Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Bakalářská práce

Forenzní analýza organizované skupiny lidí pomocí dat z elektronických komunikačních protokolů

Ondřej Nový

Vedoucí práce: doc. RNDr. Ing. Marcel Jiřina, Ph.D.

11. května 2015

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 11. května 2015 .....

Odkaz na tuto práci

Nový, Ondřej. *Forenzní analýza organizované skupiny lidí pomocí dat z elektronických komunikačních protokolů.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2015.

Abstrakt

Tato práce se pokouší vyšetřovatelům zjednodušit analýzu elektronických dat, konkrétně elektronické komunikace. Zabývá se úlohami jak najít mezi účastníky komunikace ty nejvýznamnější a jak mezi nimi najít úzce propojené skupiny. Úlohy jsou řešeny různými algoritmy pro měření centrality a detekci komunit. Algoritmy byly implementovány v programovacím jazyce Java a testovány na datech reálně použitých při vyšetřování společnosti Enron. Jak ukázalo testování, použité algoritmy jsou relativně úspěšné, nebyly ale nalezeny žádné, které by bylo možné označit jako nejlepší.

Klíčová slova forenzní analýza, analýza elektronické komunikace, analýza sociálních sítí, centralita, detekce komunit

Abstract

This work tries to make forensic analysis of electronic data, particulary electronic communication, easier. Deals with problem of finding among communication participants the most important ones and with problem of finding tightly connected groups of them. The problems are solved using various algorithms for centrality measurement and community detection. Algorithms

have been implemented in Java and tested on data used in real investigation of Enron corporation. Testing has shown that the algorithms are relatively successful, but none of them is significantly better than the others.

Keywords forensic analysis, electronic communication analysis, social network analysis, centrality, community detection

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Elektronické komunikační protokoly jsou významným zdrojem „stop“, které za sebou zanechávají jejich uživatelé. Vyšetřovatelé mohou využívat těchto protokolů, aby získávaly cenné informace o vyšetřovaných osobách, popřípadě celých skupinách páchajících organizovaný zločin. Data mohou extrahovat přímo ze zabavených zařízení, vyžádat si je od provozovatelů například sociálních sítí či získávat z internetu na vlastní pěst [1]. Jelikož ale objemy elektronické komunikace rostou, je stále těžší analyzovat všechna tato data manuálně, a proto mohou být užitečné automatizované metody.

Alzaidy v [2] mluví o nedostatku forenzních nástrojů, které by prováděly analýzu elektronických dat od extrakce samotných dat až po vizualizaci nalezených informací. Navrhuje systém, který sbírá data ze souborového systému, hledá v nich komunity a vizualizuje zjištěné závěry. V navazující práci [3] zmiňuje, že systém se setkal s pozitivními ohlasy od vyšetřovacího týmu v Kanadě.

V této práci bude snaha nalézt metody, kterými lze automatizovaně rozpoznávat strukturu organizované skupiny na základě elektronické komunikace provedené v rámci skupiny. Tyto metody mohou být základem pro systém podobný tomu Alzaidyho, který zvládne sám hromadit data a vizualizovat výsledky, specializovaný na elektronickou komunikaci, kde lze očekávat větší množství relevantních informací o vztazích v rámci skupiny, než v obecných souborových systémech.

Kapitola 1

Cíl práce

Cílem práce je nalézt a implementovat metody, které jsou schopné na základě zaznamenané elektronické komunikace poodhalit strukturu skupiny lidí v rámci níž byla provedena. Nejdříve je potřeba najít a prozkoumat pro tuto úlohu již existující nástroje. Výsledný program v programovacím jazyce Java, ve kterém budou metody implementovány, bude na vstupu vyžadovat elektronickou komunikaci zaslanou mezi členy organizované skupiny. Výstupem programu pak bude soubor s grafem nesoucím informaci o struktuře skupiny, jenž lze snadno vizualizovat.

