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](#_Toc94731144)

[1. Чишћење података 3](#_Toc94731145)

[2. Моделовање мрежа 4](#_Toc94731146)

[2.1. Subreddit network – SNet 4](#_Toc94731147)

[2.2. Filtered subreddit network – SNetF 5](#_Toc94731148)

[2.3. Targeted subreddit network – SNetT 6](#_Toc94731149)

[2.4. User network – UserNet 7](#_Toc94731150)

[3. Статистичка обрада података 8](#_Toc94731151)

[4. Основна анализа моделованих мрежа 12](#_Toc94731152)

[5. Анализа мера централности 13](#_Toc94731153)

[6. Детекција комуна 14](#_Toc94731154)

[7. Поређење SNet и SNetT мрежа 15](#_Toc94731155)

[8. Закључак 16](#_Toc94731156)

[Списак слика 17](#_Toc94731157)

[Списак табела 18](#_Toc94731158)

# 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.

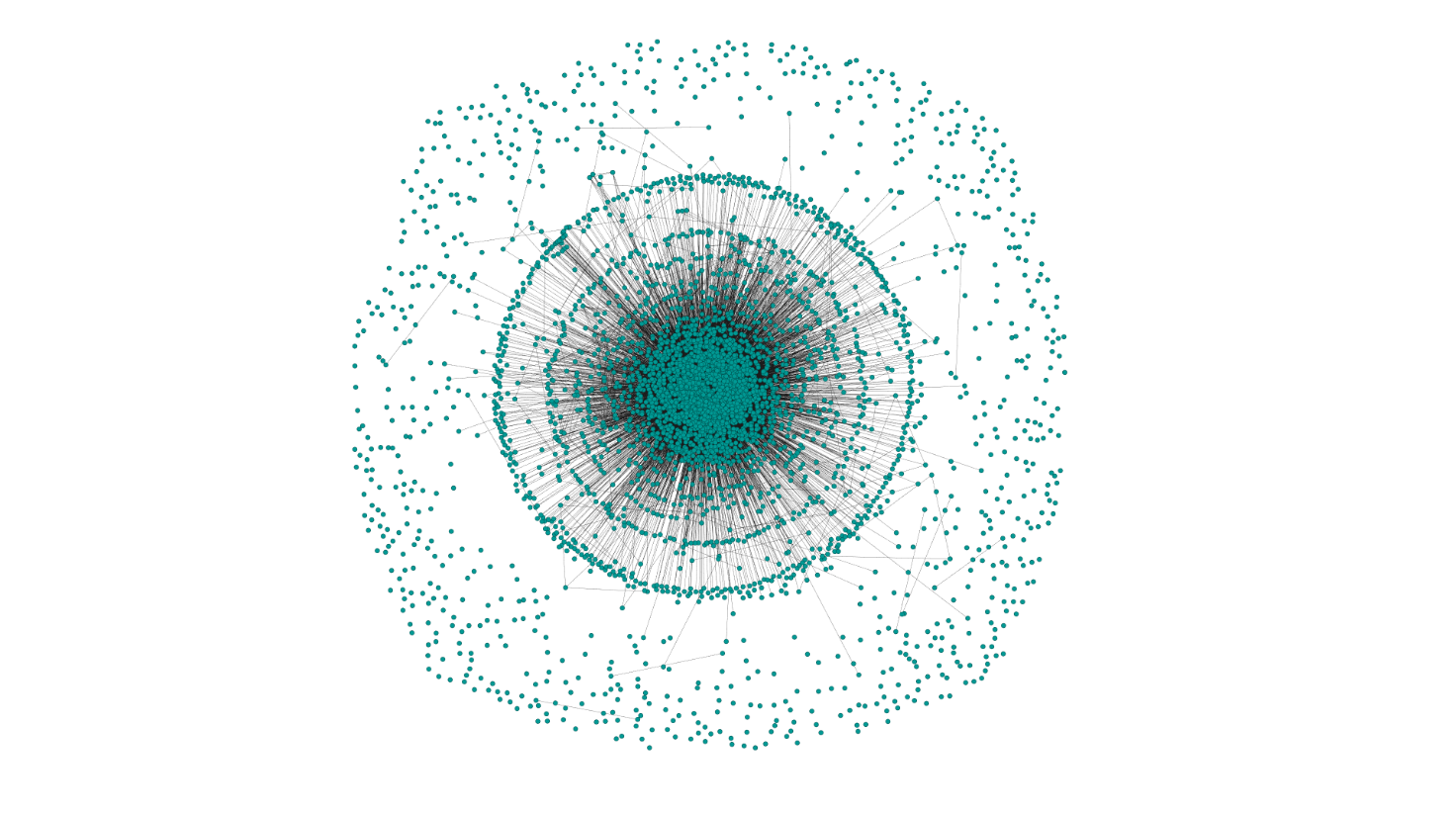
Табела 1.1 Утицај чишћења на величину скупа података

|  |  |  |
| --- | --- | --- |
|  | 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

