Универзитет у Београду

Електротехнички факултет



Анализа социјалних мрежа

Пројектни задатак

|  |  |
| --- | --- |
| Студенти: | Школска година: |
| Ана Стакић 2020/3103 Страхиња Стефановић 2020/3104 | 2021/2022. |

Београд, јануар 2022.

Садржај

[Садржај 2](#_Toc94464835)

[1. Чишћење података 3](#_Toc94464836)

[2. Моделовање мрежа 4](#_Toc94464837)

[2.1. Subreddit network – SNet 4](#_Toc94464838)

[2.2. Filtered subreddit network – SNetF 5](#_Toc94464839)

[2.3. Targeted subreddit network – SNetT 6](#_Toc94464840)

[2.4. User network – UserNet 7](#_Toc94464841)

[3. Статистичка обрада података 8](#_Toc94464842)

[4. Основна анализа моделованих мрежа 9](#_Toc94464843)

[5. Анализа мера централности 10](#_Toc94464844)

[6. Детекција комуна 11](#_Toc94464845)

[7. Поређење SNet и SNetT мрежа 12](#_Toc94464846)

[8. Закључак 13](#_Toc94464847)

[Списак слика 14](#_Toc94464848)

[Списак табела 15](#_Toc94464849)

# Чишћење података

Најпре је било потребно податке о објавама и коментарима подељене у по 12 засебних табела за сваки појединачни месец у 2008. години груписати у две засебне табеле. Прва садржи податке о свим објавама у тој години, а друга податке о свим коментарима. Приликом спајања табела, додата је и још једна колона ***month*** целобројног типа која памти редни број месеца у којем је креирана одговарајућа објава, односно коментар. Ова колона је додата да би се касније, евентуално, извршила нека анализа датих података заснована на времену настанка.

Након тога, из добијених табела су избачени редови са инвалидним идентификатором и та колона је преименована у ***submission\_id***, односно ***comment\_id***, како би постојала јединствена колона ***id*** са индексом реда. Такође, из табела су избачене колоне које нису од интереса за даљу анализу. Из табеле објава избачене су: ***created\_utc***, ***url***, ***permalink***, ***domain***, ***distinguished***. Колона ***distinguished***, иако у теорији занимљива за анализу зато што показује да ли је аутор објаве био админ одговарајућег сабредита, се није показала као корисна зато што је веома мали узорак заиста и имао неку валидну вредност те колоне (мање од 10 редова). Из тог разлога је и она уклоњена приликом стварања секундарног скупа података. Из табеле коментара су избрисане следеће колоне: ***created\_utc***, ***distinguished***, ***controversiality***. Колона ***controversiality*** је избачена из истог претходно поменутог разлога.

Током прегледа скупа података примећено је да се неретко у колони ***author*** налази вредност *[deleted]* што је вероватно последица ажурирања овог скупа података у блиској прошлости, како објаву или коментар може поставити само постојећи регистровани корисник ове платформе. Узимајући у обзир да је за даљу анализу и моделовање тражених мрежа (како мреже корисника, тако и мрежа сабредита) нужан податак о аутору, одлучено је да се сви редови са овом вредношћу уклоне. Такође, пошто сваки коментар има линк ка објави за коју је везан, било је потребно избрисати и коментаре који сада указују на објаве које више не постоје у табели објава, а које имају *[deleted]* аутора.