Kapitola 2

Značení

(a, b)	uspořádaná dvojice
$(a_1, \ldots, a_n)$	uspořádaná n-tice
$[a, b]$	neuspořádaná dvojice
$\|S\|$	mohutnost množiny S
U	množina všech uzlů v grafu
H	množina všech hran v grafu
k_u	stupeň uzlu u
w_{uv}	váha hrany mezi uzly u a v
$d(u, v)$	vzdálenost z uzlu u do uzlu v
N_u	množina všech sousedů uzlu u
$\mathbf{B}$	matice
B_{ij}	element matice $\mathbf{B}$
$\mathbf{A}$	matice sousednosti

Kapitola 3

Vymezení pojmů

3.1 Struktura organizované skupiny

Pojem „struktura organizované skupiny“ lze chápat různě. Například by jím mohly být myšleny vzájemné emočními vztahy osob ve skupině. Ty je možné zkoumat pomocí metod analýzy sentimentu, popsaných například v [4]. Tato informace ovšem nebyla pro vyšetřovatele, kterým je snaha odhalením struktury ulehčit práci při dalším jejím zkoumání, shledána jako dostatečně užitečnou, a proto se jí tato práce zabývat nebude.

Užitečnějším bylo shledáno, jak je osoba v rámci skupiny důležitá. Důležití členové skupiny obvykle mají vliv na méně významné členy a nesou tak v podstatě odpovědnost za jejich jednání. Mají také typicky nejvíce informací o skupině. Sociální statut může být významný i při utváření rozsudku v soudním procesu se členy zločinecké skupiny (minimálně v České republice):

Odnětím svobody na pět až patnáct let nebo propadnutím majetku bude pachatel potrestán, je-li vedoucím činitelem nebo představitelem organizované zločinecké skupiny určené nebo zaměřené k páchání vlastizrady (§ 309), teroristického útoku (§ 311) nebo teroru (§ 312). [5]

Problémem při zkoumání elektronické komunikace může být také to, že není známo, kdo do zločinecké skupiny vůbec patří. Navíc i v rámci nějaké skupiny může existovat několik víceméně oddělených skupin, kde každá je zodpovědná za jiné zločiny. Z toho důvodu se práce bude zabývat také dělením skupiny na menší části, pokud je to vhodné.

V této práci tedy bude strukturou organizované skupiny nadále myšleno rozdělení členů do podskupin a jejich seřazení podle významnosti, jelikož právě tyto informace jsou zde považovány za potenciálně nejužitečnější.

3.2 Data z elektronických komunikačních protokolů

Tímto pojmem (zkráceně také „elektronická komunikace“) je v této práci myšlena veškerá mezilidská komunikace, která byla provedena prostřednictvím elektronických zařízení. Typicky se jedná o SMS zprávy, telefonní hovory, komunikaci na sociálních sítích, diskuzní fóra atd.

3.3 Komunikující skupina

Pro účely této práce bude zaveden následující formální model komunikující skupiny, který představuje členy organizované skupiny a jednotlivé zprávy poslané mezi nimi, přičemž zpráva může mít i více příjemců, ale má právě jednoho autora. Více autorů nebude dovoleno, jelikož ani autoři elektronických databází, ve kterých bývá uložena elektronická komunikace, s takovým případem typicky nepočítají.

Definice 1 *Zpráva* $z = (a, P, t)$*, kde* $a \in J$*,* $P \subset J$*,* $a \in P$*,* t *je konečná posloupnost znaků.*

Ve zprávě z představuje jedinec a autora a P neprázdnou množinu příjemců, tedy jednoho či více příjemců. Autor nemůže poslat zprávu sám sobě. t představuje text zprávy.

Pokud má zpráva více příjemců, může být nazvána hromadnou zprávou.

Definice 2 *Konverzace mezi jedinci* i, j *označená jako* $k_{ij} = \{z = (a, P, t) \in Z : i = a \wedge j \in P \vee j = a \wedge i \in P\}$.

Definice 3 *Komunikující skupina* $K = (J, Z)$*, kde* J *je množina jedinců a* Z *je množina zpráv. Navíc platí, že pro* $\forall j_0, j_k \in J$ *existuje posloupnost* n *neprázdných konverzací* $k_{j_0 j_1}, k_{j_1 j_2} ... k_{j_{n-1} j_n}$.

Druhá složitější část definice komunikující skupiny jen říká, že všichni jedinci musejí být s ostatními nějak navzájem propojeni komunikací. Za komunikující skupinu se proto nepokládájí například dvě množiny jedinců, mezi nimiž neexistuje žádná zpráva, ale jedná se vlastně o dvě komunikující skupiny.

