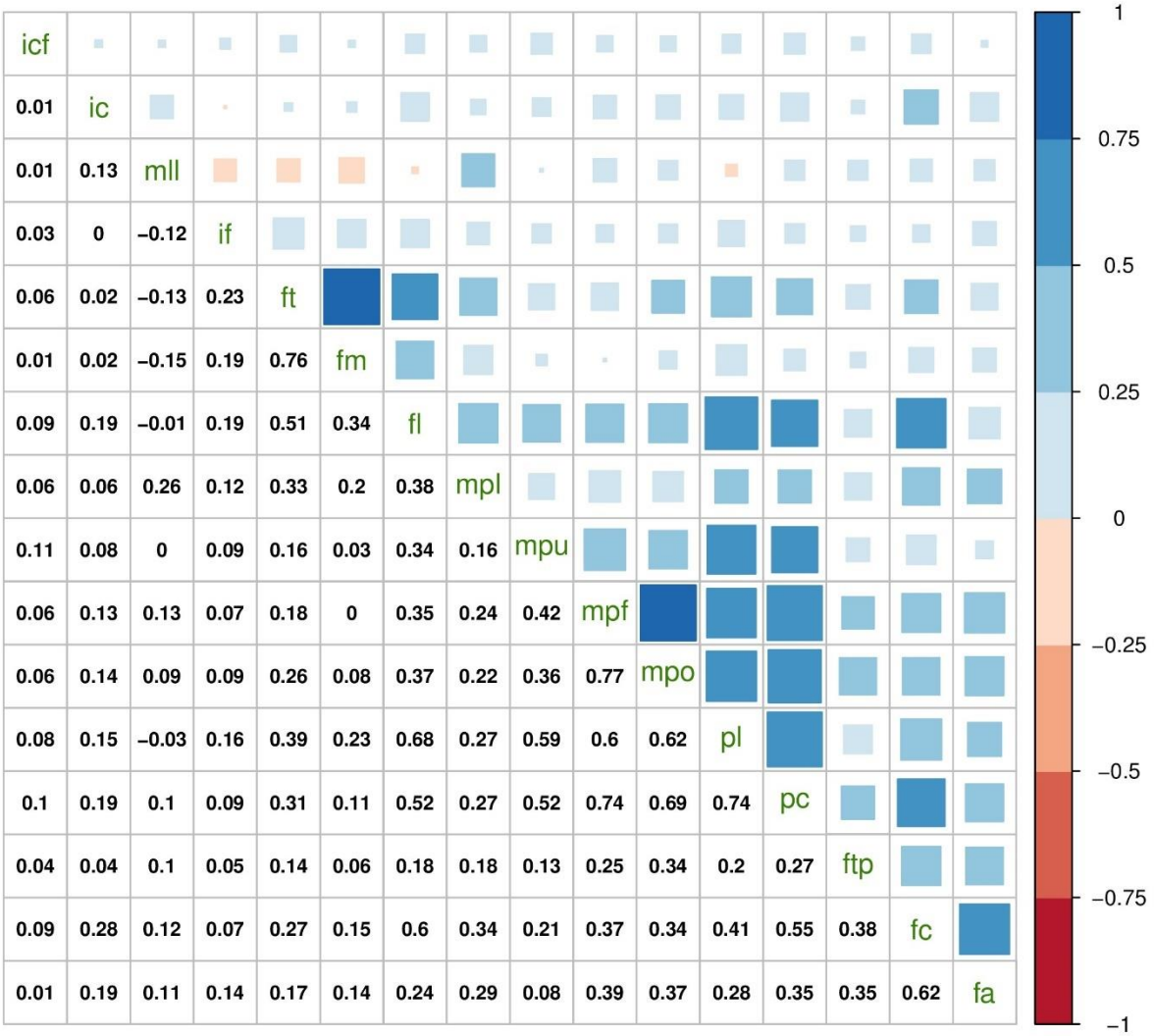
V.1 EKSPLORATORNA ANALIZA NA RAZINI ISPITANIKA

Eksploratorna analiza na razini ispitanika provedena je nad podatkovnim skupom koji u sebi sadrži atribute (interakcijske parametre + ukupan broj prijatelja) koje prikazuje Tablica V.1. U nastavku će po dijelovima biti prezentirani rezultati provedene analize.

V.1.1 Koreliranost parova atributa

S ciljem određivanja koreliranosti različitih atributa podatkovnog skupa, za sve parove atributa izračunat je Pearsonov koeficijent korelacije. Slika V.1 prikazuje izračunate vrijednosti numerički i grafički. Za atribute iz podatkovnog skupa ispisane su kratice prema Tablica V.1. U lijevom donjem dijelu slike ispisane su numeričke vrijednosti Pearsonova koeficijenta korelacije za svaki par atributa. U desnom gornjem dijelu nalaze se vizualizacije razine koreliranosti. Plavi pravokutnici prikazuju pozitivnu koreliranost, a crveni negativnu. Izostanak pravokutnika znači da promatrani atributi uopće nisu korelirani. Što su pravokutnici veći i tamniji, to je koreliranost izraženija.

[16] R je jezik i razvojno okruženje za prvenstveno statističku obradu podataka i vizualizacije



Slika V.1 Pearsonovi koeficijenti korelacije među različitim atributima u podatkovnom skupu – podatci sumirani na razini ispitanika

V.1.1.1 Rasprava

Iz prikazanih koeficijenata korelacije možemo pronaći one atribute koji su međusobno povezani. Negativno koreliranih atributa nema. Postoje neke manje negativne korelacije, ali one su značajno bliže 0 nego -1 pa ih treba tumačiti kao izostanak korelacije takvih parova atributa. Od pozitivnih korelacija najizraženija je povezanost atributa koji prikazuju broj zajedničkih slika s prijateljima, a koje je u jednom slučaju objavio prijatelj koji se nalazi na slici, a u drugom slučaju netko treći. Također, postoji visoka koreliranost između ukupnog broja prijatelja na Facebooku i ukupnog broja zajedničkih prijatelja s prijateljima na Facebooku. Visoko su korelirani i interakcijski parametri *feed_like* i *photo_like* što znači da ispitanici koji često označavaju objave svojih prijatelja sa *sviđa mi se*, ujedno i često sa *sviđa mi se* označavaju i njihove fotografije. Označavanje fotografija sa *sviđa mi se* visoko je korelirano i s komentiranjem fotografija.

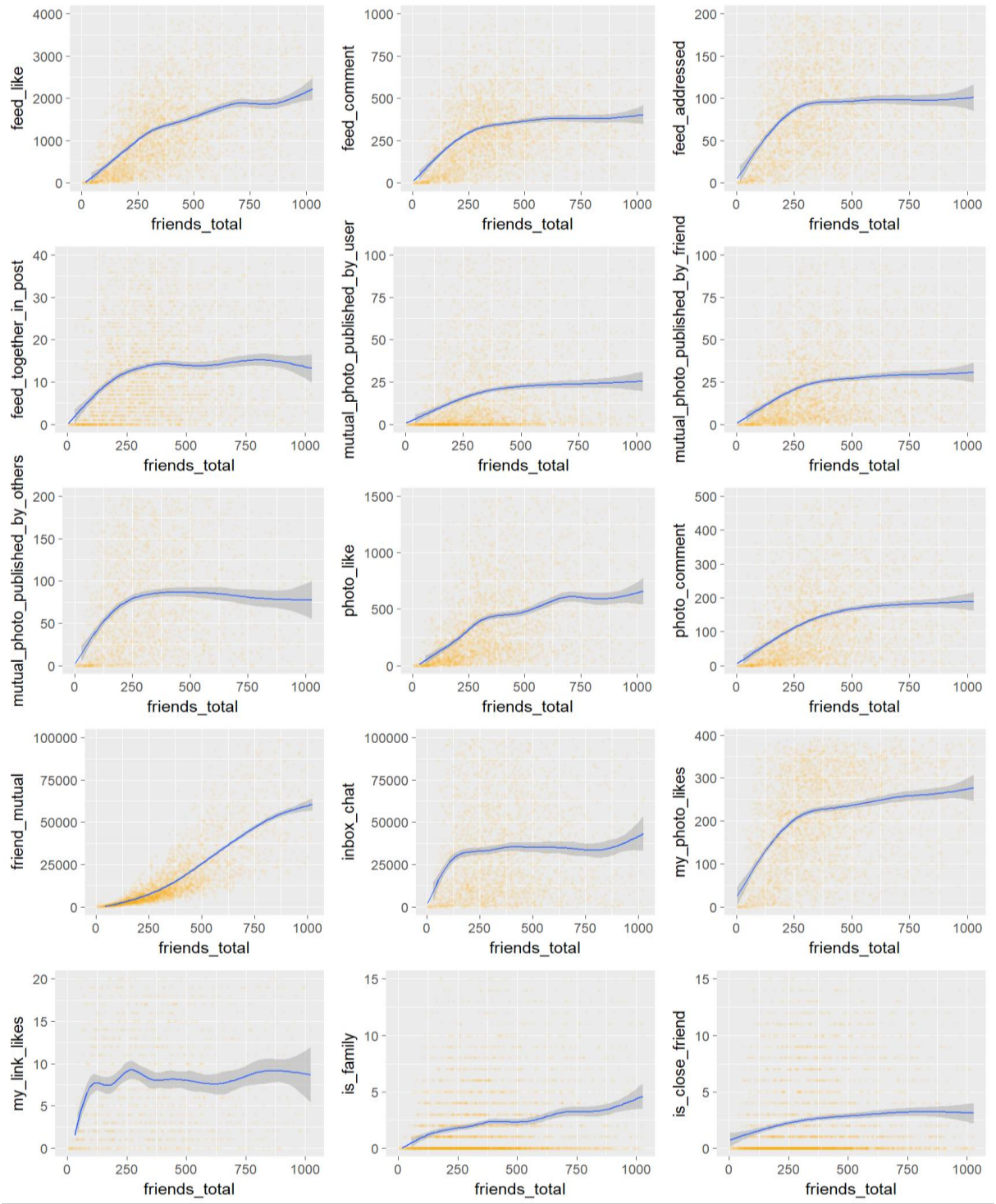
Sve navedene visoke korelacije bile su i očekivane. Ono što pomalo iznenađuje jest vrlo slaba koreliranost ukupnog broja razmijenjenih poruka s ostalim interakcijskim parametrima. Objašnjenje za to treba tražiti u odgovorima korisnika u upitniku korisničkog doživljaja interakcije na Facebooku (potpoglavlje IV.3), a u kojem se pokazalo da su ispitanici podijeljeni oko toga za što koriste Facebook. Neki ga koriste prvenstveno za razmjenu privatnih poruka, dok neki drugi više prate objave svojih prijatelja, a za razmjenu poruka koriste neki drugi od dostupnih porukatora. Posebno je zanimljivo primijetiti i da veći broj *bliskih prijatelja* ne povećava broj razmijenjenih privatnih poruka. Objašnjenje se i za to može pronaći u rezultatima spomenutog upitnika, a koji kažu da većina korisnika oznaku *bliski prijatelj* ne koristi, a i oni koji je koriste, ne koriste je samo kako bi njome označili svoje bliske prijatelje, nego i za one prijatelje čije objave žele pratiti iako možda u stvarnom životu nisu bliski prijatelji.

V.1.2 Povezanost ukupnog broja prijatelja i ukupne količine interakcije

Iz izračunatih Pearsonovih koeficijenata korelacije vidi se, između ostalog, i u kojoj je mjeri ukupan broj prijatelja povezan sa svim interakcijskim parametrima. Međutim, osim toga, činilo se zanimljivim provjeriti i kako se mijenja ukupna interakcija s porastom broja prijatelja. Kada bi korisnici jednako komunicirali sa svim prijateljima, tada bi njihova ukupna interakcija trebala linearno rasti s porastom broja prijatelja. Međutim, postoje istraživanja koja tvrde da je broj prijatelja koje čovjek ima ograničen veličinom neokorteksa [2], [56], tj. da je ukupan broj prijatelja fiksan (Dunbarov broj). Ako je to točno, većina prijatelja na Facebooku nisu pravi prijatelji pa je za očekivati da između njih i ego-korisnika nema interakcije ili je ona vrlo slaba. To bi onda za posljedicu moralo imati nelinearnost rasta ukupne interakcije s porastom broja prijatelja, tj. u nekom bi trenutku pri porastu broja prijatelja ukupna interakcija trebala prestati rasti. Odlučili smo to empirijski provjeriti nad prikupljenim podatcima.

Slika V.2 prikazuje kako ukupan broj prijatelja utječe na ostale interakcijske parametre. Žutom su bojom na slici prikazana opažanja. Plava boja prikazuje krivulju koja najbolje aproksimira opažanja. Siva sjena oko krivulje prikazuje 95-postotni interval pouzdanosti.

Sukladno početnoj pretpostavci o ograničenom ukupnom broju stvarnoživotnih prijatelja, ukupna interakcija po ispitaniku za gotovo sve parametre interakcije linearno raste samo do nekog broj prijatelja, a nakon toga se počinje asimptotski prigušivati. To potvrđuje Dunbarovu tezu o ograničenom broju prijatelja na ~150 kao i često postavljanu tezu da broj prijatelja na Facebooku značajno odstupa od broja prijatelja u realnom životu te daje još jedan poticaj izradi modela za računanje društvene udaljenosti između korisnika Facebooka, a čemu je jedan od ciljeva upravo i razlikovanje stvarnih prijatelja od onih koji su tek poznanici.



Slika V.2 Povezanost ukupne količine interakcije i broja prijatelja na Facebooku

V.1.3 Klasterizacija ispitanika s obzirom na njihove interakcijske navike

Klasterizacijom se grupiraju instance tako da za svaku instancu vrijedi da je sličnija instancama iz svoje grupe (klastera) nego instancama iz drugih grupa. Sličnost instanci, ovisno o konačnom cilju, određuje se primjenom neke od mjera za računanje:

1. sličnosti (primjerice, kosinusna sličnost)
2. udaljenosti (primjerice, euklidska udaljenost).

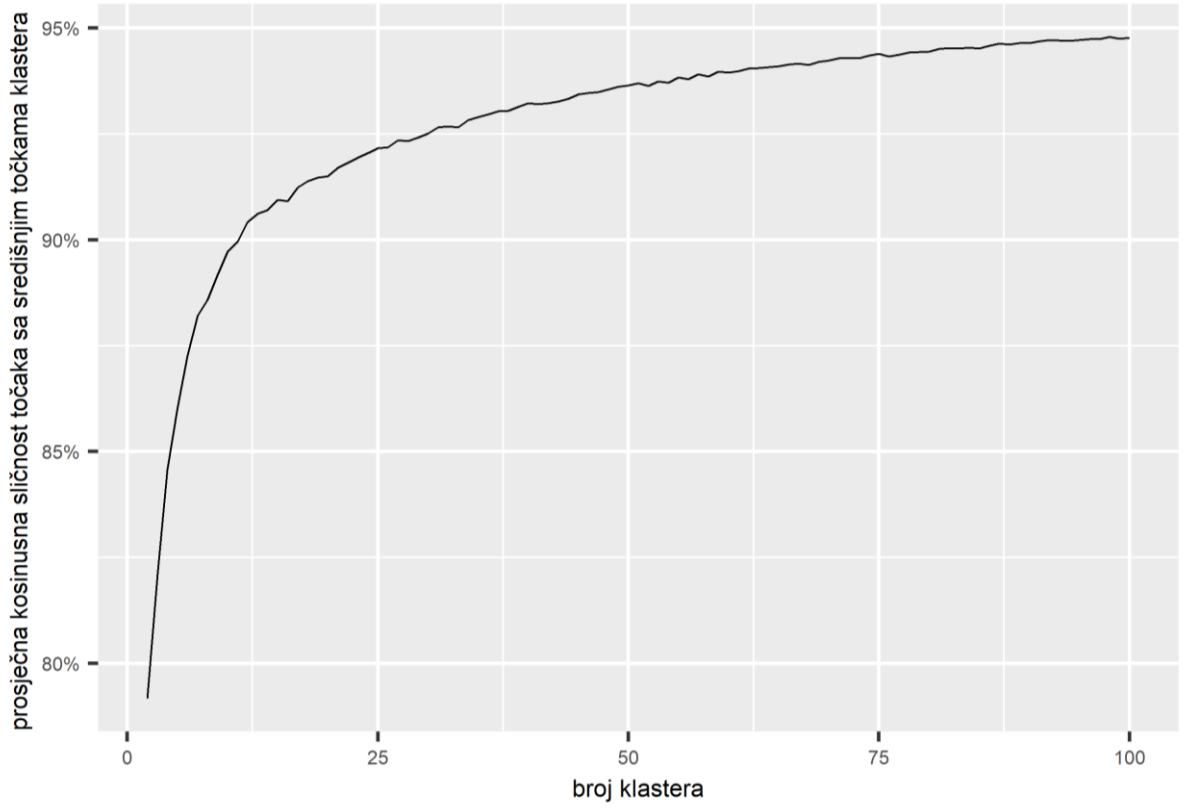
Klasterizacija, za razliku od klasifikacije, nema točno rješenje. Naime, klasifikacija je metoda nadziranog učenja kod koje postoji skup za učenje u kojem se za svaku instancu zna kojoj grupi pripada. Klasterizacija je metoda nenadziranog učenja kojoj je cilj otkriti grupe i onda u te grupe svrstati instance. Grupe dobivene klasterizacijom trebaju evaluirati domenski stručnjaci. Pogodnost rješenja ovisi o području primjene. Isto rješenje može biti različito ocijenjeno kada su uzimaju u obzir različite primjene.

U postupku klasterizacije svaka se instanca prikazuje kao točka u n-dimenzionalnom prostoru, pri čemu je *n* broj varijabli ulaznog podatkovnog skupa. Primjenom različitih algoritama klasterizacije nastoje se odrediti optimalne središnje točke klastera i onda instance promatranog podatkovnog skupa razdijeliti u klastere na način da se svaka instanca svrsta u onaj klaster čija mu je središnja točka najbliža. Uspješnost klasterizacije može se mjeriti i na način da se za svaku točku gleda njezina udaljenost u odnosu na ostale točke u tom klasteru te njezina udaljenost u odnosu na ostale točke podatkovnog skupa. Naravno, idealno bi bilo kada bi za svaku točku vrijedilo da joj je najudaljenija točka njezina klastera bliža u odnosu na najbližu točku nekog drugog klastera, ali to nije moguće uvijek postići.

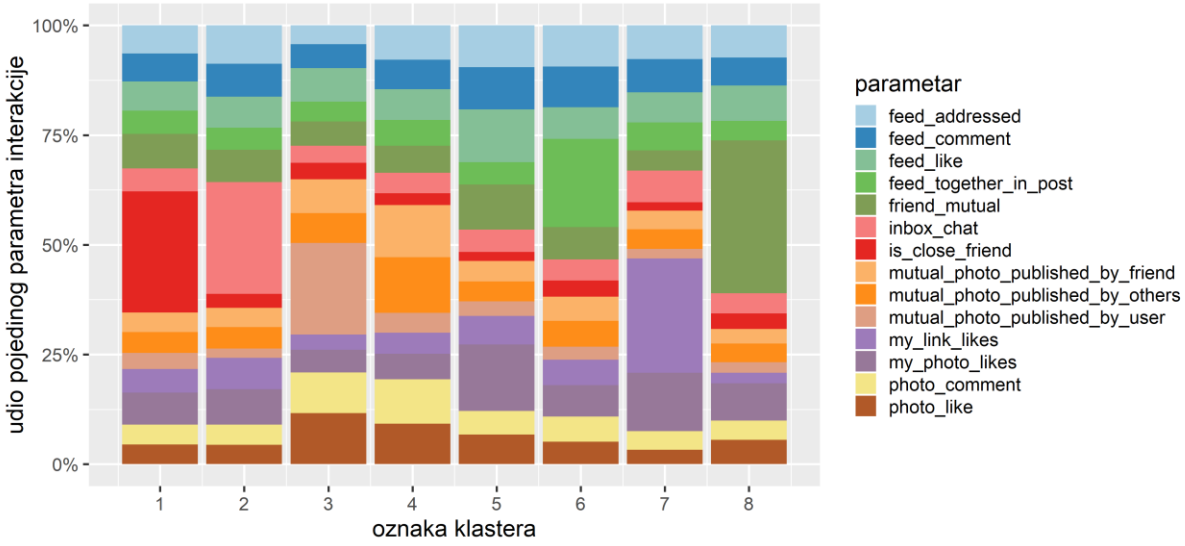
Primjenom algoritma klasteriranja *k-means* nastojalo se ispitanike podijeliti u grupe s obzirom na njihove interakcijske navike. Prethodno (u potpoglavlju IV.3) već je spomenuta tezu da se ispitanici međusobno razlikuju s obzirom na način kako koriste Facebook. Različiti korisnici skloni su korištenju različitih interakcijskih parametara različitim intenzitetom. Neki više razmjenjuju privatne poruke, neki češće tuđe objave označavaju sa *sviđa mi se*, a neki se često pojavljuju na fotografijama sa svojim prijateljima. Kako bismo razotkrili različite vrste korisnika Facebooka s obzirom na njihove interakcijske navike, algoritmom klasteriranja *k-means* korisnike Facebooka podijelili smo u klastere. Pri podjeli korisnika u klastere koristili smo kosinusnu sličnost. Kosinusna sličnost gleda relativni odnos između različitih interakcijskih parametara ili, formalnije, kut koji određena točka zauzima u n-dimenzionalnom prostoru. Primjerice, točke (3,3) i (10,10) u dvodimenzionalnom prostoru bit će, s obzirom na kosinusnu sličnost iste. Upravo to nam za ovu svrhu i odgovara. Naime, korisnike želimo klasterirati prema tome koriste li neki interakcijski parametar više u odnosu na ostale. U ovom

su kontekstu apsolutne vrijednosti nevažne. S obzirom na to, prilagođene su i vrijednosti u podatkovnom skupu. Apsolutne vrijednosti pretvorene su u relativne. Svaka apsolutna vrijednost podijeljena je s prosječnom vrijednošću za taj interakcijski parametar za sve ispitanike kako bi se dobio broj koji opisuje koliko pojedini korisnik koristi određeni interakcijski parametar u odnosu na ostale korisnike. Primjerice, ako prosječan korisnik ima ukupno 50 razmijenjenih poruka, a promatrani 100, tada će vrijednost 100 biti zamijenjena s 2 (100/50 = 2).

Algoritam klasteriranja *k-means* broj klastera ne daje kao rezultat svoga rada već ga očekuje kao ulaz. To znači da ga na neki način (iskustveno) treba odrediti analitičar. Isprobani su različiti brojevi klastera i ono što je zamijećeno jest da s povećanjem broja klastera instance svrstane u pojedini klaster postaju sve sličnije središnjoj točci klastera, ali ni za jedan broj klastera nije uočena jasna podjela iz koje bi se moglo nedvosmisleno zaključiti da u stvarnosti upravo toliko ima klastera (Slika V.3). S obzirom na to i s obzirom na to da je u prezentacijske svrhe bilo nužno odabrati neki broj klastera, odlučeno je da to bude 8. Slika V.4 vizualizira dobivene klastere.



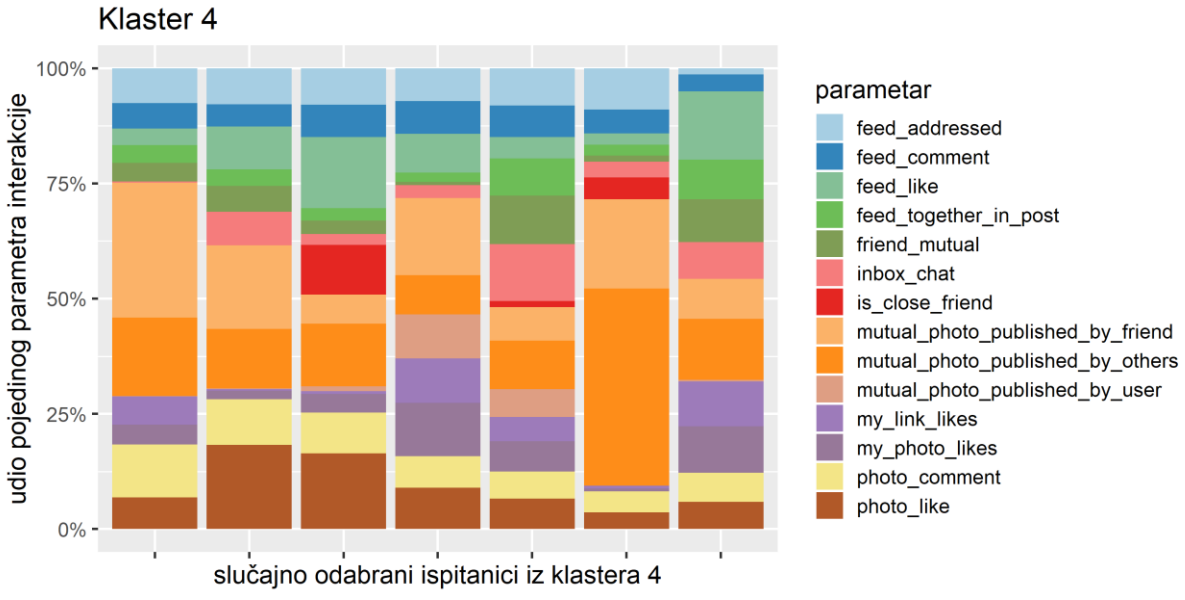
Slika V.3 Povećanje prosječne sličnosti točaka klastera sa središnjom točkom klastera s povećanjem ukupnog broja klastera



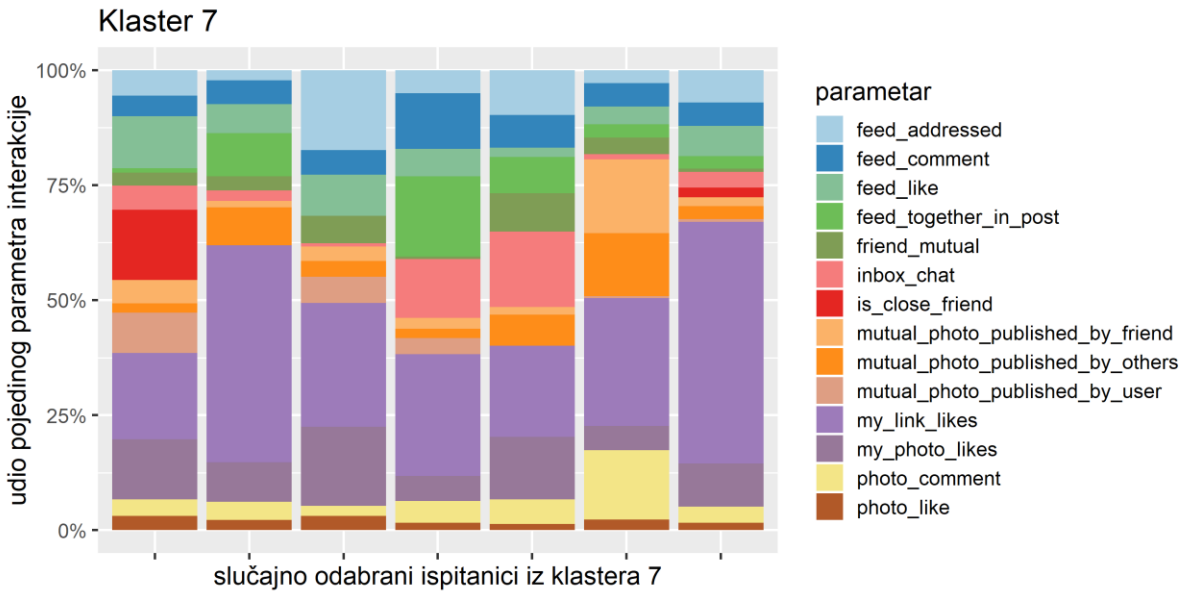
Slika V.4 Vizualizacija klastera ispitanika

Većina klastera ima neki interakcijski parametar koji je posebno izražen u odnosu na ostale. Primjerice, možemo uočiti da se u prvom klasteru nalaze korisnici koji su najskloniji označavanju prijatelja oznakom *bliski prijatelj*, a da se u sedmom klasteru nalaze korisnici koji imaju naglašenu interakciju prema svojim prijateljima, tj. puno označavaju oznakom *sviđa mi se* objave svojih prijatelja. Međutim, postoje i klasteri u kojima su svi interakcijski parametri (koliko-toliko) podjednako zastupljeni – primjerice, klasteri 4 i 5.

Kako bi se provjerilo koliko su korisnici dobro podijeljeni u klastere, tj. koliko su korisnici unutar istog klastera međusobno slični, za svaki klaster vizualizirano je 7 nasumično odabranih njegovih instanci. Slika V.5 primjer je takve vizualizacije za klaster 4, a Slika V.6 za klaster 7.



Slika V.5 Prikaz nasumično odabranih instanci klastera broj 4



Slika V.6 Prikaz nasumično odabranih instanci klastera broj 7

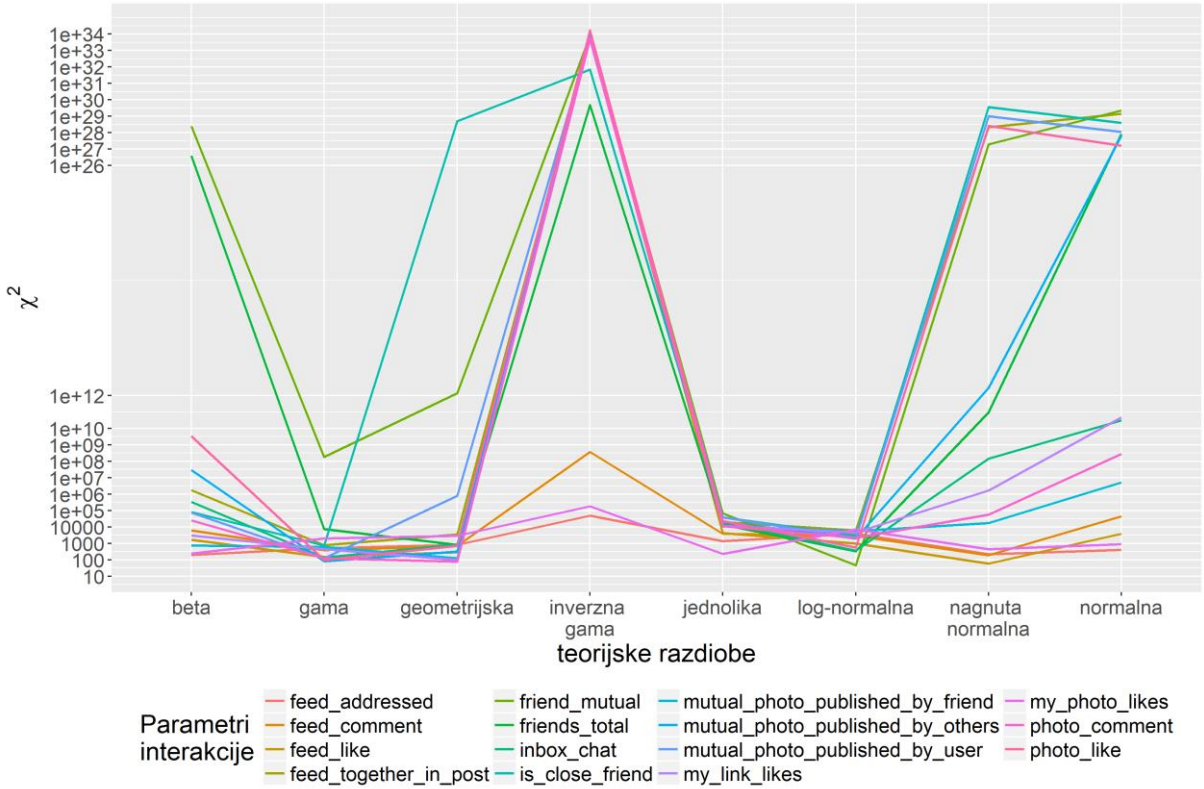
Podjela korisnika na klastere s obzirom na udio korištenja različitih interakcijskih parametara i empirijski je potvrdila pretpostavku i subjektivne izjave ispitanika da zaista različiti korisnici različito koriste Facebook. Pri izradi modela za računanje društvene udaljenosti ta će se spoznaja nastojati iskoristiti kako bi se model zasebno trenirao za različite klastere korisnika, a sve s ciljem da se dobije veća uspješnost modela u odnosu na slučaj kada se sve korisnike tretira jednako.

