

VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Analýza emailové komunikace

Analysis of Email Communication

Zadání diplomové práce

Student: **Bc. Veronika Uhrová**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 2612T025 Informatika a výpočetní technika

Téma: **Analýza emailové komunikace**
Analysis of Email Communication

Jazyk vypracování: čeština

Zásady pro vypracování:

Cílem práce je návrh a implementace systému na analýzu emailové komunikace a vizualizaci výstupů. Systém bude pracovat s reálnými daty a budou navrženy, popsány a vyhodnoceny experimenty s těmito daty. Pro implementaci je doporučen jazyk C#.

1. Rešerše obdobných řešení a analytických přístupů.
2. Návrh a implementace metody na získávání emailových zpráv z vybraného zdroje.
3. Výběr a implementace metod strojového učení a analýzy sítí vhodných pro analýzu emailové komunikace.
4. Návrh a implementace uživatelského rozhraní na analýzu emailové komunikace a vizualizaci analytických výstupů.
5. Dokumentovaná implementace systému.

Seznam doporučené odborné literatury:

- [1] S. Zehnalova, Z. Horak, M. Kudělka. Email Conversation Network Analysis: Work Groups and Teams in Organizations. ASONAM 2015.
- [2] Tang, G., Pei, J., Luk, W. S. (2014). Email mining: tasks, common techniques, and tools. Knowledge and Information Systems, 41(1), 1-31.

Další podle pokynů vedoucího práce.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **doc. Mgr. Miloš Kudělka, Ph.D.**

Datum zadání: 01.09.2017

Datum odevzdání: 30.04.2018



doc. Ing. Jan Platoš, Ph.D.
vedoucí katedry



prof. Ing. Pavel Brandštetter, CSc.
děkan fakulty

Prehlasujem, že som túto prácu vypracovala samostatne. Uviedla som všetky literárne
pramene a publikácie, z ktorých som čerpala.

V Ostrave, 20. 4. 2018

.....

Moje podakovanie patrí predovšetkým doc. Milošovi Kudělkovi, Ph.D. za odborné konzultácie a vedenie mojej diplomovej práce.

Abstrakt

Práca študuje aktuálne metódy pre analýzu emailov a detekciu sociálnych rolí v emailových dátach. Nasleduje zoznámenie sa s emailom a jeho popularitou v súčasnosti. Taktiež práca uvádza základné teoretické pojmy a teoretický náhľad na reprezentáciu siete. Uvádzajú sa tu aj základy analýzy sociálnych sietí a detekcie komúnít. Ďalej sa tu píše o frameworku pre detekciu štruktúrnych rolí a ich identifikácie. Na základe týchto poznatkov je vytvorená aplikácia pre analýzu a vizualizáciu analytických výstupov. Na záver sú uvedené prevedené experimenty týkajúce sa identifikácie sociálnych rolí v emailových dátach.

Kľúčové slová: email, sociálna sieť, sociálna rola, vizualizácia

Abstract

This paper studies current methods for analysing emails and detecting social roles in email data. This is followed by getting acquainted with the email and its popularity nowadays. Also, this thesis presents basic theoretical concepts and a theoretical overview of network representation. Here are also the basics of social networking and community detection. There is also written a framework for structural social roles detection and their identification. Based on this knowledge, an application is developed to analyze and visualize analytical outputs. Finally, experiments on the findings of the emailed data are presented.

Key Words: email, social network, social role, visualization

Obsah

Zoznam použitých skratiek a symbolov	10
Zoznam obrázkov	11
Zoznam tabuliek	12
1 Úvod	13
1.1 Motivácia	13
1.2 Vízia	13
1.3 Štruktúra práce	14
2 Súvisiace práce	15
3 Emailová komunikácia	17
3.1 Stručná história emailu	17
3.2 Štruktúra emailu	17
3.3 Emaily v súčasnosti	18
4 Definície a klasifikácie	19
4.1 Graf	19
4.2 Súvislosť grafu	20
4.3 Úplný graf	20
4.4 Stupeň vrcholu	20
4.5 Cesta	21
4.6 Uzavretá cesta	21
4.7 Komponenta grafu	21
4.8 Metriky	21
4.8.1 Closeness centrality (Centralita blízkosti)	21
4.8.2 Betweenness centrality (Centralita medziľahlosti)	21
4.8.3 Modularita	22
5 Sociálna sieť	23
5.1 História sociálnych sietí	23
5.2 Analýza sociálnych sietí	23
5.3 Komunity v sociálnych sieťach	24
5.4 Detekcia komunit	25
5.4.1 Louvainov algoritmus pre detekciu komunit	25

6	SSRM - Framework pre detekciu štruktúrálnych rolí v sociálnych sieťach	27
6.1	Rola v kontexte SSRM	27
6.2	Roly definované v SSRM	27
6.2.1	Leader	28
6.2.2	Outermost	28
6.2.3	Mediator	28
6.2.4	Outsider	28
7	Identifikácia štruktúrálnych sociálnych rolí	29
7.1	Outsider	29
7.2	Leader	29
7.2.1	Closeness centrality (Centralita blízkosti)	29
7.3	Outermost	29
7.4	Mediator	29
7.4.1	LBetweeness	30
7.4.2	CBetweeness	30
7.4.3	Normalizovaná verzia CBetweeness	30
7.4.4	Skóre rozmanitosti	31
8	Aplikácia	32
8.1	Špecifikácia	32
8.1.1	Funkčné požiadavky	32
8.2	Návrh	33
8.2.1	Návrhové vzory	34
8.3	Dôležité rozhodnutia	36
8.3.1	Dostupnosť dát	36
8.3.2	Webová vs. desktopová aplikácia	36
8.4	Použité knižnice	36
8.5	Import dát	38
8.6	Implementácia	38
8.6.1	Konštrukcia siete	38
8.6.2	Triedy pre graf, vrcholy a hrany	39
9	Experimenty	40
9.1	Analýza tímu	40
9.1.1	Základné informácie	40
9.1.2	Zmeny komunit v čase	40
9.1.3	Zmeny rolí v čase	42
9.2	Analýza jednotlivca	44
9.2.1	Základné informácie	44

9.2.2	Zmeny komunit v čase	44
9.2.3	Zmeny rôl v čase	44
10	Záver	45
10.1	Možnosti rozšírenia a zdokonalenia práce	45
	Literatura	46

Zoznam použitých skratiek a symbolov

MUA	– Mail User Agent
MTA	– Mail Transfer Agent
IMAP	– Internet Message Access Protocol
XML	– eXtensible Markup Language
SSRM	– Structural social role mining framework

Zoznam obrázkov

1	Akú formu komunikácie preferujete na formálnu komunikáciu?	18
2	Neorientovaný graf	19
3	Orientovaný graf	19
4	Súvislý (1) a nesúvislý graf (2)	20
5	Úplný graf	20
6	Graf v tvare hviezdy	22
7	Sieť s viacerými komunitami	24
8	Vizualizácia krokov Louvainovho algoritmu.	26
9	UseCase Diagram	33
10	Diagram komponent znázorňujúci jednotlivé komponenty architektúry aplikácie .	34
11	Triedny diagram - Repository pattern	35
12	Model-View-Controller	36
13	Jednoduchá sieť vytvorená s použitím knižnice vis.js	37
14	Príklad použitia knižnice vis.js	37
15	Doménový model	38
16	Vizualizácia grafu emailov s vygenerovanými komunitami celého tímu za celkový čas	41
17	Vizualizácia grafu v intervale 18.11.2010 - 31.12.2011	42
18	Vizualizácia grafu v intervale 18.11.2010 - 31.12.2013, v každej komunite detekovaný jeden leader.	43

Zoznam tabuliek

1	Informácie o členoch tímu	40
2	Základné informácie o datasete	40
3	Ďalšie informácie o datasete	40
4	Počet komunit v ročných intervaloch	41
5	Počet komunit v ročných intervaloch	42
6	Základné informácie o datasete	44
7	Ďalšie informácie o datasete	44

1 Úvod

V stručnom úvode je popísaná motivácia, ktorá viedla k vypracovaniu tejto diplomovej práce, vízia toho, čo sa malo dosiahnuť a hrubá štruktúra vypracovaného textu.

1.1 Motivácia

S cieľom uľahčiť používanie emailov a prebádať podnikateľský potenciál emailov, analýza emailov dosiahla pozoruhodný pokrok v oblasti výskumu, ale aj v praxi. Emaily teda možno považovať za zmiešanú štruktúru obsahujúcu údaje o ľuďoch zo sociálnych alebo aj organizačných aspektov.

Obsah emailu ako textové a netextové dáta

Emaily sú písané viac stručne ako väčšina ostatných dokumentov, často obsahujú hovorové výrazy a abreviácie, ktoré sa nenachádzajú v bežných slovníkoch, preto štandardné techniky analýzy textov pri práci s emailovými dátami nemusia byť efektívne.

Emaily tiež obsahujú bohatšie typy dát, ako napríklad URL linky, HTML tagy alebo obrázky. Niektoré štúdie jednoducho zjednodušia tieto netextové dátové vstupy v štádiu predpripravovania dát - vymažu ich a ďalej pracujú len s textovými dátami. Tieto netextové dáta však môžu byť užitočné v iných oblastiach, ako napríklad detekcia spamu.

Emaily reprezentujúce ľudské sociálne organizačné vzťahy

Emailová aktivita sama o sebe reprezentuje bohaté ľudské sociálne a organizačné vzťahy, ktoré spájajú ľudí do komunit a komplexných systémov. Porozumenie organizačných štruktúr alebo vzťahov naprieč ľuďmi v organizácii môže byť veľmi užitočné aj v reálnom živote. Hlavné problémy, ktoré sú investigované v analýze emailov sú detekcia spamu, kategorizácia emailov, analýza kontaktov, analýza vlastností emailových sietí a vizualizácia emailov.

1.2 Vízia

Cieľom práce je oboznámiť čitateľa s oblasťou sociálnych sietí a špeciálne s témou analýzy emailových dát a tieto znalosti demonštrovať nad reálnymi emailovými dátami. Pre uskutočnenie tohto cieľa je potrebné naštudovať informácie z oblasti analýzy emailov, reprezentácie emailu v sieti a vizualizácie sociálnych sietí vrátane aktuálnych metód publikovaných v článkoch. K tomu sa viaže tiež prieskum reprezentácie a konštrukcie emailu ako prvku sociálnej siete.

Ďalej boli vybrané metódy detekcie rolí v sociálnej sieti a navrhnutá aplikácia, ktorá umožní analyzovať a vizualizovať analytické výsledky. V tejto aplikácii s jednoduchým a použiteľným užívateľským rozhraním budú implementované vybrané metódy analýzy a bude navrhnutá prehľadná vizualizácia vzťahov. Nakoniec bude vytvorená analýza jedného prvku siete a výsledky experimentov budú zrozumiteľne prezentované.

1.3 Štruktúra práce

V prvej kapitole je uvedený prieskum o aktuálnych vedeckých článkoch, ktoré sa zaoberajú analýzou emailov a reprezentácii emailu v sociálnych sieťach.

Ďalej sa čitateľ zoznámí s emailom ako komunikačným prostriedkom a dozvie sa, ako sú na tom emaily s popularitou aktuálne.

Potom je uvedený stručný prehľad teórie grafov a definícií určitých pojmov, ktorý je nevyhnutný k porozumeniu ďalších kapitol.

V ďalšej kapitole píšem o sociálnych sieťach, ich histórii a analýze sociálnych sietí a komunitej štruktúre sociálnych sietí.