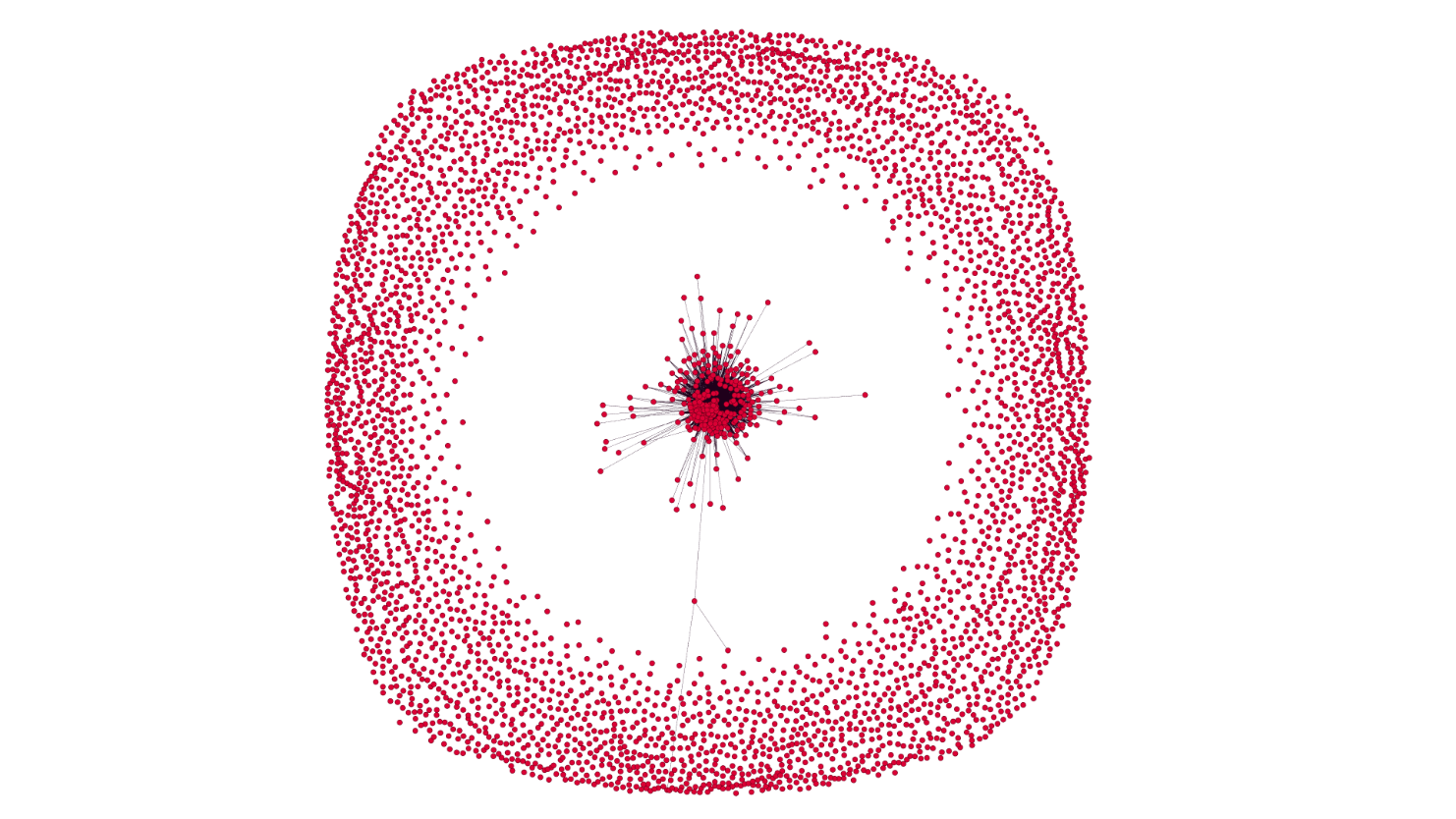
Description automatically generated

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 наведеног 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