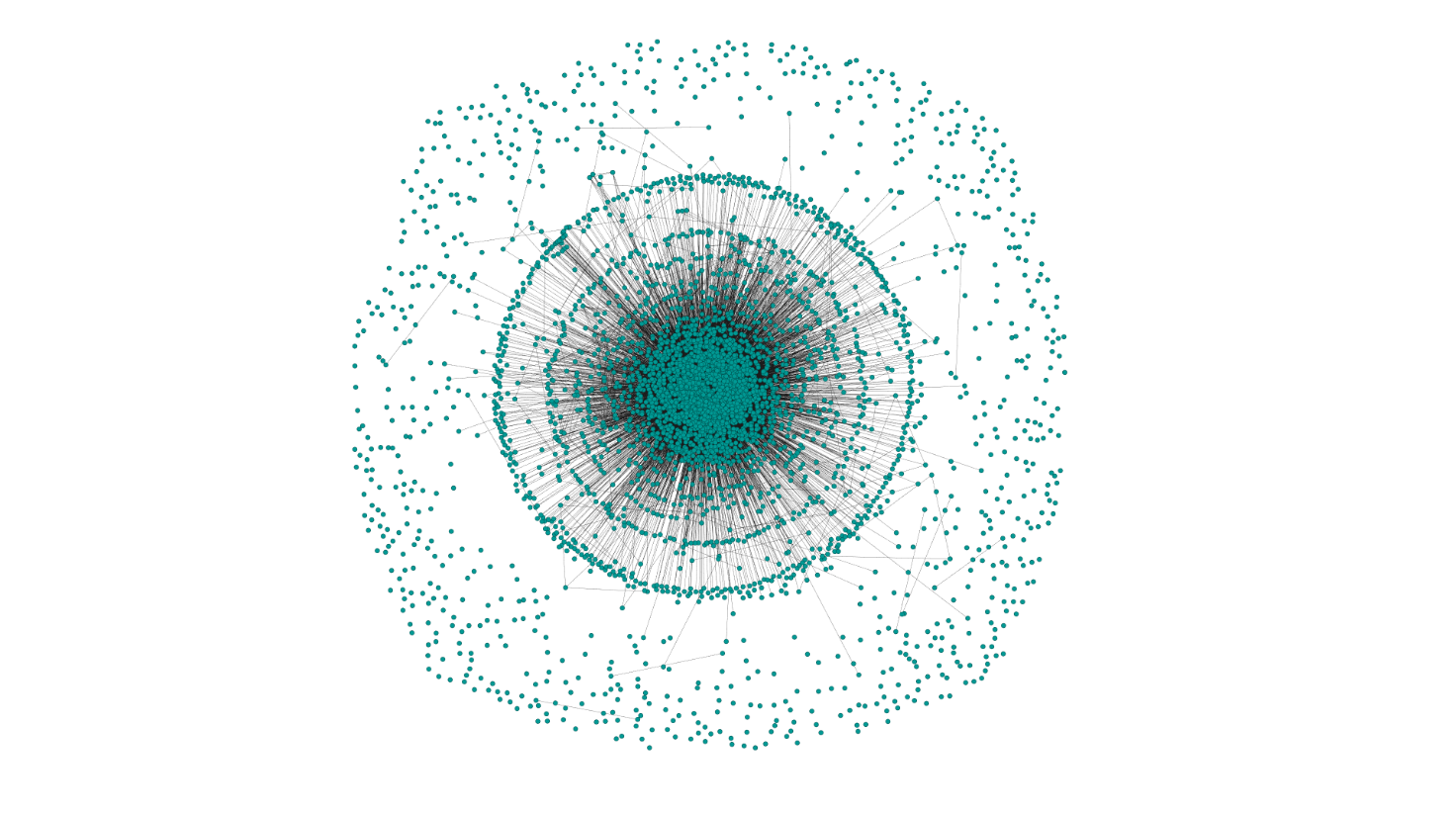
Табела 1.1 Утицај чишћења на величину скупа података

|  |  |  |
| --- | --- | --- |
|  | Примарни скуп података | Секундарни скуп података |
| Објаве | 2.519.853 × 16 | 2.044.810 × 12 |
| Коментари | 7.242.871 × 12 | 3.783.039 × 10 |

# Моделовање мрежа

## Subreddit network – SNet

Мрежа сабредита представља тежински неусмерени граф у којем су чворови сабредити, а веза између чворова се успоставља уколико постоје корисници који су били активни на оба сабредита. Тежине грана су добијене простом агрегацијом бројањем. Ова мрежа је формирана на основу комплетног скупа података.



Слика 2.1 Мрежа SNet

## Filtered subreddit network – SNetF

Мрежа добијена уклањањем грана са тежином испод одређеног прага – *w\_threshold*. Најпре је било потребно визуелизовати расподелу тежина грана у SNet мрежи.