Neskôr prechádzam k popisu a reprezentácii frameworku pre detekciu štrukturálnych rolí, popisujem sociálne roly definované v rámci tohto frameworku a následne v ďalšej kapitole referujem pomocou akých metód sa sociálne roly v rámci tohto frameworku identifikujú.

Na základe všetkých poznatkov práce je navrhnutá aplikácia vhodná k sledovaniu sociálnych rolí v emailovej sieti.

Ešte pred záverom sú uvedené výsledky prevedených experimentov týkajúcich sa poznatkami skúmanej sociálnej siete.

2 Súvisiace práce

Pre odhaľovanie vzťahov medzi ľuďmi, skupina a organizáciami z emailových sietí boli aplikované mnohé techniky a modely analýzy sociálnych sietí. Mnoho štúdií použilo maily spoločnosti *Enron* kvôli nedostatku dostupných veľkých súborov.

Napríklad Diesner, Carley a Frantz v [3] zkonštruovali z mailovej komunikácie spoločnosti Enron orientovaný graf zo vzťahu odosielateľ-prijemca, kde hrany boli vážené frekvenciou mailov, ktoré si medzi sebou poslali v čase. Potom aplikovali techniky analýzy sociálnych sietí. V práci popísali, ako vylepšili originálnu sadu a súčasné zistenia ich investigáciou vďaka analýze sociálnych sietí. Skúmajú dynamiku, štruktúru a vlastnosti organizačnej komunikačnej siete ako aj charakteristiky a vzory komunikačného správania zamestnancov z rôznych organizačných levelov. Zistili, že počas obdobia krízy sa komunikácia medzi zamestnancami stala viac rôznorodejšia v súvislosti so zavedenými kontaktami a formálnymi rolami. Taktiež počas obdobia kríz, predtým nekomunikujúci zamestnanci sa začali zapájať do vzájomného rozhovoru, takže interpersonálna komunikácia bola intenzívnejšia a sieť sa tým rozširovala. Tieto zistenia poskytli cenný pohľad do organizačnej krízy reálneho sveta, čo môže byť ďalej využité pre validáciu alebo tvorbu teórií a dynamických modelov organizačných kríz a tým to vedie k lepšiemu porozumeniu základných príčin organizačných kríz v organizáciách.

Xiaoyan Fu v [2] prezentoval rôzne metódy pre vizualizáciu emailových sietí. Vizualizácia objavuje komunikačné vzory medzi rôznymi skupinami, zobrazuje centrálnu analýzu s dôrazom na významné uzly. V práci zkonštruovali 2D vizualizáciu temporálnej emailovej siete, ktorá analyzuje vývoj emailových vzťahov, ktoré sa menia v priebehu času a zobrazenie prostredia pre nájdenie sociálnych kruhov odvodených od siete. Každá metóda bola vyhodnotená s rôznymi datasetmi od výskumnej orgnizácie. Taktiež rozšírili ich metódu pre vizuálnu analýzu siete emailových vírusov.

Ďalej Chapanond, Krishnamoorthy, Yener v [4] použil sieťové metriky a spektrálnu analýzu k analýze či už orientovaného alebo neorientovaného grafu emailov, ktorý skonštruoval zmenou prahovej hodnoty (napr. počtom vymenených emailov medzi užívateľmi). Ich výskum je postavený na vytvorení emailového grafu a štúdiu jeho vlatností či už pomocou teórie grafov alebo technikami spektrálnej analýzy. Grafová teoretická analýza zahŕňa výpočet niekoľkých grafových metrík, ako napríklad rozdelenie podľa stupňov, priemerný pomer vzdialeností, zhlukovací koeficient alebo kompaktnosť emailového grafu. Hodnoty metrík v dátovej sade emailov spoločnosti Enron porovnali aj s inými emailovými dátami.

Jednou z univerzálnejších prác je aj práca autorov Guanting Tang, Jian Pei, and Wo-Shun Luk [1]. Je to stručný prehľad hlavných výskumných snáh o analýzu mailov a popis metód, ktoré sa pri tejto analýze používajú. Nie len čo sa týka analytických alebo implemetnačných úloh, ale aj nástrojov, ktoré nám pri analýze vedia pomôcť. Aby zdôraznili rozdiely medzi analýzou mailov a bežnou analýzou textu, organizujú prieskum do piatich ťažších úloh a to: detekcia nevyžiadanej pošty, kategorizácia emailov, analýza kontaktov, analýza vlastností emailovej siete

a vizualizácia emailov. Tieto úlohy sú vlastne začlenené do rôznych spôsobov používania emailov. Systematicky preskúmajú bežne používané techniky a tiež budujú diskusiu o dostupných softwarových nástrojoch.

Na rozdiel od ostatných prác, Afra Abnar, Mansoureh Takaffoli, Reihaneh Rabbany, Omar R. Zaiane [9] definovali vlastnú metodiku pre analýzu sociálnej siete a definovali *Structural social role mining framework*, ktorý je navrhnutý pre identifikáciu štrukturálnych rolí, pre identifikáciu zmien v sieti a analýzu dopadu zmien na sieť. Definujú základné sociálne roly v sieti a navrhujú metodológie pre ich identifikáciu. Pre identifikáciu týchto rolí využívajú klasické prostriedky analýzy sociálnych sietí a tiež navrhujú nové metriky zahrňujú napríklad Betweenness centrality založenú na komunitách. Z tejto práce som vychádzala pri pomenovaní rolí zo siete a implementovala techniky pre ich identifikáciu.

Ďalšou prácou, ktorou som sa inšpirovala bola práca autorov Kudělka, Horák, Zehnalová [10], ktorá prezentuje analytický nástroj, ktorý bol vytvorený pre analýzu hlbších vzťahov v emailových dátach. Tieto vzťahy zahrňujú vzťahy založené na interakcii viacerých užívateľov v tíme. Analytické metódy popísané v práci sú založené na dvoch faktoroch. Prvým faktorom je kontext, čo je skupina viacerých užívateľov v kombinácii so slovami použitými v komunikácii. Druhým faktorom je časový interval, v ktorom bola začatá komunikácia. Práca prezentuje metódy pre váženie komunikácií, užívateľov a vzťahov, ako aj metód pre hľadanie komunit asociovaných so špecifickým kontextom.

3 Emailová komunikácia

3.1 Stručná história emailu

Za počiatky emailovej komunikácie možno považovať priližne rok 1965, kedy bola správa prenášaná medzi sálovými počítačmi pracujúcich v režime zdieľania času na univerzite *Massachusetts Institute of Technology*.

Od tejto doby prešla emailová komunikácia značným vývojom. Emaily, tak ako ich poznáme dnes, sú definované štandardom špecifikácie RFC2822 a sú prenášané pomocou komunikačných protokolov.

3.2 Štruktúra emailu

Každý email sa skladá z dvoch častí - z tzv. hlavičky (*header*) a tela emailu (*body*).

Hlavička emailu je generovaná automaticky pri vytvorení emailu a sú do nej postupne vkladané informácie zo serverov, cez ktoré správa prechádza (tzv. MTA). Pre bežných užívateľov sú z hlavičky najdôležitejšie tieto údaje: predmet správy, čas odoslania, emailová adresa odosielateľa a prijímateľa. Ostatné údaje emailoví klienti (označovaní tiež ako MUA ¹) väčšinou nezobrazujú.

Pri vytváraní emailu emailovým klientom sú väčšinou do hlavičky vložené tieto záhlavia:

- **Date** - aktuálny čas počítača, ktorý vložil záhlavie
- **From** - adresa odosielateľa
- **Cc** - špecifikuje ďalších adresátov
- **Bcc** - umožňuje rozosielanie správy medzi viacerých adresátov
- **Priotity** - priorita emailu, interpretácia sa líši vzhľadom k MUA
- **Reply-To** - špecifikuje adresu, na ktorú je zaslaná prípadná odpoveď
- **Subjekt** - predmet správy daný užívateľom
- **To** - udáva adresu príjemcu správy
- **Message-Id** unikátny identifikátor, ktorý je priradený MTA

Telo emailu obsahuje samotné dáta určené pre adresáta.

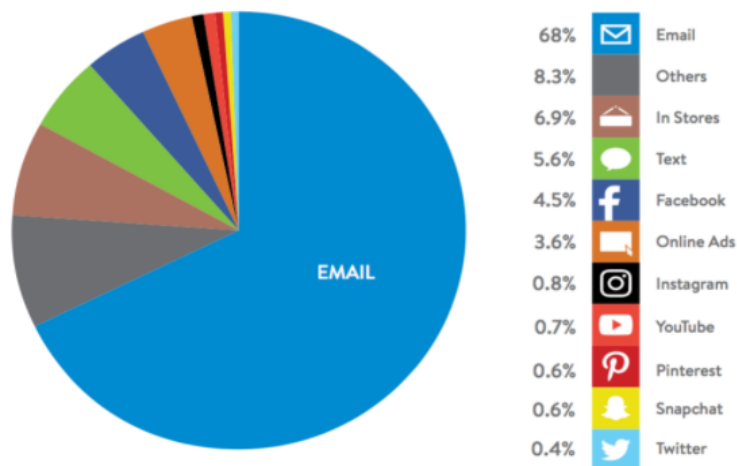
¹MUA - Mail User Agent, program, ktorý používa užívateľ na rozosielanie a prijímanie emailov (napr. Outlook), tento program komunikuje s MTA (Mail Transfer Agent), ktorý sa stará o prenos emailov v prostredí verejnej siete Internet.

3.3 Emaily v súčasnosti

Emaily teda existujú už niečo cez 50 rokov, ich popularita je však stále veľká vďaka ich efektívnosti, extrémne nízkym nákladom a kompatibilita s množstvom typov zariadení. Ako jedna z najrozšírenejších typov komunikácie v dnešnej dobe, emaily sú široko rozšírené v každodennom živote. Napríklad, spolupracovníci diskutujú prácu cez emaily, priatelia zdieľajú sociálne aktivity a skúsenosti aj cez emaily alebo veľké spoločnosti distribuujú reklamy práve pomocou emailov.

Aj keď by mnohí tvrdili, že éra emailov už je dávno preč a sú stále viac nahrádzané novými sociálnymi sieťami, nové výskumy ukazujú opak. Napríklad výskum z roku 2016 od spoločnosti Bluecore [13] ukazuje, že email je stále populárny aj u mladších generácií, hlavne na formálnu komunikáciu.

V tomto výskume boli spotrebitelia pýtaní, akú formu komunikácie preferujú pri komunikácii so značkami (internetovými obchodmi, na firemnú komunikáciu a celkovo formálnu komunikáciu). Prevažná časť opýtaných si vybrala email (68%).



Obr. 1: Akú formu komunikácie preferujete na formálnu komunikáciu?

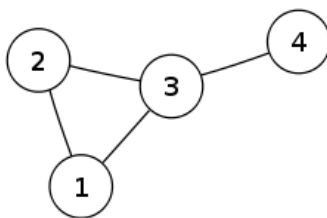
4 Definície a klasifikácie

V tejto kapitole popisujem všetky teoretické pojmy a metódy, ktoré v tejto práci spomínam a používam. V tejto kapitole budem používať matematické názvy podľa kontextu, v ktorom sa budem nachádzať.