Na obrázku 3.1 je jednoduchý příklad zpráv poslaných mezi jedinci A, B, C. Například první dvě zprávy tvoří konverzaci mezi jedinci A a B.

AUTOR	PŘÍJEMCI	OBSAH ZPRÁVY
A	B	CO DĚLÁŠ B
B	A	KRÁTÍM DANĚ JAKO VŽDY
C	A, B	CHLAPY, ODPOSLOUCHÁVAJÍ NÁS

Obrázek 3.1: Jednoduchý příklad množiny zpráv v komunikující skupině

Kapitola 4

Popis problému

V této kapitole budou formálně popsány problémy, které se v této práci budou řešit. Jedná se o problém hledání hierarchie, kde je cílem seřadit jedince podle významnosti či sociálního statutu a o problém rozdělení skupiny na podskupiny.

4.1 Problém hledání hierarchie

Mějme komunikující skupinu $K = (J, Z)$ a relaci dominance $D : J \times J$, což je binární asymetrická, nereflexivní, transitivní relace. Cílem je najít pouze na základě K funkci $h : J \to R$, takovou, aby $H = \frac{|\{(a,b) \in D : h(a) > h(b)\}|}{|D|}$, bylo co největší (ideálně 1).

Dvojice $(a, b) \in D$, znamená, že jedinec a je v hierarchii výše než jedinec b. D reprezentuje „skutečnou“ hierarchii ve skupině a pomocí ní je vyhodnocována úspěšnost funkce h, která je zjištěným odhadem hierarchie. $h(a) > h(b)$ vlastně říká, že jedinec a je v hierarchii výše než jedinec b. Úspěšnost funkce h v odhadu reálné hierarchie je označena jako H a nabývá hodnot z intervalu $\langle 0, 1\rangle$, tudíž může být vyjádřena i v procentech.

Na relaci dominance se bude v práci dále odkazovat pojmem reálná nebo skutečná hierarchie. Na funkci h se bude odkazovat pojmem zjištěná hierarchie a na H pojmem úspěšnost hledání hierarchie.

4.2 Problém rozdělení na podskupiny

Mějme komunikující skupinu $K = (J, Z)$, přirozené číslo n a funkci $c_{real} : J \to \langle 1; n\rangle$. Cílem je určit pouze na základě K přirozené číslo m funkci $c_{aprx} : J \to \langle 1, m\rangle$ tak, aby

$$C = \frac{C_p}{2} + \frac{C_n}{2}$$

, kde

$$C_p = \frac{|\{\forall a, b \in J : a \neq b \wedge c_{real}(a) = c_{real}(b) \wedge c_{aprx}(a) = c_{aprx}(b)\}|}{|\{\forall a, b \in J : a \neq b \wedge c_{real}(a) = c_{real}(b)\}|}$$

a

$$C_n = \frac{|\{\forall a, b \in J : a \neq b \wedge c_{real}(a) \neq c_{real}(b) \wedge c_{aprx}(a) \neq c_{aprx}(b)\}|}{|\{\forall a, b \in J : a \neq b \wedge c_{real}(a) \neq c_{real}(b)\}|}$$

bylo co největší (ideálně 1).

Funkce c_{real} zde představuje opravdové rozřazení jedinců do n podskupin, které má správně odhadnout i funkce c_{aprx} společně s počtem podskupin m. V ideálním případě $c_{aprx} = c_{real}$ a $m = n$.

Úspěšnost c_{aprx} a m v odhadu reálné hierarchie je označena jako C a nabývá hodnot z intervalu $\langle 0, 1\rangle$, tudíž může být vyjádřena i v procentech. Tato úspěšnost závisí z poloviny na C_p, tedy na poměru správně zařazených do stejné skupiny ku těm, kteří měli být zařazeni do stejné skupiny, a z poloviny na C_n, tedy na poměru správně zařazených do jiné skupiny ku těm, kteří měli být zařazeni do jiné skupiny. Toto měření úspešnosti je poměrně složité, ovšem nezbytné. V případě, že by se měřilo pouze C_p, pak by konstatní c_{aprx}, tedy přiřazení jedinců pouze do jedné podskupiny dosáhlo nejvyšší možné úspěšnosti. Analogický případ nastane pro C_n, pokud by každý jedinec patřil do své vlastní podskupiny.