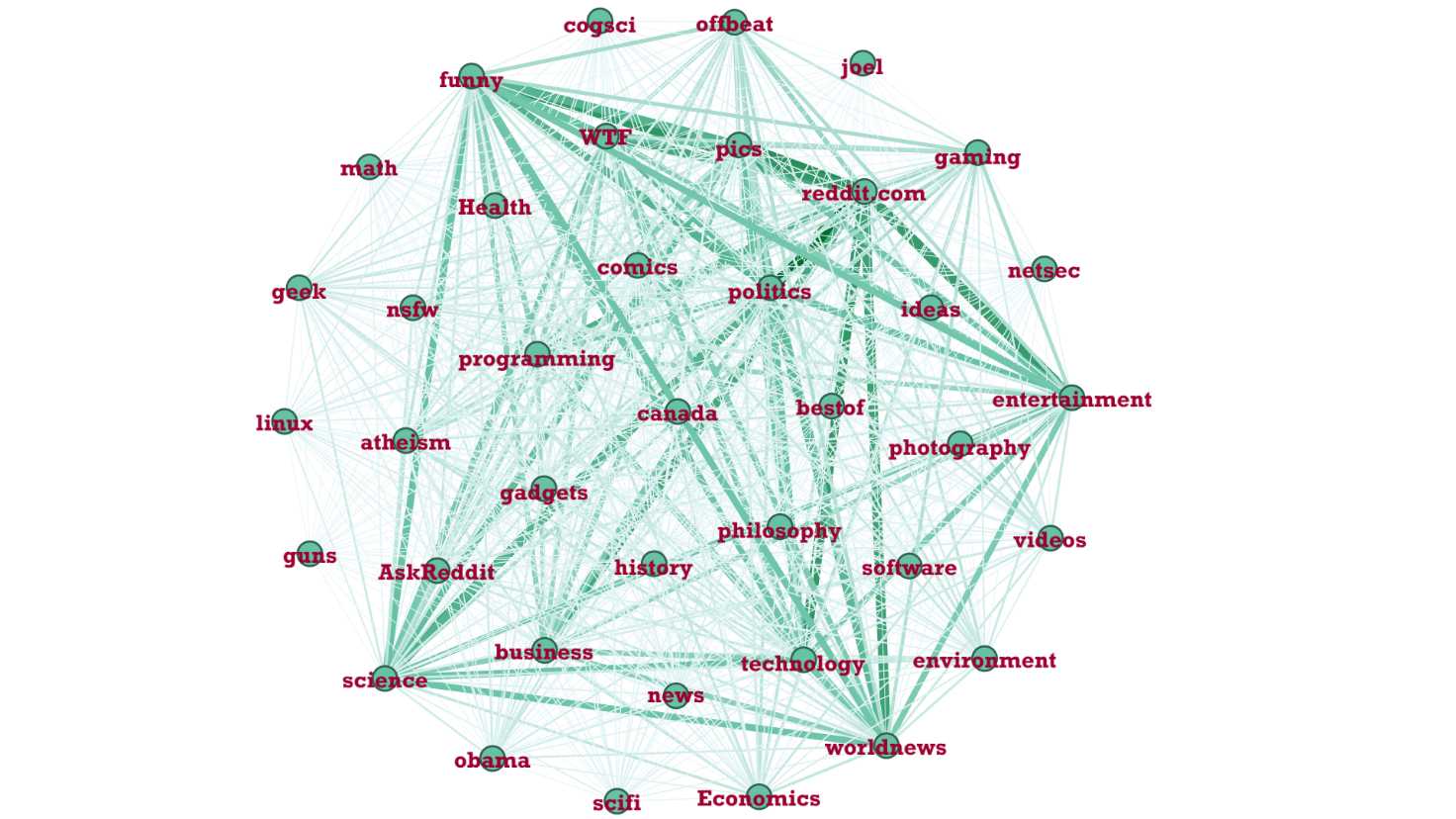
|  |  |  |
| --- | --- | --- |
|  | Пре филтрирања | После филтрирања |
| Број грана | 137.264 | 9.913 |



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**

Табела 3.1 Анализа сабредита

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?

Табела 3.2 Анализа корисника

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

Табела 3.3 Најактивнији корисници



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

Chart, scatter chart

Description automatically generated

Слика 3.1 Пирсонов коефицијент корелације

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*).