4.1 Graf

- **Neorientovaný graf**

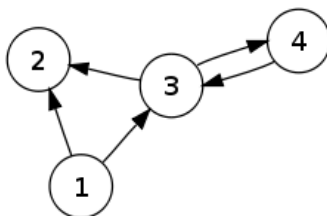
Neorientovaným grafom rozumieme usporiadanú dvojicu $G = (V, E)$, kde V je neprázdna množina *vrcholov* a E je neprázdna množina *hrán* - množina (niektorých) dvojprvkových podmnožín množiny V .



Obr. 2: Neorientovaný graf

- **Orientovaný graf**

Orientovaným grafom rozumieme usporiadanú dvojicu $G = (V, E)$, kde V je množina *vrcholov* a množina orientovaných *hrán* je $E \subseteq V \times V$.

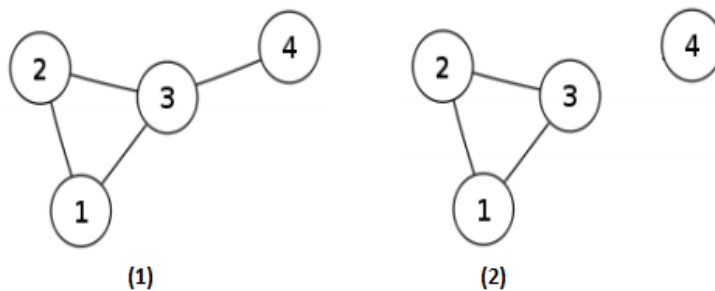


Obr. 3: Orientovaný graf

4.2 Súvislosť grafu

Hovoríme, že vrchol v je *dosiahnuteľný* z vrcholu u , ak v grafe existuje sled z vrcholu u do vrcholu v .

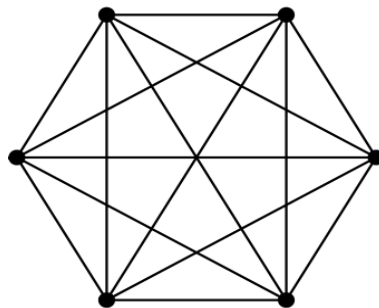
Graf nazveme *súvislý*, ak pre každé dva vrcholy u, v je vrchol v dosiahnuteľný z vrcholu u . V opačnom prípade je graf *nesúvislý*.



Obr. 4: Súvislý (1) a nesúvislý graf (2)

4.3 Úplný graf

Úplný graf na n vrcholech je neorientovaný graf, ktorý má hranu medzi každými dvoma vrcholmi. Počet jeho hrán je $m = n(n - 1)/2$.



Obr. 5: Úplný graf

4.4 Stupeň vrcholu

Stupeň vrcholu je počet vrcholov spojených s týmto vrcholom hranou, inými slovami: počet jeho susedov. V orientovanom grafe sa ešte rozlišuje vstupný a výstupný stupeň vrcholu podľa toho, koľko hrán z vrcholu vychádza alebo do neho vchádza.

4.5 Cesta

Cesta je postupnosť vrcholov v grafe taká, že medzi každými dvoma vrcholmi cesty je hrana a vrcholy sa neopakujú. V orientovaných grafoch sa ešte rozlišuje smer cesty, pričom orientácia hrán je stále rovnaká. Dĺžka cesty je počet hrán, ktoré obsahuje.

4.6 Uzavretá cesta

Uzavrená cesta, kružnica v neorientovanom a cyklus v orientovanom grafe, je cesta, ktorá začína a končí v rovnakom uzle.

4.7 Komponenta grafu

Komponenta grafu je súvislá časť grafu a medzi vrcholmi z rôznych komponent neexistuje žiadna hrana.

4.8 Metriky

V tejto časti popisujem metriky, ktoré v rámci identifikácie rolí v sieti používam. Ďalšie informácie o metrikách, ich praktickom využití a ich ďalších variantách sú zhrnuté v kapitole 7.

4.8.1 Closeness centrality (Centralita blízkosti)

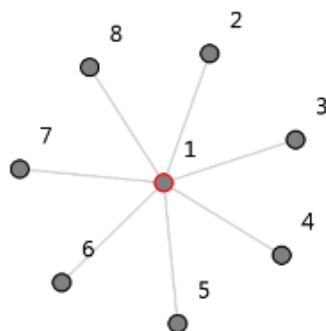
Táto centralita meria dôležitosť vrcholu grafu podľa priemernej hodnoty vzdialenosti od všetkých ostatných vrcholov v sieti. Aby dôležité vrcholy mali vyššie číslo, je táto centralita počítaná ako inverzná hodnota tohto priemeru. Vrchol dôležitý podľa tejto metriky môže mať dobrý prístup k informáciám o ostatných vrchoch alebo naopak môže ostatné vrcholy rýchlejšie ovplyvňovať.

Priemernú vzdialenosť vrcholu x_i od ostatných vrcholov možno formálne zapísať ako $l_i = \frac{1}{n} \sum_j d_{ij}$, kde n je počet vrcholov v grafe a d_{ij} je najkratšia cesta medzi vrcholmi x_i a x_j . Centralita je potom $C_i = \frac{1}{l_i}$.

4.8.2 Betweenness centrality (Centralita medzilahlosti)

Táto centralita sa odlišuje od ostatných uvedených. Jej hodnota pre vrchol je počet najkratších ciest medzi každými dvoma vrcholmi v grafe, na ktorých hodnotený vrchol leží. Pokiaľ medzi vrcholmi v sieti tečú nejaké informácie alebo sa posielajú správy, hodnota tejto metriky vyjadruje, aké množstvo informácií cez daný vrchol prejde. Táto centralita je tiež názorný príklad toho, že každá metrika počíta dôležitosť vrcholu úplne inak. Vrchol s vysokou centralitou medzilahlosti môže mať malý stupeň a nemusí ležať blízko ostatných vrcholov, stačí, keď cez neho prechádza veľa najkratších ciest. To môže nastať, pokiaľ vrchol je most medzi dvoma alebo viacerými komponentami v grafe, v extrémnom prípade pokiaľ je v strede grafu v tvare hviezdy (viď obrázok).

Betweeness vrcholu x_i možno spočítať ako $B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$, kde g_{st} je počet všetkých najkratších ciest medzi vrcholmi x_i a x_j a n_{st}^i je počet najkratších ciest, ktoré navyiac vedú cez vrchol x_i .



Obr. 6: Graf v tvare hviezd

4.8.3 Modularita

Modularita je metrika, ktorá udáva rozdiel medzi počtom existujúcich hrán medzi vrcholmi rovnakého typu a počtom takých hrán v náhodne vytvorenom grafe v pomere ku všetkým existujúcim hranám. Vrcholy rovnakého typu sú tie, ktoré patria alebo majú patriť do rovnakej skupiny alebo triedy (komunity).

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j)$$

Def: Modularita

5 Sociálna sieť

Sociálna sieť je množina sociálnych subjektov (uzly siete, spravidla jednotlivci alebo organizácie), ktoré sú prepojené jedným, alebo viacerými špecifickými druhmi vzájomnej závislosti, ako sú príbuzenstvo, priateľstvo, vzájomnosť, vízie, odpor, konflikt, obchod a pod. Sociálna sieť z pohľadu teórie grafov je definovaná ako graf $G(V, E)$, kde V je množina entít (uzlov) a E je množina vzťahov (hrán) medzi týmito entitami.

Entity grafu môžu byť rôzne (zákazníci, jednotlivci, webové stránky, bankové účty, kreditné karty, produkty). Nie je pravidlom, že len sociálna sieť ako ju pozná mnoho ľudí je sociálnou sieťou aj formálne. Prvky sociálnej siete môže mať napríklad aj skupina spolupracujúcich ľudí.

5.1 História sociálnych sietí

Pod pojmom sociálna sieť si väčšina ľudí v dnešnej dobe predstaví služby ako *Facebook*, *Twitter* a pod. Tento pojem ale vznikol dlho pred vznikom internetu a dnešných sociálnych sietí. Prívlastok sociálny, ktorý sa v dnešnej dobe často vynecháva, je dôsledkom pôvodu analýzy sociálnych sietí. V druhej polovici 20. storočia sa simultánne v rôznych oblastiach skúmania vzťahov a chovania objavil nový pohľad na vzťahy medzi sociálnymi jednotkami a to ako na sieť, graf. Preto prví predstavitelia analýzy sociálnych sietí boli pôvodne sociológovia alebo psychológovia (napríklad Moreno, Cartwright, Newcomb, Bavelas) a antropológovia (Barnes, Mitchell). Prvé použitie termínu "sociálna sieť" sa pripisuje Barnesovi (1954).

V 30. rokoch 20. storočia psychiater Moreno rozvíjal sociometriu, predchodcu dnešnej analýzy sociálnych sietí. Vypytoval sa ľudí na priateľské vzťahy a skúmal, ako tieto vzťahy ovplyvňujú ich chovanie. Potom vynašiel (sám to tvrdil) tzv. *sociogram*, čo je diagram reprezentujúci ľudí ako body a vzťahy medzi ľuďmi ako úsečky, teda dnešnú sociálnu sieť. Tento pojem sa ale začal používať až neskôr. Pomocou neho hľadal výrazné a izolované osoby v spoločnosti.

Zhruba o 20 rokov neskôr antropológ Barnes začal skúmať, ako ovplyvnia vzťahy medzi ľuďmi nielen jednotlivcov, ale aj spoločnosť ako celok a zameral sa na štúdium skupín, komún. Na práci Barnes a jeho spolupracovníkov naviazala na Univerzite na Harvarde skupina vedená Harrisom Whitom. Tá začala budovať matematickú teóriu okolo dôležitejších pojmov zo sociálnych vied a umožnila tieto javy matematicky vyjadriť, merať a modelovať.

V druhej polovici 20. storočia sa rozšírilo povedomie o sociálnych sieťach a metódy sa začali používať aj v ďalších oboroch ako ekonómia, biológia, doprava atď.

5.2 Analýza sociálnych sietí

Analýza sociálnych sietí je interdisciplinárna veda s koreňmi v sociológii, psychológii, štatistike a teórii grafov. Analýza sociálnej siete chápe sociálnu sieť ako systém prepojenia uzlov (individuálnych aktérov) prostredníctvom hrán (ich vzťahov). Možno teda povedať, že nadväzuje na matematickú teóriu grafov a metódy sieťovej analýzy. Výsledkom analýzy môže teda byť mapa

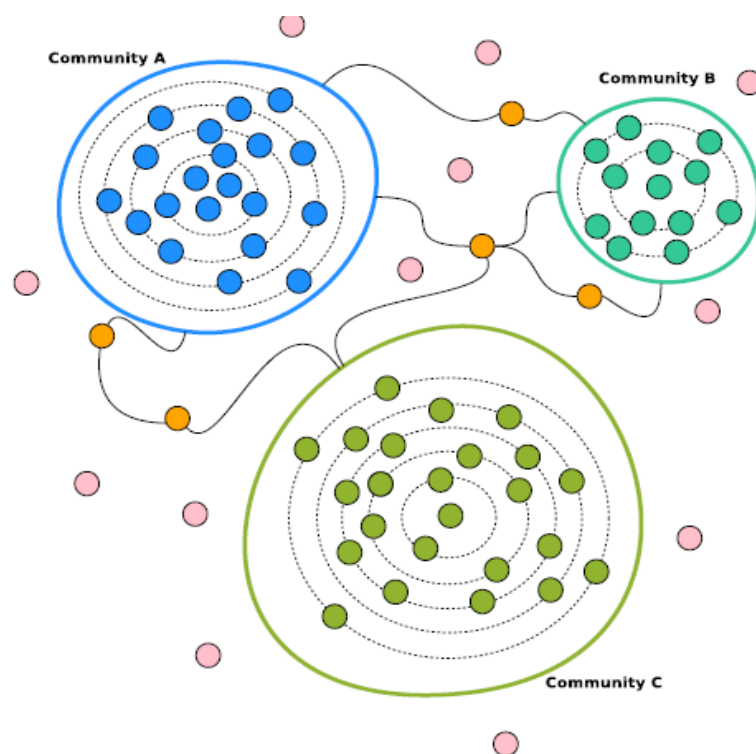
graficky znázorňujúca všetky prvky skúmaného sociálneho systému a ich vzťahy (resp. vybrané charakteristiky jednotlivých vzťahov vyjadrené vhodným spôsobom graficky). Charakteristikou môže byť napríklad vzájomná sympatia či antipatia alebo pravidelná vzájomná komunikácia alebo spolupráca.