V.1.4 Razdiobe interakcijskih parametara

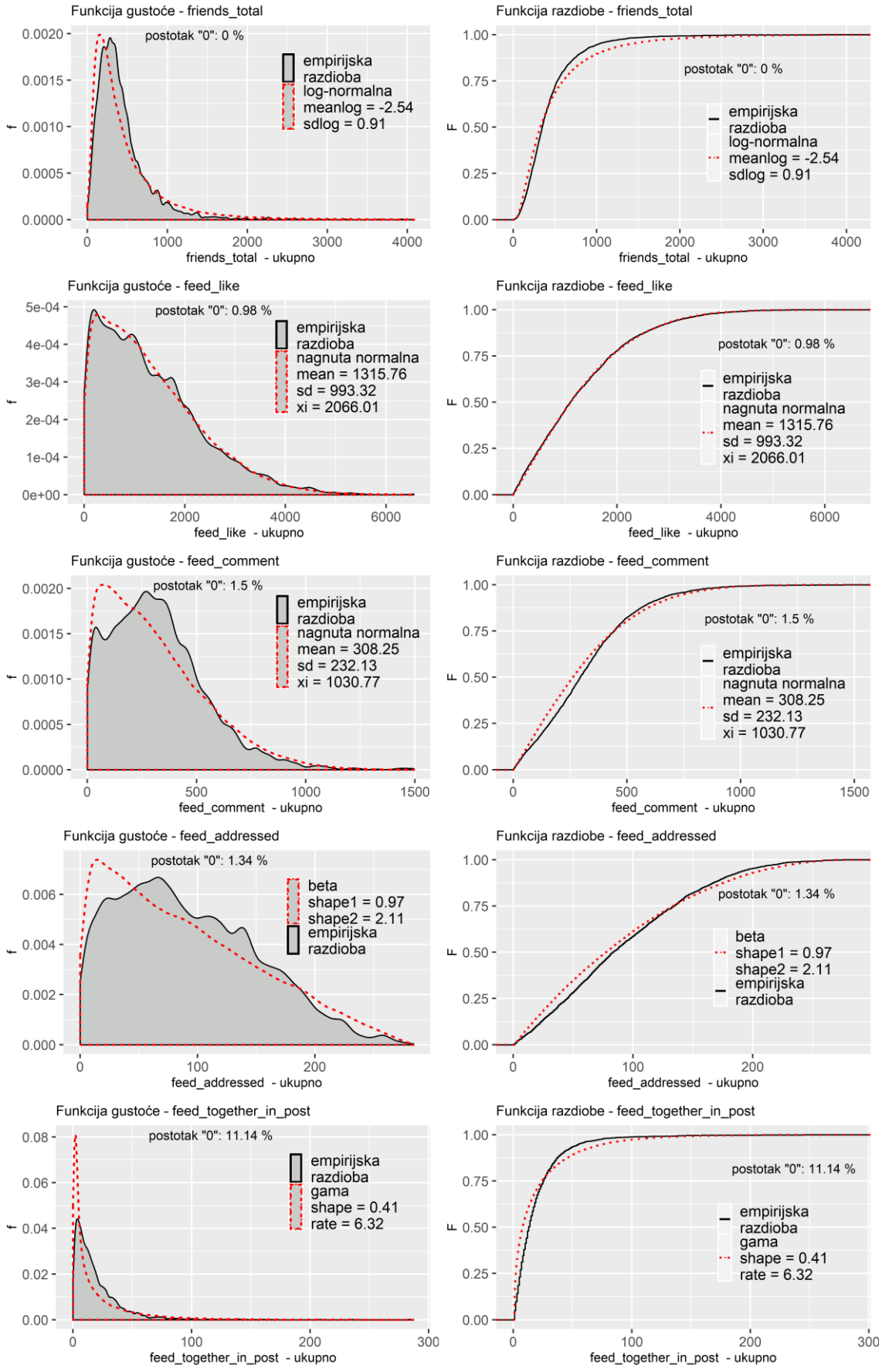
Kako bi se dobio uvid u intenzitet interakcije na Facebooku na razini ispitanika, odlučili smo pronaći empirijske razdiobe za svaki parametar interakcije. Na temelju tih empirijskih razdioba, nastojali smo pronaći teorijske razdiobe koje su najsličnije onim empirijskim. U razmatranje su uzete sljedeće teorijske razdiobe: *beta*, *gama*, *inverzna gama*, *geometrijska*, *jednolika*, *normalna*, *log-normalna* i *nagnuta normalna*. Primjenom metode MLE (engl. *maximum likelihood estimation*) pronađeni su optimalni parametri za svaku od razmatranih teorijskih razdioba za svaki parametar interakcije, tj. oni parametri za koje je razmatrana teorijska razdioba najsličnija empirijskoj razdiobi. Za teorijske razdiobe koje su definirane samo na intervalu [0,1], podatci su prvo normalizirani. Za svaku od razmatranih teorijskih razdioba s njezinim optimalnim parametrima za pojedini parametar interakcije, proveden je test χ^2 s ciljem utvrđivanja može li se odbaciti nulta hipoteza da se empirijski podatci ne ravnaju po predmetnoj teorijskoj razdiobi. Test $\chi 2$ uz 95-postotni interval povjerenja za sve je empirijske razdiobe otklonio mogućnost da se one ravnaju prema razmatranim teorijskim razdiobama.

Međutim, dobivene vrijednosti *χ2* iskoristit će se kako bi se pronašle one teorijske razdiobe koje najbolje aproksimiraju empirijske razdiobe pojedinog parametra interakcije. Slika V.7 prikazuje vrijednosti *χ2* za različite parametre interakcije za različite teorijske razdiobe. Slika V.8, Slika V.9 i Slika V.10 prikazuju funkcije gustoće i razdiobe za empirijske razdiobe svakog parametra interakcije te pridružene im teorijske razdiobe koje su prema rezultatu testa χ2 njihove najbolje aproksimacije u skupu razmatranih teorijskih razdioba. Uz svaku teorijsku razdiobu na slikama su ispisane i vrijednosti njihovih optimalnih parametara.

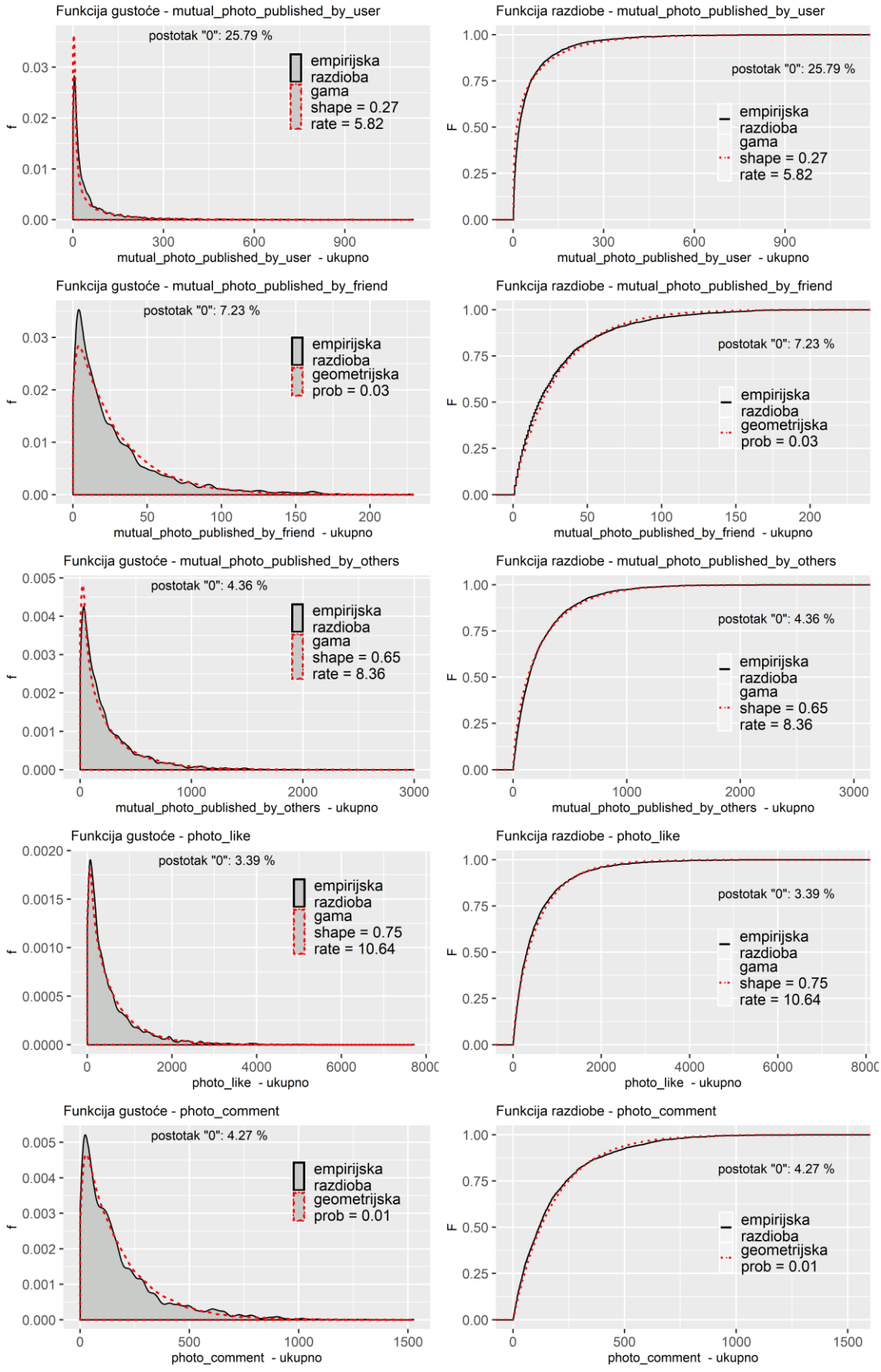
Pri određivanju aproksimacijskih teorijskih razdioba, iz empirijskog skupa podataka izostavljene su vrijednosti „0“. Uz svaku teorijsku razdiobu naveden je udio vrijednosti „0“ (koje njome nisu obuhvaćene).



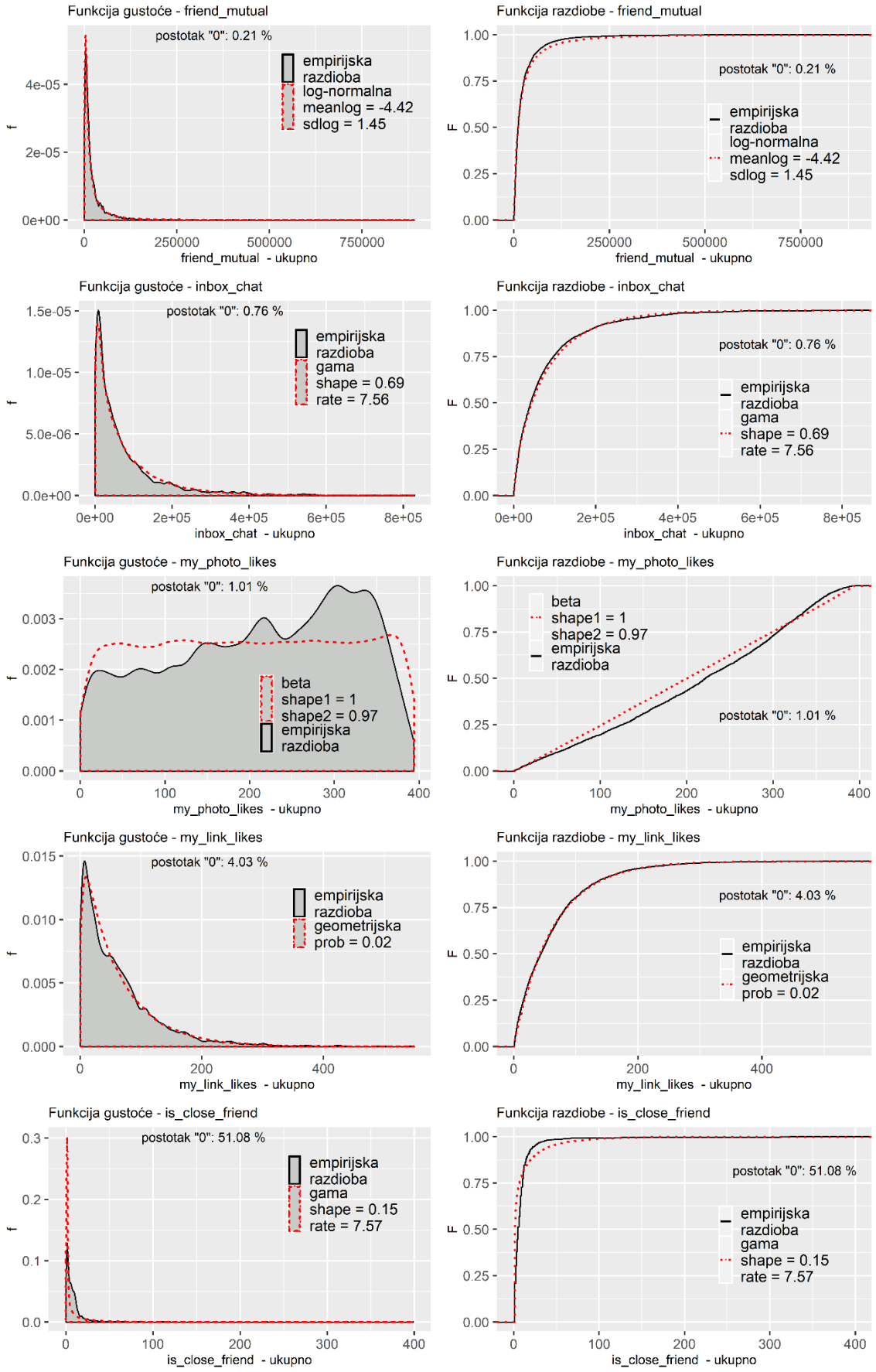
Slika V.7 Rezultati testa χ2 po parametrima interakcije za različite teorijske razdiobe s njihovim optimalnim parametrima



Slika V.8 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni ispitanika – prvi dio



Slika V.9 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni ispitanika – drugi dio



Slika V.10 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni ispitanika – treći dio

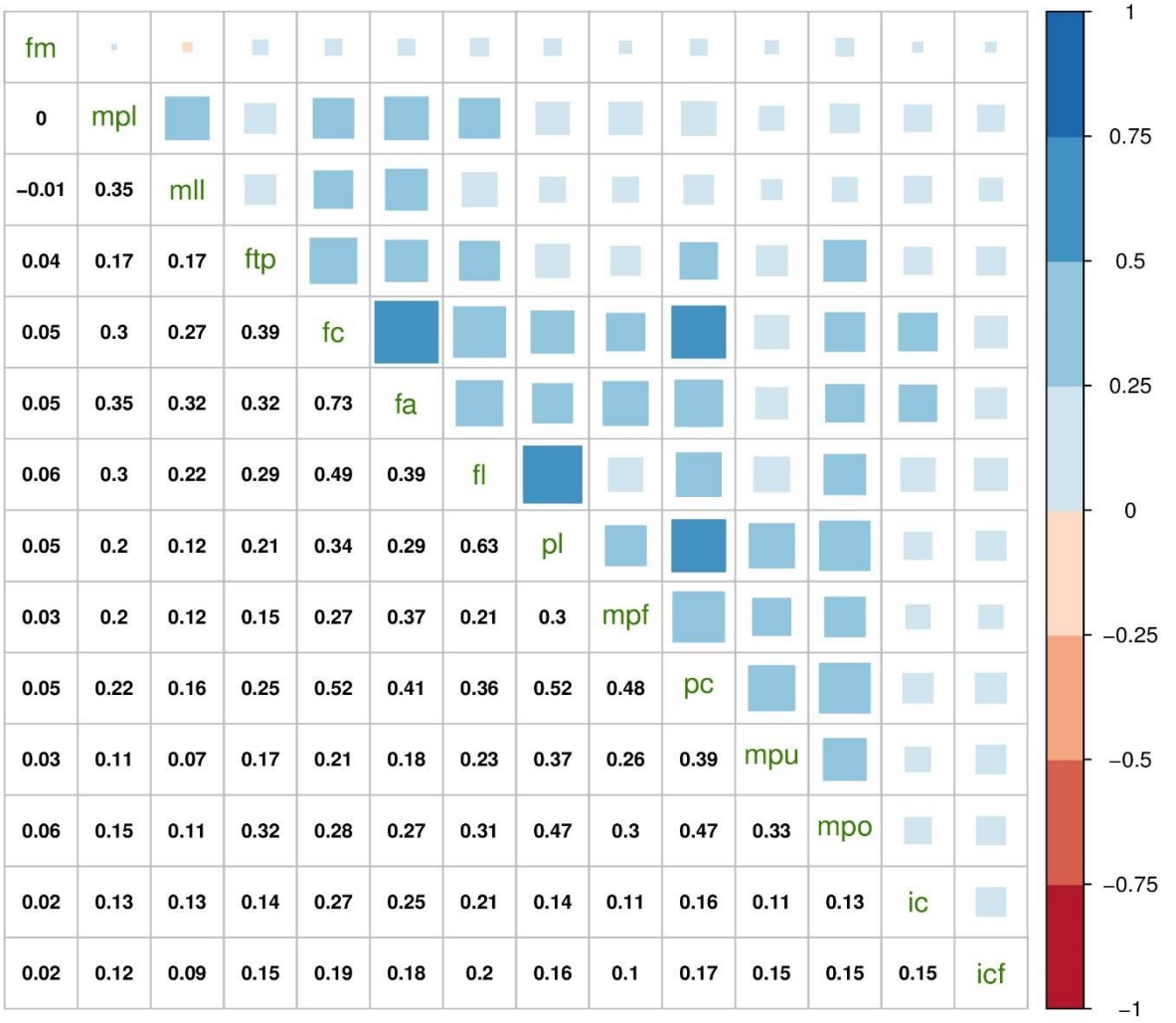
V.2 EKSPLORATORNA ANALIZA NA RAZINI PARA PRIJATELJA

Eksploratorna analiza na razini para prijatelja provedena je nad podatkovnim skupom koji u sebi sadrži interakcijske parametre koje prikazuje Tablica V.1, izuzev parametra *ukupan broj prijatelja korisnika A* koji ima smisla isključivo na razini ispitanika. U nastavku će po dijelovima biti prezentirani rezultati provedene analize

V.2.1 Koreliranost parova atributa

Slično kao i kod analize podatkovnog skupa na razini ispitanika, i u analizi podatkovnog skupa sa sumiranim podatcima na razini para prijatelja provjerili smo u kojoj su mjeri parametri interakcije međusobno korelirani. U tu smo svrhu izračunali Pearsonov korelacijski koeficijent za sve parove parametara interakcije.

Slika V.11 prikazuje rezultate numerički i grafički. Notacija je ista kao i kod prethodnog skupa pa ne će ovaj puta biti ponovo pojašnjavana.



Slika V.11 Pearsonovi koeficijenti korelacije među interakcijskim parametrima u podatkovnom skupu – podatci sumirani na razini para prijatelja

V.2.1.1 Rasprava

Slika V.11 prikazuje u kojoj su mjeri interakcijski parametri međusobno korelirani. Pokazuje se da su parametri *feed_comment* i *feed_addressed* najjače korelirani. Zanimljivo je za primijetiti da korisnici koji puno komentiraju objave svojih prijatelja često i objavljuju na zidu prijatelja. Analiza je pokazala i visoku koreliranost između parametara *photo_like* i *feed_like*, što je i posve očekivano jer se u oba slučaja radi o ostavljanju oznake *sviđa mi se* samo na različite vrste objava. Visoka koreliranost interakcijskih parametara *photo_comment* i *photo_comment* također je na tragu toga.

Niska koreliranost između različitih parametara koji označavaju zajedničko pojavljivanje na slikama, a razlikuju se tek po tome tko je objavio fotografiju, prilično je iznenađujuća. Zanimljivo je i primijetiti da broj zajedničkih prijatelja nije ni na koji način koreliran s bilo kojim drugim parametrom interakcije. To možda jest na prvu čudno, ali i sami ispitanici su u upitniku opisanom u potpoglavlju IV.3 u velikom broju rekli da broj zajedničkih prijatelja ne govori puno o njihovu odnosu s promatranom osobom.

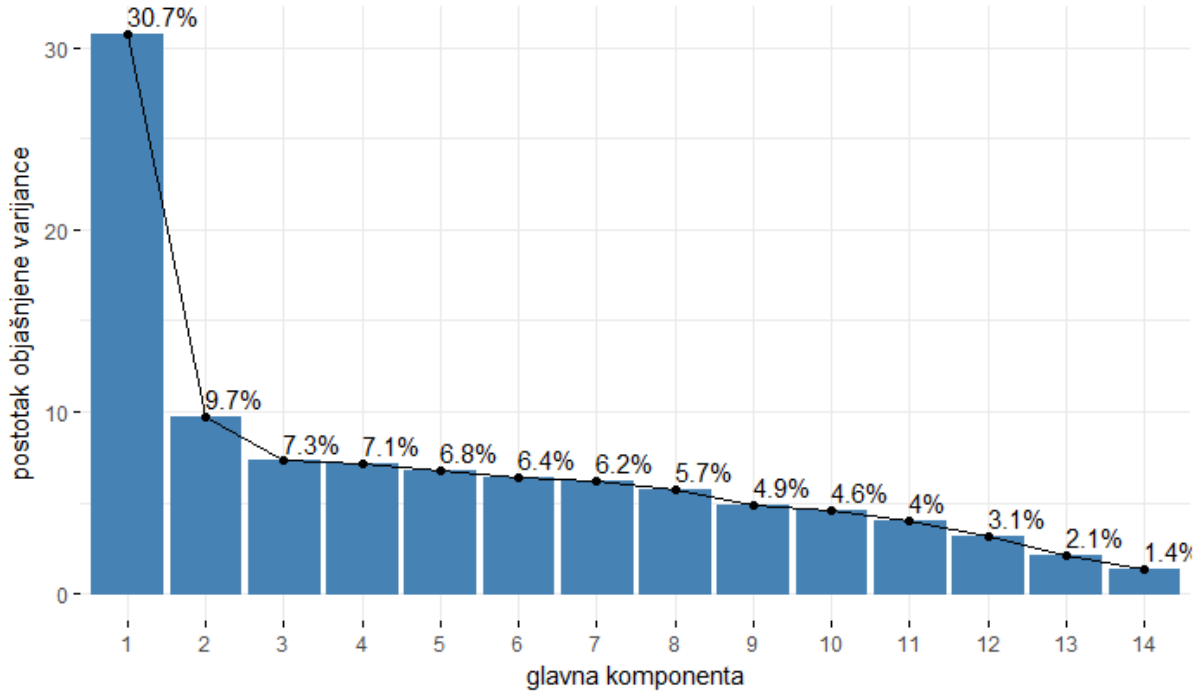
V.2.2 Analiza glavnih komponenata

Analiza glavnih komponenata (engl. *principal components analysis – PCA*) često se koristi u postupku izrade prediktivnih modela za redukciju dimenzija (broja varijabli) ulaznog podatkovnog skupa. Redukcija dimenzija može biti korisna za vizualizaciju. Manji broj dimenzija u pravilu skraćuje i vrijeme računanja kod prediktivnih modela. Također, komponente dobivene metodom PCA nisu međusobno korelirane što može biti važno za neke algoritme.

Pri analizi glavnih komponenti željeni je rezultat da prvih nekoliko komponenti objasni gotovo cijelu (>95%) varijancu podatkovnog skupa. U tom se slučaju te komponente koriste, a ostale se odbacuju čime se postiže redukcije dimenzija.

Analiza glavnih komponenata provedena je nad podatkovnim skupom s podatcima o interakciji sumiranima na razini para prijatelja, tj. nad podatcima koji imaju svojstvo prediktora u modelu za računanje društvene udaljenosti. Dobiveno je 14 komponenti. Slika V.12 prikazuje postotak objašnjenosti varijance po komponenti. Vidljivo je da prva komponenta objašnjava najviše varijance, ali tek malo više od 30% ukupne varijance. Ostale komponente objašnjavaju manje od 10% varijance. Kada se gleda kumulativna objašnjenost varijance, više od 95% postiže se tek uzimanjem u obzir prvih 12 (od 14) komponenti. To pokazuje da se primjenom metode PCA za promatrani podatkovni skup ne mogu reducirati dimenzije bez značajnijeg

gubitka varijance te će se posljedično u modelu kako ulazne varijable koristiti svi atributi podatkovnog skupa.



Slika V.12 Postotak objašnjenosti varijance po glavnim komponentama

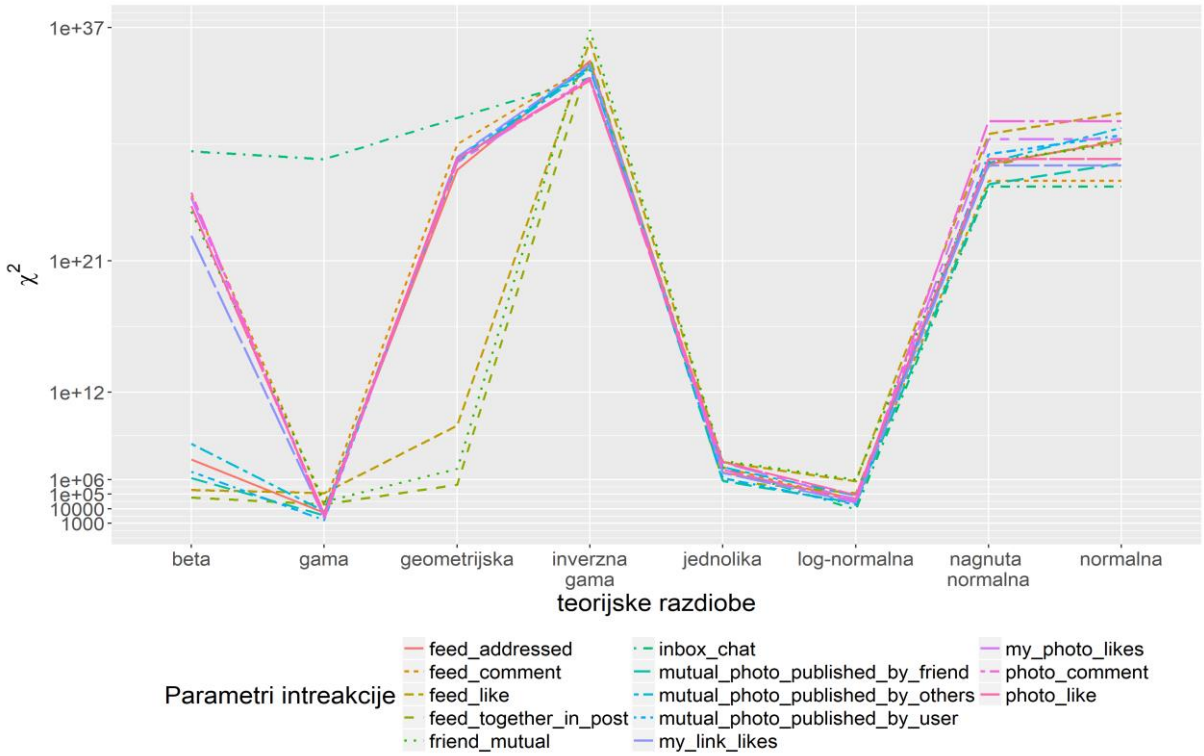
V.2.3 Razdiobe interakcijskih parametara

Kako bi se dobio uvid u to koliko i kako prijatelji na Facebooku međusobno interaktiraju, analizirane su razdiobe sumarnih vrijednosti interakcijskih parametara na razini para prijatelja. Na samom početku analize utvrđeno je da vrlo velik broj zapisa o interakciji prijatelja na Facebooku sadrži vrijednosti „0“, tj. da promatrani parovi prijatelja nikada nisu ostvarili nikakvu interakciju promatranim parametrom interakcije. S obzirom na to, prije analize empirijskih razdioba, iz svakog su skupa izbačene vrijednosti „0“ te je analizirana samo razdioba preostalih vrijednosti. Konačan rezultat za svaki interakcijski parametar prikazan je kao postotak vrijednosti „0“ te razdioba preostalih vrijednosti.

Na temelju poznatih empirijskih razdioba, nastojalo se pronaći najsličnije teorijske razdiobe. Postupak traženja odgovarajućih teorijskih razdioba isti je kao u slučaju analize podatkovnog skupa s vrijednostima sumiranim na razini ispitanika pa ne će ovdje ponovo biti objašnjavan. Slika V.13 prikazuje vrijednosti $\chi 2$ za različite parametre interakcije za različite teorijske razdiobe.

Slika V.14, Slika V.15 i Slika V.16 prikazuju funkcije gustoće i razdiobe za empirijske razdiobe svakog parametra interakcije te pridružene im teorijske razdiobe koje su prema

rezultatu testa $\chi 2$ njihove najbolje aproksimacije u skupu razmatranih teorijskih razdioba. Uz svaku teorijsku razdiobu na slikama su ispisane i vrijednosti njihovih optimalnih parametara te udio vrijednosti „0“ u izvornom skupu.