Табела 3.4 Објаве са највећим бројем коментара



# Основна анализа моделованих мрежа

1. Колика је густина мреже?
2. Колике су просечне дистанце у оквиру мреже и дијаметар мреже?

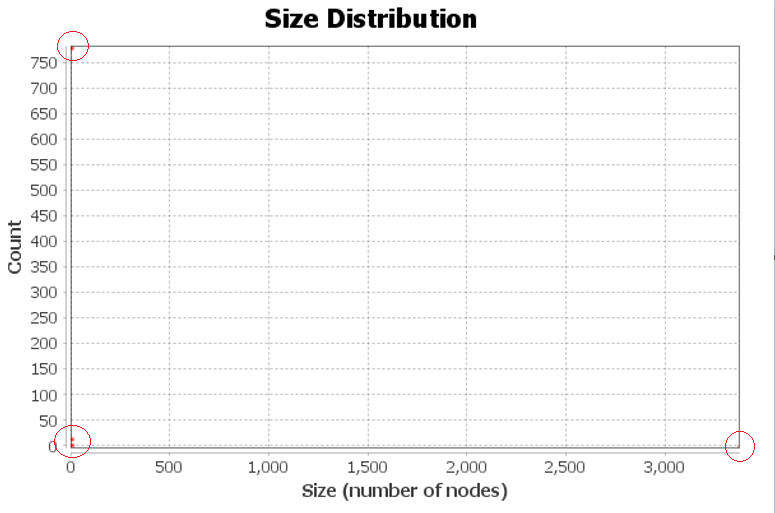
Табела 4.1 Густина и дијаметар мреже

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SNet | SNetF | SNetT | UserNet |
| Густина мреже | 0.016 | 0.001 | 1.0 | ~ 0.0 |
| Просечна дистанца | 2.1065 | 1.84 | 1.0 | 4.196 |
| Дијаметар мреже | 5.0 | 4.0 | 1.0 | 14.0 |

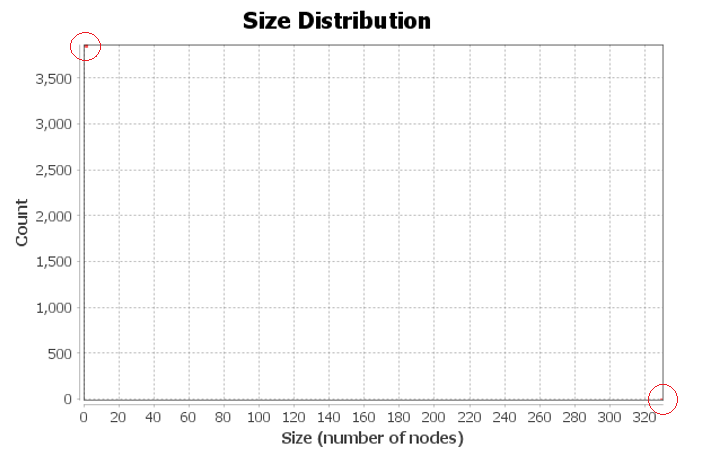
1. У којој мери је мрежа повезана и централизована? Навести број и величине повезаних компонената и проценити да ли постоји гигантска компонента.

Табела 4.2 Број повезаних компонената мреже

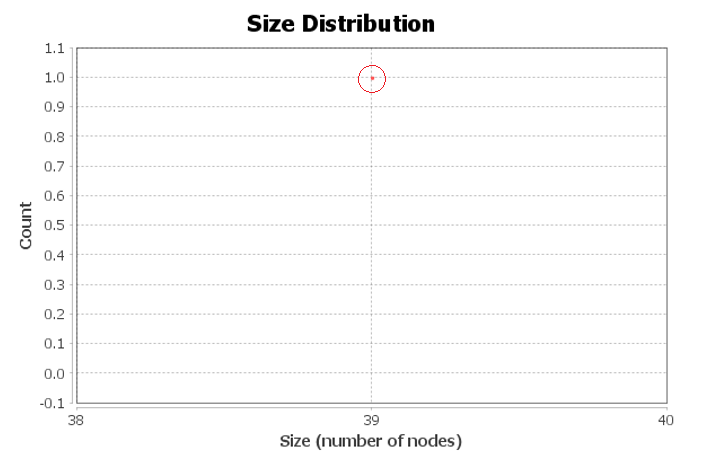
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SNet | SNetF | SNetT | UserNet |
| Број повезаних компонената | 799 | 3863 | 1 | 201 |