Analýza sociálnych sietí vystupuje napríklad ako základná technika v rámci modernej sociológie, antropológie, sociálnej lingvistiky, geografie, sociálnej psychológie, ekonómie a biológie rovnako ako populárna téma pre výskum.

5.3 Komunity v sociálnych sieťach

Sociálne siete sú riedke grafy zložené z hustých podgrfov. Tieto husté podgrafy sú nazývané komunity. Najčastejšia definícia komunity: *Komunita je zhuk uzlov, kde počet vnútorných hrán v komunite je väčší ako počet vnokajších hrán – mimo komunity.* [8]

Algoritmy pre dolovanie komunit sú založené na spojoch medzi uzlami, ktoré naznačujú spojenie dvoch entít. Napr. SCAN (Structural Clustering Algorithm for Networks) je metóda pre detekciu komunit v súvislosti na to, ako uzly zdieľajú svojich susedov len s ohľadom na priame spojenie. Teda ak sú dva uzly spojené a tiež zdieľajú rozumné množstvo ich susedov, patria do rovnakej komunity.



Obr. 7: Sieť s viacerými komunitami

5.4 Detekcia komunít

Detekcia komunít je proces identifikácie zhlukov uzlov siete silne prepojených medzi sebou a menej silne prepojených so zvyškom siete. Detekcia komunít v grafoch má za cieľ identifikovať moduly a ich prípadnú hierarchickú organizáciu.

Problém detekcie komunít vyžaduje rozdelenie siete do komunít husto prepojených uzlov, pričom uzly patriace do odlišných komunít sú len slabo prepojené. Presné formulácie tohto optimalizačného problému sú známe ako výpočtovo neriešiteľné. Vyhľadávanie rýchlych algoritmov pritiahlo veľký záujem vďaka zvyšujúcej sa dostupnosti rozsiahlych sieťových dátových súborov a vplyvu sietí na každodenný život. Môžeme rozlišovať niekoľko typov algoritmov detekcie komunít: *rozdeľovacie* algoritmy - tie detekujú spojenie vnútri siete a postupne ich odstraňujú zo siete, *aglomeratívne* algoritmy - zlučujú podobné uzly a postupne komunity podľa spoločných črt a *optimalizačné* metódy sú postavené na maximalizácii objektívnej funkcie. Kvalita rozdielov vyplývajúcich z týchto metód sa často meria takzvanou modularitou. Je to hodnota v intervale od -1 do 1, ktorá meria hustotu spojov vnútri komunít v porovnaní s prepojeniami medzi komunitami.

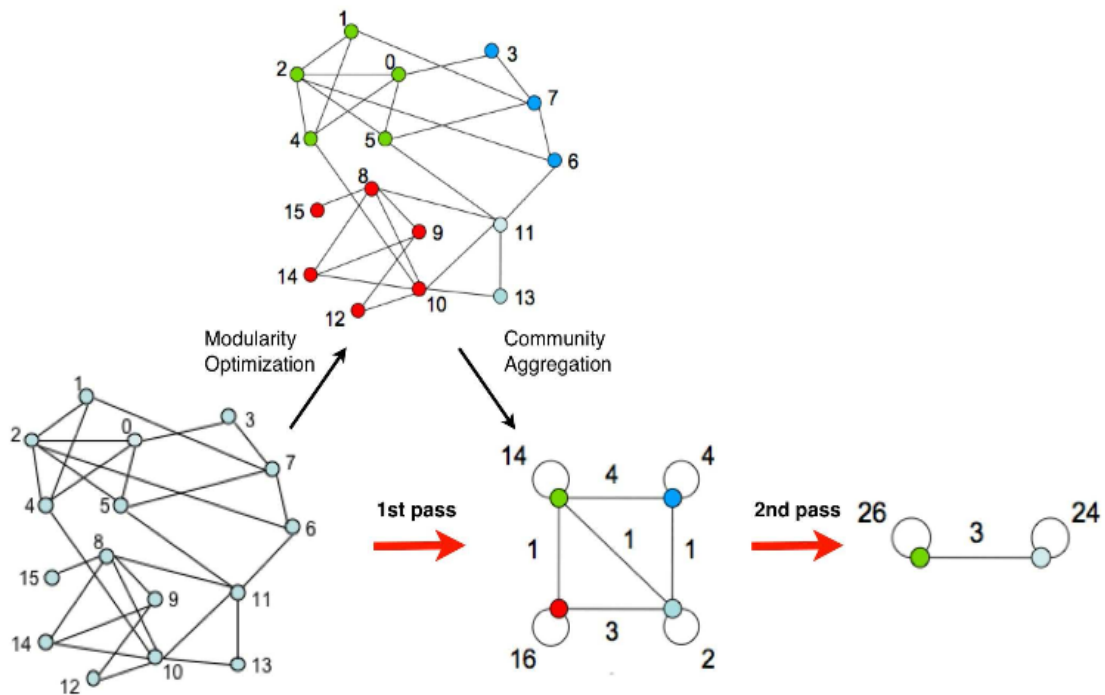
5.4.1 Louvainov algoritmus pre detekciu komunít

Veľmi obľúbeným a rýchlym algoritmom pre detekciu komunít je Louvainova metóda, ktorú navrhli Blondel, Guillaume, Lambiotte a Lefebvre [12]. Je to jednoduchá metóda pre extrakciu komunitnej štruktúry veľkých sietí. Je to heuristická metóda, ktorá je postavená na optimalizácii modularity. Je preukázané, že prekoná všetky ostatné známe metódy detekcie komunít, pokiaľ ide o čas výpočtu. Navyše kvalita detekovaných komunít je veľmi dobrá.

Výpočet algoritmu je rozdelený do dvoch fáz, ktoré sa iteratívne opakujú. Predpokladajme, že začíname s váženou sieťou s N uzlami. Ako prvé označíme každý uzol siete inou komunitou. Takže v tomto prvotnom rozdelení je toľko komunít, ako je uzlov. Potom pre každý uzol i uvažujeme susedov j a vyhodnotíme prírastok modularity, ktorý by nastal, ak z i sme odstránili uzol i z jeho komunity a priradili by sme ho do komunity uzla j . Uzol i je potom vložený do komunity, pre ktorú je tento prírastok najvyšší, ale len ak je tento prírastok kladný. Ak nie je možný žiadny kladný prírastok, uzol i ostáva vo svojej komunite. Tento proces je aplikovaný opätovne a sekvenčne pre všetky uzly kým sa nedosiahne žiadne zlepšenie a prvá fáza je kompletná. Prvá fáza končí, keď je dosiahnuté lokálne maximum modularity, keď žiadny uzol už nemôže zlepšiť modularitu. Je taktiež dôležité, že výstup algoritmu závisí na postupe, v ktorom sú uzly brané do úvahy. Výsledky algoritmu ale naznačujú, že usporiadanie uzlov nemá významný vplyv na získanú modularitu. Zoradenie však môže ovplyvniť výpočtový čas. Problém pri výbere objednávky preto stojí za to študovať, pretože by mohol poskytnúť dobrú heuristiku na zvýšenie výpočtového času.

Druhá fáza algoritmu spočíva vo vytvorení novej siete, ktorej uzly sú komunity nájdené počas prvej fázy algoritmu. K tomu, aby sa nová sieť vytvorila, váhy spojení medzi novými uzlami sú

dané sumou váh prepojení medzi uzlami korešpondujúcich dvoch komunít. Spojenia medzi uzlami tej istej komunity vedú k slučkám v novej sieti. Keď je druhá fáza kompletná, je možné znovu aplikovať prvú fázu algoritmu na výslednú váženú sieť a proces opakovať. Pri konštrukcii sa počet komunít znižuje pri každom priechode. Proces sa opakuje, kým nie sú žiadne ďalšie zmeny a dosiahne sa maximálna modularita.



Obr. 8: Vizualizácia krokov Louvainovho algoritmu.

Každý priechod je tvorený dvomi fázami: prvá, kde je modularita optimalizovaná tým, že umožňuje len miestne zmeny komunít a druhá, kde nájdené komunity sú agregované tak, aby bolo možné vytvoriť sieť komunít. Priechody sú opakované iteratívne kým nie je možný žiadny nárast modularity.

6 SSRM - Framework pre detekciu štrukturálnych rolí v sociálnych sieťach

Afra Abnar, Mansoureh Takaffoli, Reihaneh Rabbany, Osmar R. Zaiane [9] definovali *Structural social role mining framework*, ktorý je navrhnutý pre identifikáciu štrukturálnych rolí, pre identifikáciu zmien v sieti a analýzu dopadu zmien na sieť. Definujú základné sociálne roly v sieti (menovite Leader, Outermost, Mediator, Outsider).

6.1 Rola v kontexte SSRM

Sociálna rola je síce základný sociologický pojem, ale stále neexistuje žiadny konsenzus v jej definícii. Podľa SSRM je rola je považovaná za pozíciu jednotlivca v spoločnosti. Informácie o sociálnej sieti sú kategorizované do štrukturálnych a neštrukturálnych vlastností. Štrukturálne vlastnosti sú príbuzné ku konštrukcii grafu ako sú spojenia entít (hrany), štruktúra susedov a pozícia entity v tejto štruktúre. Ale neštrukturálne vlastnosti sú ostatné informácie, ktoré neodrážajú konštrukciu grafu ako atribúty entít a spojení. SSRM definuje rolu v sieti ako: Rola entity v sieti je to, ako sa entita správa voči ostatným a jej vplyv na atribúty a štruktúry ostatných entít.

6.2 Roly definované v SSRM

Ludské siete sú vnútorne zložené z viacerých komunít. V sociálnej sieti s viacerými komunitami, vlastnosti uzlov kolíšu podľa toho, či je existencia komunít dostatočná alebo zanedbateľná. Z pohľadu sociálnej siete, uzol môže byť centrom celej siete, ale nie centrom v jeho komunite. SSRM sa teda zameriava na štúdium sociálnych sietí s predpokladom existencie komunít v sieti, ako jej základnej črty.

V sociálnych sieťach môžu byť komunity explicitné alebo implicitné. Explicitné komunity sú postavené nezávisle na jej členoch a sú založené na množine pravidiel. V tomto prípade, ľudia sa stanú členmi tejto komunity častejšie až po zformovaní komunity. Zamestnanci firmy alebo študenti sú príkladom dvoch explicitných komunít. Zatiaľ čo formácia implicitných komunít ťažko závisí na jej členoch a spojeniach. Tým pádom neexistuje žiadna externá podmienka na vybudovanie implicitnej komunity. Implicitné komunity sú postavené postupne ako sa ľudia spoločne stretávajú. Napríklad, skupina priateľov, v ktorej nie je žiadne pravidlo pre správanie sa jednotlivcov, je príklad implicitnej komunity. V oboch prípadoch explicitnej aj implicitnej komunity, by mali existovať aj špeciálni jednotlivci, ktorí tieto komunity manažujú a kontrolujú. Napríklad v školskej triede je to učiteľ alebo inštruktor. Pre firmu to je manažér vo vedení a pre skupinu priateľov je to zase človek, ktorého komunikačné schopnosti prinášajú ďalších členov alebo posilňujú vzťahy medzi tými stálymi. Títo dôležití jednotlivci sú ešte viac výrazní, keď je komunita obrovská.