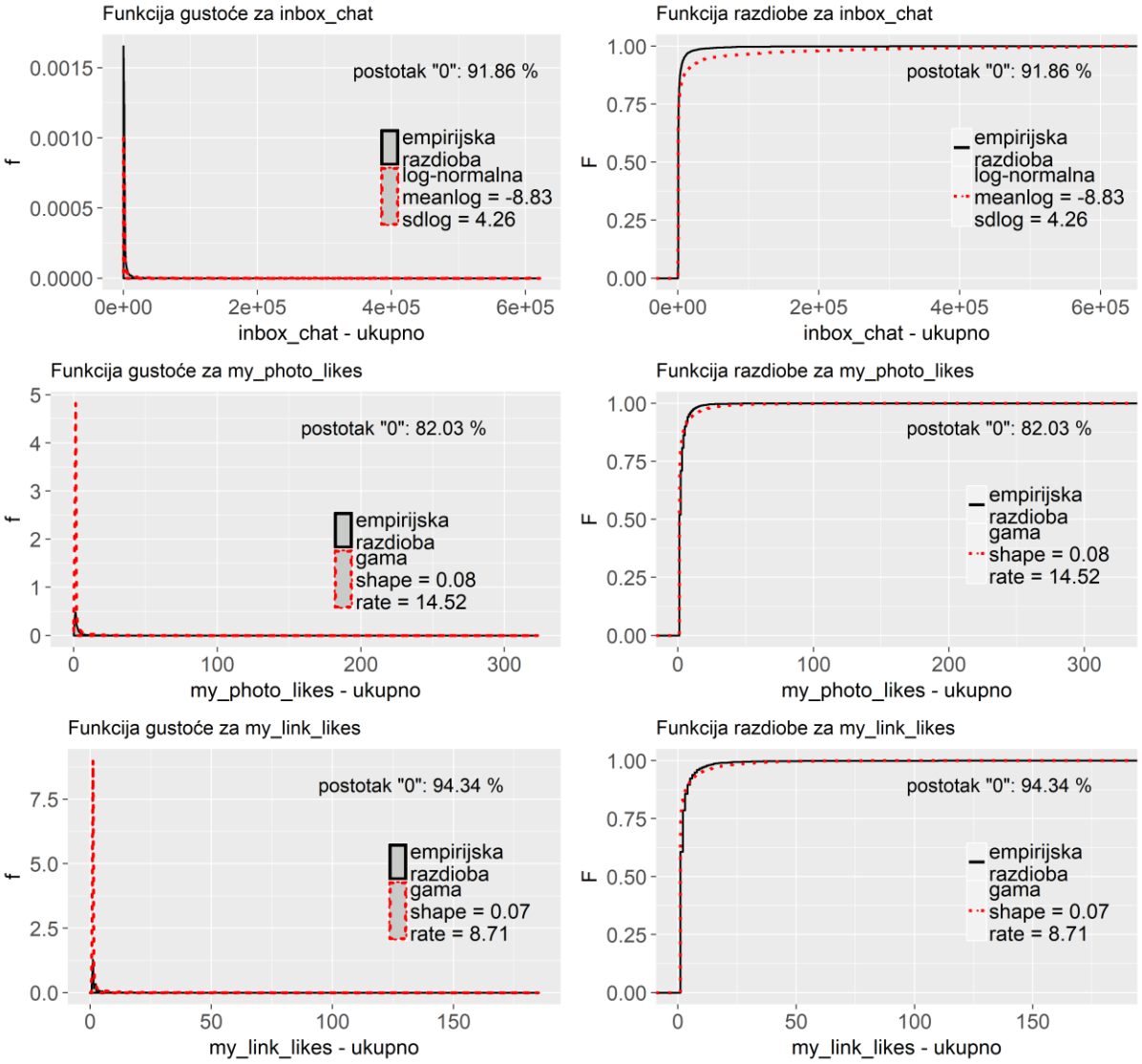


Slika V.13 Rezultati testa χ2 po parametrima interakcije za različite teorijske razdiobe s njihovim optimalnim parametrima

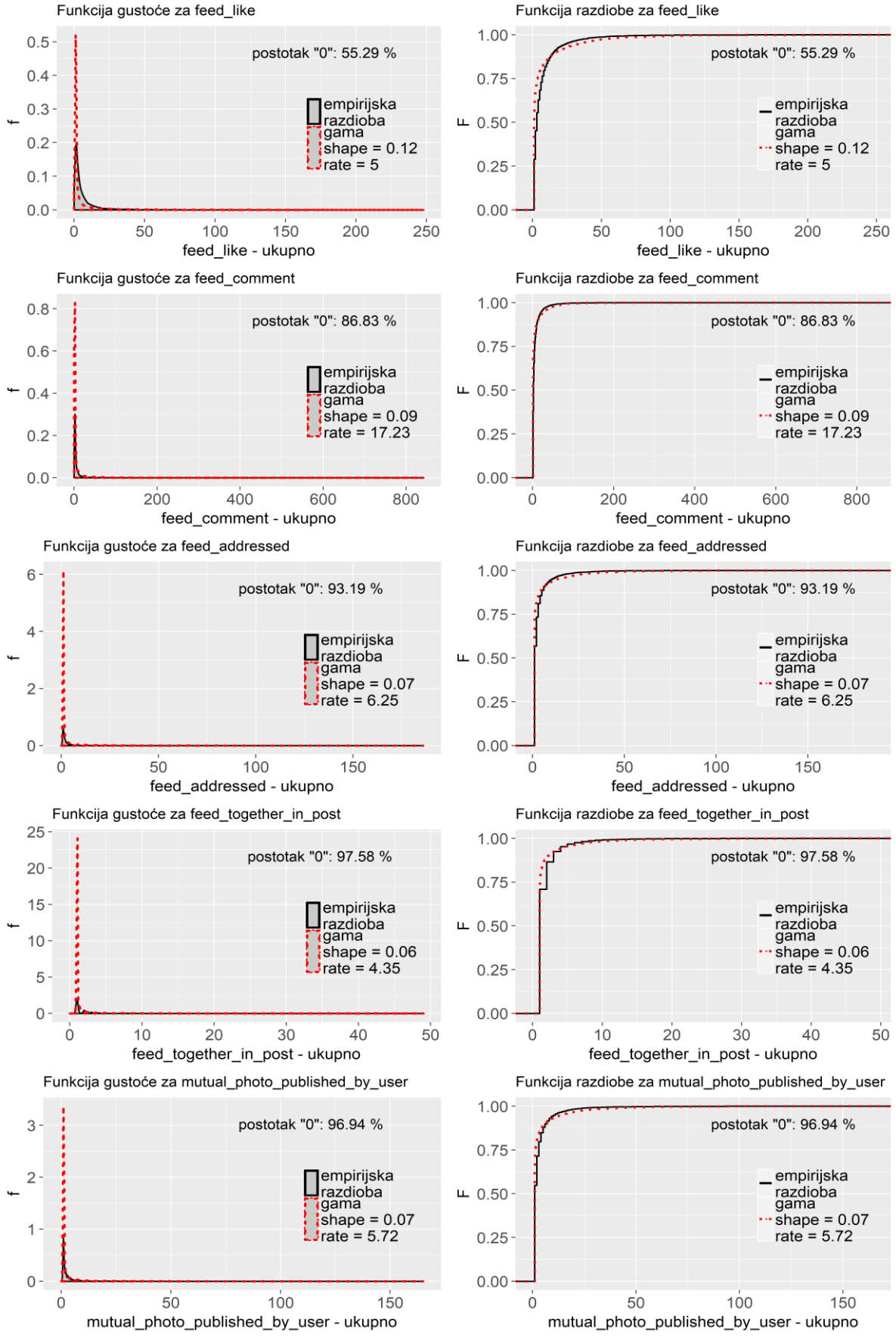
V.2.3.1 Rasprava

Pri analizi razdioba interakcijskih parametara, najupečatljivija je činjenica da gotovo svi interakcijski parametri u sebi imaju preko 90% vrijednosti „0“, što znači da tek rijetki parovi prijatelja na Facebooku imaju interakciju promatranim parametrom interakcije. Čak 42,4% parova prijatelja na Facebooku nema baš nikakvu interakciju za cijelo vrijeme trajanja njihova fejsbukovskog prijateljstva, a 2,14%, osim što nemaju nikakvu interakciju, nemaju niti jednog zajedničkog prijatelja. Iako je dobro poznata činjenica da na Facebooku većinu „prijatelja“ čine tek poznanici, vrlo je iznenađujuće da korisnici Facebooka gotovo sa svakim drugim fejsbukovskim *prijateljem*, otkako su sklopili prijateljstvo, nemaju baš nikakvu interakciju. To, naravno, ide u prilog u Dunbarovoj teoriji [2], [56] o veličini neokorteksa u mozgu kao ograničavaču broja prijatelja na okvirnoj brojci od oko 150 prijatelja, ali i motivira izradu modela za računanje društvene udaljenosti, a koji upravo omogućava razlikovanje stvarnih prijatelja na Facebooku od poznanika.

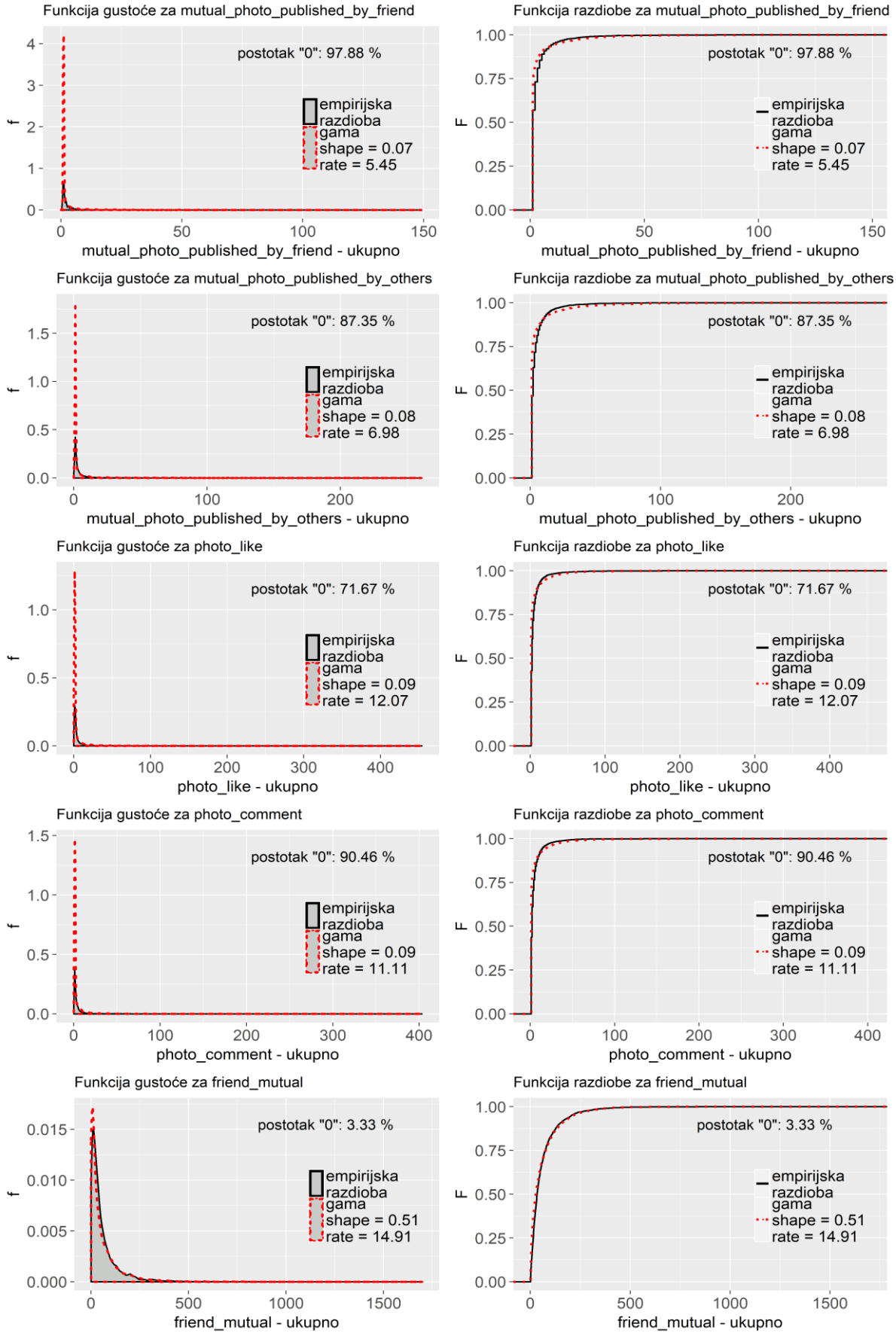
Što se tiče teorijskih razdioba koje najbolje aproksimiraju one empirijske, najbolja teorijska razdioba za gotovo sve interakcijske parametre jest gama-razdioba, dok je broj razmijenjenih privatnih poruka iznimka s log-normalnom razdiobom.



Slika V.14 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni para prijatelja – prvi dio



Slika V.15 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni para prijatelja – drugi dio



Slika V.16 Prikaz funkcija gustoće i razdiobe parametara interakcije sumiranih na razni para prijatelja – treći dio

VI. DRUŠTVENA UDALJENOST

Portal hrvatskog strukovnog nazivlja Struna, kojim upravlja Institut za hrvatski jezik i jezikoslovlje, definira *društvenu udaljenost* kao „mjeru spremnosti pojedinca ili skupine za međuljudsku komunikaciju i međudjelovanje s drugim pojedincima ili skupinama“ [57]. Dodaje da „na mjernoj ljestvici društvene udaljenosti koja se rabi u istraživanjima brak kao najintimniji odnos među ljudima označuje jedan kraj (potpuno prihvaćanje zajedničkoga života), a na drugome je kraju odsustvo svakoga odnosa ili potpuno odbijanje doticaja“ [57].

Sociolozi i socijalni psiholozi trude se objasniti kako drugi ljudi utječu na naše osjećaje, razmišljanja i ponašanja [20], [22]. Poznavanje povezanosti među ljudima ključno je pri tome. Desetljećima je veliki uspornik takvih istraživanja bio je nedostatak relevantnog podatkovnog skupa. Pojavom sustava za društveno umrežavanje i digitalnim (strukturiranim i računalno čitljivim) pohranjivanjem podataka o interakciji milijardi ljudi, taj je problem uvelike ublažen. Pokazalo se da interakcija ljudi na sustavima za društveno umrežavanje u velikoj mjeri opisuje njihove odnose u stvarnom životu [3], [4], [20] pa se upravo podatci o povezanosti ljudi preuzeti sa sustava za društveno umrežavanje mogu iskoristiti kao temelj za određivanje intenziteta odnosa među ljudima.

Određivanje intenziteta odnosa među ljudima na temelju njihove interakcije na sustavima za društveno umrežavanje tema je brojnih istraživanja [3]–[19]. Cilj nekih istraživanja jest iz skupa *mrežnih prijatelja* promatranog korisnika izlučiti one njemu bliske [14], [16], [18], druga istraživanja nastoje *mrežne prijatelje* promatranog korisnika klasificirati u više razreda [3], [17], dok treća kategorija istraživanja nastoji numerički opisati intenzitet odnosa između promatranog korisnika i svih njegovih mrežnih prijatelja [4], [6], [9]–[15], [19].

Poznavanje intenziteta odnosa među ljudima može imati različite primjene. Jedna od primjena jest u preporučiteljskim sustavima [34]–[37]. Preporučiteljski sustavi u osnovi se mogu podijeliti na one *zasnovane na sadržaju* i one *zasnovane na suradnji*. Preporučiteljski sustavi *zasnovani na sadržaju* rade na način da analiziraju prethodno iskazan interes korisnika kroz prethodne konzumacije (ranije kupovine i slično) i preporučuje im se sadržaj sličan prethodno konzumiranom. Takvi sustavi rijetko korisnicima nude sadržaj koji ne odgovara njihovim interesima, ali nerijetko se javlja problem monotonije u listama preporuke, tj. korisniku se stalno preporučuje sadržaj sličan onome koji je do tada konzumirao. Međutim, korisnik često ima i brojne druge interese uz one koje je iskazao prethodno konzumiranim sadržajima. Najjednostavniji način otkivanja tih neotkrivenih interesa jest kroz nasumično nuđenje korisniku vrsta sadržaja kakve ranije nije konzumirao. Na taj je način velika

vjerojatnost da će korisnik prije ili kasnije dobiti neki novi njemu zanimljiv sadržaj, ali velik dio tako ponuđenog sadržaja biti će promašen. Pametniji način traženja korisnikovih još neiskazanih interesa jest analizirati kakav sadržaj konzumiraju prijatelji predmetnog korisnika i korisniku, osim sadržaja koji je sličan njegovim ranije iskazanim interesima, nuditi i sadržaj sličan onome koji su prethodno konzumirali njegovi prijatelji. Takav se pristup oslanja na načelo homofilije [58] koje kaže da bliski ljudi često imaju slična razmišljanja i interese, tj. da se ljudi često druže s ljudima sličnima sebi. Preporučiteljski sustavi koji se pri preporuci sadržaja oslanjaju na prethodno iskazane interese korisnikovih prijatelja nazivaju se preporučiteljskim sustavima *zasnovanim na suradnji*. Kako bi preporučiteljski sustavi zasnovani na suradnji mogli raditi, nužno je znati tko su prijatelji ego-korisnika. Na sustavima za društveno umrežavanje lista prijatelja promatranog korisnika nerijetko je javno dostupna[17]. Međutim, korisnici su skloni na društvenim mrežama povezivati se s velikim brojem ljudi koji tek u manjem dijelu predstavljaju njihove stvarne prijatelje. Primjerice, u istraživanju *NajFrend* provedenom na Facebooku (opisanom u potpoglavlju IV.2) prosječan broj prijatelja jest 429. Ako uzmemo u obzir da prosječna osoba prema *Dunbarovu broju* može imate tek okvirno 150 prijatelja [2], posve je jasno da *prijatelje* na Facebooku ne čine samo pravi prijatelji, već i poznanici pa čak i ljudi koje ego-korisnik možda nikada uživo nije vidio niti imao ikakvu formalnu ili neformalnu interakciju s njima, izuzev uspostave prijateljstva na Facebooku [11]. U tom su kontekstu modeli koji na temelju interakcije korisnika na sustavima za društveno umrežavanje (posredno) mogu odrediti intenzitet odnosa između korisnika vrlo korisni[18]. Znanstveni izazov jest kako iz velikog skupa *mrežnih prijatelja* otkriti one istinske prijatelje, tj. ljude za koje je velika vjerojatnost da su u interesima i razmišljanjima slični promatranom korisniku.

Posredno određivanje intenziteta odnosa među ljudima može se koristiti i u korporacijama za analiziranje odnosa među zaposlenicima s ciljem uočavanja *strukturnih rupa*[19] i sličnih anomalija te unaprjeđenje poslovanja kroz unaprjeđenje odnosa među ljudima [27], [28], [31], [32].

[17] Trend javne dostupnosti liste prijatelja takav je da se ta dostupnost, uslijed povećanog fokusa na zaštitu osobnih podataka i privatnosti, smanjuje, ali i dalje se taj popis nerijetko može pronaći javno objavljen.
[18] Naravno, preduvjet za rad takvih algoritama jest privola korisnika za pristup njihovim osobnim podatcima. Sustavi za društveno umrežavanje osobne podatke svojih korisnika ne smiju ustupiti trećoj strani bez prethodno dobivene suglasnosti korisnika.
[19] U organizacijama koja posluju na širokom području često se događa da se zaposlenici međusobno ne poznaju. Ti ljudi neupitno imaju zajedničke interese, ali još nisu uspjeli uspostaviti kontakt, tj. među njima postoje potencijalne veze. Zbog nepostojanja stvarnih veza među njima, dolazi do tzv. *strukturalnih rupa* (engl. *structural holes*) koje sprječavaju kvalitetan protok znanja kroz organizaciju.

Moguća primjena modela za posredno određivanje intenziteta odnosa među ljudima jest i u sprječavanju odlazaka (važnih) korisnika (engl. *churn prevention*) [24]–[26], što se posebno primjenjuje u telekomunikacijskom sektoru. Slično kao i kod preporučiteljskih sustava, sprječavanju odlazaka korisnika može se pristupiti izolirano na razini pojedinog korisnika ili uzimajući u obzir njegov širi društveni kontekst. Kroz gledanje šireg društvenog konteksta promatranog korisnika, tj. njegove mreže prijatelja, mogu se uočiti neki važni korisnici ili korisnici s visokim rizikom od odlaska koji po individualnim parametrima to ne bi bili. Primjerice, korisnici koji možda malo troše, ali su snažno povezani sa značajnim korisnicima (čvorovi s velikim utjecajem) ili korisnici koji svojim ponašanjem ne odaju dojam da bi mogli otići, ali nekoliko njihovih dobrih prijatelja u prethodnom je razdoblju otišlo. Kod odlaska korisnika s velikim utjecajem kod konkurentskog ponuđača postoji rizik da će za sobom povući i neke druge značajne korisnike pa je bitno zadržati i njih iako sami za sebe nekad nisu značajni. Naravno, kako bi se mogli analizirati odnosi među korisnicima nužno je prethodno generirati njihov društveni graf. Kako bi to bilo moguće, nužno je imati modele koji na temelju dostupne interakcije korisnika (one kojoj korisnici dozvole pristup) mogu odrediti intenzitet odnosa.

Mogućnost posrednog određivanja intenziteta odnosa među ljudima može se primijeniti i u obrazovnom sektoru. Generiranje društvenih grafova studenata može olakšati razumijevanje ponašanja studenata [38]. Primjerice, na temelju izgrađenog društvenog grafa nastavnik može odrediti najutjecajnije studente u grupi, tj. središnje čvorove analiziranog društvenog grafa. Ako takvi studenti usvoje gradivo, na taj će način usvajanje gradiva potencijalno biti olakšano i ostalim studentima.

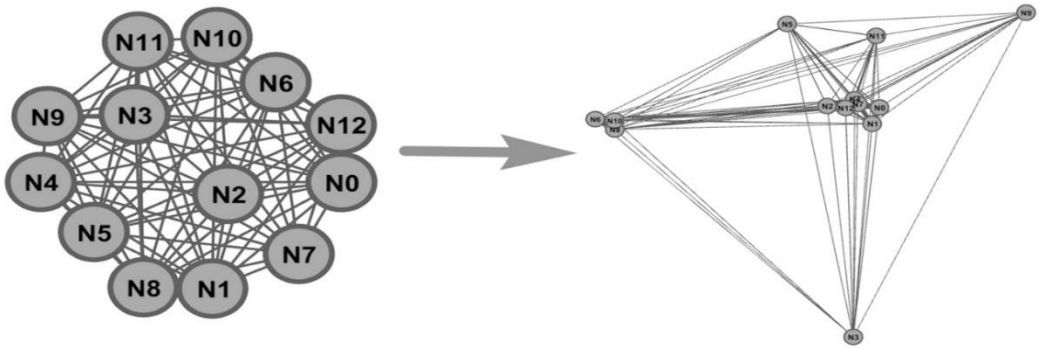
U dosadašnjim istraživanjima intenzitet povezanosti pretežito se izražavao primjenom koncepta *težine* ili *snage* veze, tj. primjenom koncepta sličnosti što znači da veći broj povlači veći intenzitet povezanosti. Za takav način bilježenja intenziteta odnosa nije moguće koristiti mnoge uobičajene algoritme matematičke teorije grafova u svrhu analize društvenih grafova (ili mreža) – primjerice, Dijkstrin algoritam za pronalaženje najkraćeg puta. U ovom se radu uvodi mogućnost izračuna i bilježenja intenziteta odnosa među ljudima na temelju koncepta udaljenosti. Koncept udaljenosti podrazumijeva bilježenje intenziteta odnosa između čvorova na način da manji broj podrazumijeva veći intenzitet odnosa, a veći podrazumijeva manji intenzitet, tj. da su bliskiji oni korisnici koji su manje udaljeni. Upravo taj način bilježenja intenziteta odnosa među čvorovima vrlo je pogodan za daljnje primjene. Primjerice, udaljenost između čvorova proporcionalna je vremenu propagacije informacije. Također, istraživanja [59], [60] pokazuju da ljudi u pravilu imaju više zajedničkih interesa s ljudima koji su im geografski bliže. Naime, geografska udaljenost jedno je od najznačajnijih ograničenja pri izgradnji

društvenih mreža [61]. Posljedično se pokazuje da ljudi imaju intenzivnije odnose s drugim ljudima koji su od njih geografski manje udaljeni pa se prikazivanje intenziteta odnosa kroz koncept udaljenosti može iskoristiti i za određivanje (približne) geografske lokacije korisnika [59]. Osim toga, prikaz intenziteta odnosa među čvorovima primjenom koncepta udaljenosti vrlo je pogodan za vizualizaciju društvenog grafa. Koncept udaljenosti omogućuje i korištenje standardnih algoritama za određivanje najkraćeg puta u društvenom grafu, a primjenom kojih se može razviti sustav za preporučivanje novih prijatelja koji može preporučivati i prijatelje u trećem ili četvrtom koraku ako je udaljenost između ego-korisnika i njih mala.

Slika VI.1[20] ilustrira transformaciju binarnog društvenog grafa u obogaćeni društveni graf primjenom koncepta udaljenosti. Analizirao se odnos prijateljstva. Kroz vizualizaciju se može vidjeti kako je siromašan eksplicitni binarni graf prijateljstva na Facebooku, na temelju analize interakcije korisnika, pretvoren u obogaćeni društveni graf u kojem su korisnici kod kojih je intenzitet prijateljstva izraženiji međusobno bliže u grafu.

Zbog svega pobrojanog, ovom se disertacijom predlaže i verificira inovativan pristup računanju intenziteta odnosa među ljudima primjenom koncepta udaljenosti. Tako izračunat intenzitet odnosa među ljudima zvat će se njihovom **društvenom udaljenošću**.

Različiti ljudi povezani su različitim vezama različitih intenziteta – najbolji prijatelji ili poznanici, bliža ili dalja rodbina, kolege iz iste sobe na poslu, kolege s kata, suradnici na projektu ili tek sustručnjaci u različitim organizacijama i slično. Društvena udaljenost može se računati za različite vrste odnosa na temelju podataka za koje se prethodno pokaže da imaju ulogu prediktora za tu vrstu odnosa.



Slika VI.1 Primjer transformacije eksplicitnog binarnog grafa u obogaćeni društveni graf primjenom koncepta udaljenosti za odnos prijateljstva

[20] Za izradu slike korišten je alat *Gephi* te njegove nadogradnje *MDS Statistics* i *MDS Layout*

VI.1 MODELI ZA RAČUNANJE DRUŠTVENE UDALJENOSTI

VI.1.1 Formalni opis modela za računanje društvene udaljenosti

Društvena udaljenost numerička je reprezentacija intenziteta odnosa između odabranog para ljudi (*A* i *B*) za koju vrijedi da je to manja što je intenzitet odnosa veći. Društvena se udaljenost definirana je na intervalu [0, ∞> pri čemu je društvena udaljenost *0* kada je *A=B* dok sa slabljenjem odnosa među ljudima, društvena udaljenost raste.

U okviru ovoga rada društvena će se udaljenost računati na temelju interakcije korisnika na sustavu za društveno umrežavanje Facebook. **Model za računanje društvene udaljenosti** između korisnika *A* i *B* matematička je funkcija koja kao ulazne parametre prima podatke o interakciji između korisnika *A* i *B* pojedinim interakcijskim parametrom (ip_i) te koeficijente značajnosti pojedinog parametra interakcije (W_{ipi}), a kao izlaz vraća numeričku reprezentaciju društvene udaljenosti korisnika *A* i *B*:

$$društvena\ udaljenost_{A,B}(ip_1, ip_2, ip_3, \dots ip_n, W_{ip_1}, W_{ip_2}, W_{ip_3}, \dots, W_{ip_n}) \qquad \text{(VI.1)}$$

Parametar ip_i može pohranjivati *količinu* interakcije između ego-korisnika i njegovog promatranog prijatelja ili *rang* koji, s obzirom na razinu interakcije pojedinim interakcijskim parametrom s ego-korisnikom, zauzima neki njegov prijatelj. Primjerice, ako pretpostavimo da ego-korisnik ima 3 prijatelja: *A*, *B* i *C* te da je s korisnikom *A* razmijenio 57 poruka, s korisnikom *B* 86, a s korisnikom *C* 19 poruka. Tada su brojevi 57, 86 i 19 količine interakcije parametrom *inbox_chat* između ego-korisnika i njegovih prijatelja. Kada se to pretvori u rangove, prijatelju *B* pridružuje se rang 1, prijatelju *A* rang 2, a prijatelju *C* rang 3.

S obzirom na to, modeli za računanje udaljenosti u osnovi se dijele na:

1. modele koji kao ulaz uzimaju **količinu interakcije** ego-korisnika s promatranim prijateljem po promatranim parametrima interakcije
2. modele koji uzimaju **rang** koji promatrani prijatelj zauzima s obzirom na količinu interakcije s ego-korisnikom.

U nastavku će odvojeno biti prikazani i analizirani modeli zasnovani na količini i modeli zasnovani na rangu kako bi se vidjelo je li zaista bitna konkretna količina ili je dovoljno promatrati samo rang prijatelja.

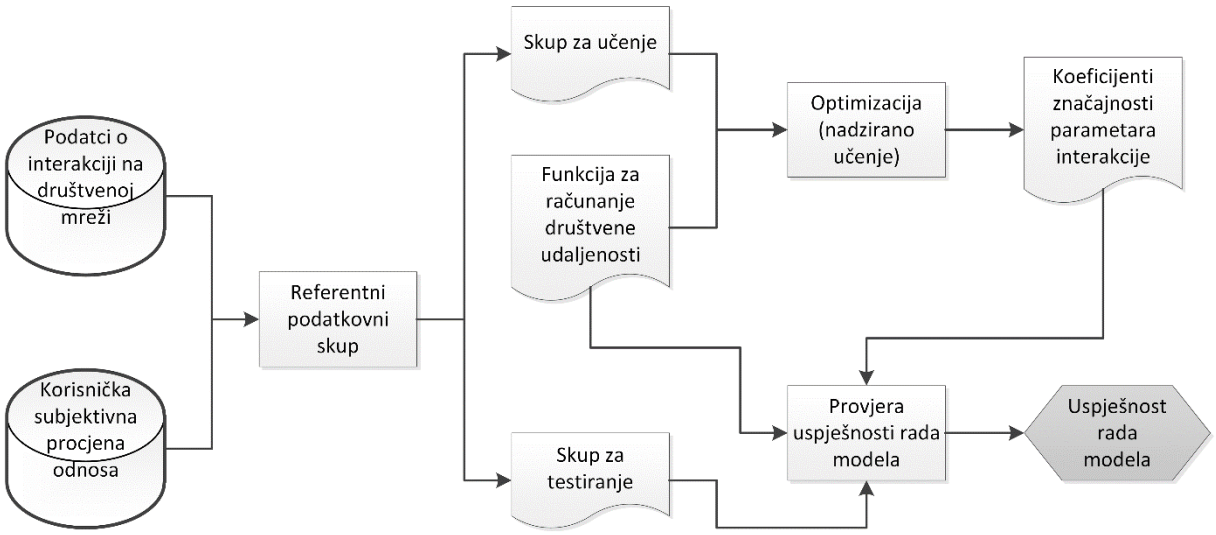
VI.1.2 Idejno rješenje za postupak izgradnje i treniranja modela za računanje društvene udaljenosti

Slika VI.2 prikazuje idejno rješenje procesa izgradnje, treniranja i testiranja modela za računanje društvene udaljenosti. Model za računanje društvene udaljenosti u osnovi koristi neku od funkcija za računanje udaljenosti koje su opisane u potpoglavlju II.2, ali uz određene prilagodbe. Cilj je pronaći optimalnu funkciju za računanje društvene udaljenosti te optimalne koeficijente značajnosti pojedinih parametara interakcije. U tom procesu treba isprobati rad različitih funkcija za računanje udaljenosti te za svaku funkciju pronaći optimalne koeficijente značajnosti pojedinih parametara interakcije. Također, treba za sve funkcije za računanje udaljenosti isprobati računanje udaljenosti *po količini* i *po rangu* interakcije. Za pronalaženje optimalne funkcije za računanje udaljenosti te optimalnih koeficijenata značajnosti parametara interakcije koriste se podatci o interakciji korisnika na sustavu za društveno umrežavanje Facebook te podatci o korisničkoj subjektivnoj procjeni njihova odnosa s njihovim mrežnim prijateljima, koji zajedno čine referentni podatkovni skup. Referentni podatkovni skup prikupljen je istraživanjem *NajFrend* koje je opisano u potpoglavlju IV.2. Iz referentnog skupa izuzeti su svi ispitanici koji nisu predali ispunjeni upitnik ili ga nisu ispunjavali barem 5 minuta, s ciljem da se eliminiraju oni korisnici koji nisu ozbiljno pristupili popunjavanju upitnika pa samim time ni njihovi odgovori nisu relevantni. Na taj način dobije se referentni podatkovni skup koji sadrži 2.273 ispitanika.

Referentni podatkovni skup dijeli se na *skup za učenje* (250 ispitanika i 106.654 njihovih prijatelja, tj. 106.654 zapisa o interakciji između ispitanika i njihovih prijatelja) i *skup za testiranje* (2.023 ispitanika i 845.150 njihovih prijatelja, tj. 845.150 zapisa o interakciji između ispitanika i njihovih prijatelja)[21]. Podjela na skup za učenje i skup za testiranje napravljena je slučajnim odabirom ispitanika. Na skupu za učenje za svaku od razmatranih funkcija za računanje društvene udaljenosti traže se optimalni koeficijenti značajnosti parametara interakcije. Optimalni koeficijenti značajnosti parametara interakcije jesu oni za koje model odnose među ljudima procjenjuje najsličnije moguće subjektivnoj korisničkoj procjeni. Postupak pronalaženja optimalnih koeficijenata parametara interakcije u osnovi slijedi principe

[21] Uobičajeno je referentni podatkovni skup podijeliti na skup za učenje i skup za testiranje na način da skup za učenje bude veći. U ovom slučaju je skup za testiranje višestruko veći. Razlog za to jest činjenica da je referentni podatkovni skup vrlo velik i podskup od 250 ispitanika vrlo je relevantan za učenje modela, iako čini tek 10-ak% od ukupnog broja ispitanika. Veći podatkovni skup za učenje nije odabran zbog prevelike računalne složenosti procesa traženja optimalnih koeficijenata značajnosti interakcijskih parametara. U skupu za testiranje nalaze se svi podatci iz referentnog podatkovnog skupu koji nisu u skupu za učenje. Skup za testiranje mogao je biti i značajno manji, ali, s obzirom na to da je postupak testiranja rada modela računalno značajno jednostavniji od postupka učenja modela, u skupu za testiranje ostavljeno je preko 2.000 ispitanika i gotovo milijun zapisa o interakciji s prijateljima ispitanika.

nadziranog strojnog učenja. Naime, postoji skup poznatih ulaza i poznatih (željenih) izlaza iz modela. Na temelju takvog skupa treba izraditi model koji će na temelju ulaza moći sam odrediti izlaze. Takav model u osnovi je funkcija za računanje udaljenosti koja se parametrizira dodavanjem koeficijenata značajnosti parametara interakcije. Cilj učenja jest pronaći one vrijednosti koeficijenata značajnosti parametara interakcije uz koje će model davati najbolje izlaze, tj. za koje će izlazi modela biti čim sličniji subjektivnoj procjeni ispitanika. S obzirom na to da je skup potencijalnih rješenja nije u potpunosti pretraživ, optimalni koeficijenti značajnosti parametara interakcije tražit će se primjenom heurističkih metoda, tj. konkretno primjenom genetskih algoritama (detaljnije opisani u odvojku VI.1.2.1). Optimalni parametri određeni na skupu za učenje trebaju biti provjereni na skupu za testiranje. Rezultat dobiven na skupu za testiranje smatra se uspješnošću rada modela. U konačnici će biti odabrana ona funkcija udaljenosti koja uz za nju pronađene najbolje koeficijente značajnosti parametara interakcije daje najveću uspješnost rada modela.