A picture containing chart

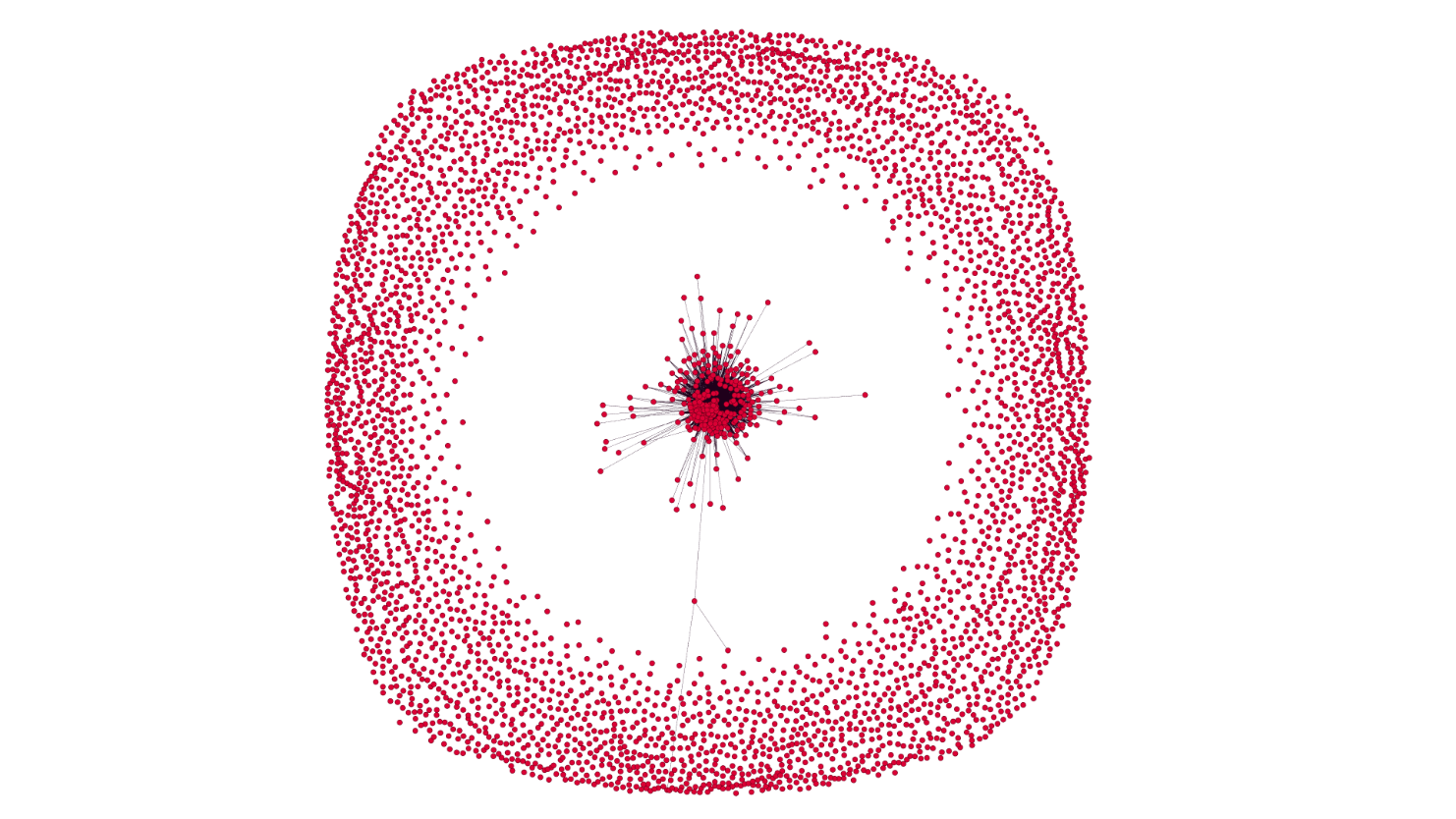
Description automatically generated

Слика 2.2 Расподела тежина грана

Са графика се може видети да највећи проценат грана има минималну тежину (један), док са повећањем тежине број грана драстично опада све до тзв. *outlier*-а који представљају гране огромне тежине. Из претходно наведенеог може се рећи да тежине грана, као и тежине чворова прате *power law* расподелу. За потребе генерисања SNetF мреже, као тежински праг изабрана је вредност 20 што драстично редукује број грана у мрежи.