Podľa toho SSRM framework definuje pre jednotlivcov v sociálnej sieti určité roly podľa ich vzťahov a pozícií v komunitách až po ich interakcie s ostatnými jednotlivcami. Z perspektívy komunit, v sieti existujú jednotlivci niekoľkých typov:

- so žiadnym vzťahom ku nejakej komunite
- so spojením s viacerými komunitami
- dôležitý členovia komunity
- bežný členovia komunity, ktorí formujú väčšinu
- nedôležitý členovia komunity, ktorí nemajú na komunitu pozorovateľný efekt

Na základe týchto poznatkov SSRM definuje štyri základné roly - **leader**, **mediator**, **outermost** a **outsider**.

6.2.1 Leader

Sú mimoriadni jednotlivci v zmysle centrality alebo významu v každej komunite. V reálnom svete bývajú títo členovia siete veliteľmi, riaditeľmi, manažérmi, vládcami, prezidentami, autoritami, administrátormi atd.

6.2.2 Outermost

Je to časť menej dôležitých jednotlivcov v každej komunite, ktorých vplyv a efekt na komunitu sú nižšie ako vplyv väčšiny členov komunity. Miesta, kde sa môže outermost v sieti nachádzať sú periférie alebo hranice grafu.

6.2.3 Mediator

Sú to jednotlivci, ktorí zohrávajú dôležitú rolu v spojení komunit v medzi sebou. Fungujú ako mosty medzi odlišnými komunitami. Do tejto skupiny patria vyjednávači, sprostredkovatelia alebo aj rozbočovače v sieti.

6.2.4 Outsider

Sú to jednotlivci, ktorí nie sú spojení so žiadnou komunitou v sieti. Buď majú takmer rovnaké prepojenie k rôznym komunitám alebo majú len veľmi slabé väzby na komunity.

7 Identifikácia štrukturálnych sociálnych rolí

Majúc sieť s komunitami explicitne známymi alebo extrahovanými nejakým dolovacím algoritmom, následne popisujem metodológie pre identifikovanie definovaných štrukturálnych rolí.

7.1 Outsider

Najviac priamočiarou rolou pre identifikáciu je outsider. Je to jednotlivec, ktorý v sieti nepatrí do žiadnej komunity. Identifikácia tejto roly je tak celkom priamočiara.

7.2 Leader

Leader je v každej komunite výnimočný centrálny člen. Pre identifikovanie takýchto uzlov SSRM využíva metriku *closeness centrality*.

7.2.1 Closeness centrality (Centralita blízkosti)

V súvislosti s grafom *closeness centrality* uzlu je metrika centrality v sieti, vypočítaná ako súčet dĺžok najkratsích ciest medzi uzlom a všetkými ostatnými uzlami v grafe. Čiže čím viac je uzol centrálnější, tým bližšie je k ostatným uzlom.

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}}$$

Def: Closeness centrality

Pre každý uzol sa stanoví hodnota *closeness centrality*. Hodnoty *closeness centrality* sú blízke normálnemu rozdeleniu, v ktorom 95% populácie dát patrí do intervalu $[\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma]$

Leadri ležia na hornom chvoste distribučnej funkcie, a teda horný interval použijeme pre identifikovanie leadrov. A teda uzly, ktoré majú väčšiu hodnotu *closeness centrality* ako krajná hodnota tohto intervalu, sú identifikované ako leadri.

7.3 Outermost

Podobne ako pri role *Leader* pre identifikovanie outermostov sa využíva metrika *closeness centrality*. Outermosti budú ležať však na spodnom chvoste distribučnej funkcie *closeness centrality*.

A tak teda uzly, ktoré majú hodnotu *closeness centrality* nižšiu ako $[\mu - 2 \cdot \sigma]$, sú outermosti.

7.4 Mediator

Rolu mediator zastávajú tí jednotlivci, ktorí spájajú viacero komunít a sú tzv. spojmy medzi komunitami.

Pre identifikáciu mediátorov sa definujú metriky založené na metrike betweeness centrality a to: *LBetweeness* - *LBC* a *CBetweeness* - *CBC* a ďalej metriky, ktoré vyjadrujú koľko rozdielnych komunit uzol spája: *DSCount* a *DSPair*.

7.4.1 LBeweeness

LPath - Pred definíciou *LBetweeness* je potrebné definovať *LPath* a to nasledovne: *LPath* je množina všetkých najkratších ciest medzi lídrami dvoch rozdielnych komunit.

$$LPath = l | startNode(l) \in leaderSet(c_i) \wedge endNode(l) \in leaderSet(c_j) \wedge c_i \neq c_j$$

Def: Lpath

LBetweeness centralita pre uzol v - $LBX(v)$ je počet jedinečných *LPath* ktoré obsahujú v .

Ak pre každú cestu $p \in LPath$ definujeme $I_l(p, v) = 1$ ak v leží na p , inak $I_l(p, v) = 0$ potom:

$$LB(v) = \sum_{p \in LPath} I_l(p, v)$$

Def: LBbetweeness

7.4.2 CBetweeness

CBetweeness počítá počet najkratších ciest medzi rozdielnymi komunitami. s_p a e_p označujú štartovací a koncový uzol najkratšej cesty p . Taktiež c_v označuje komunitu, do ktorej uzol v patrí. Množina všetkých najkratších ciest, ktoré spájajú rozdielne komunity: $CPaths = \{p | c_{s_p} \neq c_{e_p}\}$. Taktiež definujeme $I_p(p, v) = 1$ ak v leží na ceste p a 0 keď neleží.

$$CB(v) = \frac{1}{2} \sum_{p \in CPaths} I_p(p, v)$$

Def: CBetweeness

7.4.3 Normalizovaná verzia CBetweeness

Pravdepodobnosť nájdenia viac viditeľných mediátorov vo väčších komunitách je väčšia v porovnaní s menšími komunitami. Táto situácia sa stáva, pretože vo väčších komunitách je pochopteľne viac uzlov, čo vedie k viacerým najkratším cestám medzi nimi. Pre kompenzáciu tohoto efektu je definovaná normalizovaná verzia *CBC*:

Navrhnuté metriky *CBC* a *LBC* sú nevyhnutné pre identifikovanie mediátorov, ale nie sú dostatočné. Napríklad pre sieť pozostávajúcu z desiatich komunit a dvoch mediátorov M_1 a M_2 , kde oba ležia na sto najkratších cestách medzi komunitami majú oba rovnaké hodnoty *CBC*.

$$NBC(v) = \frac{1}{2} \sum_{p \in CPaths} \frac{I_p(p, v)}{\min(|c_{s_p}|, |c_{e_p}|)}$$

Def: Normalizovaná verzia CBetweeness

Kdežto M_1 spája dve rozdielne komunity, kým M_2 spája všetkých 10. Pri takomto scenári M_2 spája komunity viac globálne a mal by byť skôr posudzovaný ako mediátor ako M_1 . A tak *SSRM* definuje tzv. metriku **skóre rozmanitosti**, ktorá označuje rozdielne komunity, ktoré sú prepojené cez uzol.

7.4.4 Skóre rozmanitosti

Táto metrika ukazuje koľko rozdielnych komunít je spojených cez špecifický uzol v . Túto metriku definujeme v dvoch variantach:

1. **DSCount** - je definovaný ako počet rozdielnych komunít, ktoré sú spojené daným uzlom. Nech $I_d(c_i, v) = 1$ ak $\exists p \in CPaths : s_p \in c_i \wedge v \in p$. Potom DSCount uzla v je definované ako:

$$DS_{count}(v) = \frac{1}{2} \sum_{c_i} I_d(c_i, v)$$

Def: DSCount

2. **DSPair** - Skóre rozmanitosti môže byť definované ako počet párov komunít, ktoré majú najmenej jednu najkratšiu cestu medzi ich členmi, ktoré prechádzajú uzlom v . Definujeme $I_d(c_i, c_j, v) = 1$ ak $\exists p \in CPaths : s_p \in c_i \wedge e_p \in c_j \wedge v \in p$

$$DS_{pair}(v) = \frac{1}{2} \sum_{c_i} \sum_{c_j \neq c_i} I_d(c_i, c_j, v)$$

Def: DSPair

Aj keď viac mediátorov môže mať rovnaké hodnoty jednotlivých metrík, môžu sa odlišovať napríklad v počte komunít, ktoré spájajú. *SSRM* to berie do úvahy a definuje tzv. *mediacy score* ako násobok normalizovanej CBetweeness a skóra rozmanitosti:

$$MS(v) = NBC(v) \cdot DS_{count}(v)$$

Def: Mediacy score

8 Aplikácia

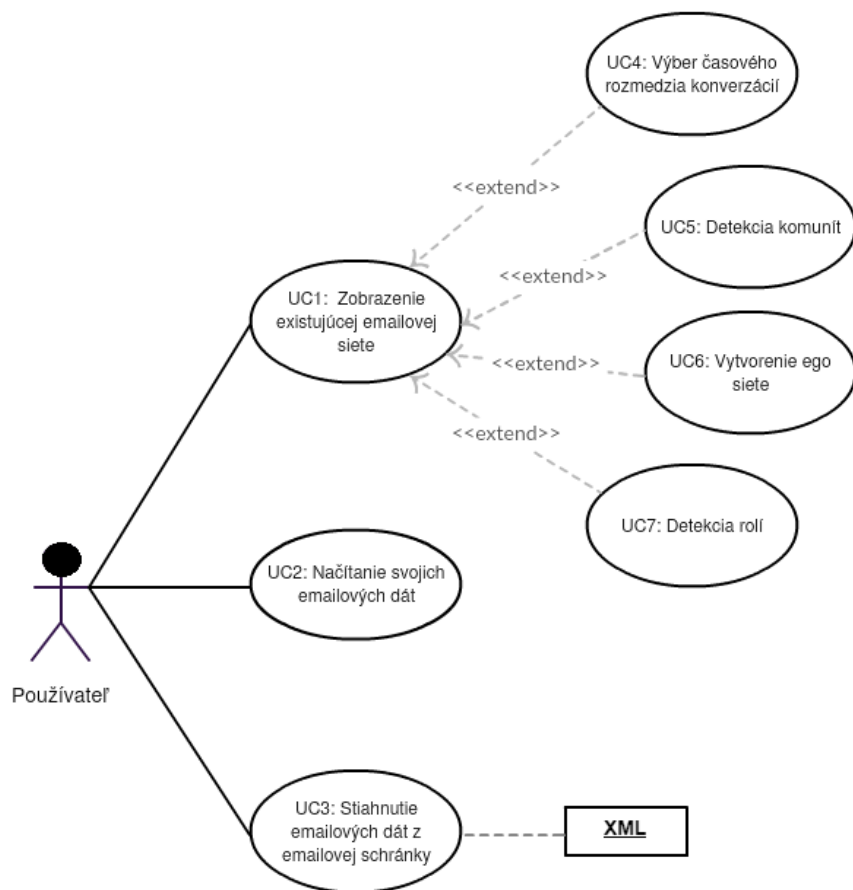
Táto kapitola obsahuje všetky podrobnosti o vývoji aplikácie, návrhu a ďalej špecifikáciách požiadavkov. Sú tu uvedené informácie o implementácii, návrhu, návrhových vzoroch, ale aj konštrukcii siete, predpríprave dát. Táto časť taktiež obsahuje diagramy najdôležitejších tried aplikácie alebo diagramy prípadov použitia.