Slika VI.2 Proces izrade i odabira optimalnog modela za računanje društvene udaljenosti

VI.1.2.1 Genetski algoritmi

Genetski algoritmi, kao podvrsta evolucijskih algoritama, heuristička su metoda optimiranja koja imitira prirodni evolucijski proces. Predstavljaju model strojnog učenja čije ponašanje potječe iz procesa evolucije koji se neprekidno odvija u prirodi. Po načinu djelovanja, genetski algoritmi ubrajaju se u *metode slučajnog pretraživanja prostora rješenja* u potrazi za globalnim optimumom. U tu skupinu ubrajaju se još i evolucijske strategije, simulirano kaljenje te genetsko programiranje. Snaga tih metoda, a ponajviše genetskih algoritama, leži u tome što su u stanju odrediti položaj globalnog optimuma u višedimenzionalnom prostoru s više lokalnih ekstrema. Klasične determinističke metode tražit će lokalni ekstrem za koji se ne zna je li ujedno

i globalni. Stohastičke metode, kojima pripadaju i genetski algoritmi, mogu s nekom vjerojatnošću odrediti globalni ekstrem. Temeljna razlika između ova dva pristupa jest u tome da determinističke metode daju siguran rezultat, ali manje značajan (lokalni ekstrem), a uz rezultate dobivene stohastičkim metodama uvijek stoji postotak koji označava vjerojatnost da je to rješenje uistinu ono za koje ga se smatra, no, osim lokalnih, mogu dati i globalne optimume. Vjerojatnost ispravnosti rješenja kod stohastičkih metoda povećava se s brojem ponavljanja procesa rješavanja.

Osnovna karakteristika evolucije u prirodi jest prilagođavanje živih bića uvjetima u prirodi. Kako u evoluciji preživljavaju najjači, tako će i kod genetskih algoritama preživjeti skup najjačih, tj. najboljih rješenja. Analogija evolucije kao prirodnog procesa i genetskog algoritma kao metode optimiranja očituje se u procesu *selekcije (odabira)* i *genetskim operatorima*. Mehanizam odabira u prirodnom evolucijskom procesu čine okolina i uvjeti u prirodi. Kod genetskih algoritama ključ selekcije jest *funkcija cilja (ili dobrote)* koja na odgovarajući način predstavlja problem koji se rješava. U prirodi jedinka koja je najbolje prilagođena uvjetima i okolini ima najveću vjerojatnost preživljavanja i parenja pa tako i prenošenja vlastitog genetskog materijala. Kod genetskih algoritama jednu jedinku predstavlja jedno rješenje. *Selekcijom* se odabiru dobre jedinke (rješenja) koje se prenose u sljedeću generaciju, a *manipulacijom* genetskog materijala stvaraju se nove jedinke. U prirodi pri reprodukciji dolazi do izmjene gena kroz proces manipulacije genetskog materijala koji se provodi *križanjem* i *mutacijom*. Isto tako u prijelazu s generacije na generaciju mijenjaju se i rješenja kod genetskih algoritama. Svojstva postojećih rješenja (roditelja) miješaju se kako bi se stvorila djeca (nova generacija). Taj proces naziva se *križanjem*. Uz križanje, izmjena gena obavlja se i znatno rjeđim procesom mutacije. *Mutacija* je proces slučajne promjene genetskog materijala do kojeg u prirodi dolazi pod utjecajem vanjskih uzroka. Križanja i mutacija u terminologiji GA-a naziva se *genetskim operatorima*, a proces izdvajanja najboljih jedinki unutar svake generacije *odabirom* ili *selekcijom*.

Ciklus *reprodukcije*, *selekcije* i *manipulacije* ponavlja se sve dok nije zadovoljen uvjet zaustavljanja genetskog algoritma.

U nastavku će detaljnije biti objašnjeni pojmovi *populacije*, *funkcije cilja (ili dobrote)*, *selekcije (ili odabira)* te *elitizma*.

Populacija

Populacija je skup jedinki iste vrste smještenih na nekom području. Kako dio populacije stari i umire, tako se razmnožavanjem stvaraju novi potomci i veličina populacije u svakoj

generaciji ostaje približno nepromijenjena. U genetskom algoritmu populacija je skup potencijalnih rješenja zadanog problema. Početna populacija može biti odabrana slučajnim odabirom ili nekim drugim optimizacijskim postupkom. Odumiranje slabijih jedinki, ili lošijih potencijalnih rješenja, koje se nisu uspjele prilagoditi novim životnim uvjetima i opstanak jedinki koje su to uspjele (selekcija) u genetskim se algoritmima određuje uporabom *funkcije cilja ili dobrote*. Razmnožavanje jedinki koje su, zahvaljujući svojim dobrim svojstvima, uspjele preživjeti provodi se operacijom križanja.

Svakim novim ponavljanjem (korakom algoritma – prelaženjem iz generacije u generaciju), jedinke u populaciji poprimaju sve bolja svojstva. Algoritam obično završava dosezanjem zadanog broja ponavljanja, tj. broja generacija ili istekom prethodno određenog vremena za rad algoritma. Kada je uvjet završetka ispunjen, iz dobivene populacije odabire se najbolja jedinka i ona predstavlja rješenje optimizacijskog problema.

Funkcija cilja (ili dobrote)

Funkcija cilja imitira prirodnu okolinu koja obavlja selekciju nad jedinkama. Što je jedinka bolje prilagođena okolini u kojoj živi, tj. što je rješenje bolje, veća je vjerojatnost njegova preživljavanja, tj. prenošenja u novu generaciju. Funkcija cilja, kao i svaka druga funkcija, ima određene ulazne parametre i određeni izlaz. Izlaz funkcije cilja treba biti broj. U ovisnosti o tome traži li se minimum ili maksimum, zadaća genetskog algoritma jest tako mijenjati ulazne parametre da izlaz iz funkcije cilja bude čim manji (ako se traži minimum) ili čim veći (ako se traži maksimum). Funkcijom cilja mjeri se kvaliteta rješenja, a konačno rješenje su oni ulazni parametri u funkciju cilja koji daju najmanji ili najveći izlaz iz funkcije cilja, ovisno o tome koji se ekstrem traži..

Odabir odgovarajuće funkcije cilja ključan je problem kod implementacije genetskog algoritma. Također, s obzirom na to da se radi o funkciji koja se u algoritmu najviše koristi, funkcija cilja treba biti što je moguće jednostavnija i brža.

Selekcija (ili odabir)

Svrha selekcije jest čuvanje i prenošenje dobrih svojstava na sljedeću generaciju te izbacivanje loših svojstva. Selekcijom se biraju dobre jedinke koje se onda prenose na novu populaciju. Najjednostavniji postupak selekcije bio bi takav da se jedinke sortiraju po dobroti (rezultatu funkcije dobrote ili cilja) te da se izbaci ranije definiran broj najlošijih jedinki. Na taj bi način proces optimiranja vrlo brzo završio. Međutim, takav se pristup ipak ne primjenjuje. Razlog je u tome što i one najlošije jedinke imaju neka dobra svojstva koja treba iskoristiti. Kad se iz populacije izbace najlošije jedinke, ostaju samo najbolje, no te najbolje ne predstavljaju

ujedno i skup svih najboljih svojstava. Čak i najlošije jedinke često neka svojstva imaju bolja od najboljih. Tu se činjenicu ne smije zanemariti. No, nije dobro niti ako je vjerojatnost preživljavanja svih jedinki podjednaka. Na taj se način gubi velik broj dobrih jedinki. Prema tome, problem je jedino smisleno riješiti dodjeljivanjem niske vjerojatnosti preživljavanja (ali veće od 0) lošim jedinkama, a visoke (no manje od 1) dobrim jedinkama. Određenom malom broju jedinki s najboljim svojstvima, *elitnim jedinkama*, može se dodijeliti vjerojatnost preživljavanja 1, tj. osigurati da će one sigurno nepromijenjene preći u novu generaciju.

Elitizam

Kako se trenutno najbolja rješenja ne bi izgubila upotrebom genetskih operatora ili eliminacijom, tijekom selekcije javlja se potreba za zaštitom najbolje ili nekoliko najboljih jedinki od izmjena (manipulacije) ili eliminacije. Zaštita najboljih jedinki naziva se elitizam i osigurava da se svaka nova generacija kreće prema globalnom optimumu. Zaštićene jedinke zovemo elitnim jedinkama.

Pregled genetskih algoritama preuzet je uz prilagodbe iz ranijeg seminarskog rada izrađenog u sklopu studija na temelju istraživanja genetskih algoritama [62] te na temelju skripte [63].

VI.1.3 Računanje uspješnosti rada modela za računanje društvene udaljenosti

Zadaća modela za računanje društvene udaljenosti jest numerički odrediti društvenu udaljenost između parova prijatelja. Međutim, uspješnost rada modela nije moguće izravno odrediti jer u referentnom podatkovnom skupu ne postoje odgovori kojima korisnici određuju točnu brojčanu vrijednost društvene udaljenosti između sebe i svojih prijatelja niti je razumno takav odgovor od ispitanika tražiti. Uspješnost rada modela određuje se posredno. Provjera uspješnosti provodi se na temelju odgovora ispitanika na pitanje u kojem je ispitanik trebao odlučiti koji mu je prijatelj bolji u ponuđenom paru prijatelja (opisano u odvojku IV.2.2.1) te društvenih udaljenosti koje je model izračunao između promatranih ispitanika i njegovih prijatelja. Model je ispravno odgovorio ako je izračunao manju društvenu udaljenost za onog prijatelja u paru za kojeg je ispitanik rekao da mu je bolji. Uspješnost rada modela računa se na sljedeći način:

$$uspješnost = \frac{broj\ ispravnih\ odgovora\ modela}{ukupno\ promatranih\ slučajeva} \times 100\ \% \qquad \text{(VI.2)}$$

U skupu zapisa na temelju kojih se računa uspješnost modela postoje sljedeći karakteristični slučajevi:

1. Ispitanik ne može odlučiti koji mu je prijatelj u ponuđenom paru bolji
2. Model ne može odlučiti koji je prijatelj u ponuđenom paru bolji jer, najčešće zbog nedostatka podataka o interakciji, računa jednaku udaljenost za oba prijatelja.

Slučajevi u kojima ispitanik ne može odlučiti koji mu je prijatelj bolji izuzimaju se iz referentnog podatkovnog skupa jer im nedostaje informacija o činjeničnom stanju koje bi model trebao *pogoditi*. Slučajeve u kojima model ne može odlučiti ima smisla i ostaviti i izuzeti. Ako se takvi slučajevi ostave, oni se moraju klasificirati kao slučajevi u kojim model nije dao ispravan odgovor čime postaju isti kao i slučajevi u kojima je model dao neispravan odgovor te na taj način negativno utječu na uspješnost rada modela. To ima smisla ako želimo da uspješnost mjeri postotak slučajeva u kojima model daje ispravan odgovor pa je posve svejedno je li model dao pogrešan odgovor ili nije uopće odgovorio. Na taj je način uspješnost i definirana u (VI.2).

Međutim, osim što je zanimljivo vidjeti postotak slučajeva u kojima se odluka modela poklapa s odlukom ispitanika, zanimljivo je vidjeti i postotak slučajeva u kojima se odluka modela poklapa s odlukom ispitanika, ali samo za slučajeve u kojima model može dati odgovor. S obzirom na to, uz uspješnost definiranu izrazom (VI.2), uvest ćemo i **pouzdanost** koja je u osnovi definirana istim izrazom, ali pri njezinu računanju zanemaruju se slučajevi u kojima model ne može odlučiti. **Uspješnost** možemo opisati kao postotak slučajeva u kojima model može donijeti ispravnu odluku, a **pouzdanost** kao postotak slučajeva u kojima će odluka modela biti ispravna, ako je model mogao donijeti odluku. Odnos uspješnosti i povezanosti može se iskazati sljedećim izrazom:

$$uspješnost = pouzdanost \times udio\ slučajeva\ u\ kojima\ model\ može\ donijeti\ odluku \tag{VI.3}$$

U odjeljku VI.1.4 vidjet će se da postoje modeli za računanje udaljenosti vrlo niske uspješnosti i vrlo visoke pouzdanosti. To su modeli koji relativno rijetko mogu donijeti odluku, ali kada je donesu, velika je vjerojatnost da je ta odluka ispravna.

Pri računanju uspješnosti (i pouzdanosti) zasebno se razmatra ego-graf[22] svakog ispitanika.

[22] Ego-graf čine jedan ispitanik (ego) te svi njegovi prijatelji.

VI.1.4 Modeli za računanje društvene udaljenosti zasnovani na samo jednom parametru interakcije

Kako bismo ispitali prediktivna svojstva pojedinih parametara interakcije, provjerit ćemo koliko dobro rade modeli za računanje društvene udaljenosti na temelju samo jednog odabranog parametra interakcije. Kasnije ćemo usporediti uspješnost takvih modela s modelima koji uzimaju u obzir više parametara interakcije kako bi se utvrdilo postoji li potreba za kompleksnijim modelima ili je dovoljno računati udaljenost na temelju samo jednog parametara interakcije. S obzirom na to da se modeli za računanje društvene udaljenosti mogu podijeliti na one koji rade na temelju *količine interakcije* i one koji rade na temelju *ranga interakcije*, bit će zasebno razmatrana ta dva slučaja. S obzirom na to da modeli u sebi sadrže tek jedan parametar interakcije, koeficijenti značajnosti parametara nisu potrebni pa time niti proces učenja modela. Uspješnost (neparametriziranih) modela bit će ispitana na skupu za testiranje.

VI.1.4.1 Modeli zasnovani na rangu interakcije

Model za računanje društvene udaljenosti na temelju samo jednog parametra interakcije koji radi na temelju *ranga interakcije* definira se na sljedeći način:

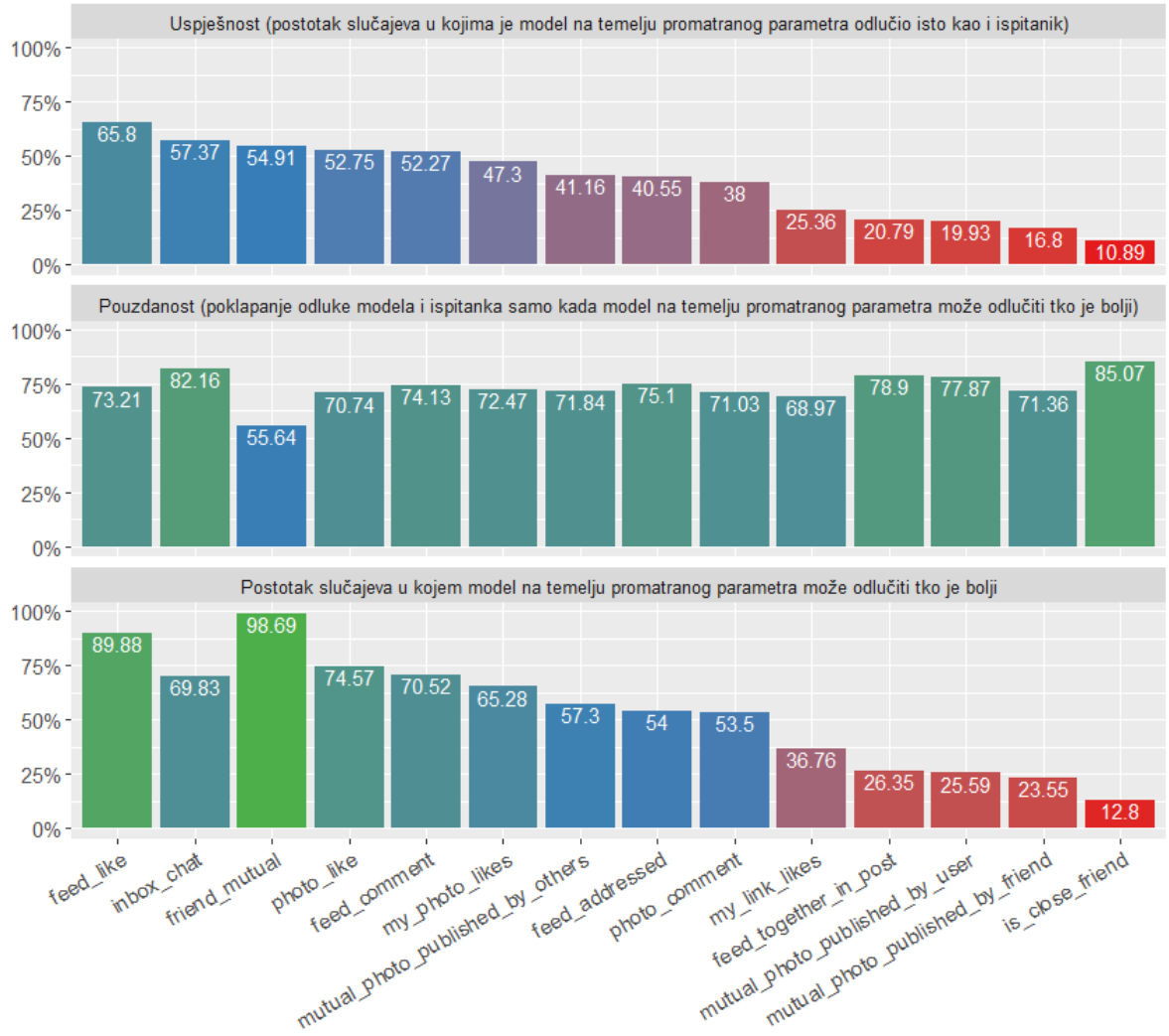
$$društvena\ udaljenost(ispitanik, prijatelj, parametar) = \frac{rang(ispitanik, prijatelj, parametar)}{\max\left(rang(ispitanik, parametar)\right)} \tag{VI.4}$$

Prijatelje ispitanika treba poredati po intenzitetu komunikacije pojedinim parametrom interakcije na način da prvo mjesto zauzima onaj prijatelj s kojim je ispitanik imao najviše interakcije korištenjem promatranog parametra interakcije. Gornji izraz dobiven je prema izrazu za računanje udaljenosti između poredanih varijabli koji je predstavljen u odjeljku II.2.4.

U nastavku ćemo provjeriti uspješnost i pouzdanost modela za računanje društvene udaljenosti za sve parametre interakcije koji se uzimaju u obzir (popisani u Tablica V.1). Svi rezultati koji će biti prezentirani u nastavku dobiveni su na skupu za testiranje. S obzirom na to da se koristi samo jedan parametar, težine nisu važne pa se nije provodio ni proces pronalaženja odgovarajućih težina.

Slika VI.3 prikazuje uspješnost, pouzdanost i pokrivenost modela koji rade na temelju samo jednog parametra. Pokazalo se da se na temelju parametra *feed_like* u najvećem postotku (65,8%) može odrediti tko je bolji prijatelj u promatranom paru prijatelja (uspješnost). Kada je na temelju toga parametra moguće odrediti tko je bolji prijatelj (pouzdanost), odluka će u 73,21% slučajeva biti ispravna. Naime, na temelju parametra *feed_like* odluku o tome tko je

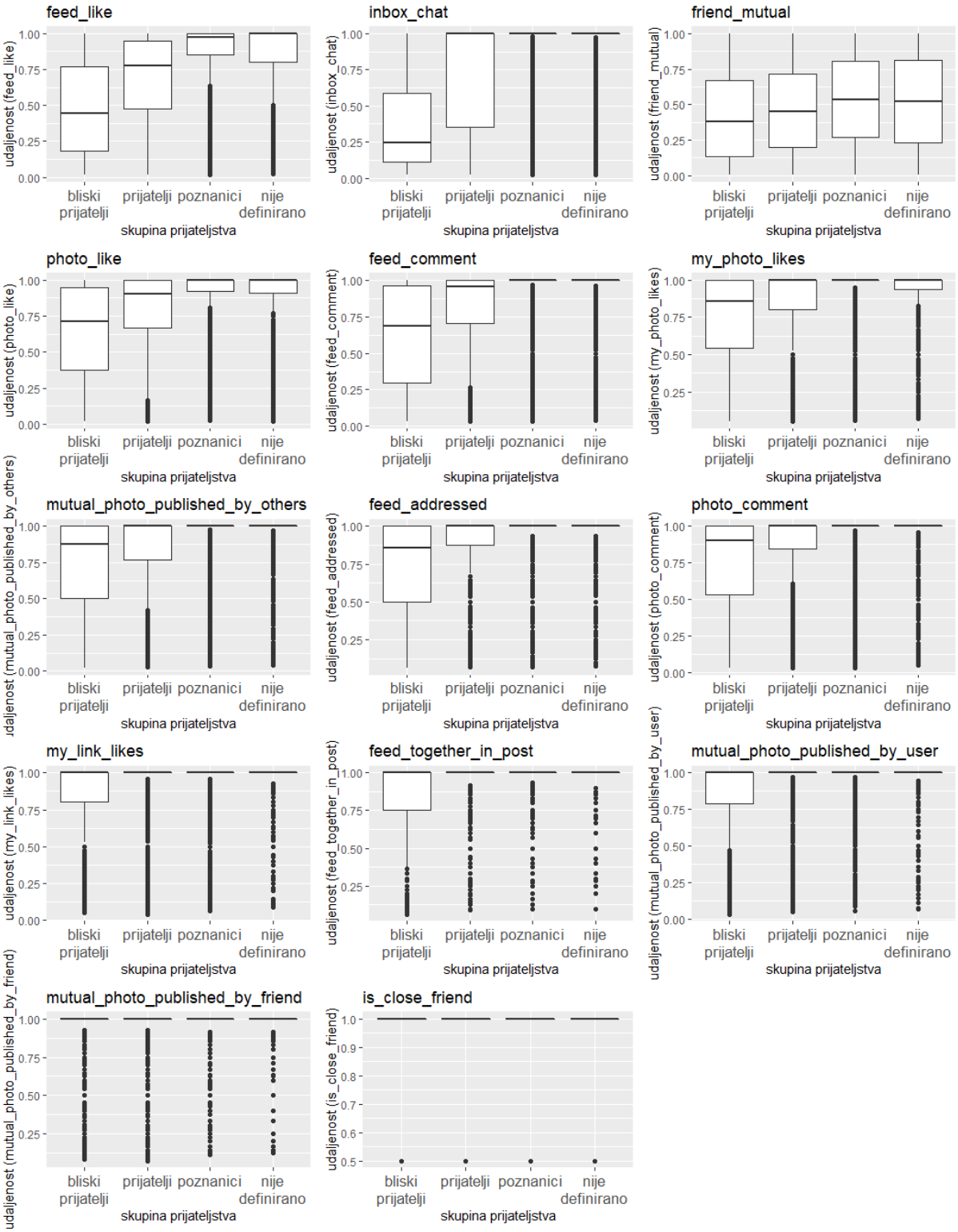
bolji u promatranom paru prijatelja moguće je donijeti u 89,88% slučajeva. Parametar *is_close_friend* najpouzdaniji je s postotkom ispravnog određivanja boljeg prijatelja u paru od 85,07%, ali odluku može donijeti tek u 12,8% slučajeva pa mu je uspješnost tek 10,89%. Na temelju parametra *friend_mutual* moguće je u gotovo 100% slučajeva odlučiti tko je bolji prijatelj u paru, ali uz vrlo nisku pouzdanost od svega 55,64%.



Slika VI.3 Prikaz uspješnosti, pouzdanosti i pokrivenosti modela koji rade na temelju samo jednog parametra interakcije

U drugoj skupini pitanja u anketnom upitniku u sklopu istraživanja *NajFrend* (opisano u odvojku IV.2.2.2) ispitanici su trebali svoje djelomično nasumično odabrane prijatelje rasporediti u tri skupine: *bliski prijatelji*, *prijatelji* i *poznanici*. Također, mogli su nekog prijatelja ostaviti neraspoređenim. Slika VI.4 kutijastim dijagramima prikazuje u kojim se intervalima nalaze udaljenosti za prijatelje koje je ispitanik svrstao u određene skupine prijateljstva. Debela vodoravna crta prikazuje srednju vrijednost udaljenosti po skupini, crta na dnu pravokutnika vrijednost 1. kvartila, a crta na vrhu pravokutnika vrijednost 3. kvartila. Kod

većine parametara interakcije može se primijetiti da udaljenost raste što je skupina prijateljstva *slabija*.



Slika VI.4 Udaljenosti po skupinama prijateljstva za modela zasnovane na rangu interakcije te samo jednom parametru interakcije

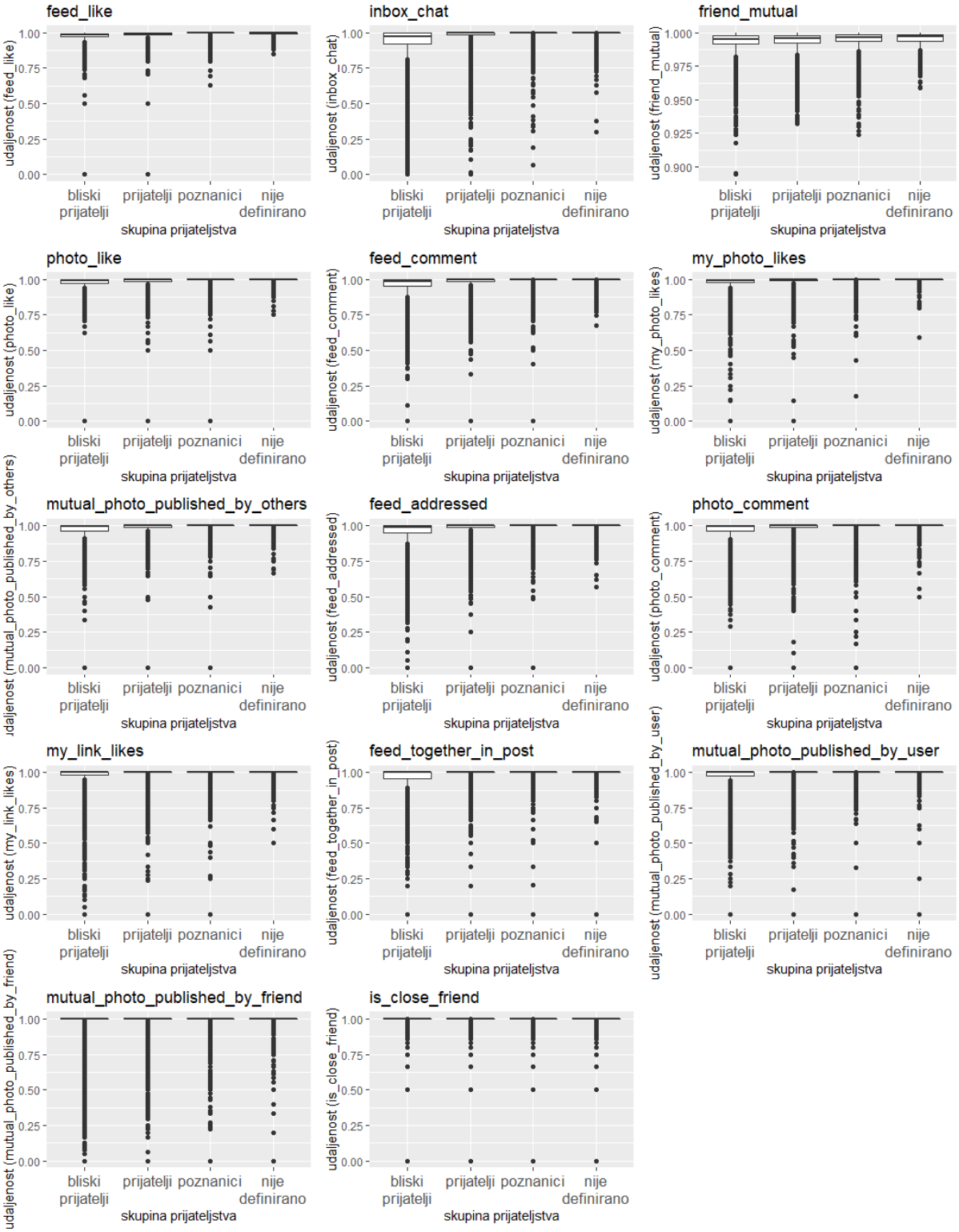
VI.1.4.2 Modeli zasnovani na količini interakcije

Model za računanje društvene udaljenosti na temelju samo jednog parametra interakcije koji radi na temelju *količine interakcije* definira se na sljedeći način:

$$društvena\ udaljenost(ispitanik, prijatelj, parametar) = \sqrt{1 - \frac{količina\ interakcije(ispitanik, prijatelj, parametar)}{\sum_{prijatelji}(količina\ interakcije(ispitanik, parametar))}} \quad \text{(VI.5)}$$

S obzirom na to da se uspješnost modela provjerava na skupu na kojem je potrebno odrediti boljeg prijatelja u paru prijatelja, a pri čemu korisnik s više interakcije ima i viši rang, uspješnost oba pristupa jednaka je i prikazuje je Slika VI.3. Ipak, razdiobe udaljenosti općenito i po skupinama različite su. Slika VI.5 kutijastim dijagramima prikazuje u kojim se intervalima nalaze udaljenosti za prijatelje koje je ispitanik svrstao u određene skupine prijateljstva, a koje su izračunate na temelju količine interakcije. U odnosu na udaljenosti izračunate po rangu interakcije (Slika VI.4), može se uočiti da su udaljenosti izračunate po količini interakcije puno zgusnutije. Međutim, u oba se slučaja naziru razlike između skupina prijateljstva.

Slika VI.3 pokazuje da je na temelju izračuna društvene udaljenosti na temelju bilo kojeg parametra interakcije zasebno moguće odrediti boljeg prijatelja u paru s pouzdanošću većom od referentne pouzdanosti (nasumično pogađanje – 50%). U odjeljku VI.1.5 ispitat će se različiti modeli za računanje društvene udaljenosti zasnovani na kombinaciji više interakcijskih parametara s ciljem da se pronađe pouzdaniji i uspješniji model od zasebno najuspješnijeg (model na temelju parametra *feed_like* – 65,67%) i najpouzdanijeg modela (model na temelju parametra *is_close_friend* – 84,71%).