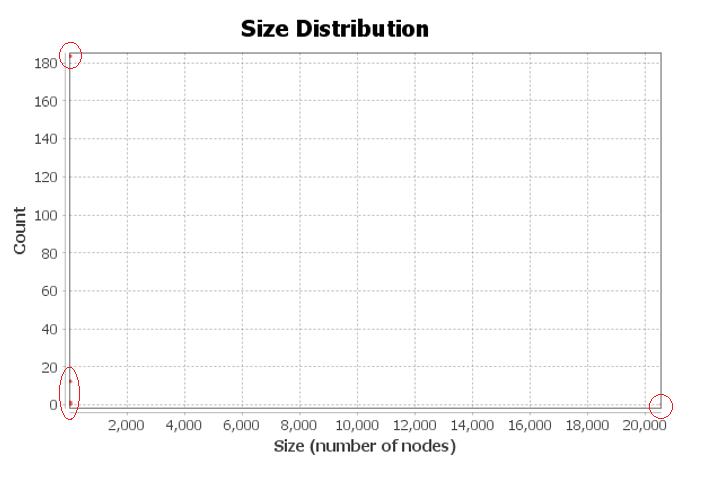
Слика 4.1 Величина компонената мреже SNet



Слика 4.2 Величина компонената мреже SNetF



Слика 4.3 Величина компонената мреже SNetT



Слика 4.4 Величина компонената мреже UserNet

Све мреже осим мреже SNetT поседују једну гигантску компоненту. Мрежа SNetT је специфична по томе што представља комплетан граф, односно има једну повезану компоненту максималне величине.

1. Колики је просечни, а колики глобални коефицијент кластеризације мреже? Каква је расподела локалног коефицијента кластеризације њених чворова? Да ли је кластерисање изражено или не? Одговор дати упоређивањем са случајно генерисаном *Erdos-Renyi* мрежом истих димензија.

Табела 4.3 Коефицијент кластеризације мреже

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Просечни | Глобални | Просечни | Глобални |
| Са утицајем тежина грана | | Без утицаја тежина грана | |
| SNet | 0.000244 | 0.001199 | 0.618883 | 1.0 |
| SNetF | 0.000446 | 0.013474 | 0.066099 | 1.0 |
| SNetT | 0.0621021 | 0.1411455 | 1.0 | 1.0 |
| UserNet | 0.000320 | 0.094210 | 0.028932 | 1.0 |

У наставку следе визуелизовани прикази расподела коефицијената кластеризације за сваку од мрежа, као и за случајно генерисане *Erdos-Renyi* мреже истих димензија.

Chart

Description automatically generatedChart, histogram

Description automatically generated

Слика 4.5 Расподела коефицијента кластеризације мреже SNet и случајне мреже истих димензија

Са левог графика се може видети да највећи број чворова има максималан коефицијент кластеризације (један) у оригиналној мрежи. Тај скуп чворова може представљати један добро повезани кластер унутар саме мреже, централни скуп са слике **Слика 2.1 Мрежа Snet**, па самим тим кластерисање ове мреже јесте изражено.

Chart

Description automatically generatedChart

Description automatically generated

Слика 4.6 Расподела коефицијента кластеризације мреже SNetT и случајне мреже истих димензија

Приликом креирања случајне мреже, прослеђена су два параметра – број чворова у мрежи и вероватноћа да између два чвора постоји грана. Управо учествовање ове вероватноће, која за потпуно повезани мрежни граф SNetT има вредност 1 (100%), доводи до тога да и оригинална и случајно генерисана мрежа изгледају идентично. Самим тим и њихова распорела коефицијента кластеризације је потпуно иста и сви чворови поседују коефицијент 1 (један), а мрежа представља тачно један кластер.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Слика 4.7 Расподела коефицијента кластеризације мреже SNetF и случајне мреже истих димензија

Слично као код расподеле коефицијента кластеризације SNet мреже, са левог графика се може видети да највећи број чворова има максималан коефицијент (један) у оригиналној мрежи. У примеру ове мреже, скок тог дела графика још је израженији у односу на сличан скок код SNet мреже, разлог тога видљив је и као изолованији подскуп чворова на слици **Слика 2.3 Мрежа SnetF**, а самим тим кластерисање и ове мреже јесте изражено.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Слика 4.8 Расподела коефицијента кластеризације мреже UserNet и случајне мреже истих димензија

Расподела коефицијента кластеризације за мрежу корисника представља доминацију чворова са јако малим вредностима коефицијента (између 0 и 0.2), што говори о слабој повезаности чворова, него што је то случај у осталим мрежама. Са друге стране, уочава се велика сличност расподеле оригиналне и случајно генерисане мреже. Да ли је кластерисање изражено?