Na c_{real} se bude v práci dále odkazovat pojmem reálné nebo skutečné rozdělení na podskupiny a na m jako na počet reálných skupin. Na c_{aprox} se bude dále odkazovat pojmem zjištěné rozdělení na podskupiny, na m jako na počet zjištěných skupin a na C pojmem úspěšnost rozdělení na podskupiny.

4.2.1 Poznámka k úspěšnosti

Důležité je zmínit, že úspěšnost řešení předchozích problémů na jedné komunikující skupině neimplikuje úspěšnost i na jiných komunikujících skupinách. Naopak, pokud je snaha o co nejlepší úspěšnost na jedné komunikující skupině, kdy h/c_{aprx} je příliš složitá funkce, může dojít k takzvanému „přeučení“ modelu (kterým je zde logika generující funkce h/c_{aprx}) a úspěšnost na jiných komunikujících skupinách může být horší, než v případě jednoduššího h/c_{aprx}.

Úspěšnost řešení problému je také pevně vázána na „reálnou hierarchii“/„reálné rozdělení na podskupiny“, na základě nichž je vypočítávána úspěšnost. Co je „reálná hierarchie“ či „reálné rozdělení na podskupiny“ už ale není nijak pevně dáno. Může se například v rámci společnosti jednat o formální hierarchii, která je pevně daná a v rámci společnosti známá. Ta ovšem nemusí korespondovat s jakousi přirozenou hierarchií (jak je uvedeno i v [6]), která bude vyšetřovatele typicky zajímat.

Kapitola 5

Analýza sociálních sítí

Tato kapitola se bude zabývat takzvanou analýzou sociálních sítí, jelikož se ukázalo, že se jedná o obor, který se velmi intenzivně používá k řešení problémů podobných těm zavedeným v 4.

Podle [7] je analýza sociálních sítí, dále jen SNA (z anglického „social network analysis“), věda zabývající se studiem sociálních entit (dále jen *aktér*, v originále *actor*) a vazbami mezi nimi. Aktérem může být jedinec, ale také organizace. Vazba je velice široký pojem, který může představovat názory, transakce, asociace, komunikaci, fyzickou blízkost, formální vztah, příbuzenství atd.

Sociální sítě se typicky reprezentují jako graf. Aktéři jsou reprezentováni uzly a vazby jsou reprezentovány hranami, přičemž graf může být orientovaný či neorientovaný a mívá různé vlastnosti, podle toho, jaký model se pro zkoumání konkrétní modelované sociální sítě hodí [7].

5.1 Sociální síť

Následující definice sociální sítě není jediná možná, ale v této práci se bude uvažovat přávě tato, jenž byla převzata z [7]

Definice 4 *Sociální síť je konečná množina aktérů N a množina vztahů L. Vztah L_i, kde $i \in \langle 1, |L| \rangle$ je množina orientovaných vazeb či množina neorientovaných vazeb. Orientovaná vazba je uspořádaná dvojice aktérů (N_i, N_j), zatímco neorientovaná vazba je neuspořádaná dvojice aktérů $[N_i, N_j]$, kde $i, j \in \langle 1, |N| \rangle \wedge i \neq j$*

- Pojmem vazba se rozumí buď orientovaná vazba či neorientovaná vazba. Pokud se mluví o vazbě nějakého vztahu, pak se implicitně sděluje informace, zda je vazba orientovaná či neorientovaná.
- Definice nepovoluje vazby aktéra se sebou. Nelze tedy do sociální sítě zanést informaci například o tom, že aktér má ráda sám sebe. Tato

informace by nejspíše byla pro účely analýzy sociálních sítí irelevantní, jelikož nemá za úkol zkoumat aktéry samotné, ale pouze jejich vlastnosti dané pozicí v sociální síti.

- Sociální síť lze snadno reprezentovat grafem, kdy aktéři jsou reprezentováni uzly, orientované vazby jsou orientované hrany a neorientované vazby jsou neorientované hrany.

- Sociální síť může být výhodné obohatit váhami vazeb, které vypovídají o relativní intezitě vazeb v rámci dané sociální sítě.

V rámci SNA se obvykle měří vlastnosti jednotlivých uzlů či celé sítě. Tyto vlastnosti budou popsány ve zbytku této kapitoly s důrazem na ty, které jsou užitečné k hledání hierarchie či k rozdělování na podskupiny v sociální síti.