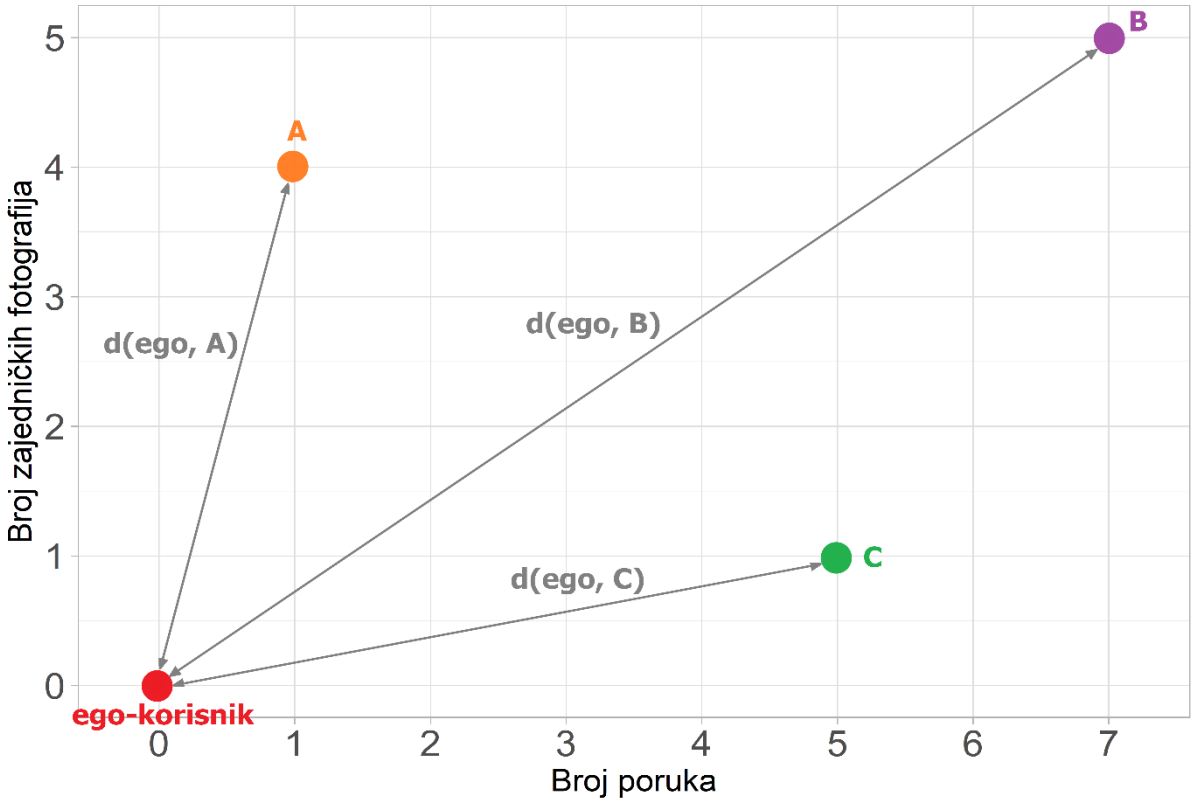
Slika VI.5 Udaljenosti po skupinama prijateljstva za modele zasnovane na količini interakcije te samo jednom parametru interakcije

VI.1.5 Modeli za računanje društvene udaljenosti zasnovani na više parametara interakcije

Modeli za računanje društvene udaljenosti na temelju više parametara interakcije oslanjaju se na modele za računanje udaljenosti na temelju jednog parametra interakcije.

Pretpostavlja se *n*-dimenzionalan prostor u kojem svaka od *n* dimenzija predstavlja jedan parametar interakcije. Ispitanik (ego-korisnik) se nalazi u središtu koordinatnog sustava. Svaki njegov prijatelj prikazuje se kao točka u tom *n*-dimenzionalnom prostoru, pri čemu je vrijednost odsječka na svakoj koordinatnoj osi jednaka društvenoj udaljenosti između ispitanika i njegova prijatelja izračunatoj na temelju parametra interakcije pridruženog odnosnoj osi (prema izrazima uvedenima u odjeljku VI.1.4). Društvena udaljenost između ispitanika i njegova prijatelja jednaka je udaljenosti između točaka *n*-dimenzionalnog prostora koje predstavljaju ispitanika i njegova prijatelja. Udaljenost između točaka n-dimenzionalnog prostora može se računati primjenom različitih načina računanja udaljenosti.

Slika VI.6 prikazuje ego-korisnika i njegova tri prijatelja: *A*, *B* i *C*. Zbog jednostavnosti vizualizacije, razmatra se interakcija posredstvom samo 2 parametra interakcije: broja razmijenjenih privatnih poruka i broja zajedničkih fotografija. Kako bi se ilustrirali različiti načini računanja udaljenosti, u nastavku će se za primjer izračunati udaljenosti između ego-korisnika i njegovih prijatelj na više načina: Minkowskijevom udaljenošću za parametre p=1 (blokovska udaljenost), za p=2 (euklidska udaljenost) te za p=5; prosječna udaljenost za parametar p=2 te Mahalanobisova udaljenost.



Slika VI.6 Vizualizacija točaka koje predstavljaju ego-korisnika i njegove prijatelje

Blokovska udaljenost

$$d_{blokovska}(ego, A) = (|0-1| + |0-4|) = 5$$
$$d_{blokovska}(ego, B) = (|0-7| + |0-5|) = 12$$
$$d_{blokovska}(ego, C) = (|0-5| + |0-1|) = 6$$

Euklidska udaljenost

$$d_{euklidska}(ego, A) = (|0-1|^2 + |0-4|^2)^{\frac{1}{2}} = \sqrt{1+16} = 4{,}123$$
$$d_{euklidska}(ego, B) = (|0-7|^2 + |0-5|^2)^{\frac{1}{2}} = \sqrt{49+25} = 8{,}602$$
$$d_{euklidska}(ego, C) = (|0-5|^2 + |0-1|^2)^{\frac{1}{2}} = \sqrt{25+1} = 5{,}099$$

Minkowskijeva udaljenost za p=5

$$d_{minkowskijeva(p=5)}(ego, A) = (|0-1|^5 + |0-4|^5)^{\frac{1}{5}} = \sqrt[5]{1+1024} = 4$$
$$d_{minkowskijeva(p=5)}(ego, B) = (|0-7|^5 + |0-5|^5)^{\frac{1}{5}} = \sqrt[5]{16807+3125} = 7{,}243$$
$$d_{minkowskijeva(p=5)}(ego, C) = (|0-5|^5 + |0-1|^5)^{\frac{1}{5}} = \sqrt[5]{3125+1} = 5$$

Prosječna udaljenost

$$d_{prosječna}(ego, A) = \sqrt{\frac{1}{2}[(0-1)^2 + (0-4)^2]} = \sqrt{\frac{1+16}{2}} = 2{,}915$$
$$d_{prosječna}(ego, B) = \sqrt{\frac{1}{2}[(0-7)^2 + (0-5)^2]} = \sqrt{\frac{49+25}{2}} = 6{,}083$$
$$d_{prosječna}(ego, C) = \sqrt{\frac{1}{2}[(0-5)^2 + (0-1)^2]} = \sqrt{\frac{25+1}{2}} = 3{,}606$$

Mahalanobisova udaljenost

Kako bi se mogla izračunati Mahalanobisova udaljenost između bilo koje dvije točke, potrebno je prvo na temelju svih poznatih točaka izračunati matricu kovarijance u skladu s naputkom iz odvojka II.2.5.1. Ako 4 poznate točke zapišemo matrično, dobivamo sljedeće:

	X	**Y**
A	1	4
B	7	5
C	5	1
ego	0	0

Na temelju navedenih poznatih točaka dobivamo sljedeći matricu kovarijance:

$$S = \begin{bmatrix} 10{,}9167 & 3{,}8333 \\ 3{,}8333 & 5{,}6667 \end{bmatrix}$$

Međutim, za izračun Mahalanobisove udaljenosti potreban nam je inverz matrice kovarijance:

$$S^{-1} = \begin{bmatrix} 0{,}1201 & -0{,}0813 \\ -0{,}0813 & 0{,}2314 \end{bmatrix}$$

Sada imamo sve potrebno kako bismo mogli računati Mahalanobisove udaljenost:

$$d_{mahalanobisova}(A, B) = \sqrt{(A-B)^T S^{-1} (A-B)}$$

$$d_{mahalanobisova}(ego, A) = \sqrt{\begin{bmatrix} -1 \\ -4 \end{bmatrix} \begin{bmatrix} 0{,}1201 & -0{,}0813 \\ -0{,}0813 & 0{,}2314 \end{bmatrix} [-1 \quad -4]} = 1{,}7813$$

$$d_{mahalanobisova}(ego, B) = \sqrt{\begin{bmatrix} -7 \\ -5 \end{bmatrix} \begin{bmatrix} 0{,}1201 & -0{,}0813 \\ -0{,}0813 & 0{,}2314 \end{bmatrix} [-7 \quad -5]} = 2{,}4462$$

$$d_{mahalanobisova}(ego, C) = \sqrt{\begin{bmatrix} -5 \\ -1 \end{bmatrix} \begin{bmatrix} 0{,}1201 & -0{,}0813 \\ -0{,}0813 & 0{,}2314 \end{bmatrix} [-5 \quad -1]} = 1{,}5564$$

U gornjim primjerima (radi jednostavnosti) oba promatrana parametra interakcije jednako su tretirana, a odsječak po odgovarajućim osima nije bio izračunata udaljenost na temelju pridruženog parametra interakcije, nego ukupna količina interakcije promatranim parametrom.

U postupku traženja dobrog modela za računanje društvene udaljenosti potrebno je, u skladu s idejnim rješenjem opisanim u odjeljku VI.1.2, ispitati rad različitih načina za računanje udaljenosti te za svaki taj način kroz proces strojnog učenja, a na temelju eksperimentalno prikupljenih podataka, odrediti čim bolje koeficijente značajnosti pojedinih parametara interakcije. U sklopu ove disertacije razmatrani su sljedeći načini računanja udaljenosti: Minkowskijeve udaljenosti za vrijednosti parametra *p=1, 2, 3, 4, 5, 10, 20 i 100*, prosječna Minkowskijeva udaljenost za iste vrijednosti parametra *p* te Mahalanobisova udaljenost. Svaki način računanja udaljenosti isproban je za slučaj kada su udaljenosti po pojedinim parametrima interakcije računani na temelju *količine* interakcije i kada su računani na temelju *ranga* interakcije.

Optimalni koeficijenti značajnosti pojedinog parametra interakcije određuju se zasebno za svaki način računanja udaljenosti te zasebno za slučaj kada se udaljenosti po svakom parametru interakcije računaju na temelju *količine* interakcije i zasebno kada se računaju na

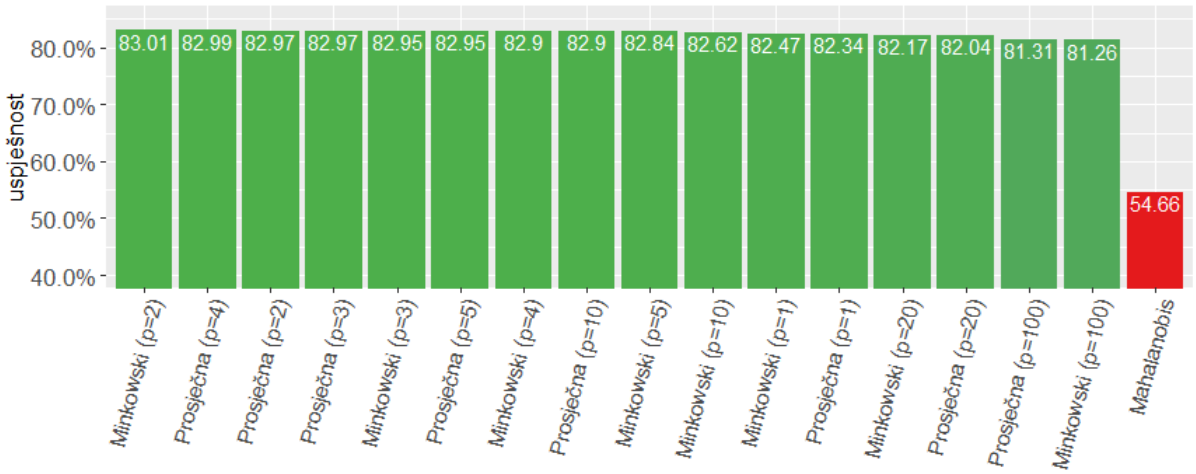
temelju *ranga* interakcije. Za svaki promatrani slučaj koeficijenti značajnosti promatranih parametara interakcije određuju se primjenom genetskih algoritama nad skupom za učenje modela uz pretpostavku da se svi koeficijenti nalaze u intervalu [-5, 15] pa se pretraživanje ograničava na taj interval. Za genetske algoritme postavljene su sljedeće početne postavke:

- Veličina populacije – 100 jedinki
- Mutacija – 0,3
- Križanje – 0,9
- Najveći broj koraka algoritma – 50
- Prekid algoritma nakon 10 koraka bez poboljšanja.

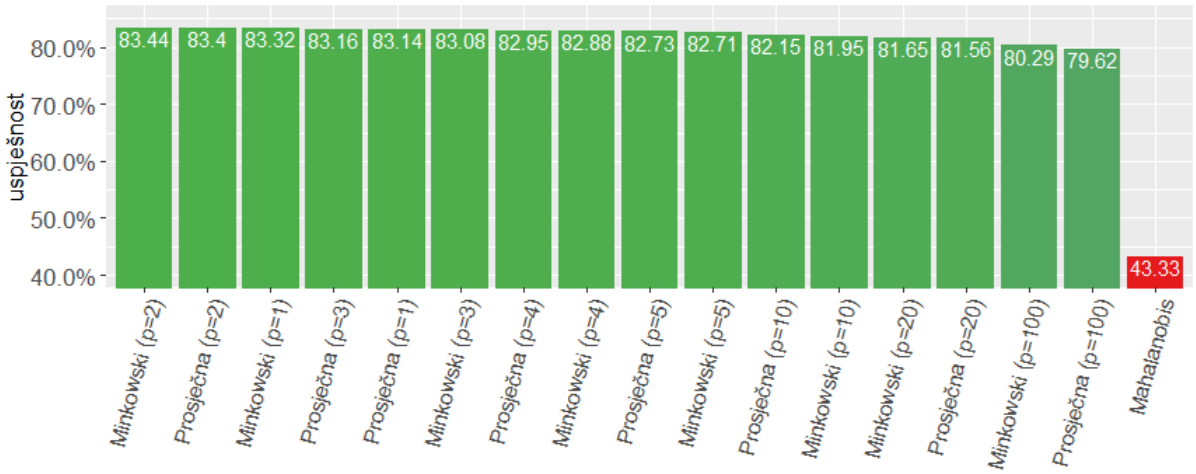
Za svaki promatrani slučaj genetski se algoritam pokreće 20 puta te se uzima najbolje rješenje, tj. koeficijenti značajnosti koji su za određeni slučaj dali najbolje rezultate. U nastavku će prvo biti prikazani rezultati procesa učenja, a zatim i verifikacija naučenih modela.

VI.1.5.1 Rezultati procesa učenja

Na skupu na kojem se provodilo učenje modela postignute su najbolje uspješnosti prikazane na Slika VI.7 za modele zasnovane na *količini* interakcije, a na Slika VI.8 za modele zasnovane na *rangu* interakcije.



Slika VI.7 Najbolje uspješnosti modela za računanje društvene udaljenosti na temelju količine *interakcije dobivene na skupu za učenje uz optimalne koeficijente značajnosti parametara interakcije*



Slika VI.8 Najbolje uspješnosti modela za računanje društvene udaljenosti na temelju ranga *interakcije dobivene na skupu za učenje uz optimalne koeficijente značajnosti parametara interakcije*

Minkowskijeva i prosječna udaljenost daju podjednake rezultate neovisno o vrijednosti parametra *p*, dok se Mahalanobisova udaljenost pokazala značajno lošijom za računanje društvene udaljenosti.

Tablica VI.1 i Tablica VI.2 prikazuju koeficijente značajnosti pojedinih parametara interakcije uz koje određeni način računanja udaljenosti postiže najveću uspješnost na skupu za učenje. Tablica VI.1 pokriva slučaj kada se udaljenosti računaju na temelju *količine* interakcije, a Tablica VI.2 kada se udaljenosti računaju na temelju *ranga* interakcije.

Tablica VI.1 Eksperimentalno dobiveni optimalni koeficijenti značajnosti parametara interakcije po načinima računanja udaljenosti za računanje udaljenosti na temelju količine *interakcije*

Optimalni koeficijenti značajnosti parametara interakcije														
Načini računanja udaljenosti		fl	fc	fa	ftp	mpu	mpf	mpo	pl	pc	fm	ic	mpl	mll	icf
	Mahalanobisova udaljenost	3,26	4,85	5,10	2,24	-3,45	4,55	6,16	4,28	6,58	11,97	4,96	5,42	4,47	5,92
	Minkowskijeva udaljenost (p=1)	6,69	3,91	4,76	2,04	3,84	0,84	5,08	-0,03	-0,46	1,87	12,88	9,29	1,20	9,54
	Minkowskijeva udaljenost (p=2)	5,55	2,88	0,97	2,38	3,84	-0,19	2,86	4,02	2,11	2,38	11,61	8,94	2,21	4,80
	Minkowskijeva udaljenost (p=3)	7,70	5,00	2,42	3,69	6,13	2,82	4,23	3,77	2,97	3,18	11,70	7,81	3,50	6,34
	Minkowskijeva udaljenost (p=4)	5,33	4,20	1,51	3,10	3,69	2,27	4,26	2,13	2,16	3,13	7,75	5,68	3,11	6,92
	Minkowskijeva udaljenost (p=5)	8,97	2,25	7,00	6,57	3,70	5,79	7,52	4,61	5,07	3,26	11,24	8,40	5,31	8,52
	Minkowskijeva udaljenost (p=10)	7,22	6,31	4,68	6,38	6,83	2,40	3,73	5,23	4,65	3,18	8,02	7,20	4,37	7,64
	Minkowskijeva udaljenost (p=20)	8,00	5,32	3,56	2,44	7,98	3,55	3,09	7,08	4,26	5,71	8,79	8,19	7,28	8,33
	Minkowskijeva udaljenost (p=100)	9,04	4,70	3,81	1,96	2,73	-0,78	6,23	8,24	8,21	8,04	9,24	8,48	4,71	9,22
	Prosječna udaljenost (p=1)	6,77	2,98	3,36	2,28	2,82	1,44	6,10	-0,23	-1,02	1,44	11,70	8,47	1,32	8,81
	Prosječna udaljenost (p=2)	4,92	1,20	2,90	2,81	3,46	0,92	2,42	3,64	2,17	1,01	11,26	8,10	0,86	8,86
	Prosječna udaljenost (p=3)	6,01	3,71	2,17	3,74	4,67	2,01	4,64	5,20	0,83	2,64	10,51	8,41	3,34	11,64
	Prosječna udaljenost (p=4)	7,73	5,38	2,23	4,57	7,41	2,87	4,40	6,26	3,30	3,09	11,02	8,04	4,35	6,89
	Prosječna udaljenost (p=5)	7,60	2,65	2,90	5,56	5,75	2,42	5,32	4,25	3,72	3,90	9,11	8,10	4,42	6,84
	Prosječna udaljenost (p=10)	5,84	5,60	4,84	5,32	5,29	2,32	4,81	4,70	3,53	2,75	6,82	5,97	5,17	6,30
	Prosječna udaljenost (p=20)	4,40	8,19	3,40	2,98	7,90	1,22	2,25	4,82	3,32	3,91	8,72	5,04	4,08	8,28
	Prosječna udaljenost (p=100)	6,16	7,85	6,88	5,64	1,23	4,89	4,26	3,58	6,33	6,67	7,94	7,87	2,48	7,81

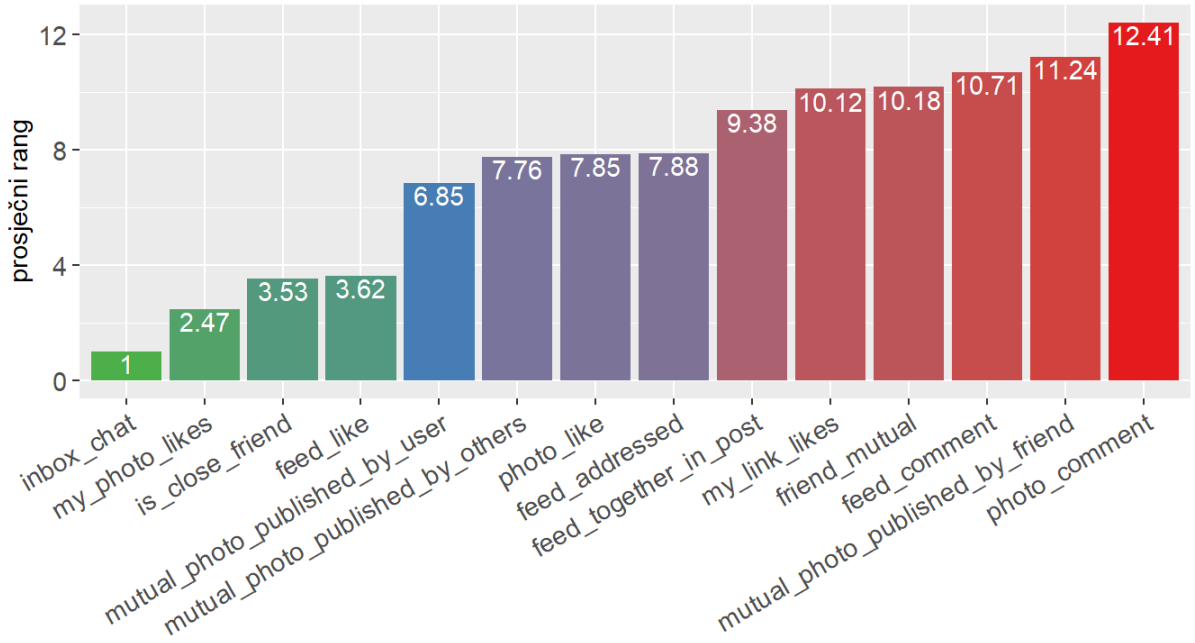
Tablica VI.2 Eksperimentalno dobiveni optimalni koeficijenti značajnosti parametara interakcije po načinima računanja udaljenosti za računanje udaljenosti na temelju ranga *interakcije*

Optimalni koeficijenti značajnosti parametara interakcije														
Načini računanja udaljenosti		fl	fc	fa	ftp	mpu	mpf	mpo	pl	pc	fm	ic	mpl	mll	icf
	Mahalanobisova udaljenost	1,11	6,21	6,98	6,56	6,23	4,96	6,97	6,04	-4,94	5,05	3,28	8,15	5,65	5,16
	Minkowskijeva udaljenost (p=1)	3,18	4,02	-0,56	4,27	5,57	-0,34	3,31	0,98	1,42	0,44	10,36	6,55	2,91	10,82
	Minkowskijeva udaljenost (p=2)	4,69	3,40	0,21	6,29	6,90	2,96	3,66	4,19	0,94	0,98	10,68	7,83	1,36	8,81
	Minkowskijeva udaljenost (p=3)	5,17	6,63	4,44	7,57	7,34	3,24	4,91	4,53	3,13	2,36	10,80	7,93	3,79	9,51
	Minkowskijeva udaljenost (p=4)	4,90	6,50	4,73	7,52	7,78	0,61	4,59	1,57	4,67	4,05	10,89	7,96	1,76	10,05
	Minkowskijeva udaljenost (p=5)	4,83	4,97	4,14	5,96	6,35	4,13	5,20	0,52	0,95	2,12	8,64	6,16	4,57	8,23
	Minkowskijeva udaljenost (p=10)	4,47	5,52	5,20	5,41	5,79	4,71	1,60	4,30	3,35	2,06	7,58	5,79	2,96	7,56
	Minkowskijeva udaljenost (p=20)	4,05	5,64	3,92	7,44	8,64	3,86	1,84	5,00	2,66	4,40	9,96	6,77	2,85	9,94
	Minkowskijeva udaljenost (p=100)	5,40	6,45	5,59	6,02	4,95	0,70	6,08	5,59	-0,42	5,83	7,21	6,08	1,71	6,46
	Prosječna udaljenost (p=1)	5,33	1,82	2,48	4,18	6,78	0,07	3,20	-0,54	-0,10	0,62	10,47	6,21	2,56	9,97
	Prosječna udaljenost (p=2)	4,08	3,90	1,13	6,09	6,56	2,15	4,68	4,03	1,97	1,90	10,51	8,33	2,06	10,48
	Prosječna udaljenost (p=3)	5,44	4,34	-0,52	5,68	6,34	4,04	5,11	4,97	2,01	3,73	9,64	6,90	3,21	8,46
	Prosječna udaljenost (p=4)	5,47	7,61	3,07	8,00	9,00	1,72	6,33	4,34	4,69	-0,69	12,20	9,13	4,15	11,15
	Prosječna udaljenost (p=5)	5,71	6,03	1,70	6,67	7,17	7,52	5,99	4,88	3,27	0,33	9,99	7,31	5,40	9,35
	Prosječna udaljenost (p=10)	1,70	4,70	4,54	4,86	4,80	3,74	3,86	3,82	1,75	2,78	6,29	4,80	2,07	6,31
	Prosječna udaljenost (p=20)	4,35	4,96	4,25	6,73	6,24	4,48	1,79	4,58	2,43	4,60	8,34	5,73	4,26	8,33
	Prosječna udaljenost (p=100)	5,22	5,38	3,78	5,36	3,33	1,76	5,06	4,41	1,45	5,42	6,21	5,57	5,03	5,65

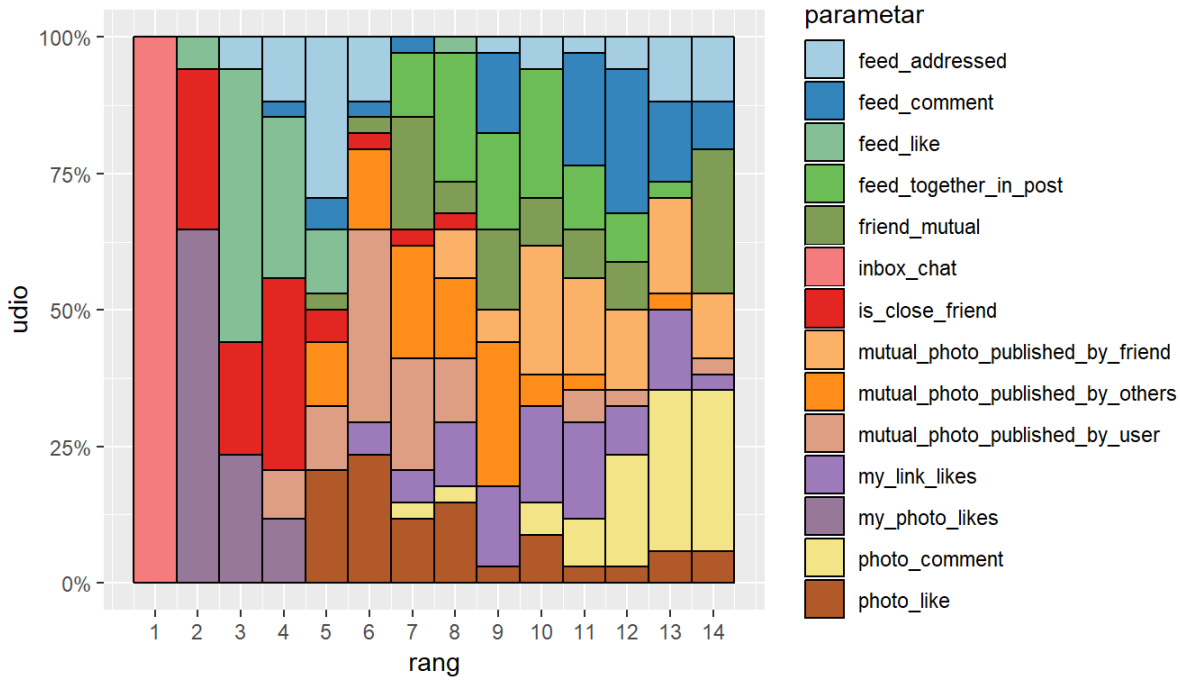
U tablicama su prikane kombinacije koeficijenata značajnosti uz koje je dobivena najveća uspješnost na skupu za učenje. Međutim, kako je proces učenja proveden genetskim algoritmima kao heurističkom metodom, različita pokretanja algoritma vraćala su različite skupove koeficijenata značajnosti koji imaju sličnu uspješnost (različiti lokalni ekstremi). Za 20 pokretanja genetskog algoritma za svaki način računanja udaljenosti dobiveno je i 20 (često) različitih rješenja[23], tj. različitih kombinacija vrijednosti koeficijenata globalne značajnosti. Za svako pokretanje koeficijenti su rangirani te se u nastavku grafički prikazuje koliko je često koeficijent značajnosti pojedinog parametra rangiran na nekom mjestu i koji je prosječan rang koeficijenta značajnosti pojedinog parametra interakcije. S obzirom na to da su isprobane mogućnosti velikog broja načina računanja udaljenosti i da bi detaljno prezentiranje rezultata za svaki model zauzelo bi previše prostora, u nastavku će biti prikazano kako se mijenjaju rangovi koeficijenata značajnosti pri različitim pokretanjima genetskog algoritma za model koji se pokazao najboljim na skupu za učenje – za slučaj kada se promatrala *količina* interakcije i kada se promatrao *rang* interakcije. Radi se modelu zasnovanom na Minkowskijevoj udaljenosti uz vrijednosti parametara $p=2$, tj. o euklidskoj udaljenosti. Slika VI.9 prikazuje prosječan rang parametara interakcije pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir *količine* interakcije. Slika VI.10 prikazuje učestalost pojavljivanja pojedinog parametra interakcije na određenom rangu pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir *količine* interakcije. Slika VI.11 prikazuje prosječan rang parametara interakcije pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir *rang* interakcije. Slika VI.12 prikazuje učestalost pojavljivanja pojedinog parametra interakcije na određenom rangu pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir *rang* interakcije. Iz spomenutih slika može se vidjeti da se rangovi koeficijenata značajnosti pojedinih parametara interakcije mijenjaju od pokretanja do pokretanja genetskog algoritma. Međutim, i iz prosječnog ranga, a i iz pregleda učestalosti zauzimanja pojedinog ranga može se uočiti uzak raspon rangova u kojem se određeni koeficijent značajnosti nalazit. Ako se rangovi koeficijenata značajnosti parametara interakcije

[23] Ponekad ima i više od 20 rješenja jer jedno pokretanje genetskog algoritma može završiti i s više jednako vrijednih rješenja. To znači da je algoritam završio s više kombinacija vrijednosti koeficijenata značajnosti pojedinih parametara interakcije koji daju istu uspješnost, a ta je uspješnost ujedno i najbolja dobivena uspješnost.

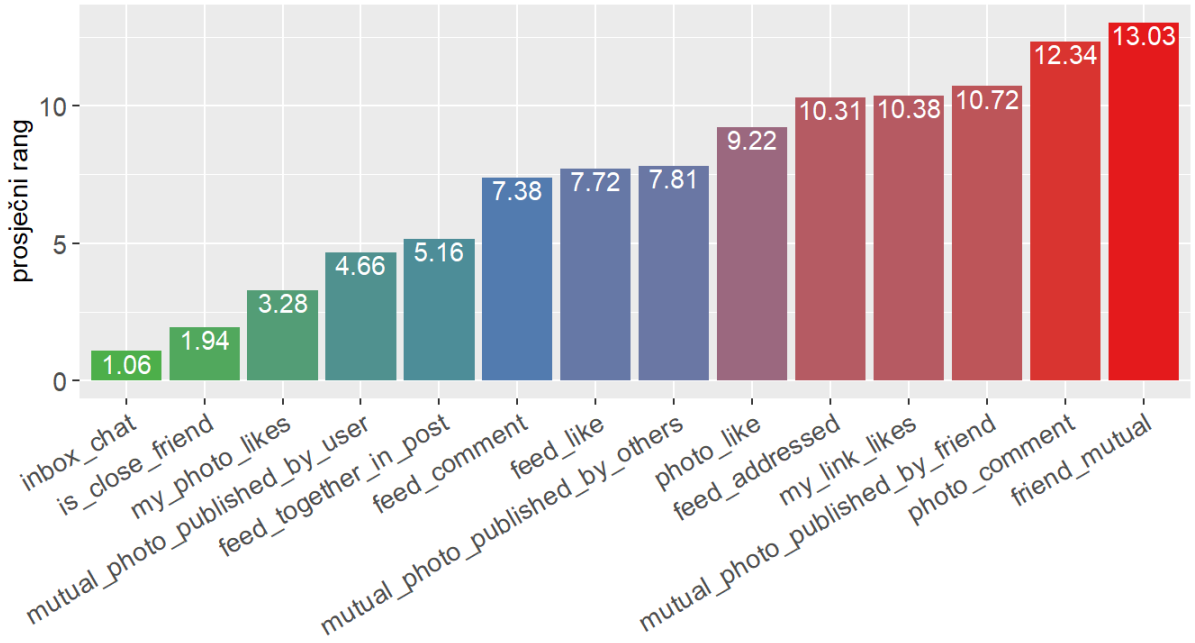
usporede s pouzdanošću modela za računanje udaljenosti na temelju jednog parametra interakcije (Slika VI.3), može se uočiti sličan redoslijed parametara.