8.1 Špecifikácia

Aplikácia slúži ako užívateľské rozhranie na analýzu emailovej komunikácie a vizualizáciu analytických výstupov. Aplikácia umožňuje exportovať dáta z emailovej schránky alebo importovať vlastný XML súbor s emailovými dátami a ďalej s týmito dátami pracovať a zobrazovať sieť emailovej komunikácie. Umožňuje vytvorenie ego-siete alebo detekovať vo vytvorenej sieti komunity. Najdôležitejšou časťou aplikácie je detekcia štrukturálnych rolí v sieti, čiže detekcia dôležitých a nedôležitých členov emailovej komunikácie.

8.1.1 Funkčné požiadavky

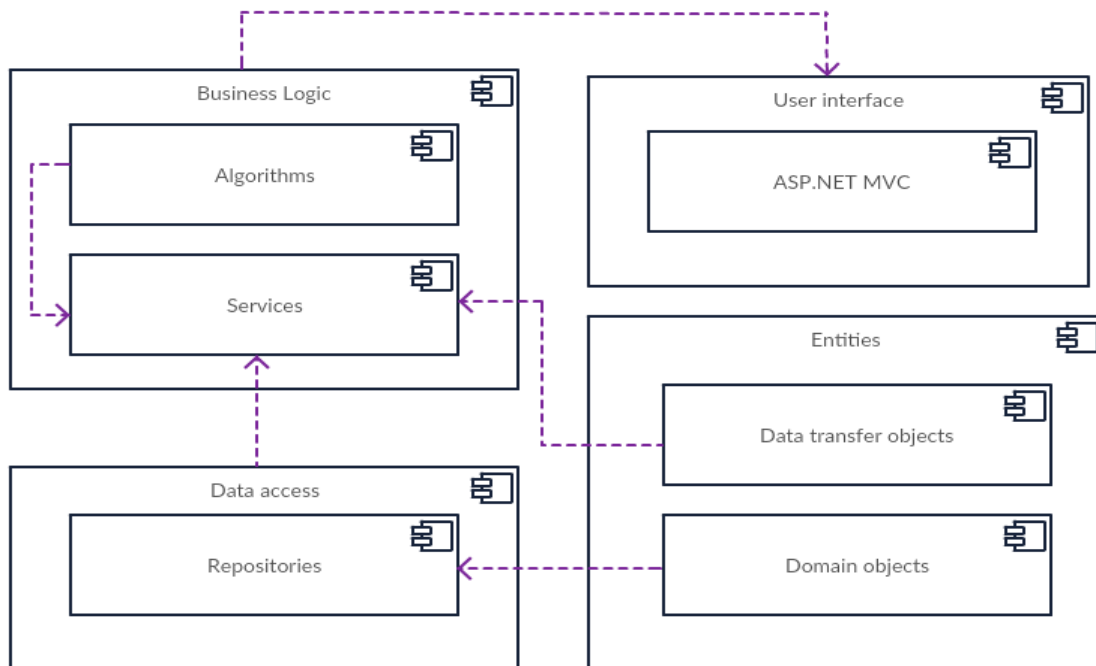
- Export dát z emailovej schránky
- Import vlastného XML súboru s emailovými dátami
- Zobrazenie informácií o emailovej sieti
- Vizualizácia emailovej siete
- Vytvorenie ego-siete
- Detekcia komunít
- Detekcia štrukturálnych rolí v sieti
- Výber časového rozmedzia emailových konverzácií



Obr. 9: UseCase Diagram

8.2 Návrh

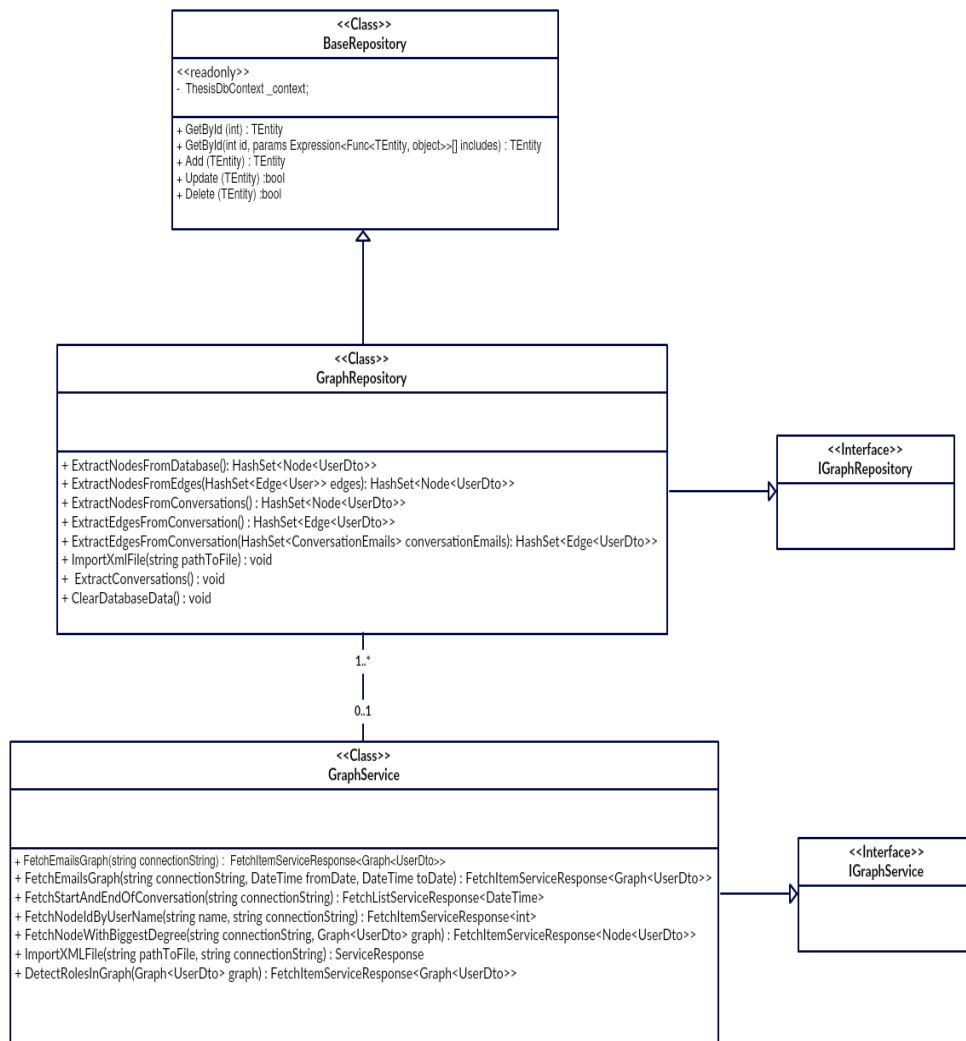
Aplikácia je vytvorená ako .NET aplikácia (veria .NET Frameworku 4.6). Je vytvorená ako trojvrstvová, pre uloženie dát sa používa SQL databáza. Najnižšia vrstva aplikácie slúži na získavanie dát z databázy, pre prepojenie s databázou a posielanie dát z aplikácie do databázy používam Entity Framework a používam tu návrhový vzor Repository. Od tejto časti je oddelená časť s business logikou a na najvyššej časti, ktorá slúži len na zobrazenie dát a komunikáciu s užívateľom, používam známy prístup Model-View-Controller.



Obr. 10: Diagram komponent znázorňujúci jednotlivé komponenty architektúry aplikácie

8.2.1 Návrhové vzory

Repository Návrhový vzor Repository je základným kameňom doménou riadeného návrhu. Model aplikácie teda nemá poňatie o tom, akým spôsobom je perzistovaný. O to sa stará práve Repository. Navyiac práve vďaka tomu, že sa o persistenciu stará cudzí objekt, stačí poznať len jeho rozhranie a v prípade potreby ho ľahko nahradiť iným. [11]

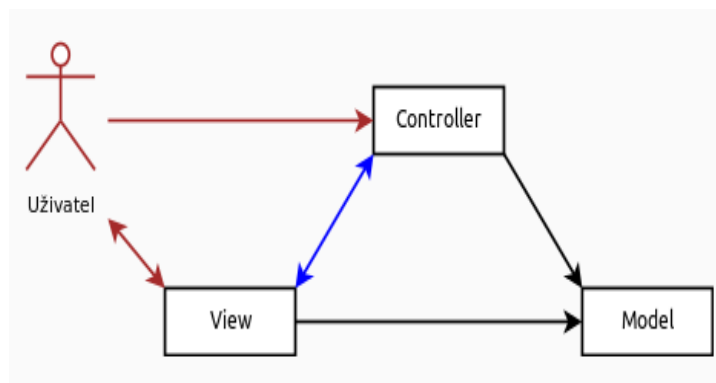


Obr. 11: Triedny diagram - Repository pattern

Model-View-Controller V aplikácii je použitý tradičný vzor Model View Controller (MVC). Je to jeden z najpoužívanejších a najobecnejších architektonických vzorov.

MVC rozdeľuje program do troch hlavných častí:

- **Model** - dáta a súvisiace operácie
- **View** - prezentácia dát (užívateľské rozhranie), obsahuje priamy odkaz na model, aby mohol jeho dáta prezentovať vonkajšiemu svetu
- **Controller** - riadi tok udalostí v programe, konkrétne v tejto aplikácii kontrolery obsahujú len volanie metód z inej vrstvy aplikácie



Obr. 12: Model-View-Controller

8.3 Dôležité rozhodnutia

Pri navrhovaní aplikácie bolo potrebné urobiť niekoľko dôležitých rozhodnutí.

8.3.1 Dostupnosť dát

Pôvodne sa zvažovalo použitie aplikácie a analýzy dát nad verejne dostupnou anonymizovanou emailovou sadou. Emailových dát je ale veľmi málo a chcela som, aby sa výsledky práce dali overiť nie len inými analytickými nástrojmi, ale aj empiricky. Takže som využila to, že pracujem a moja emailová schránka teda nie je chudobná na maily. Navyše mi radi pomohli aj moji kolegovia a poskytli mi svoje emailové dáta. Takto som zozbierala reálne emailové dáta štyroch ľudí, o ktorých je známe ich postavenie v tíme alebo aj dátum nástupu do práce. Takže výsledky daných algoritmov som vedela porovnať s reálnou situáciou v kolektíve.