Табела 2.1 Промена броја грана приликом филтрације

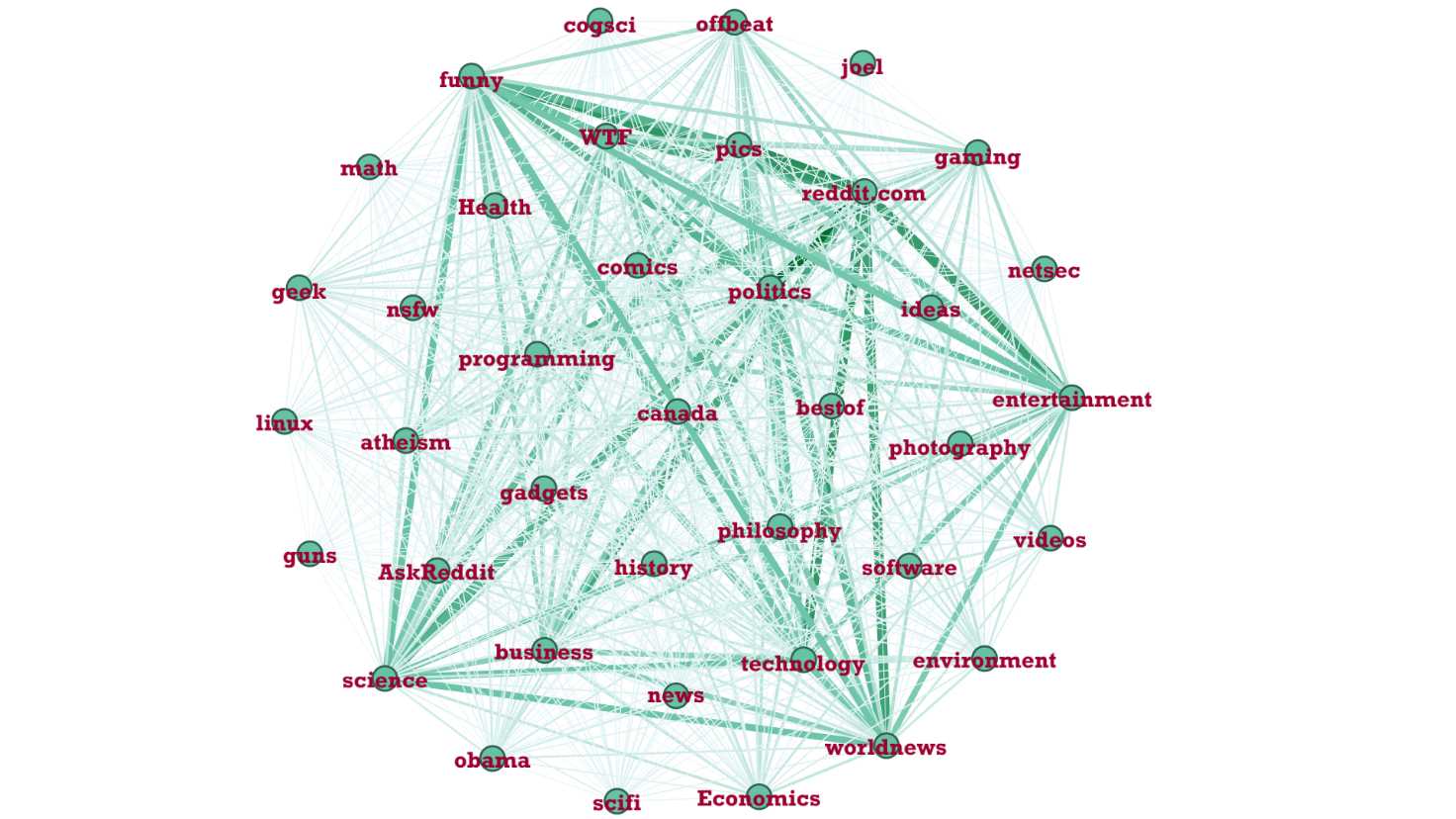
|  |  |  |
| --- | --- | --- |
|  | Пре филтрирања | После филтрирања |
| Број грана | 137.264 | 9.913 |



Слика 2.3 Мрежа SNetF

## Targeted subreddit network – SNetT

Још једна подмрежа првобитне SNet мреже с тим да се сада уклањају сви сабредити који не припадају задатом скупу сабредита чија тематика је блиска теми економске кризе. Укупан број сабредита је **4191**, док подскуп који се посматра броји **39** сабредита.



Слика 2.4 Мрежа SNetT

## User network – UserNet

Мрежа интеракције између корисника *Reddit* платформе у којој су корисници повезани уколико су коментарисали једни другима објаве или коментаре.

A picture containing accessory

Description automatically generated

Слика 2.5 Мрежа UserNet

# Статистичка обрада података

1. Колико постоји различитих сабредита који се појављују у посматраном периоду? Који су најважнији по броју корисника, а који по броју коментара?