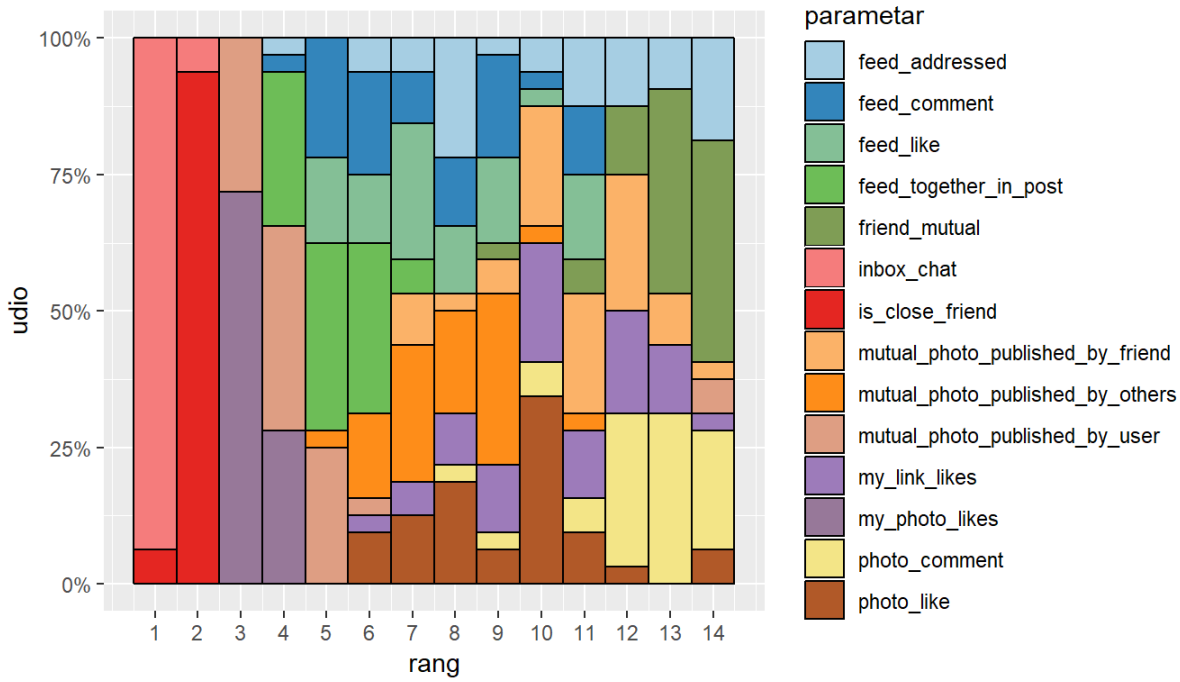
Slika VI.9 Prosječan rang parametara interakcije pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir količine *interakcije*



Slika VI.10 Učestalost pojavljivanja pojedinog parametra interakcije na određenom rangu pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir količine *interakcije*



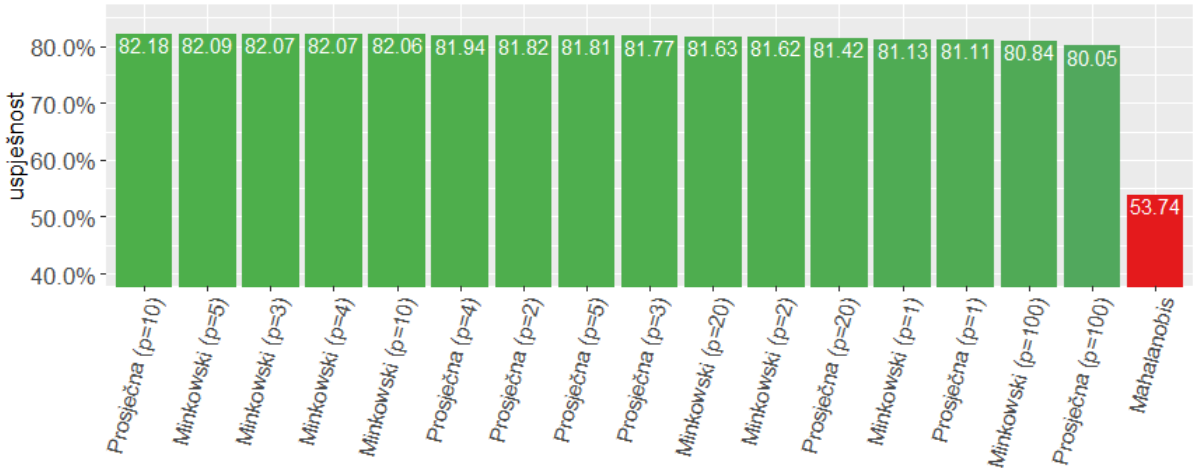
Slika VI.11 Prosječan rang parametara interakcije pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir ranga *interakcije*



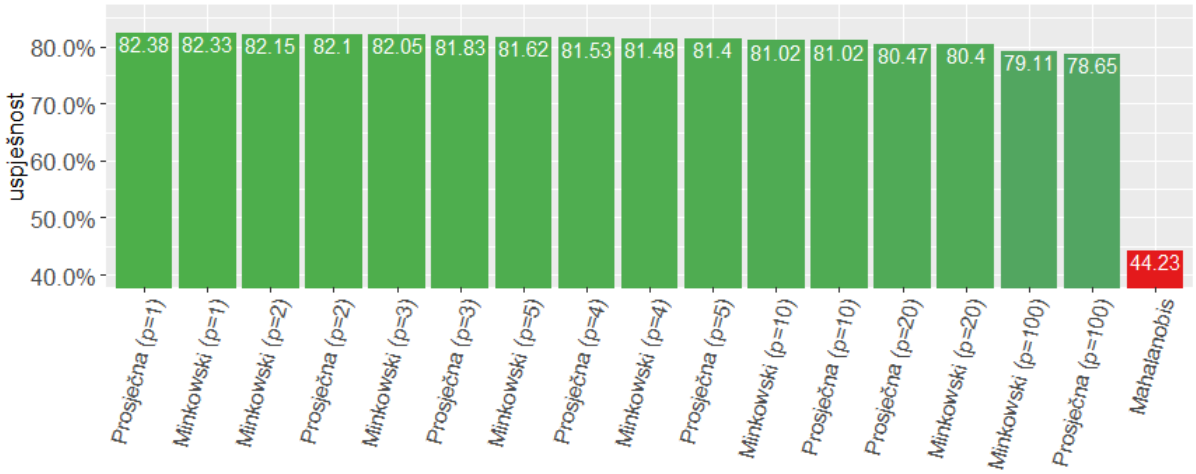
Slika VI.12Učestalost pojavljivanja pojedinog parametra interakcije na određenom rangu pri višestrukom pokretanju genetskog algoritma za računanje udaljenosti primjenom euklidske udaljenosti uz uzimanje u obzir ranga *interakcije*

VI.1.5.2 Verifikacija rezultata dobivenih na skupu za učenje nad skupom za testiranje

Svrha skupa za učenje jest naučiti model, tj. pronaći one vrijednosti parametara značajnosti za koje modela najbolje radi. Međutim, uspješnost rada modela ne može se mjeriti na skupu za učenje već na nekom novom skupu, tj. na primjerima koji nisu korišteni u procesu učenja modela. Takav se skup zove skup za testiranje. Na skupu za testiranje dobivene uspješnosti prikazane su na Slika VI.13 za modele zasnovane na *količini* interakcije te na Slika VI.14 za modele zasnovane na *rangu* interakcije.



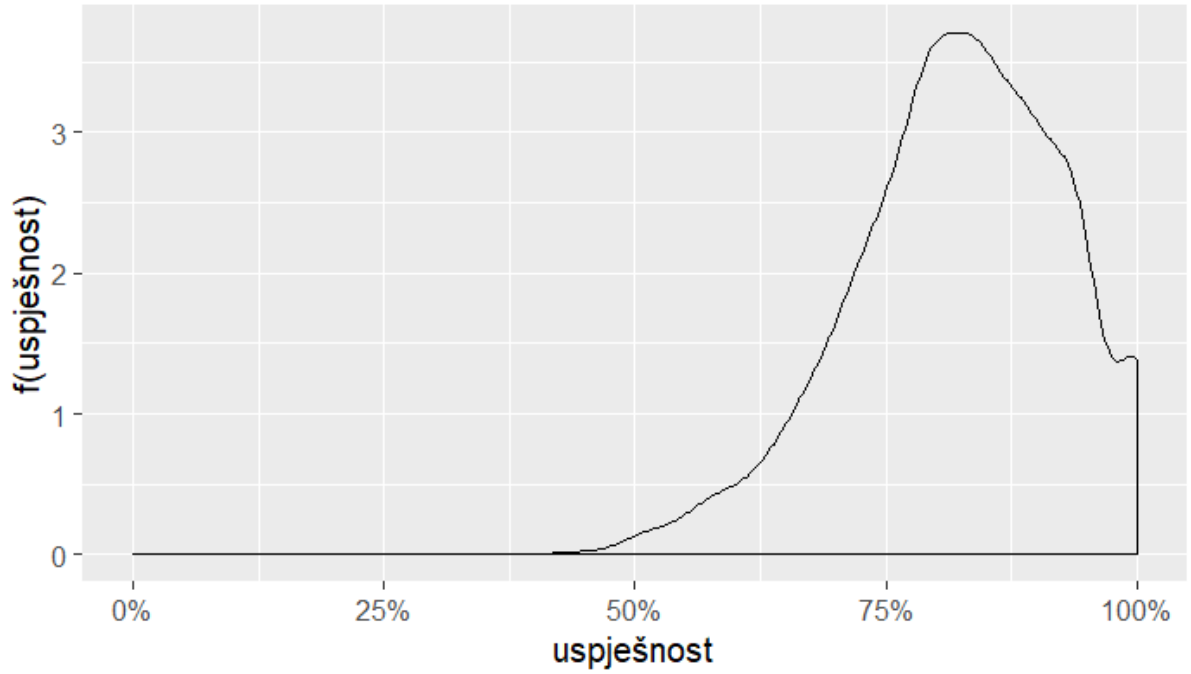
Slika VI.13 Najbolje uspješnosti modela za računanje društvene udaljenosti na temelju količine *interakcije dobivene na skupu za testiranje uz optimalne koeficijente značajnosti parametara interakcije*



Slika VI.14 Najbolje uspješnosti modela za računanje društvene udaljenosti na temelju ranga *interakcije dobivene na skupu za testiranje uz optimalne koeficijente značajnosti parametara interakcije*

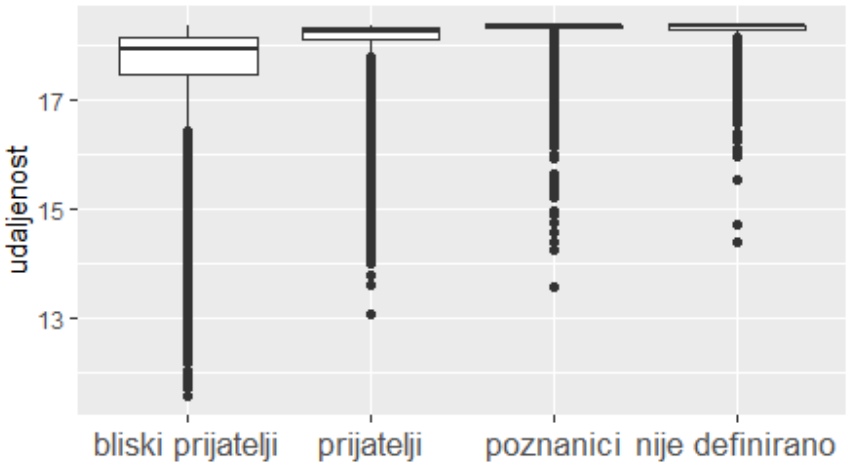
Uspješnosti na skupu za testiranje ponešto su niže od onih na skupu za učenje, ali te su razlike vrlo male i zanemarive. Uspješnosti različitih načina računanja udaljenosti međusobno su slične, izuzev uspješnosti za Mahalanobisovu udaljenost koja je značajno lošija od ostatka.

Uspješnost rada modela može se ispitati i za svakog ispitanika u skupu za testiranje zasebno. Slika VI.15 prikazuje funkciju gustoće uspješnosti po korisnicima.

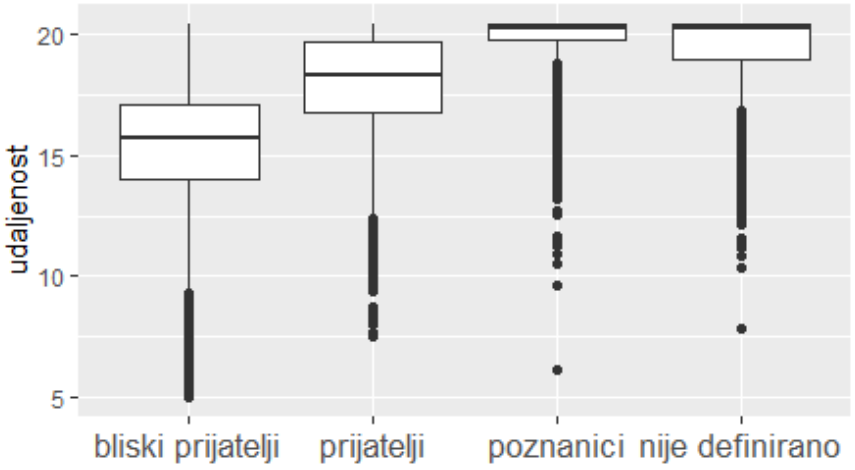


Slika VI.15 Funkcija gustoće uspješnosti rada modela zasnovanog na količini *interakciji i euklidskoj udaljenosti po korisniku*

Ako se pogleda kako su udaljenosti raspoređene po skupinama prijateljstva, može se uočiti da su od ispitanika najmanje udaljeni njegovi *bliski prijatelji*, nešto su udaljeniji *prijatelji*, a najudaljeniji su *poznanici*. Slika VI.16 kutijastim dijagramom prikazuje razdiobu udaljenosti po skupinama prijateljstva za zasnovan na *količini* interakcije i primjeni euklidske udaljenosti. Slika VI.17 prikazuje razdiobu udaljenosti za model za računanje udaljenosti zasnovan na *rangu* interakcije i primjeni euklidske udaljenosti. U oba se slučaja uočava postojanje udaljenosti karakterističnih za neku vrstu prijateljstva, ali su one u različitim intervalima. Kod modela zasnovanog na *količini* interakcije udaljenosti su zgusnutije u odnosu na model zasnovan na *rangu* interakcije.

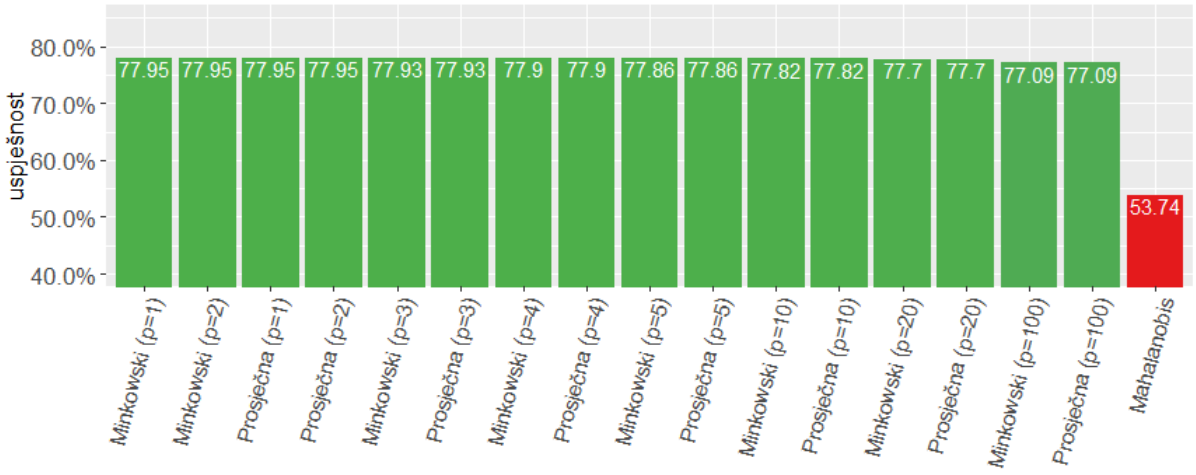


Slika VI.16 Udaljenosti po skupinama prijateljstva za model zasnovane na količini *interakcije te primjeni euklidske udaljenosti*

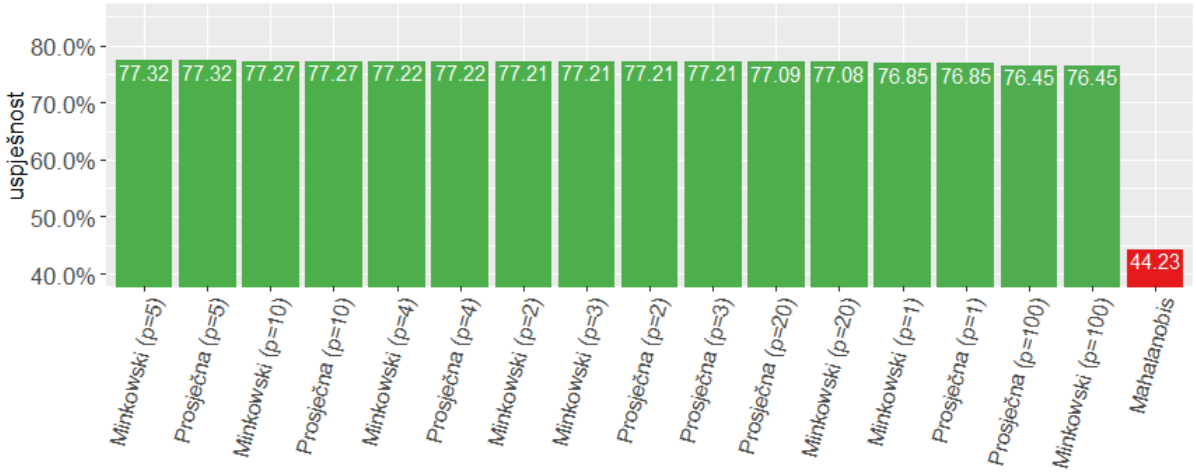


Slika VI.17 Udaljenosti po skupinama prijateljstva za model zasnovane na rangu *interakcije te primjeni euklidske udaljenosti*

Kako bismo provjerili koliko je važno (i je li uopće važno) odrediti koeficijente značajnosti pojedinih parametara interakcije, na skupu za testiranje provjerili smo uspješnosti rada modela za slučaj kada su koeficijenti značajnosti svih modela postavljeni na jednaku vrijednost, tj. kada ih se ne uzima u obzir. Slika VI.18 prikazuje uspješnosti za modele koji ne uzimaju obzir koeficijente značajnosti, a rade na temelju *količine* interakcije, a Slika VI.19 modele koji rade na temelju *ranga* interakcije. Može se uočiti da modeli koji ne uzimaju u obzir koeficijente značajnosti parametara interakcije rade lošije u odnosu na modele koji te parametre uzimaju u obzir, ali i da istodobno, u smislu uspješnosti, rade bolje od modela koji rade na temelju samo jednog parametra interakcije. Takvi rezultati potvrđuju potrebu korištenja više parametara interakcije pri računanju društvene udaljenosti kao i potrebu za određivanjem odgovarajućih koeficijenata značajnosti pojedinih parametara interakcije.



Slika VI.18 Uspješnosti modela koji ne uzimaju u obzir koeficijente značajnosti parametara interakcije, a udaljenosti određuju na temelju količine *interakcije*



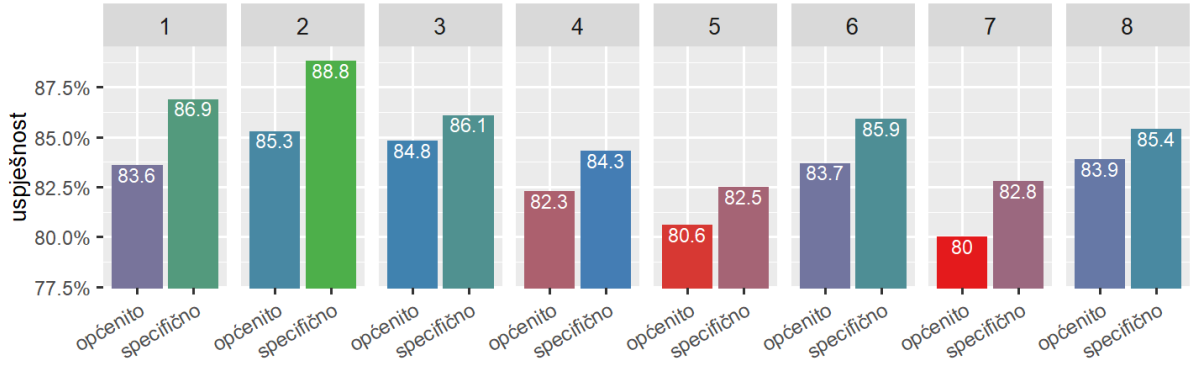
Slika VI.19 Uspješnosti modela koji ne uzimaju u obzir koeficijente značajnosti parametara interakcije, a udaljenosti određuju na temelju ranga *interakcije*

Kod modela zasnovanih na samo jednom parametru interakcije zasebno je analizirana uspješnost modela i zasebno njegova pouzdanost. S obzirom na to da su modeli radili na temelju samo jednog parametra interakcije, često na temelju izračuna modela nije bilo moguće odrediti boljeg prijatelja u paru. Zbog toga je često bila velika razlika između uspješnosti i pouzdanosti. Za modele koji uzimaju u obzir više parametara interakcije, pouzdanost se ne analizira zasebno jer je broj slučajeva u kojima na temelju izračuna modela nije moguće odabrati boljeg prijatelja u paru, s obzirom na velik broj razmatranih parametara, vrlo mali, tj. vrijednost pouzdanosti je (gotovo) u potpunosti jednaka vrijednosti uspješnosti.

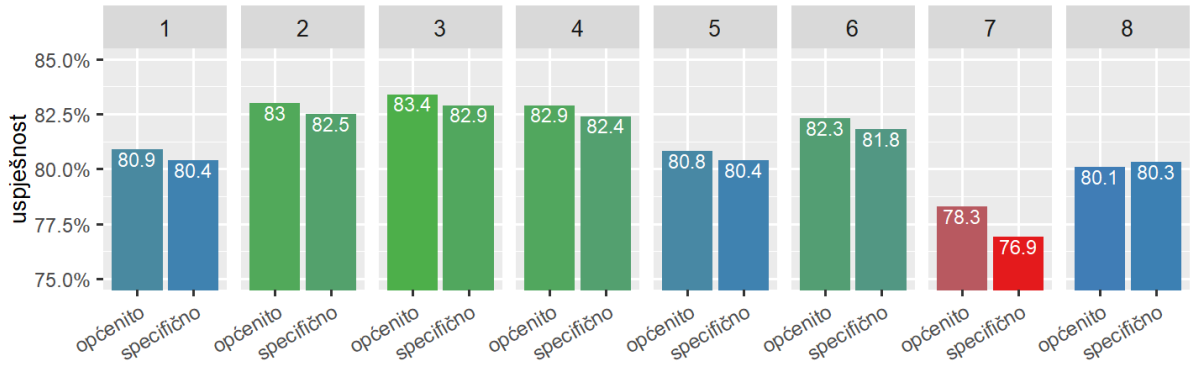
Kada se uspoređuju uspješnosti modela s jednim parametrom i modela s više parametara, može se uočiti da su modeli s više parametara značajno uspješniji. Međutim, kada se gleda pouzdanost, pokazuje se da su modeli koji rade na temelju parametra koji bilježi je li

promatrani prijatelj označen kao blizak s ispitanikom (*is_close_friend*) te modeli koji računaju udaljenost na temelju broja razmijenjenih privatnih poruka (*inbox_chat*) pouzdaniji od modela koji uzimaju u obzir više parametara interakcije. Međutim, problem sa spomenutim modelima koji rade na temelju samo jednog parametara interakcije jest u tome što odluku mogu donijeti u svega 11,76% ili 70,04%, za razliku od modela koji rade na temelju više parametara interakcije, a koji odluku mogu donijeti u gotovo 100% slučajeva.

U istraživanju *BestFriends* (opisano u potpoglavlju IV.1) uvedena je ideja da parametri interakcije nemaju jednak značaj za sve korisnike Facebooka, već da se koeficijent značajnosti sastoji od globalnog (univerzalnog) koeficijenta značajnosti pojedinog ispitanika i za ispitanika specifičnog koeficijenta značajnosti. Naknadne provjere modela predloženog u istraživanju *BestFriends* pokazale su da je doprinos specifičnog koeficijenta značajnosti uspješnosti rada modela zanemariv, tj. da je uspješnost rada modela ista i kada se iz modela u potpunosti isključi model specifične značajnosti. Ipak, ideja da parametri interakcije nisu jednako značajni za sve korisnike i da ovise o navikama korisnika, koje su različite o čemu je više bilo riječi u potpoglavlju IV.3, i dalje je zanimljiva. Umjesto da se traže specifični koeficijenti interakcije za svakog korisnika zasebno, odlučeno je da se korisnici podijele u skupine (klastere) s obzirom na svoje interakcijske navike (podjela korisnika u skupine detaljno je prezentirana u odjeljku V.1.3). Za svaku skupinu korisnika zasebno se proveo postupak učenja modela, tj. tražili su se parametri interakcije uz koje se dobiva najveća uspješnost na skupu za učenje za pojedini klaster. U nastavku će biti prikazani rezultati dobiveni primjenom modela koji se zasniva na euklidskoj udaljenosti i *količini* interakcije. Univerzalna uspješnost toga modela na skupu za učenje (prema Slika VI.7) iznosi 83,01%, a na skupu za testiranje (prema Slika VI.13) iznosi 81,66%. Slika VI.20 prikazuje rezultate za pojedine klastere dobivene na skupu za učenje, a Slika VI.21 rezultate dobivene na skupu za testiranje.



Slika VI.20 Uspješnosti po klasterima za računanje društvene udaljenosti primjenom euklidske udaljenosti na temelju količine *interakcije na skupu za učenje*

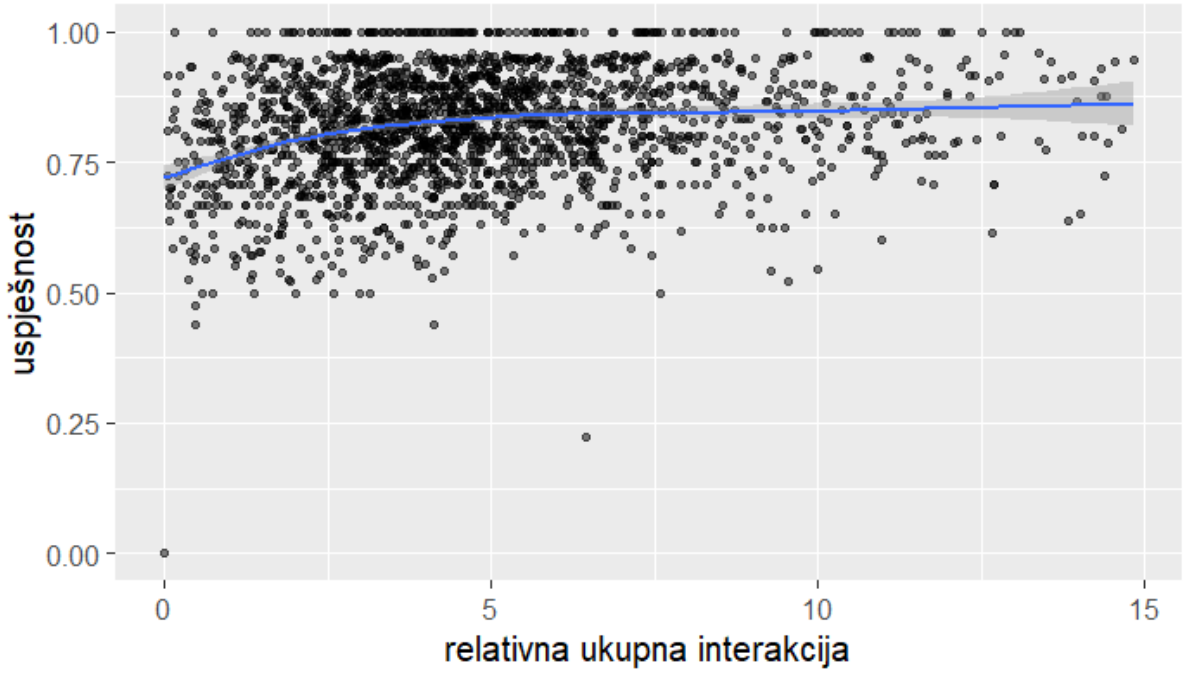


Slika VI.21 Uspješnosti po klasterima za računanje društvene udaljenosti primjenom euklidske udaljenosti na temelju količine *interakcije na skupu za testiranje*

Iako su rezultati na skupu za učenje djelovali ohrabrujuće, tj. davali su naznake da bi se hipoteza da će modeli izrađeni specifično za pojedini klaster za taj klaster raditi bolje od univerzalnih modela, na skupu za testiranje dobiveni su suprotni rezultati. Pokazalo se da je porast uspješnosti na skupu za učenje rezultat je pretreniranosti modela. Pretreniranost modela podrazumijeva da su izgrađeni modeli prilagođeni točno onim jedinkama nad kojima je provođeno učenje, a ne grupi jedinki koju su jedinke u skupu za učenje predstavljale. Dakle, iz dobivenih rezultata može se zaključiti da izrada zasebnih modela za svaku skupinu korisnika ne daje željene rezultate, tj. da ne dovodi do povećanja uspješnosti rada modela. Međutim, iako se pokazalo da društvenu udaljenost za sve korisnike, neovisno o klasteru u koji su svrstani s obzirom na svoje interakcijske navike, treba računati istim modelom, odvojeno računanje uspješnosti za različite klastere pokazalo je da je društvenu udaljenost za različite klastere moguće izračunati uz različitu razinu prosječne uspješnosti. Najniža uspješnost dobivena je za korisnike koji pripadaju klasteru 7. Kada se pogleda Slika V.4, može se uočiti da su u klaster 7 razvrstani korisnici koji imaju naglašeniju interakciju prema svojim prijateljima, u odnosu na intenzitet interakcije njihovih prijatelja prema njima. Možda se lošija uspješnost rada modela za taj klaster može objasniti time što je vjerojatno da se radi o korisnicima koji neselektivno označavaju objave drugih ljudi sa *sviđa mi se* pa je i u podatcima o interakciji manje prisutna informacija o intenzitetu njihova odnosa s prijateljima, a više je prisutan šum. Model je ispodprosječno uspješan i za 5. i 8. klaster kod kojih su također u manjoj mjeri zastupljene interakcije potaknute od strane prijatelja. S druge strane, najveća je uspješnost modela za klaster 2, u kojem je dominantan parametar interakcije broj razmijenjenih privatnih poruka, te za klaster 3 u kojem se korisnik vrlo često pojavljuje na zajedničkim fotografijama sa svojim prijateljima.

Analiza uspješnosti rada modela provedena je u ovisnosti o ukupnoj interakciji korisnika. Početno postavljena hipoteza bila je da će uspješnost rada modela biti veća za one

korisnike (ispitanike) koji više koriste Facebook, tj. kod kojih je veća ukupna interakcija. Kako bi se ukupna interakcija korisnika svela na jedan broj, odlučeno je da će se ukupna interakcija pojedinog korisnika prikazati točkom N-dimenzionalnog prostora, pri čemu će svaka dimenzija predstavljati njegovu ukupnu interakciju jednim interakcijskim parametrom. Ukupna interakcija pojedinim parametrom interakcije prikazuje se relativno, tj. kao omjer ukupne interakcije promatranog korisnika predmetnim parametrom interakcije i prosječne ukupne interakcije promatranim parametrom za sve ispitanike u skupu. Ukupna interakcija pojedinog korisnika jednaka je euklidskoj udaljenosti točke koje predstavlja korisnika i ishodišta koordinatnog sustava. Ako na osi *x* prikažemo ukupnu interakciju korisnika (izračunatu na gore opisan način), a na osi *y* uspješnost rada modela za promatranog korisnika, dobivamo graf prikazan na Slika VI.22. Na grafu su ispitanici prikazani točkicama. Plava crta predstavlja trend promjene uspješnosti u ovisnosti o promjeni relativne ukupne interakcije. Sivi omot plave crte predstavlja 95% interval pouzdanosti.