8.3.2 Webová vs. desktopová aplikácia

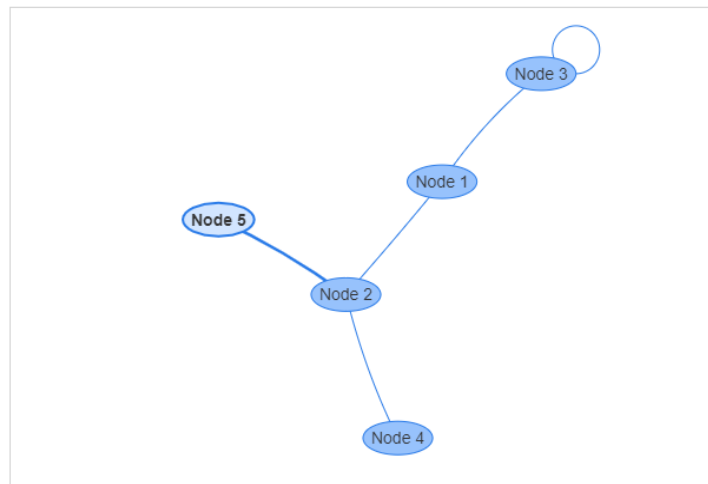
Bolo nutné sa rozhodnúť, či vyvíjať aplikáciu ako webovú alebo desktopovú. Ako platforma bola zvolená Microsoft .Net a programovací jazyk C#. Jednou z variant bola desktopová aplikácia vyvíjaná vo Windows Forms. WinForms je osvedčená technológia, je vyladená, v základe obsahuje veľké množstvo grafických prvkov. Toto sú výhody rozšírenej a dlho používanej technológie. Nevýhoda je ale práve zastaralosť a ťažkopádnosť v kreslení a spravovaní grafického rozhrania. Keďže ale doba ide dopredu a web a webové aplikácie sú stále viac používané a v súčasnosti existuje mnoho grafických knižníc pre vizualizáciu grafického rozhrania, rozhodla som sa aplikáciu vyvíjať ako webovú.

8.4 Použité knižnice

vis.js

Vis.js je dynamická vizualizačná knižnica. Knižnica je navrhnutá tak, aby bola ľahko ovládateľná a aby mohla spracovať veľké množstvo dynamických dát a umožňovala manipuláciu s dátami a

interakciu s nimi. Knižnica sa skladá z častí *DataSet*, *Timeline*, *Network*, *Graph2d* a *Graph3d*. Pre moju aplikáciu som používala len časť *Network*.



Obr. 13: Jednoduchá sieť vytvorená s použitím knižnice vis.js

```
<style type="text/css">
#mynetwork {
width: 600px;
height: 400px;
border: 1px solid lightgray;
}
</style>

<script type="text/javascript">
// create an array with nodes
var nodes = new vis.DataSet([
{id: 1, label: 'Node 1'},
{id: 2, label: 'Node 2'},
{id: 3, label: 'Node 3'},
{id: 4, label: 'Node 4'},
{id: 5, label: 'Node 5'}
]);

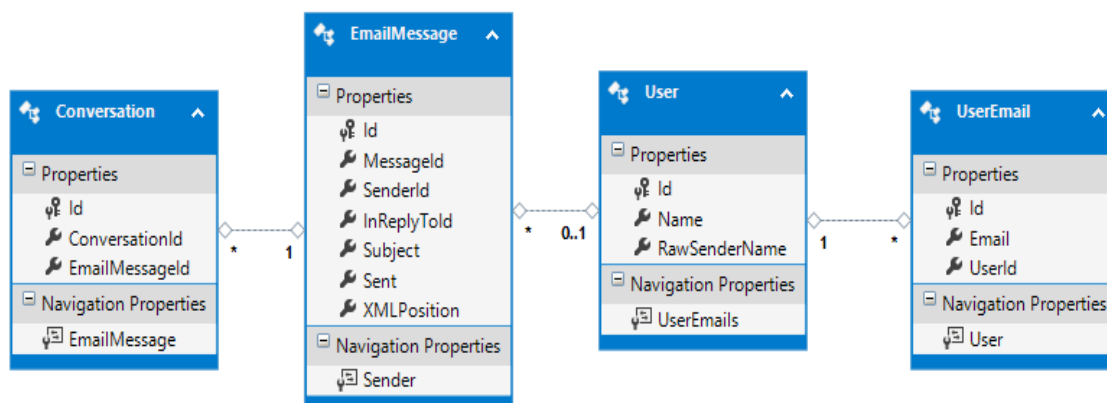
// create an array with edges
var edges = new vis.DataSet([
{from: 1, to: 3},
{from: 1, to: 2},
{from: 2, to: 4},
{from: 2, to: 5},
{from: 3, to: 2}
]);

// create a network
var container = document.getElementById('mynetwork');
var data = {
nodes: nodes,
edges: edges
};
var options = {};
var network = new vis.Network(container, data, options);
</script>
```

Obr. 14: Príklad použitia knižnice vis.js

8.5 Import dát

Do aplikácie je možné nahrať XML súbor, ktorý je spracovaný uloženou SQL procedúrou, ktorá rozparsuje emailové dáta na jednoduché entity - *User*, *EmailMessage*, *UserEmail* a *Conversation* a uloží ich do SQL databázy.



Obr. 15: Doménový model

8.6 Implementácia

Aplikácia je napísaná v jazyku C#, grafické rozhranie je nainplementované pomocou návrhového vzoru Model View Controller a graf bol vizualizovaný pomocou knižnice vis.js. Aplikácia bola vyvíjaná vo Visual Studiu 2017.

8.6.1 Konštrukcia siete

Rozdiel medzi prístupom rôznych štúdií a mojím prístupom pri konštrukcii grafu z emailového datasetu je v konštrukcii komunikačnej siete. Ako základnú stavebnú jednotku siete som si zvolila **konverzáciu**. Inšpirovala som sa prácou autorov Kudělka, Horák, Zehnalová [10]. Konverzácia je teda súbor emailov, ktorá začína jediným emailom, obsahuje najmenej 2 emaily a dvoch rôznych odosielateľov. Vrcholom siete (grafu) sa teda stane užívateľ, ktorý bol ako odosielateľ aspoň v jednej takejto konverzácii. Hrana medzi užívateľmi je zostrojená medzi užívateľmi, ktorí boli spolu v jednej konverzácii ako odosielatelia. Pre konverzáciu ešte ukladám čas jej začiatku, užívateľ si následne v aplikácii môže zvoliť časový rozsah konverzácií.

8.6.2 Triedy pre graf, vrcholy a hrany

Pre uloženie siete v pamäti slúži generická trieda `Graph<T>`. Vrcholy a hrany drží ako `Dictionary<int, HashSet<Node<T>>>`, čiže ako mapu vrcholov s ich susednými vrcholmi. Pre uloženie vrcholov a hrán grafu slúžia zoznamy hrán a vrcholov uložené ako `HashSet<Node<T> >` a `HashSet<Edge<T>>`. Trieda je generická preto, aby bolo možné vytvoriť graf pre rôzne entity. Triedy reprezentujúce vrcholy a hrany sú tiež generické a predstavujú ich `Node<T>` a `Edge<T>`. Pre identifikáciu komunity, bola vztvorená trieda `Comunity<T>`.

9 Experimenty

Pri práci som využila svoju emailovú sadu a tiež emailové sady, ktoré mi poskytli moji kolegovia. Takže som zozbierala emaily od štyroch ľudí, o ktorých viem ich postavenie v tíme alebo aj dátum nástupu do práce. Rozhodla som sa tieto emaily agregovať a analyzovať emaily v rámci celého tímu a potom analyzujem vybraného jednotlivca.

9.1 Analýza tímu

Náš tím pozostáva zo štyroch ľudí, o ktorých bližšie informácie, ktoré môžu pomôcť pri analýze rolí a komunit, sú v nasledujúcej tabuľke.

Meno	Pozícia	Dátum nástupu do firmy
Andrej Parimucha	Test engineer	1.7.2011
Veronika Uhrová	Junior Developer	1.9.2016
Andrej Matejčík	Medior Developer	1.6.2017
Tibor Palatka	Lead Developer	1.11.2010

Tabuľka 1: Informácie o členoch tímu

9.1.1 Základné informácie

	počet
Emaily	21743
Konverzácie	11545
Poživatelia (emailové adresy)	370

Tabuľka 2: Základné informácie o datasete

	počet
Používatelia, ktorí sú aspoň v jednej konverzácii	273
Emaily poslané z pracovného emailu	20367
Konverzácie iniciované niekým z kolegov	10307

Tabuľka 3: Ďalšie informácie o datasete

9.1.2 Zmeny komunit v čase

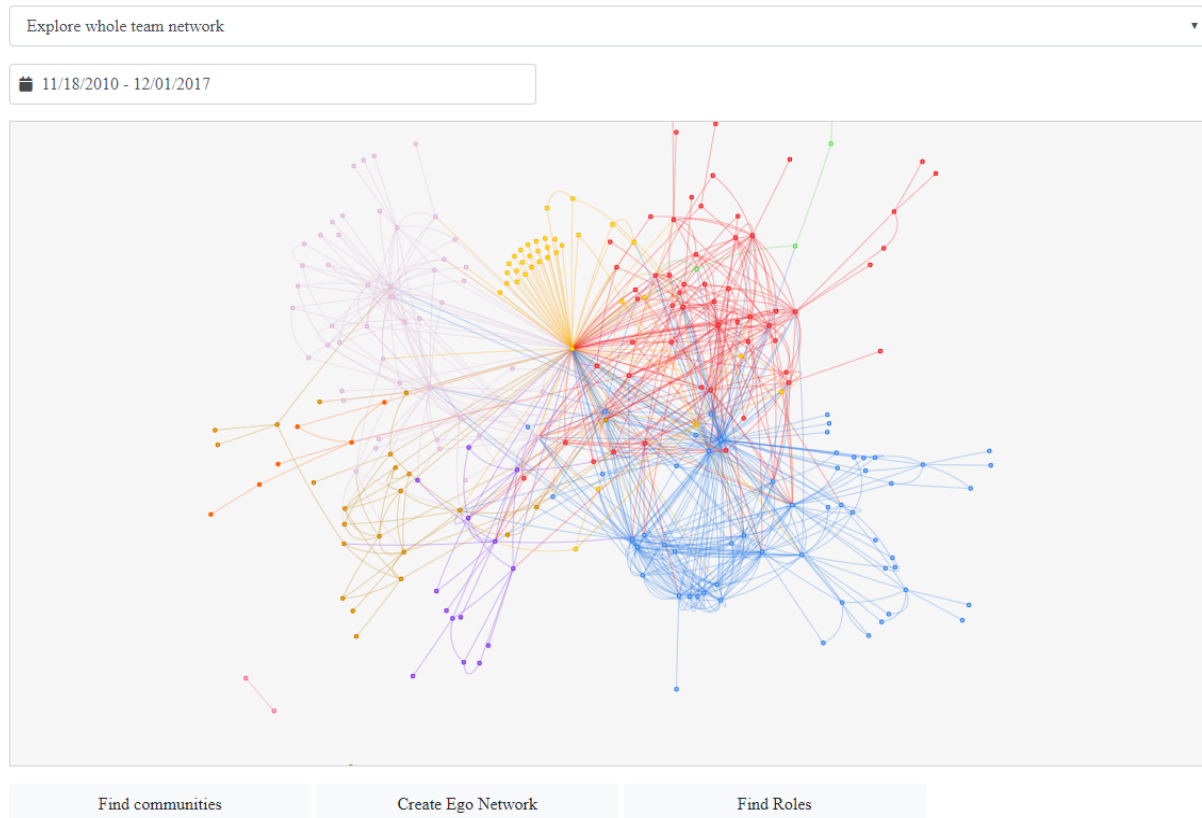
Celkový časový rozsah všetkých konverzácií je od 18. Novembra 2010 do 1. Decembra 2017. Za túto dobu sa v rámci tímu vytvorilo celkovo 13 komunit.