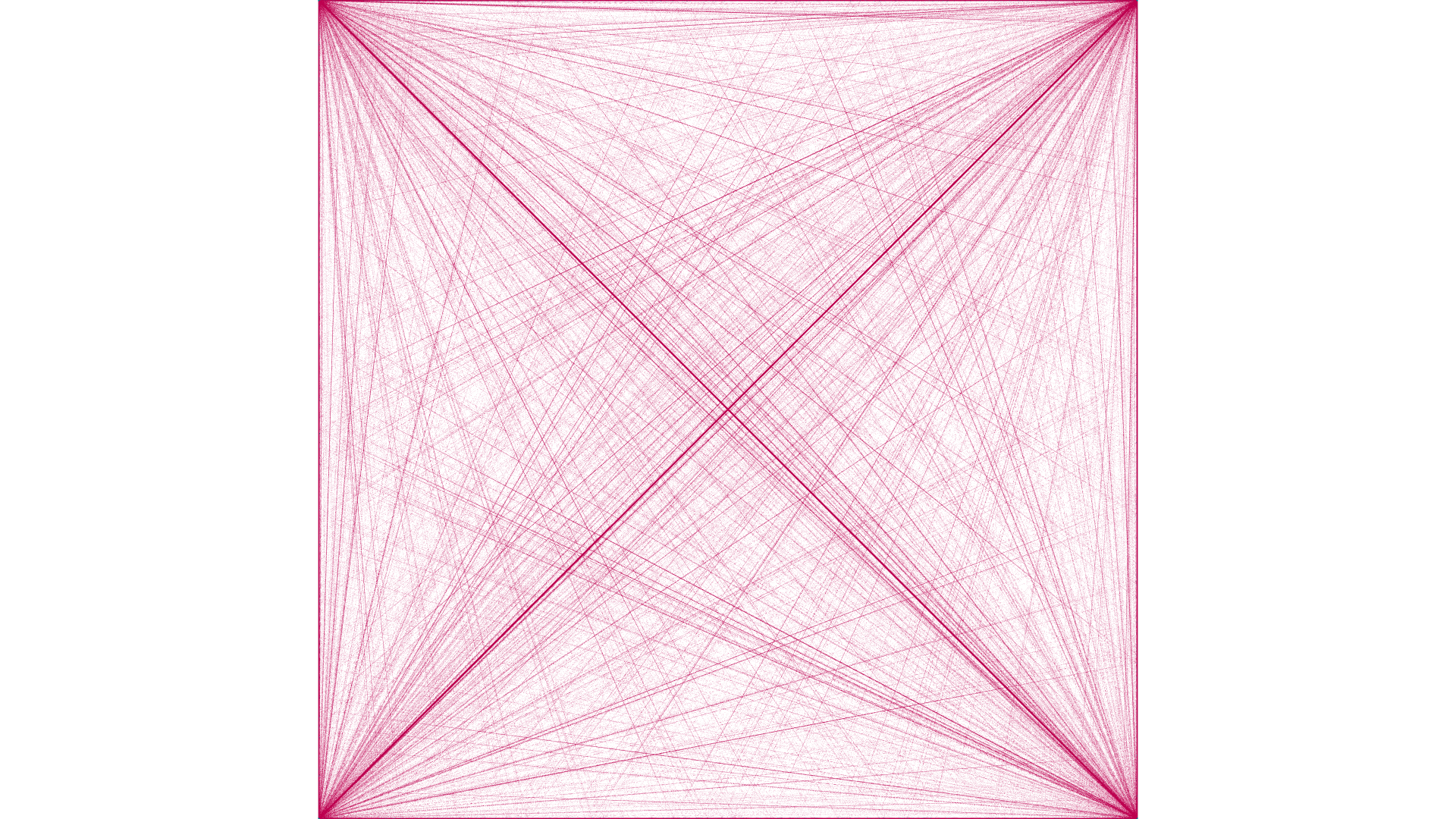
Укупан број сабредита: **4191**

1. Какав је просечан број забележених корисника активних у посматраном периоду по сабредиту? Корисник се сматра активним на сабредиту ако је забележен барем један коментар или објава тог корисника.

Просечан број корисника по сабредиту: **144.4063**

1. Ко су корисници са највећим бројем објава, а ко корисници са највећим бројем коментара?



1. Који корисници су активни на највећем броју сабредита? На колико су сабредита активни?



1. Како су корелисани бројеви објава и бројеви коментара корисника? Одредити Пирсонов коефицијент корелације и извршити визуелизацију.

Chart, scatter chart

Description automatically generated

1. Које објаве поседују највећи број коментара и на којим су сабредитима постављене? Приказати податке о тим објавама, укључујући то на којем су сабредиту постављене и шта им је садржај (ако је поље објаве ***over 18*** постављено на *false*).



# Основна анализа моделованих мрежа

# Анализа мера централности

# Детекција комуна

# Поређење SNet и SNetT мрежа

# Закључак

Списак слика

[Слика 2.1.1 Мрежа SNet 4](#_Toc94464801)

[Слика 2.2.1 Расподела тежина грана 5](#_Toc94464802)

[Слика 2.2.2 Мрежа SNetF 6](#_Toc94464803)

[Слика 2.3.1 Мрежа SNetT 7](#_Toc94464804)

Списак табела

[Табела 1.1 Утицај чишћења на величину скупа података 3](#_Toc94464809)

[Табела 2.1 Промена броја грана приликом филтрације 5](#_Toc94464810)