Slika VI.22 Promjena uspješnosti s promjenom ukupne interakcije

Na Slika VI.22 uočava se blagi porast uspješnosti rada korisnika s porastom relativne ukupne interakcije, tj. potvrđuje se hipoteza da će modeli bolje raditi za korisnike koji više koriste Facebook.

Zaključak

Verifikacija modela za računanje društvene udaljenosti očekivano je dala nešto niže uspješnosti u odnosu na one dobivene na skupu za učenje. Pokazalo se da različite vrste

Minkowskijevih udaljenosti i prosječnih Minkowskijevih udaljenosti daju slične rezultate neovisno o vrijednosti parametara *p*. Pokazalo se i da modeli koji uzimaju u obzir više parametara interakcije bolje rade kada im se pridruže eksperimentalno dobiveni koeficijenti značajnosti parametara interakcije, nego kada ih nema, tj. kada su svi postavljeni na jednaku vrijednost. Dodatno je provjereno treba li za različite skupine korisnika s obzirom na njihove interakcijske navike (u kojem udjelu koriste koji interakcijski parametar) tražiti specifične koeficijente značajnosti parametara interakcije. Pokazalo se da to nije potrebno, tj. da takvi modeli ne rade ništa bolje od modela koji koriste općenite koeficijente značajnosti. Međutim, pokazalo se da modeli daju različite prosječne uspješnosti za različite skupine korisnika pa se na temelju pripadnosti korisnika klasteru može preciznije odrediti očekivana uspješnost rada modela za njega. Provjereno je utječe li ukupna količina interakcije korisnika na Facebooku na uspješnost rada modela za njega. Pokazalo se da modeli slabije rade za korisnike koji imaju manje ukupne interakcije na Facebooku, a bolje za one korisnike koji više koriste Facebook.

VI.2 VERIFIKACIJA MODELA OBOGAĆENOG DRUŠTVENOG GRAFA

Model obogaćenog društvenog grafa uveden je i formalno opisan u poglavlju III. Izgradnja obogaćenog društvenog grafa zasniva se na izračunu društvene udaljenosti između ljudi za različite vrste njihovih odnosa. Najvažniji dio procesa izrade obogaćenog društvenog grafa jest izrada modela za računanje društvene udaljenosti. Društvena udaljenost u kontekstu ovog doktorskog rada računa se za odnos *prijateljstva* na temelju interakcije korisnika na sustavu za društveno umrežavanje Facebook. Osnovni preduvjet za izradu modela za računanje društvene udaljenosti jest prikupljanje referentnog podatkovnog skupa. Referentni podatkovni skup u sebi treba sadržavati podatke o interakciji korisnika na Facebooku te korisnikovu vlastitu (subjektivnu) procjenu njegova odnosa s njegovim mrežnim prijateljima. Korisnikova procjena odnosa s njegovim prijateljima smatra se činjeničnim stanjem (engl. *ground truth*) te je cilj izraditi model koji će na temelju podataka o interakciji moći davati procjene čim sličnije onima koje je dao korisnik.

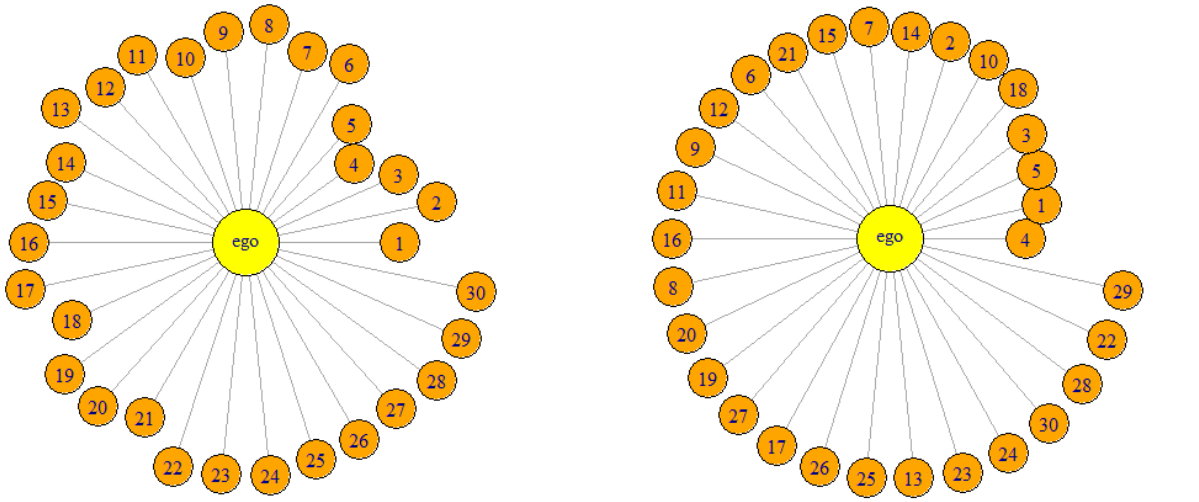
U cilju boljeg razumijevanja odnosa ljudi na Facebooku i prikupljanja referentnog podatkovnog skupa, provedeno je nekoliko društvenih istraživanja. Provedena društvena istraživanja opisana su u poglavlju IV. Za izradu modela za računanje društvene udaljenosti korišten je referentni podatkovni skup prikupljen istraživanjem *NajFrend*, tj. prvenstveno njegov dio u kojem su korisnici trebali odabrati boljeg u ponuđenom paru njihovih prijatelja. U potpoglavlju VI.1 opisan je proces izrade različitih modela za računanje društvene udaljenosti.

Rad svih izrađenih modela verificiran je na skupu za testiranje koji je disjunktan sa skupom nad kojim se provodilo učenje modela.

S obzirom na prirodu referentnog podatkovnog skupa koji je skup ego-grafova, verifikacija modela obogaćenog društvenog grafa također je provedena kroz ego-grafove.

Vizualizacija ego-grafova obavljena je pomoću paketa *igraph* u R-u. U svrhu vizualnog dočaravanja udaljenosti, paket *igraph* proširen je funkcijom koja na temelju udaljenosti određuje položaj svakog prijatelja u odnosu na ispitanika. Udaljenost prijatelja od ego-korisnika (ispitanika) proporcionalna je njihovoj društvenoj udaljenosti. Omogućene su dvije vrste poretka prijatelja ispitanika: nasumičan i poredan s obzirom na društvenu udaljenost.

Slika VI.23 prikazuje slučajno odabranog ispitanika iz skupa za testiranje te 30 njegovih nasumično odabranih prijatelja. Lijevo na slici prijatelji su poredani nasumično, a desno je poredak određen na temelju izračunate društvene udaljenosti. Udaljenost svakog prijatelja od ego-korisnika (ispitanika) proporcionalna je njihovoj modelom izračunatoj društvenoj udaljenosti. Za računanje društvene udaljenosti korišteni su podatci o *rangu* interakcije te euklidska udaljenost. U primjeru sa slike, prijatelj koji je društveno najmanje udaljen od ispitanika jest čvor 4, dok je čvor 29 najudaljeniji.



Slika VI.23 Vizualizacija ego-grafa

VII. ULOGA SINTETIČKIH PODATAKA U BUDUĆIM ISTRAŽIVANJIMA DRUŠTVENIH MREŽA

U ovom će poglavlju biti dan osvrt na trend povećanog fokusa na zaštitu osobnih podataka koji je sveprisutan u zadnje vrijeme. Bit će objašnjeno što su to osobni podatci, zašto je bitno njima pažljivo upravljati te kakve posljedice obveza povećanog opreza pri radu s osobnim podatcima može imati na analizu društvenih mreža. Kako bi se uklonio rizik rada s osobnim podatcima, tj. s podatcima stvarnih osoba, razmatra se mogućnost korištenja umjetnih (sintetičkih) podataka te se predlaže idejno rješenje generatora sintetičkih podataka o interakciji korisnika na sustavu za društveno umrežavanje, tj. generiranje sintetičkog proširenog društvenog grafa.

VII.1 ŠTO SU TO OSOBNI PODATCI I ZAŠTO JE BITNO S NJIMA PAŽLJIVO UPRAVLJATI?

Tema zaštite osobnih podataka (marginalno) je prisutna u zakonodavstvu više od dva desetljeća, a u svijesti ljudi možda nekoliko godina. Veliki pomak u tom području dogodio se kada je, nakon višegodišnje rasprave, 27. travnja 2016. godine donesena *Uredba (EU) 2016/679 Europskog parlamenta i vijeća o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka te o stavljanju izvan snage Direktive 95/46/EZ (Opća uredba o zaštiti podataka – GDPR)* [40]. Nakon proteka jednogodišnjeg pripremnog razdoblja, uredba je 25. svibnja 2018. stupila na snagu[24]. Opća uredba o zaštiti podataka propisuje kazne do 20 milijuna eura ili 4% godišnjeg prometa (što je veće). Ilustracije radi, prethodno važeći hrvatski Zakon o zaštiti osobnih podataka propisivao je kazne do 40.000 kuna, a u praksi se gotovo uopće nije primjenjivao. Kazne koje propisuje GDPR bile su prvi pravi poticaj svima koji se bave obradom osobnih podataka da se zamisle nad time što rade, da prouče sadržaj Uredbe i nastoje joj se prilagoditi.

Opća uredba o zaštiti podataka prije svega ima zadaću promijeniti našu svijest o, kako obradi podataka općenito, tako i o obradi osobnih podataka[25]. Osobni podatak svaki je podatak

[24] Uredbe Europske unije imaju zakonodavnu snagu u svim zemljama članicama. Provedba Uredbe u Republici Hrvatskoj definirana je Zakonom o provedbi Opće uredbe o zaštiti podataka, koji je Hrvatski sabor donio 27. travnja 2018.

[25] Prema Općoj uredbi o zaštiti podataka [40], obrada podataka znači svaki postupak ili skup postupaka koji se obavljaju na osobnim podacima ili na skupovima osobnih podataka, bilo automatiziranim bilo neautomatiziranim sredstvima kao što su prikupljanje, bilježenje, organizacija, strukturiranje, pohrana, prilagodba ili izmjena, pronalaženje, obavljanje uvida, uporaba, otkrivanje prijenosom, širenjem ili stavljanjem na raspolaganje na drugi način, usklađivanje ili kombiniranje, ograničavanje, brisanje ili uništavanje.

koji se može dovesti u vezu s identificiranom osobom ili osobom koju se s velikom vjerojatnošću može identificirati. Osobu se može identificirati na temelju njezinih identifikatora (ime, prezime, OIB, broj bankovnog računa i slično), ali i iz konteksta. Primjerice, ako gledamo snimku s vjenčanja ženske osobe A i muške osobe B i na njoj se u daljini vidi kako osoba u velikoj bijeloj haljini pleše, s velikom vjerojatnošću možemo zaključiti da je to osoba A iako to nigdje ne piše, a ne vidimo čak ni crte njezina lica. Tada je i taj isječak snimke osobni podatak jer otkriva da je osoba A (iako tek neizravno identificirana) plesala.

Opća uredba o zaštiti podataka potaknula je veliku promjenu razmišljanja i shvaćanja. Proživljavamo transformaciju iz svijeta u kojem je cilj bio prikupiti čim više (često osobnih) podataka, u kojem iz dana u dan rastu kapaciteti pohrane i u kojem pohranjujemo sve što možemo jer bi nam možda jednom moglo zatrebati, u svijet u kojem pohranjujemo samo ono za što u trenutku pohrane znamo svrhu pohrane i to čuvamo samo dok ta svrha pohrane traje. To iziskuje velike promjene u našem shvaćanju jer postupci koji su do sada bili uobičajeni i općeprihvaćeni postaju neprihvatljivi ili značajno otežani. Naravno, to ruši i same temelje dubinske analize podataka koja je po svojoj definiciji sekundarna obrada i kojoj je cilj *izvući* neke informacije iz podataka koji su prikupljeni u neku posve drugu svrhu. Ta je transformacija svima nama koji se bavimo obradom podataka vrlo teška, ali nužna. Godinama se radilo na tome da se osmisle principi i metode koje će omogućiti izlučivanje čim više informacija iz dostupnih podataka i u tome nije postojala izraženija svijest o opasnostima koje to može generirati. U cilju postizanja napretka u velikoj su mjeri ignorirane moguće zlouporabe. Razvijene metode, pogotovo metode umjetne inteligencije, s vremenom su postale toliko moćne da su postale opasne. Velike korporacije poput Facebooka ili Googlea došle su u posjed ogromnih količina podataka o ljudima diljem svijeta. Do zastrašujućih podataka. Znaju što čitamo na Internetu, gdje boravimo, što volimo, čime se bavimo i slično. To samo po sebi nije problem dok se koristi za poštene namjene. Primjerice, prikupljeni podatci koji se koriste kako bi se korisnicima preporučivali sadržaji koji ih zanimaju korisni su i za korisnika i za onoga tko mu sadržaj nudi. Problem nastaje kada se te podatke počinje zloupotrebljavati, tj. kada ih se počinje koristiti na štetu korisnika kojem pripadaju. Primjer takve zlouporabe jest slučaj *Cambridge Analytice* u kojem su spoznaje o korisnicima Facebooka iskorištene kako bi ih se kroz ciljano plasirane dezinformacije potaknulo da svoj glas daju određenom kandidatu na izborima u SAD-u. Postoji u svijetu i vrlo uhodana praksa prodaje osobnih podataka. To samo po sebi opet možda nije nužno problem, ali postaje kada se tako prikupljeni podatci koriste protivno interesima korisnika. Primjerice, osiguravajuća kuća od zdravstvenih ustanova dobije podatke o zdravstvenom stanju njezinih pacijenata i na temelju toga bolesnijim ljudima nudi

skuplje police osiguranja. Tada se nečiji osobni podatci koriste protivno njegovim interesima. To je vrlo ozbiljan problem jer ruši temelje osiguranja i njegova osnovna načela solidarnosti u kojoj osiguranje plaćaju svi kako bi se skupilo dovoljno novaca za liječenje onih koji su ozbiljno bolesni, a koji sami toliki trošak ne bi mogli podnijeti.

S obzirom na sve raširenu praksu zlouporabe osobnih podataka, bilo je nužno donijeti zakonsku regulativu koja vrlo striktno propisuje kako se osobni podatci smiju obrađivati. Takva zakonska regulativa, nažalost, jednako se odnosi na one koji osobne podatke zloupotrebljavaju kao i na one koji to ne rade. Uredba s jedna strane štiti pojedince čije se osobni podatci obrađuju, a s druge strane jako komplicira procese što je (bespotreban) problem za sve one koji osobne podatke ne zloupotrebljavaju.

VII.2 ANALIZA DRUŠTVENIH MREŽA I GDPR

U kontekstu analize društvenih mreža, zaštita osobnih podataka znači uvođenje puno restriktivnijih mjera vezanih za prikupljanje i obradu osobnih podataka. Društvene se mreže nakon stupanja na snagu GDPR-a, naravno, i dalje smiju analizirati, ali za to mora postojati eksplicitna privola ispitanika. Osim toga, posebnu pažnju treba posvetiti čuvanju osobnih podataka što podrazumijeva poduzimanje razumnih sigurnosnih mjera kako bismo povjerene nam osobne podatke sačuvali od mogućeg kompromitiranja, tj. stavljanja na raspolaganje osobama koje bi ih iskoristile suprotno namjeni za koju su prikupljani. To prije svega znači da se prikupljeni skupovi osobnih podataka ne smiju činiti javno dostupnima. Međutim, u znanstvenim istraživanjima često je prisutna potreba za takvim podatcima – bilo za potrebe razvoja novih algoritama bilo za potrebe proučavanja svojstava društvenih mreža u sociologiji bilo za neke treće istraživačke potrebe. Jedno moguće, ali ujedno i vrlo složeno rješenje jest u prikupljanju zasebnih podatkovnih skupova u svakoj istraživačkoj grupi. Srećom, postoje i ponešto izvedivije alternative kojima se može doskočiti spomenutom problemu:

1. provođenjem posebnih mjera anonimizacije podataka prije njihova dijeljenja [64]–[68]
2. generiranjem sintetičkih podatkovnih skupova [69]–[76].

U praksi uobičajeno korišteni pristupi anonimizacije podataka podrazumijevaju samo uklanjanje imena i demografskih podataka, što se pogrešno smatra uklanjanjem identifikatora [65], [77]. To nije dovoljno jer je korisnike moguće reidentificirati na temelju poznavanja njihove povezanosti s ostalim ljudima [65] ili kroz povezivanje s nekim drugim javno dostupnim podatcima [77]. Primjer takve deanonimizacije jest poznat Netflixov slučaj [78].

Naime, Netflix je 2006. godine javno objavio tobože anonimizirani podatkovni skup s podatcima o svojih 100 milijuna korisnika. Iz tog su skupa bili uklonjeni tek atributi za koje se mislilo da su jedini identifikatori u skupu. Ubrzo nakon objave skupa, Narayanan i Shmatikov uspjeli su reidentificirati podskup toga skupa povezujući podatke s Netflixa s javnim ocjenama na IMDB.comu [79]. Nakon što su uspješno deanonimizirali podskup Netflixovih podataka, Narayanan i Shmatikov još su davne 2009. predložili generički pasivni algoritam deanonimizacije anonimiziranih društvenih grafova kroz povezivanje društvenih grafova s različitih sustava za društveno umrežavanje [77]. U narednim su godinama predlagani mnogi novi pristupi anonimiziranju podataka sa sustava za društveno umrežavanje koji bi trebali otežati njihovu deanonimizaciju. Izvrstan pregled različitih pristupa anonimizaciji i pregled napada kojima je cilj deanonimizacija podataka može se pronaći u [80]. Pri anonimiziranju podataka treba biti svjestan da snažna anonimizacija dovodi do veće zaštite privatnosti, ali i do smanjenja korisnosti podataka [80].

Međutim, iako postoje brojni pristupi anonimiziranju podataka, napadači su uvijek korak ispred i neodgovorno bi bilo tvrditi da postoji neprobojna anonimizacija podataka. S obzirom na to, umjesto (anonimiziranih) izvornih podataka, ponekad se koriste sintetički podatci. Sintetički podatci svakako su sigurnije rješenje jer se izbjegava rad s osobnim podatcima i ne postoji rizik od narušavanja privatnosti. Međutim, izazov je izvesti generator koji će biti u mogućnosti generirati sintetičke podatkovne skupove s jednakim svojstvima i mogućnostima primjene kao empirijski podatci.

Kroz godine su implementirani različiti generatori sintetičkih društvenih grafova [69]–[76]. Zadaća tih generatora jest generirati sintetički društveni graf koji je po svojim svojstvima čim sličniji empirijskom društvenom grafu. Zajedničko svim generatorima društvenih grafova dostupnim u literaturi jest generiranje binarnih društvenih grafova, tj. društvenih grafova koji bilježe tek informaciju o tome koji su čvorovi međusobno povezani. Nije pronađen ni jedan rad koji bi predlagao kako proširiti sintetički binarni društveni graf s granama koje opisuju interakciju između korisnika. Zbog toga će u nastavku biti predloženo idejno rješenje za generator sintetičkog proširenog društvenog grafa i na taj način napravljen prvi korak u njegovu stvaranju. U kontekstu povećanog fokusa na zaštitu osobnih podataka, izrada generatora sintetičkih podataka vrlo je važna za daljnji razvoj ovog područja.

VII.3 GENERIRANJE SINTETIČKOG PROŠIRENOG DRUŠTVENOG GRAFA

U ovom će odjeljku prvo biti dan pregled dosadašnjih istraživačkih nastojanja na području generiranja sintetičkih podatkovnih skupova, a zatim će biti predstavljeno idejno rješenje generatora sintetičkog proširenog društvenog grafa, koje kombinira i nadograđuje postojeće pristupe generiranju sintetičkih podatkovnih skupova.

VII.3.1 Generiranje sintetičkih podatkovnih skupova

Analiza društvenih mreža nije jedino područje u kojem podatkovni skupovi potrebni u svrhu provođenja istraživanja nisu u dovoljnoj mjeri dostupni znanstvenoj zajednici. Primjerice, u medicini je tek vrlo ograničena količina podataka za potrebe dubinske analize podataka javno dostupna zbog zaštite privatnosti pacijenata [81]. U vojnom sektoru također se najveći dio podataka smatra vojnom tajnom i ne smije se javno objavljivati [82]. Općenito gledano, razlozi za nedostupnost podatkovnih skupova mogu biti sljedeći:

1. Podatci su inherentno rijetki (rijetke bolesti, rijetke gramatičke strukture i slično)
2. Podatci nisu dostupni zbog vlasničkih sustava, povjerljivosti poslovnih ugovora, privatnosti zapisa i slično
3. Podatci su skupi (mogu se dobiti korištenjem skupe opreme koja zahtijeva značajna ulaganja ljudskih ili materijalnih resursa)
4. Distribucija događaja od interesa vrlo je neuravnotežena (otkrivanje prijevara, otkrivanje iskočnica u podatcima, razdiobe s dugim repovima i slično) [83].

Iako neki problemi (primjerice, inherentno rijetki podatci) ne mogu biti jednostavno riješeni, nekim drugima može se doskočiti kroz generiranje sintetičkih podatkovnih skupova sa svojstvima čim sličnijima empirijskim podatcima. Kada empirijski podatkovni skupovi nisu javno dostupni ili nisu dostupni u dovoljnoj količini, a generiranje sintetičkih podatkovnih skupova je moguće, tada sintetički podatkovni skupovi jesu izvediva i nužna alternativa.

Pri generiranju sintetičkih podatkovnih skupova u literaturi se mogu pronaći tri osnovna pristupa:

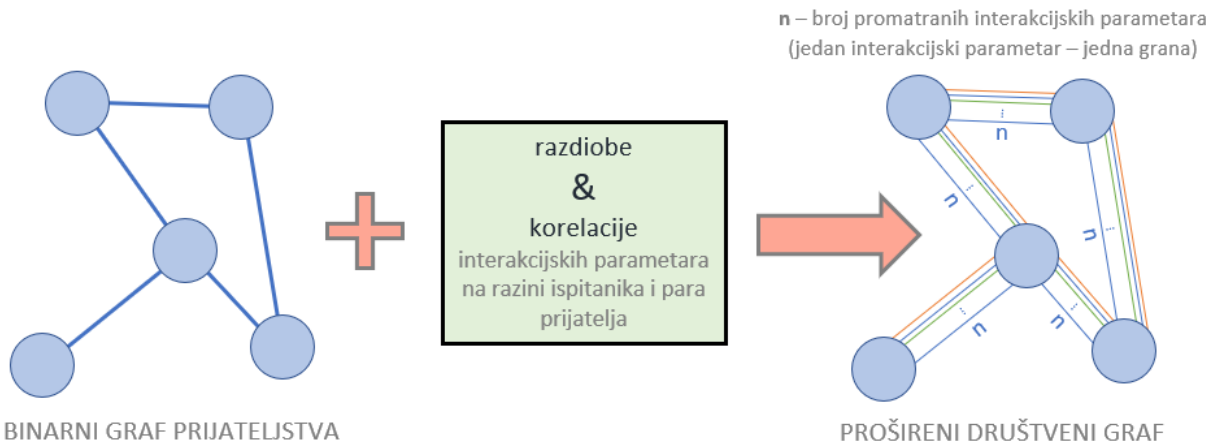
1. Koristiti mali uzorak dostupnih podataka i umnožiti ga pomoću slučajne permutacije ili sličnog algoritma uz dodavanje šuma u podatke [84], [85]
2. Iskoristiti dostupan podatkovni skup kao ulaz u (ne)nadzirane algoritme strojnog učenja kao što su, primjerice, neuronske mreže, stabla odlučivanja, skriven

Markovljev model, koji otkrivaju sakrivena svojstva podataka i generiraju prošireni podatkovni skup kao rezultat [81], [83], [86]

3. Generirati novi podatkovni skup temeljem poznatih ključnih svojstava empirijskih podataka [82], [87], [88].

Generiranje sintetičkih društvenih grafova ima nekih posebnosti u odnosu na generiranje "običnih" sintetičkih podatkovnih skupova. Kada bi se graf društvene mreže generirao samo na temelju poznate razdiobe stupnja čvora (broja prijatelja), dobili bismo prilično homogenu mrežu bez nekih karakterističnih obilježja empirijskih društvenih mreža. Mnoga se istraživanja zato bave izradom modela za generiranje društvenih grafova [69]–[76]. Zajednički cilj svima jest generirati sintetički društveni graf koji ima karakteristike čim sličnije empirijskom društvenom grafu, ali pristupi su različiti. Neki modeli uzimaju u obzir demografske značajke kao ključne za povezivanje čvorova, neki geografsku blizinu, a neki dinamiku mrežnih struktura kao što su trijade. Međutim, svi u literaturi pronađeni modeli generiraju tek binarne društvene grafove.

Izrada generatora sintetičkog proširenog društvenog grafa vrlo je izazovan i složen proces. U nastavku će biti predstavljeno idejno rješenje generatora sintetičkog fejsbukovskog proširenog društvenog grafa koji kao ulaz koristi binarni graf prijateljstva te informacije o razdiobama i korelacijama fejsbukovskih interakcijskih parametara na razini ispitanika i para prijatelja. Na temelju tih ulaza, predloženo idejno rješenje generira prošireni društveni graf u kojem je, uz postojeće binarne grane prijateljstva, između svaka dva povezana čvora dodana po jedna nova grana za svaki od promatranih interakcijskih parametara. Uz 1 binarnu vezu između dva čvora, generirani sintetički graf imat će N veza s kontinuiranim težinama, pri čemu je N broj promatranih parametara interakcije (Slika VII.1).



Slika VII.1 Osnovna ideja predloženog generatora proširenog društvenog grafa

VII.3.2 Idejno rješenje za generator proširenog društvenog grafa

Na početku poglavlja obrazložili smo potrebu za stvaranjem sintetičkih podatkovnih skupova, pogotovo u slučajevima kada izvorni podatkovni skupovi sadrže osobne podatke. Prezentirali smo i različite pristupe generiranju sintetičkih podatkovnih skupova. Sintetički podatkovni skup s podatcima o interakciji na sustavu za društveno umrežavanje, ili sintetički prošireni društveni graf, specifičan je slučaj sintetičkih podataka pa se ne može u potpunosti primijeniti neki od postojećih pristupa generiranju sintetičkih podatkovnih skupova. U ovom odjeljku prezentirat ćemo idejno rješenje generatora sintetičkog proširenog fejsbukovskog društvenog grafa. Predloženo rješenje u sebi na odgovarajući način kombinira postojeća rješenja i nadodaje dijelove specifične za proširenih društveni graf. Proces generiranja sintetičkog proširenog društvenog grafa može se podijeliti u dva osnovna dijela:

1. Prikupljanje i analiziranje empirijskih podataka s ciljem izlučivanja njihovih ključnih svojstava
2. Generiranje sintetičkog proširenog društvenog grafa na temelju prethodno izlučenih ključnih svojstava empirijskih podataka

Slika VII.2 prikazuje korake idejnog rješenja za proces izrade generatora sintetičkog proširenog fejsbukovskog društvenog grafa.

Početni korak u osmišljavanju idejnog rješenja za izradu spomenutog generatora jest prikupljanje empirijskih podataka čiji se sintetički pandan planira generirati. U razmatranom se slučaju radi o podatcima o interakciji korisnika na Facebooku koji su prikupljeni u prethodno opisanom istraživanju *NajFrend* (potpoglavlje IV.2) i pohranjeni u obliku proširenog društvenog grafa. Sljedeći korak jest analiza prikupljenih podataka i ekstrakcija njihovih ključnih svojstava. S tim ciljem od prikupljenih podataka treba stvoriti dva podatkovna skupa – podatkovni skup s ukupnom interakcijom za svaki parametar interakcije na razini ispitanika te podatkovni skup s ukupnom interakcijom na razini para prijatelja. Naime, prvi skup pokazuje kako je ukupna interakcija raspoređena po korisnicima (neki korisnici koriste Facebook više, a neki manje), dok drugi skup pokazuje kako je raspoređena interakcija pojedinog korisnika s njegovim prijateljima (korisnici ne komuniciraju jednako frekventno sa svim svojim prijateljima). Za svaki od navedenih podatkovnih skupova potrebno je izlučiti razdiobe i korelacije za sve parametre odnosno parove parametara interakcije, tj. napraviti eksploratornu analizu podatkovnih skupova. Provedena eksploratorna analiza opisana je u poglavlju V.

Generator proširenog društvenog grafa na ulazu bi trebao primiti:

1. Binarni graf prijateljstva
2. Listu s ukupnim količinama interakcije za svaki interakcijski parametar, sumiranu na razini ego-korisnika (broj zapisa u listi treba biti jednak broju čvorova u binarnom grafu prijateljstva)
3. Informaciju o razdiobama i korelacijama između parametara interakcije na razini para prijatelja.

Prije generiranja sintetičkog proširenog društvenog grafa, treba odlučiti koliko čvorova graf koji se generira treba imati. Zatim treba korištenjem nekog od postojećih pristupa [69]–[76], [89] generirati binarni graf prijateljstva odabrane veličine. U [80] Siddula i Yingshu posebno predlažu algoritam topologije malog svijeta (engl. *Small World Topology algorithm*) [76] kao vrlo dobru zamjenu za empirijski binarni društveni graf sa sustava za društveno umrežavanje.

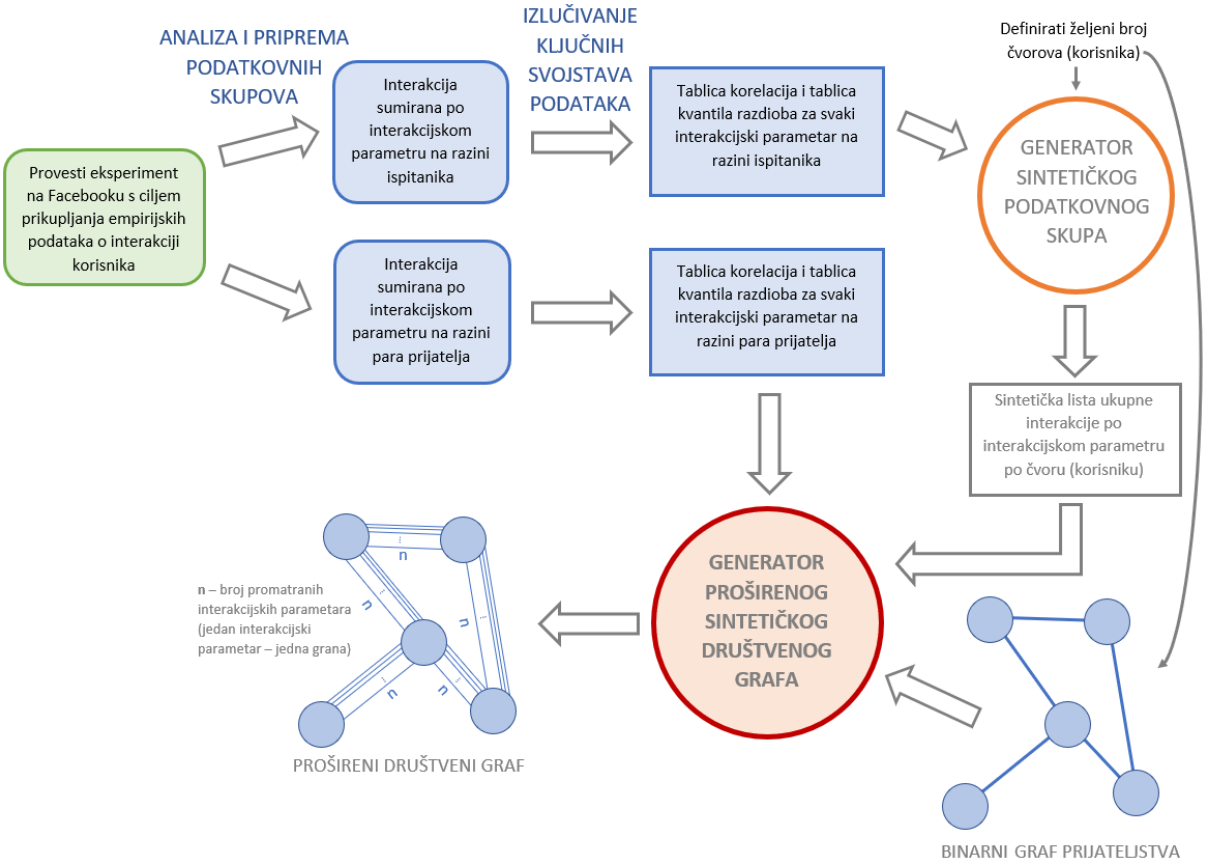
Kako bismo binarni graf prijateljstva pretvorili u prošireni društveni graf, potrebno je postojeće binarne grane zamijeniti novim granama koje će opisivati intenzitet interakcije među promatranim čvorovima. To treba napraviti u dva koraka. U prvom koraku potrebno je odrediti ukupne iznose interakcije po čvorovima u grafu, dok je u drugom te ukupne iznose potrebno razdijeliti na veze s prijateljima promatranog čvora. Kako bi u konačnici dobiveni sintetički prošireni društveni graf mogao po svojim svojstvima biti sličan empirijskom grafu, nužno je pri određivanju intenziteta interakcije među čvorovima voditi se svojstvima empirijskog skupa. U tu su svrhu izlučene razdiobe i korelacije dva prethodno opisana podatkovna skupa.