Časový rozsah, ktorý som volila neskôr bol po jednom roku a sledovala som koľko komunit sa v tomto roku nachádzalo a či nejaká komunita pribudla. V nasledujúcej tabuľke je zobrazený časový interval a počet komunit nachádzajúcich sa v sieti za tento časový interval.

Časový interval	Počet komunit
18.11.2010 - 31.12.2010	1
01.01.2011 - 31.12.2011	4
01.01.2012 - 31.12.2012	5
01.01.2013 - 31.12.2013	4
01.01.2014 - 31.12.2014	9
01.01.2015 - 31.12.2015	7
01.01.2016 - 31.12.2016	13
01.01.2017 - 01.12.2017	9

Tabuľka 4: Počet komunit v ročných intervaloch

Táto tabuľka ukazuje nárast komunit v rámci každého roka. Najväčší rast komunit je medzi rokom 2015 a 2016, čo odzrkadľuje reálnu situáciu, keďže firmu na prelome týchto rokov kúpil korporát, takže sa najviac naberali ľudia, čiže nárast emailov a komunit je prirodzený. V ďalšom roku sa situácia konsolidovala takže nastal menší pokles komunit.



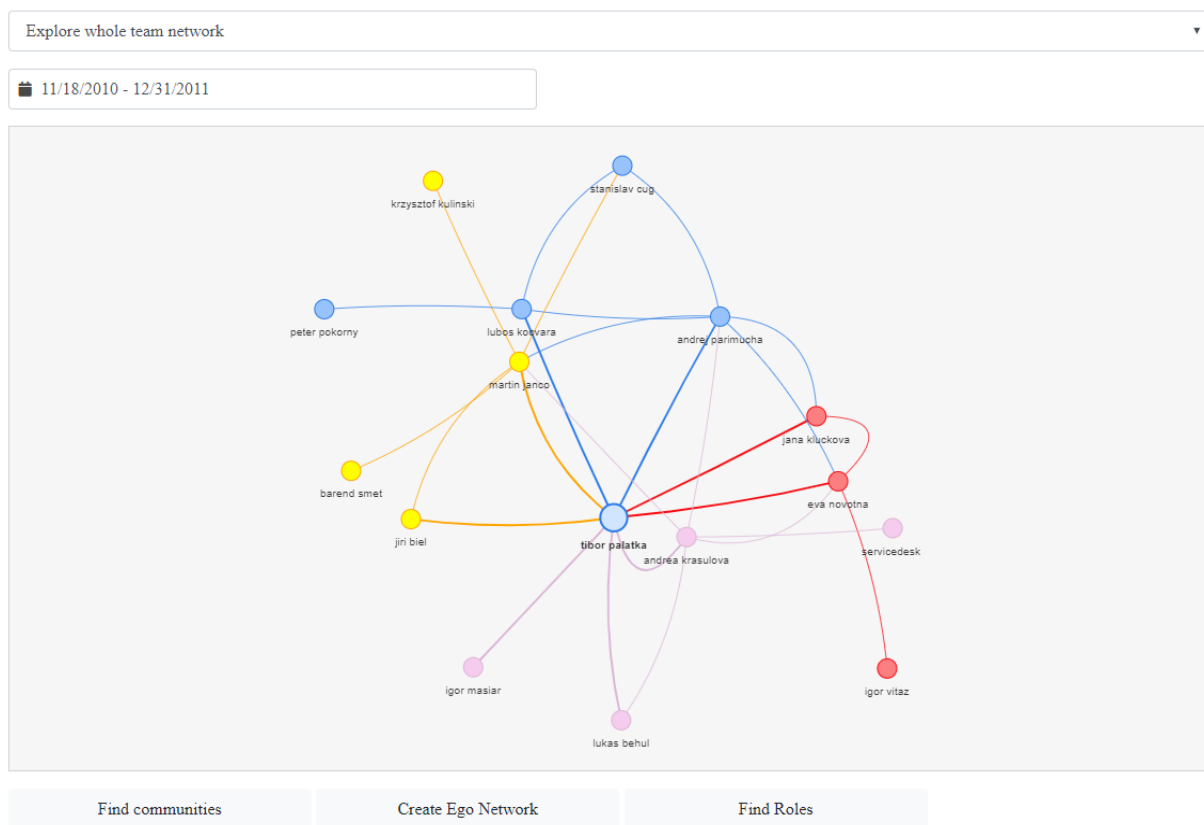
Obr. 16: Vizualizácia grafu emailov s vygenerovanými komunitami celého tímu za celkový čas

9.1.3 Zmeny rôl v čase

Tak ako sa menia v čase komunity, tak sa určite menia aj roly, ktoré sa vyskytujú v tíme a tiež sa menia ľudia, ktoré tieto roly zastupujú. Pre analýzu rôl som použila ročné intervaly, ktoré pridávam k počiatku všetkých konverzácií.

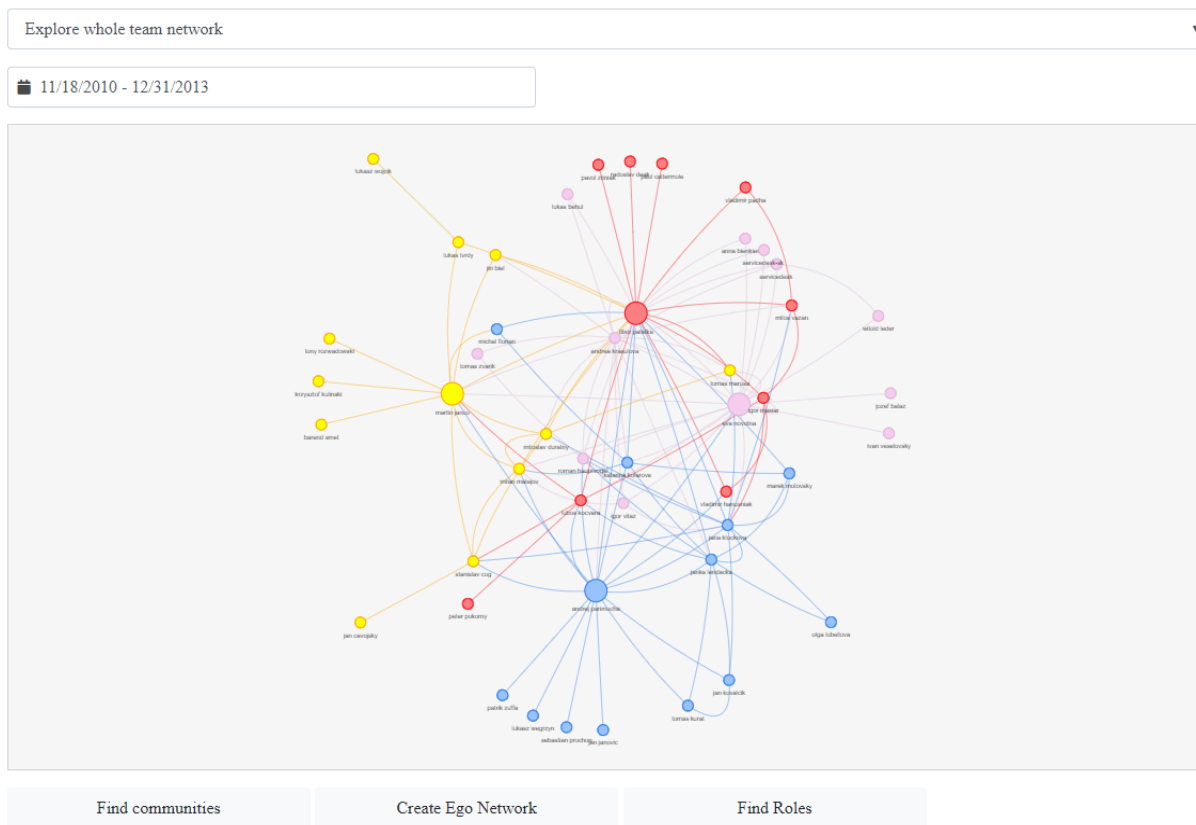
Časový interval	Leader	Mediator	Outsider	Outermost	Počet komunit
18.11.2010 - 31.12.2010	0	1	0	0	1
18.11.2010 - 31.12.2011	0	1	0	0	4
18.11.2010 - 31.12.2012	3	1	1	1	5
18.11.2010 - 31.12.2013	4	0	0	0	4
18.11.2010 - 31.12.2014	4	3	0	0	8
18.11.2010 - 31.12.2015	4	3	0	0	8
18.11.2010 - 31.12.2016	7	4	0	0	13
18.11.2010 - 01.12.2017	8	3	0	2	13

Tabuľka 5: Počet komunit v ročných intervaloch



Obr. 17: Vizualizácia grafu v intervale 18.11.2010 - 31.12.2011

Na obrázku 17 vidieť pekný príklad roly *Mediator* (uzol *Tibor Palatka*). V tomto prípade uzol, ktorý je označený ako *Mediator*, spája všetky komunity grafu.



Obr. 18: Vizualizácia grafu v intervale 18.11.2010 - 31.12.2013, v každej komunite detekovaný jeden leader.

9.2 Analýza jednotlivca

Ako jednotlivca, ktorého budem analyzovať som si vybrala *lead developera*, pretože je vo firme najdlhšie a tým pádom sa budú dobre rozoznávať či už zmeny v komunitách alebo rolách.

9.2.1 Základné informácie

	počet
Emaily	10740
Konverzácie	4713
Poživatelia (emailové adresy)	309

Tabuľka 6: Základné informácie o datasete

	počet
Používatelia, ktorí sú aspoň v jednej konverzácii	229
Emaily poslané z pracovného emailu	9850
Konverzácie iniciované niekým z kolegov	4016

Tabuľka 7: Ďalšie informácie o datasete

9.2.2 Zmeny komunit v čase

9.2.3 Zmeny rôl v čase

- porovnanie toho, či sa rola zmenila počas každého roku stráveného vo firme

10 Záver

10.1 Možnosti rozšírenia a zdokonalenia práce

Literatura

- [1] Guanting Tang, Jian Pei, and Wo-Shun Luk
Email Mining: Tasks, Common Techniques, and Tools
- [2] Xiaoyan Fu
Visualization and Analysis of Email Networks
- [3] J. Diesner, T. L. Frantz, and K. M. Carley
Communication networks from the enron email corpus it's always about the people. Enron is no different
- [4] A. Chapanond, M. S. Krishnamoorthy, and B. Yener
Graph Theoretic and Spectral Analysis of Enron Email Data
- [5] TeamNETData
<http://inflex.cz:8075/TeamNETdata/>
- [6] N. Crossley, E. Bellotti, G. Edwards, M. G. Everett, J. Koskinen, M. Tranmer
Social network analysis for Ego-Nets
- [7] Petr Kovář
Úvod do Teorie grafů - skripta VŠB
- [8] M.E.J. Newman
The structure and function of complex networks
- [9] Afra Abnar, Mansoureh Takaffoli, Reihaneh Rabbany, Osmar R. Zaiane
SSRM: Structural Social Role Mining for Dynamic Social Networks
- [10] Zehnalová, Horák, Kudělka
Email Conversation Network Analysis: Work Groups and Teams in Organizations
- [11] Repository pattern
<https://www.rarous.net/weblog/271-active-record-vs-repository-pattern.aspx>
- [12] Vincent D. Blondel, Jean-Loup Guillaume¹, Renaud Lambiotte and Etienne Lefebvre
Fast unfolding of communities in large networks
- [13] Do Millennial and Gen Z Consumers Still Use Email?
<https://www.bluecore.com/blog/do-millennials-use-email/>
- [14] vis.js
<http://visjs.org/>