1. На основу одговора на питања 8 и 10, проценити да ли мрежа исказује особине малог света.

Да мрежа исказивала особине малог света треба да има малу просечну удаљеност било која два чвора у односу на величину мреже и висок степен кластеризације.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Просечна дистанца | Просечан степен кластеризације | Особина малог света |
| SNet | 2.1065 | 0.618883 | ДА (?) |
| SNetF | 1.84 | 0.066099 | НE (низак степен кластеризације) |
| SNetT | 1.0 | 1.0 | ДА |
| UserNet | 4.196 | 0.028932 | НE (низак степен кластеризације) |

1. Извршити асортативну анализу по степену чвора и дати одговор да ли је изражено асортативно мешање. У случају да је мрежа усмерена, анализу извршити и по улазном и по излазном степену чвора. Приложити и визуелизацију.
2. Да ли мрежа испољава феномен клуба богатих (енг. *rich club phenomenon*)?
3. Каква је дистрибуција чворова по степену и да ли прати *power law* расподелу?
4. Одредити најважније хабове и ауторитете у мрежи. Како су они распоређени и уграђени у мрежу, да ли су на периферији или у језгру мреже?

# Анализа мера централности

1. Спровести анализе централности по степену, блискости и релационој централности. Дати преглед најважнијих актера по свакој од њих.
2. Ко су најважнији актери по централности по сопственом вектору? Шта нам то говори о њима?
3. Рангирати чворове по Кацовој централности (енг. *Katz centrality*) са варијацијом параметара. При рачунању Кацове централности, експериментисати са додељивањем другачије вредности параметра β за сабредит који се у приложеним CSV фајловима идентификује врeдношћу колоне ***subreddit*** једнаком *reddit.com*. Дати преглед најважнијих актера у случају да је β исто за све сабредите и у случају да је β наведеног сабредита значајно веће.
4. На основу претходна три питања предложити и конструисати хеуристику (композитну меру централности) за проналажење најважнијих актера и пронаћи их. Обратити пажњу на тип мреже који се анализира (усмерена или неусмерена) и, сходно томе, прилагодити колико различите мрежне метрике утичу на хеуристику.

# Детекција комуна

1. Ако величина мреже дозвољава, спектралном анализом или анализом дендрограма проценити потенцијалне кандидате за број комуна у мрежи.
2. Спровести кластерисање Лувенском методом (максимизацијом модуларности) у алату *Gephi* за три различите вредности параметра резолуције. Конструисати визуелизације и дискутовати избор параметра резолуције на добијено кластерисање (број и величина кластера).
3. Које заједнице (комуне) се могу уочити приликом анализе мреже? Да ли постоји неко објашњење за детектоване комуне?
4. Ко су актери који се могу окарактерисати као кључни брокери (мостови) у мрежи? Шта их чини брокерима?

# Поређење SNet и SNetT мрежа

1. Упоредити карактеристике SNet и SNetT мрежа. Коментарисати потенцијалне разлике и проценити да ли су сабредити из SNetT активнији и боље повезани од остатка мреже.
2. Како су распоређени чворови из SNetT у оквиру SNet мреже? Да ли припадају језгру или периферији или су мешовито распоређени?

# Закључак

Списак слика

[Слика 2.1 Мрежа SNet 4](#_Toc94806725)

[Слика 2.2 Расподела тежина грана 5](#_Toc94806726)

[Слика 2.3 Мрежа SNetF 6](#_Toc94806727)

[Слика 2.4 Мрежа SNetT 7](#_Toc94806728)

[Слика 2.5 Мрежа UserNet 7](#_Toc94806729)

[Слика 3.1 Пирсонов коефицијент корелације 9](#_Toc94806730)

Списак табела

[Табела 1.1 Утицај чишћења на величину скупа података 3](#_Toc94806733)

[Табела 2.1 Промена броја грана приликом филтрације 5](#_Toc94806734)

[Табела 3.1 Анализа сабредита 8](#_Toc94806735)

[Табела 3.2 Анализа корисника 8](#_Toc94806736)

[Табела 3.3 Најактивнији корисници 9](#_Toc94806737)

[Табела 3.4 Објаве са највећим бројем коментара 10](#_Toc94806738)