Na temelju poznate korelacijske matrice i kvantila distribucija interakcijskih parametara na razini ego-korisnika potrebno je generirati sintetički podatkovni skup koji će sadržavati podatke o ukupnoj interakciji za svakog ego-korisnika, tj. za svaki čvor u prethodno generiranom binarnom grafu prijateljstva. Dakle, treba generirati podatkovni skup koji će imati po jedan stupac za svaki interakcijski parametar te onoliko redaka koliko ima čvorova u prethodno generiranom binarnom grafu prijateljstva. Razdiobe svakog interakcijskog parametra, tj. svakog stupca u skupu trebaju biti jednake onima koje su prethodno eksperimentalno utvrđene, kao i korelacije među različitim interakcijskim parametrima, tj. stupcima u podatkovnom skupu. U tu svrhu predlažemo da se iskoristi algoritam koji je prezentiran i detaljno opisan u [87], a koji omogućuje generiranje sintetičkog podatkovnog skupa na temelju poznatih distribucija svakog atributa u skupu te poznate matrice korelacija

među atributima u skupu. Nakon što se generira opisani podatkovni skup, njegove retke treba nasumično pridružiti čvorovima binarnog grafa prijateljstva.

Poznate ukupne količine interakcije za svaki čvor u grafu treba razdijeliti na prijatelje čvora. Najjednostavniji pristup bio bi napraviti to jednoliko, ali posve je jasno da je to neživotno jer ljudi ne komuniciraju jednako sa svim svojim mrežnim prijateljima. Kako bi razdioba bila čim realističnija, predlažemo da se ukupna interakcija razdijeli na prijatelje promatranog korisnika poštujući eksperimentalno utvrđene (prezentirane u poglavlju V) distribucije pojedinih parametara i njihove međusobne korelacije između promatranog korisnika i svih njegovih prijatelja.

Primjenom opisanih koraka dobit ćemo sintetički prošireni društveni graf koji će imati svojstva interakcije među korisnicima slična onima koja imaju empirijski društveni grafovi.



Slika VII.2 Skica predloženog idejnog rješenja generatora sintetičkog proširenog društvenog grafa

Idejno rješenje generatora proširenog društvenog grafa predstavljeno je u radu [54].

ZAKLJUČAK

U sklopu ovog doktorskog rada pojam obogaćenog društvenog grafa uveden je, formalno opisan i verificiran. U odnosu na dosadašnje radove koji intenzitet odnosa između pojedinaca opisuju *težinom* ili *snagom* veze, u sklopu ovoga je rada uveden novi koncept *društvene udaljenosti* koji poznati koncept udaljenosti primjenjuje za izračun intenziteta odnosa među ljudima.

Posebna je pozornost pri izradi doktorskog rada posvećena provođenju društvenih istraživanja. Provedena su tri društvena istraživanja od koji se istraživanje *NajFrend* iz 2015. godine posebno ističe svojom sveobuhvatnošću (širinom postavljenih pitanja) i brojem ispitanika (više od 3.000). Podatci prikupljeni provedenim istraživanjima iskorišteni su za bolje razumijevanje interakcije ljudi na sustavima za društveno umrežavanje i povezanosti interakcije na tim sustavima s intenzitetom odnosa ljudi (u nevirtualnom životu) te za izradu i testiranje modela za računanje društvene udaljenosti.

Predloženi modeli za računanje društvene udaljenosti zasnivaju se na poznatim načinima računanja udaljenosti: Minkowskijevoj udaljenosti, prosječnoj udaljenosti te Mahalanobisovoj udaljenosti. Spomenuti, kao i mnogi drugi načini računanja udaljenosti, detaljno su opisani u početnom dijelu rada. Modeli za računanje društvene udaljenosti izrađeni su i testirani na temelju referentnog podatkovnog skupa oblikovanog na temelju podataka prikupljenih istraživanjem *NajFrend*. Pri izgradnji modela korišten je pristup nadziranog strojnog učenja kroz primjenu heurističke metode genetskih algoritama. Uspješnost rada modela prikazana je i analizirana.

U završnom poglavlju rada nalazi se komentar na trend povećanog fokusa na zaštitu osobnih podataka koji je prije svega potaknut nedavnim početkom pune primjene *Opće uredbe o zaštiti podataka (GDPR)*. S obzirom na porast rizičnosti rada sa stvarnim osobnim podatcima, predloženo je idejno rješenje za generiranje sintetičkog proširenog društvenog grafa, tj. za generiranje sintetičkog podatkovnog skupa s podatcima o interakciji korisnika na sustavima za društveno umrežavanje.

LITERATURA

[1] "Global social media ranking 2018 | Statistic," 2018. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed: 29-Dec-2018].

[2] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *J. Hum. Evol.*, vol. 22, no. 6, pp. 469–493, Jun. 1992.

[3] J. Ilić, L. Humski, D. Pintar, M. Vranić, and Z. Skočir, "Proof of Concept for Comparison and Classification of Online Social Network Friends Based on Tie Strength Calculation Model," in *ICIST 2016 Proceedings*, 2016, pp. 159–164.

[4] M. Majić, J. Skorin, L. Humski, and Z. Skočir, "Using the interaction on social networks to predict real life friendship," in *2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2014, pp. 378–382.

[5] M. Diaby, E. Viennet, and T. Launay, "Toward the next generation of recruitment tools: An Online Social Network-based Job Recommender System Mamadou," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013, pp. 821–828.

[6] L. Pappalardo, G. Rossetti, and D. Pedreschi, "'How Well Do We Know Each Other?': Detecting Tie Strength in Multidimensional Social Networks," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012, pp. 1040–1045.

[7] S. M. U. Khan and J. M. Shaikh, "Predicting students blood pressure by Artificial Neuron Network: Facebook predict students blood pressure," in *2014 Science and Information Conference*, 2014, pp. 430–437.

[8] D. Sun and W. C. Lau, "Social relationship classification based on interaction data from smartphones," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2013, pp. 205–210.

[9] M. Min, D. Choi, J. Kim, and J.-H. Lee, "The identification of intimate friends in personal social network," in *2011 International Conference on Computational Aspects of Social Networks (CASoN)*, 2011, pp. 233–236.

[10] M. Gupte and T. Eliassi-Rad, "Measuring tie strength in implicit social networks," in *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, 2012, pp. 109–118.

[11] R. Xiang, J. Neville, and M. Rogati, “Modeling Relationship Strength in Online Social Networks,” in *Proceedings of the 19th international conference on World wide web - WWW ’10*, 2010, pp. 981–990.

[12] X. Li, Q. Yang, X. Lin, S. Wu, and M. Wittie, “Itrust: interpersonal trust measurements from social interactions,” *IEEE Netw.*, vol. 30, no. 4, pp. 54–58, Jul. 2016.

[13] S. Krakan, L. Humski, and Z. Skočir, “Determination of Friendship Intensity between Online Social Network Users Based on Their Interaction,” *Teh. Vjesn. / Tech. Gaz.*, vol. 25, no. 3, pp. 655–662, 2018.

[14] I. Kahanda and J. Neville, “Using Transactional Information to Predict Link Strength in Online Social Networks,” in *Proceedings of the Third International Conference on Weblogs and Social Media - ICWSM 2009*, 2009, pp. 74–81.

[15] V. Podobnik, D. Štriga, A. Jandras, and I. Lovrek, “How to calculate trust between social network users?,” in *SoftCOM 2012, 20th International Conference on Software, Telecommunications and Computer Networks*, 2012, pp. 1–6.

[16] M. Stupalo, J. Ilić, L. Humski, Z. Skočir, D. Pintar, and M. Vranić, “Applying the binary classification methods for discovering the best friends on an online social network,” in *2017 14th International Conference on Telecommunications (ConTEL)*, 2017, pp. 155–162.

[17] N. Sever, L. Humski, J. Ilić, Z. Skočir, D. Pintar, and M. Vranic, “Applying the multiclass classification methods for the classification of online social network friends,” in *2017 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2017, pp. 1–6.

[18] E. Gilbert and K. Karahalios, “Predicting tie strength with social media,” in *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 2009, pp. 211–220.

[19] R. Michalski, P. Kazienko, and D. Krol, “Predicting Social Network Measures Using Machine Learning Approach,” in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012, pp. 1056–1059.

[20] A. Biancini, “Social Psychology Testing Platform Leveraging Facebook and SNA Techniques,” in *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, 2012, pp. 776–783.

[21] T. Correa, A. W. Hinsley, and H. G. de Zúñiga, “Who interacts on the Web?: The intersection of users’ personality and social media use,” *Comput. Human Behav.*, vol. 26, no. 2, pp. 247–253, Mar. 2010.

[22] P. B. Smith, M. H. Bond, and C. Kâğıtçıbaşı, *Understanding social psychology across cultures: living and working in a changing world*. USA: SAGE Publications, 2006.

[23] U. Droftina, M. Štular, and A. Košir, "Predicting Influential Mobile-Subscriber Churners using Low-level User Features," *Automatika*, vol. 56, no. 4, pp. 522–534, 2015.

[24] K. Dasgupta *et al.*, "Social ties and their relevance to churn in mobile telecom networks," in *Proceedings of the 11th international conference on Extending database technology Advances in database technology - EDBT '08*, 2008, pp. 668–677.

[25] C. Phadke, H. Uzunalioglu, V. B. Mendiratta, D. Kushnir, and D. Doran, "Prediction of Subscriber Churn Using Social Network Analysis," *Bell Labs Tech. J.*, vol. 17, no. 4, pp. 63–75, Mar. 2013.

[26] G. Benedek, Á. Lublóy, and G. Vastag, "The Importance of Social Embeddedness: Churn Models at Mobile Providers," *Decis. Sci.*, vol. 45, no. 1, pp. 175–201, Feb. 2014.

[27] L. Humski *et al.*, "Building implicit corporate social networks: The case of a multinational company," in *Proceedings of the 12th International Conference on Telecommunications*, 2013, pp. 31–38.

[28] G. Di Tommaso, G. Stilo, P. Velardi, D. Informatica, and S. Universita, "Women leadership in enterprise social networks," in *International Conference on Information Society (i-Society 2015)*, 2015, pp. 73–78.

[29] J. Sena and M. Sena, "Corporate Social Networking," *Issues Inf. Syst.*, vol. 9, no. 2, pp. 227–231, 2008.

[30] J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller, "Motivations for social networking at work," in *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08*, 2008, pp. 711–720.

[31] C.-Y. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges, "SmallBlue: People Mining for Expertise Search," *IEEE Multimed.*, vol. 15, no. 1, pp. 78–84, 2008.

[32] C.-Y. Lin *et al.*, "Social Network Analysis in Enterprise," *Proc. IEEE*, vol. 100, no. 9, pp. 2759–2776, 2012.

[33] B. Nie, H. Zhang, and Y. Liu, "Social interaction based video recommendation: Recommending YouTube videos to Facebook users," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2014, pp. 97–102.

[34] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "'Make New Friends, but Keep the Old' – Recommending People on Social Networking Sites," in *Proceedings of the*

27th international conference on Human factors in computing systems - CHI 09, 2009, pp. 201–210.

[35] L. Chen, C. Shao, and P. Zhu, "Social recommendation using quantified social tie strength," in *2015 Seventh International Conference on Advanced Computational Intelligence (ICACI)*, 2015, pp. 84–88.

[36] O. Oechslein and T. Hess, "The Value of a Recommendation: The Role of Social Ties in Social Recommender Systems," in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 1864–1873.

[37] I.-H. Ting, P. S. Chang, and S.-L. Wang, "Understanding Microblog Users for Social Recommendation Based on Social Networks Analysis," *J. Univers. Comput. Sci.*, vol. 18, no. 4, pp. 554–576, 2012.

[38] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, Jan. 2013.

[39] R. Rabbany, M. Takaffoli, and O. R. Zaiane, "Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques," in *In Proceedings of educational data mining*, 2011, pp. 21–30.

[40] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - GDPR." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=HR. [Accessed: 30-May-2018].

[41] J. Travers and S. Milgram, "An Experimental Study of the Small World Problem," 1969.

[42] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks," *Science (80-.).*, vol. 301, no. 5634, pp. 827–829, Aug. 2003.

[43] "Kako su nastale, a kako će se razvijati društvene mreže? - tportal." [Online]. Available: https://www.tportal.hr/tehno/clanak/kako-su-nastale-a-kako-ce-se-razvijati-drustvene-mreze-20151210. [Accessed: 28-Jan-2019].

[44] G. Gan, C. Ma, and J. Wu, *Data clustering : theory, algorithms, and applications*. Philadelphia, USA: SIAM, Society for Industrial and Applied Mathematics, 2007.

[45] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*. Waltham, USA: Morgan Kaufmann, 2011.

[46] M. Greenacre and R. Primicerio, *Multivariate analysis of ecological data*. Bilbao, Spain: Fundación BBVA, 2013.

[47] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A Comparison Study on

Similarity and Dissimilarity Measures in Clustering Continuous Data," *PLoS One*, vol. 10, no. 12, pp. 1–20, Dec. 2015.

[48] H. F. Lai, "Identify Implicit Social Network by RST/FL Framework," in *2009 International Conference on Advances in Social Network Analysis and Mining*, 2009, pp. 362–363.

[49] V. Podobnik, "Multi-Agent System for Telecommunication Service Provisioning based on User Profiles, PhD Thesis," Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2010.

[50] "Top 10 Most Popular Social Networks 2013 | Visual.ly," *Visually*, 2013. [Online]. Available: https://visual.ly/community/infographic/social-media/top-10-most-popular-social-networks-2013. [Accessed: 14-Oct-2013].

[51] J. Kincaid, "EdgeRank: The Secret Sauce That Makes Facebook's News Feed Tick | TechCrunch," *TechCrunch*, 2010. [Online]. Available: https://techcrunch.com/2010/04/22/facebook-edgerank/?guccounter=1. [Accessed: 14-Oct-2013].

[52] E. Khadangi and A. Bagheri, "Comparing MLP, SVM and KNN for predicting trust between users in Facebook," in *ICCKE 2013*, 2013, pp. 466–470.

[53] L. Humski, D. Pintar, and M. Vranić, "Exploratory Analysis of Pairwise Interactions in Online Social Networks," *Automatika*, vol. 58, no. 4, pp. 422–428, 2018.

[54] L. Humski, D. Pintar, and M. Vranic, "Analysis of Facebook Interaction as Basis for Synthetic Expanded Social Graph Generation," *IEEE Access*, vol. 7, pp. 6622–6636, 2019.

[55] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," in *Proceedings of the Royal Society of London*, 1895, vol. 58, pp. 240–242.

[56] R. I. M. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, "The structure of online social networks mirrors those in the offline world," *Soc. Networks*, vol. 43, 2015.

[57] "društvena udaljenost | Struna | Hrvatsko strukovno nazivlje." [Online]. Available: http://struna.ihjj.hr/naziv/drustvena-udaljenost/25369/. [Accessed: 06-Jan-2019].

[58] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 415–444, Aug. 2001.

[59] J. Chen, Y. Liu, and M. Zou, "From tie strength to function: Home location estimation in social network," in *2014 IEEE Computers, Communications and IT Applications Conference*, 2014, pp. 67–71.

[60] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical

prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, pp. 61–70.

[61] N. Noë, R. M. Whitaker, and S. M. Allen, "Personality Homophily and Geographic Distance in Facebook," *Cyberpsychology, Behav. Soc. Netw.*, vol. 21, no. 6, pp. 361–366, 2018.

[62] L. Humski, "Razvojno okruženje za genetske algoritme," Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2008.

[63] M. Golub, *Genetski algoritam*. Zagreb, Hrvatska: Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2004.

[64] L. N. Ni, C. Li, X. Wang, H. L. Jiang, and J. G. Yu, "DP-MCDBSCAN: Differential Privacy Preserving Multi-Core DBSCAN Clustering for Network User Data," *IEEE Access*, vol. 6, pp. 21053–21063, 2018.

[65] D. Yin, Y. Shen, and C. Liu, "Attribute Couplet Attacks and Privacy Preservation in Social Networks," *IEEE Access*, vol. 5, pp. 25295–25305, 2017.

[66] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[67] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115.

[68] P. Mittal, C. Papamanthou, and D. Song, "Preserving Link Privacy in Social Network Based Systems," Aug. 2012.

[69] P. Erdös and A. Rényi, "On random graphs, I," *Publ. Math.*, vol. 6, pp. 290–297, 1959.

[70] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[71] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–12, Oct. 1999.

[72] L. H. Wong, P. Pattison, and G. Robins, "A spatial model for social networks," *Phys. A Stat. Mech. its Appl.*, vol. 360, no. 1, pp. 99–120, Jan. 2006.

[73] J. Badham and R. Stocker, "A Spatial Approach to Network Generation for Three Properties: Degree Distribution, Clustering Coefficient and Degree Assortativity," *J. Artif. Soc. Soc. Simul.*, vol. 13, no. 1, 2010.

[74] M. L. de Almeida, G. A. Mendes, G. Madras Viswanathan, and L. R. da Silva, "Scale-free homophilic network," *Eur. Phys. J. B*, vol. 86, no. 2, p. 38, Feb. 2013.

[75] M. Q. Pasta, F. Zaidi, and C. Rozenblat, "Generating online social networks based on

socio-demographic attributes," *J. Complex Networks*, vol. 2, no. 4, pp. 475–494, Dec. 2014.

[76] A. R. Puniyani, R. M. Lukose, and B. A. Huberman, "Intentional Walks on Scale Free Small Worlds." 2001.

[77] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings - IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[78] Julianne Pepitone, "Netflix accidentally reveals rental histories," *CNN*, 2010. [Online]. Available: http://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/index.html. [Accessed: 30-May-2018].

[79] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[80] M. Siddula, L. Li, and Y. Li, "An Empirical Study on the Privacy Preservation of Online Social Networks," *IEEE Access*, vol. 6, pp. 19912–19922, 2018.

[81] S. Yang, Y. Zhou, Y. Guo, R. A. Farneth, I. Marsic, and B. S. Randall, "Semi-Synthetic Trauma Resuscitation Process Data Generator," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 573–573.

[82] J. Lee, J. Hong, B. Hong, and J. Ahn, "A generator of test data set for tactical moving objects based on velocity," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 4011–4013.

[83] M. Robnik-Šikonja, "Data Generators for Learning Systems Based on RBF Networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 5, pp. 926–938, May 2016.

[84] Syahaneim, R. A. Hazwani, N. Wahida, S. I. Shafikah, Zuraini, and P. N. Ellyza, "Automatic Artificial Data Generator: Framework and implementation," in *2016 International Conference on Information and Communication Technology (ICICTM)*, 2016, pp. 56–60.

[85] N. Iftikhar, X. Liu, F. E. Nordbjerg, and S. Danalachi, "A Prediction-Based Smart Meter Data Generator," in *2016 19th International Conference on Network-Based Information Systems (NBiS)*, 2016, pp. 173–180.

[86] R. Liu, B. Fang, Y. Y. Tang, and P. P. K. Chan, "Synthetic Data Generator for Classification Rules Learning," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 2016, pp. 357–361.

[87] J. Ruscio and W. Kaczetow, "Simulating Multivariate Nonnormal Data Using an

Iterative Algorithm," *Multivariate Behav. Res.*, vol. 43, no. 3, pp. 355–381, Sep. 2008.

[88] C. D. Vale and V. A. Maurelli, "Simulating multivariate nonnormal distributions," *Psychometrika*, vol. 48, no. 3, pp. 465–471, Sep. 1983.

[89] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph," *arXiv*, 2011.

ŽIVOTOPIS

Luka Humski rođen je 13. kolovoza 1987. u Zagrebu. U Zagrebu je završio Osnovnu školu Vladimira Nazora te prirodoslovno-matematički smjer III. gimnazije. Preddiplomski studij računarstva te diplomski studij informacijske i komunikacijske tehnologije završio je pri zagrebačkom Fakultetu elektrotehnike i računarstva (FER). Po završetku diplomskog, na istom je fakultetu upisao i poslijediplomski doktorski studij. Od 2011. godine zaposlen je pri Zavodu za osnove elektrotehnike i električka mjerenja FER-a.

Kroz osnovnoškolsko i srednjoškolsko obrazovanje sudjelovao je na brojnim natjecanjima i smotrama radova od općinske do državne razine iz čak 8 različitih predmeta, za što je višestruko nagrađivan. U okviru studija na FER-u nagrađen je trećom nagradom na natječaju za najbolji računalni program FER-a, priznanjem Josip Lončar za prvu godinu diplomskog studija, brončanom plaketom Josip Lončar za posebno uspješno studiranje na cijelom diplomskom studiju te diplomom s najvećom pohvalom (lat. *summa cum laude*) kakvu dobiva tek 1% najboljih studenata u generaciji.

Kroz sve tri razine studija bio je vrlo aktivan u zajednici. Bit će pobrojane neke od zadaća koje je pri tome obavljao ili ih još uvijek obavlja: studentski predstavnik u Fakultetskom vijeću FER-a i više odbora, član Studentskog zbora FER-a, potpredsjednik Odbora za znanstvena i sportska natjecanja FER-a, predstavnik studenata FER-u u Skupštini Studentskog zbora Sveučilišta u Zagrebu, predstavnik studenata u Vijeću tehničkog područja Sveučilišta u Zagrebu, pomoćnik prorektorice za istraživanje i tehnologiju Sveučilišta u Zagrebu, a kraće je vrijeme sudjelovao i u radu Rektorskog kolegija Sveučilišta u Zagrebu u širem sastavu. Aktivan je i izvan akademske zajednice. Odbojkaški je sudac nacionalnog ranga (dvoranska odbojka), član je Upravnog odbora Kluba odbojke na pijesku Siget, a obnašao je i dužnost člana Upravnog odbora Hrvatskog odbojkaškog saveza.

U istraživačkom smislu bavi se analizom društvenih mreža, analizom i obradom podataka, strojnim učenjem te elektroničkim poslovanjem. Sudjelovao je i sudjeluje na više znanstvenih projekata i projekata suradnje s gospodarstvom. Objavio je 3 rada u znanstvenim časopisima kategorije A (CC i SCI Expanded), 14 radova u zbornicima međunarodnih znanstvenih skupova te još nekoliko sažetaka u zbornicima skupova uz nekoliko neobjavljenih predavanja održanih na različitim stručnim skupovima. Recenzirao je više radova za međunarodne znanstvene časopise i konferencije, a član je programskog odbora jedne međunarodne znanstvene konferencije.

Aktivno sudjeluje u provođenju nastave iz četiri predmeta na preddiplomskom, diplomskom i poslijediplomskom specijalističkom studiju FER-a. Bio je izravan voditelj pri izradi nekoliko desetaka završnih i diplomskih radova.

Član je IEEE.

Popis radova objavljenih u časopisima i na međunarodnim konferencijama:

1. **Humski, Luka**; Pintar, Damir; Vranić, Mihaela. Analysis of Facebook Interaction as Basis for Synthetic Expanded Social Graph Generation. // IEEE access. 7 (2019) ; 6622-6636. **(časopis kategorije A)**
2. **Humski, Luka**; Pintar, Damir; Vranić, Mihaela. Exploratory analysis of pairwise interactions in online social networks. // Automatika : časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije. 58 (2018) , 4; 422-428. **(časopis kategorije A)**
3. Krakan, Sanja; **Humski, Luka**; Skočir, Zoran. Determination of Friendship Intensity between Online Social Network Users Based on Their Interaction. // Tehnički vjesnik : znanstveno-stručni časopis tehničkih fakulteta Sveučilišta u Osijeku. 25 (2018) , 3; 655-662. **(časopis kategorije A)**
4. Blašković, Bruno; Skočir, Zoran; **Humski, Luka**. Scenario Modeling and Verification for Business Processes. // Lecture Notes in Artificial Intelligence. 7327 (2012) ; 414-423. **(međunarodna konferencija)**
5. Sever, Nikolina; **Humski, Luka**; Ilić, Juraj; Skočir, Zoran; Pintar, Damir; Vranić, Mihaela. Applying the Multiclass Classification Methods for the Classification of Online Social Network Friends // 2017 International Conference on Software, Telecommunications and Computer Networks / Rožić, Nikola ; Lorenz, Pascal (ur.). Split : FESB, Sveučilište u Splitu, 2017. **(međunarodna konferencija)**
6. Stupalo, Maja; Ilić, Juraj; **Humski, Luka**; Skočir, Zoran; Pintar, Damir; Vranić, Mihaela. Applying the Binary Classification Methods for Discovering the Best Friends on an Online Social Network // Proceedings of the 14th International Conference on Telecommunications ConTEL 2017 / Dobrijević, Ognjen ; Džanko, Matija (ur.). Zagreb : Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2017. 155-161. **(međunarodna konferencija)**
7. Ilić, Juraj; **Humski, Luka**; Pintar, Damir; Vranić, Mihaela; Skočir, Zoran. Proof of Concept for Comparison and Classification of Online Social Network Friends Based on Tie Strength Calculation Model // Proceedings ICIST 2016 / Zdravković, M. ;

Trajnović, M. ; Konjović, Z. (ur.). Beograd : Society for Information Systems and Computer Networks, 2016. 159-164. **(međunarodna konferencija)**

8. Vranić, Mihaela; Pintar, Damir; **Humski, Luka**. Automated extraction and visualization of learning concept dependencies using Q-matrices and exam results // 24th International Conference on Software, Telecommunications and Computer Networks - SoftCOM 2016 / Rožić, Nikola ; Begušić, Dinko (ur.). Split : FESB, 2016. **(međunarodna konferencija)**
9. Vranić, Mihaela; Pintar, Damir; **Humski, Luka**; Skočir, Zoran; Škopljanac-Mačina, Frano; Brstilo, Ivana; Đuho, Nika; Klasnić, Kristina; Mališa, Snježana; Žuković, Ana Marija. University Social Network Benefits Analysis and Proposed Framework // 24th International Conference on Software, Telecommunications and Computer Networks - SoftCOM 2016 / Rožić, Nikola ; Begušić, Dinko (ur.). Split : FESB, 2016. **(međunarodna konferencija)**
10. Majić, Maja; Skorin, Jurica; **Humski, Luka**; Skočir, Zoran. Using the interaction on social networks to predict real life friendship // 22th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2014) : proceedings. Split : FESB, 2014. 1-1. **(međunarodna konferencija)**
11. **Humski, Luka**; Štriga, Darko; Podobnik, Vedran; Vrdoljak, Boris; Banek, Marko; Skočir, Zoran; Lovrek, Ignac. Building Implicit Corporate Social Networks: the Case of a Multinational Company // Proceedings of the 12th International Conference on Telecommunications / Pripuzic, Kresimir ; Banek, Marko (ur.). Zagreb : University of Zagreb, 2013. 31-38. **(međunarodna konferencija)**
12. **Humski, Luka**; Lažegić, Igor; Skočir, Zoran. Data Warehouse for FER e-Invoice System // Mipro 2012 - 35. međunarodni skup / Biljanović, Petar (ur.). Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2012. 1987-1992. **(međunarodna konferencija)**
13. **Humski, Luka**; Vrdoljak, Boris; Skočir, Zoran. Concept, Development and Implementation of FER e- Invoice System // SoftCOM 2012 - 20. International Conference od Software, Telecomunications & Computer Networks / Rožić, Nikola ; Begušić, Dinko (ur.). Split : FESB, 2012. 1-5. **(međunarodna konferencija)**
14. Čičak, Alan; Halambek, Ivan; Hari, Ivo; **Humski, Luka**; Lažegić, Igor; Previšić, Marko; Rukavina, Goran; Šulc, Matija; Vrdoljak, Boris; Skočir, Zoran. Sustav za razmjenu e-računa korištenjem komunikacijske mreže // 34th international convention on information and communication technology, electronics and microelectronics

(MIPRO 2011) : proceedings / Golubić, Stjepan ; Mikac, Branko ; Hudek, Vlasta ; Šimunić, Dina (ur.). Zagreb : Denona, 2011. 187-192. **(međunarodna konferencija)**

15. **Humski, Luka**; Skočir, Zoran. Volleyball Information System // 11the international conference on Telecommunications (Contel 2011) : proceedings / Plank, Thomas ; Loschnigg, Markus (ur.). Graz : Graz University of Technology, 2011. 117-124. **(međunarodna konferencija)**
16. Ptiček, M.; Vrdoljak, B**.; Humski, L**.; Skočir, Z.; Bolanča, G.; Gašparić, Ž. The potential of SEPA Credit Transfer implementation in Croatia // Mipro Proceedings / Biljanović, P. (ur.). Rijeka : Grafik, 2015. 1823-1828. **(međunarodna konferencija)**
17. **Humski, Luka**; Skočir, Zoran; Vrdoljak, Boris. Sigurnost sustava FER e-račun // MIPRO 2012. - 35. međunarodni skup / Biljanović, Petar (ur.). Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2012. 1858-1863. **(međunarodna konferencija)**

BIOGRAPHICAL NOTE

Luka Humski was born in Zagreb on 13th Oct. 1987. In Zagreb, he attended "Vladimir Nazor" elementary school and "III. gimnazija Zagreb" high school, where he graduated from science and mathematics programme. In Faculty of Electrical Engineering and Computing in Zagreb he graduated from undergraduate computing programme and obtained his Bachelor's degree, after which he enrolled in information and communication technologies graduate programme. He graduated and obtained his Master's degree in 2011. After finishing his graduate studies, he enrolled into doctoral study at the same faculty. Since 2011 he is employed as an assistant at the Department of Electrical Engineering Fundamentals and Measurements of the same faculty.

During his elementary and high school education he participated in many competitions and musters held on levels spanning from county level to state level in 8 different subjects, for which he has won multiple awards. During his faculty studies he has won 3rd place award in faculty's best software competition, "Josip Lončar" award for the first year of graduate study, bronze "Josip Lončar" plaque for excellence in entire graduate studies, and Master's diploma *summa cum laude* (with greatest honours) received by top 1% students of the generation.

He was very active in the community during all three levels of his faculty studies. Some of the many activities and duties he performed or is still performing are: students' representative in Faculty Council and multiple boards, member of the faculty's Student Council, vice-president of the faculty's Council for the scientific and sporting competitions, faculty's students' representative in the Assembly of Student Council of the University of Zagreb, students' representative in the Technical Area Council of the University of Zagreb, student helper of the vice-rector for science and technology of the University of Zagreb, and as a member of expanded assembly of the University of Zagreb Rector's Council. He is also active outside the academic community. He is a referee at the national level for the indoor volleyball, a member of the steering committee of the "Siget" beach volleyball club, and he has also been a member of the Croatian Volleyball Federation steering committee.

His research interests are social network analysis, data processing and analysis, machine learning and e-business. He has participated or is still participating in multiple science projects and industry cooperation projects. He has published 3 papers in A-category scientific journals (CC and SCI Expanded), 14 papers in international conferences and a few abstracts in proceedings of conferences, and he also has a few unpublished presentations which he held in various technical symposiums. He has reviewed multiple papers for international scientific

journals and conferences, and he is an international program committee member of one scientific conference.

He is active in performing teaching duties in four courses in the undergraduate, graduate and postgraduate specialist studies at the Faculty of Electrical Engineering and Computing. He helped mentor a few dozen Bachelor's and Master's thesis.

He is a member of the IEEE.