5.2 Komunikační síť

Komunikační sítí zde bude chápán speciální případ sociální sítě, která má pouze jeden vztah, a to sociální interakci, tedy výměnu zpráv. Tento vztah bude neorientovaný, jelikož jednostranné posílání zpráv nejspíš nebude příliš častým jevem. Vazby budou mít váhy, aby bylo možné vyjádřit intenzitu interakce mezi jedinci, zjištěnou na základě vyměněných zpráv. Tyto váhy budou pozitivní. Váha vazby nula bude ekvivalentní s tím, že vazba neexistuje.

Zde specifikovanou komunikační síť, jelikož má pouze jeden vztah, je tedy možné modelovat jako neorientovaný graf s váženými hranami a nadále se proto budou zaměňovat pojmy jako „sociální síť“ s „graf“, „aktér“ s „uzel“, „vazba“ s „hrana“ a „účastnit se“ s „incidovat“.

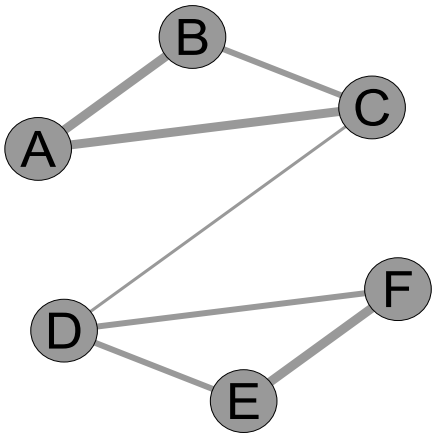
Na obrázku 5.1 je zobrazena jednoduchá komunikační síť s aktéry A, B, C, D, E, F, kde váha vazby je znázorněna tloušťkou čáry. Jak je zřejmé, síť je tvořena dvěma podskupinami (klikami), A, B, C a D, E, F, v rámci nichž aktéři hojně komunikují. Mezi podskupinami také probíhá komunikace, a to díky aktérům C a D, ovšem ne příliš velká, jak je zřejmé z váhy vazby mezi C a D.

5.3 Stupeň

Stupěň uzlu v neorientovaném grafu je počet hran, které s ním incidují. [8]

Vstupní stupeň uzlu v orientovaném grafu je počet hran, které do něj vstupují. Výstupní stupeň uzlu v orientovaném grafu je počet hran, které z něj vystupují.[9]

Stupeň je základní atribut, kterým lze odhadovat významnost aktéra v sítí, a je na něm založena centralita podle stupně.



Obrázek 5.1: Příklad komunikační sítě

5.4 Hustota

Hustota grafu je $\frac{počet\ hran}{počet\ všech\ možných\ hran}$ [10], tedy $\frac{|H|}{\frac{|U|(|U|-1)}{2}}$ pro neorientované grafy a $\frac{|H|}{|U|(|U|-1)}$ pro orientované.

5.5 Klika

Podle [10] je klika množina nejméně dvou ulzů, mezi nimiž existují všechny možné hrany. Jiná definice podle [11] říká, že klika grafu je jeho maximální úplný podgraf.

V kontextu sociálních sítí dává klika ideál toho, jak by mělo vypadat něco, co lze intuitivně nazvat skupinou či podskupinou vzhledem k síti, jelikož se zde vyskytuje vazba mezi všemi dvojicemi aktérů. Druhá definice zajišťuje slůvkem „maximální“, že nebyl vynechán žádný aktér, který by do skupiny také mohl patřit. První definice zase odmítá říkat klika samotnému aktérovi, což nezní úplně přirozeně, stejně jako nezní přirozeně říkat mu skupina.

5.6 Průměr

Průměr grafu je maximální $d(u, v)$, $\forall u, v \in U$. [11]

Tato vlastnost říká, jak daleko to od sebe mají v síti dva nejvzdálenější aktéři, což může například vypovídat o tom, jak dlouho trvá šíření informace.

5.7 Excentricita

Excentricita uzlu u je $\max d(u, v) \forall v \ \in U$. [11]

Je to tedy vzdálenost uzlu k od něj nejvzdálenějšímu uzlu.

5.8 Shlukovací koeficient

Definice 5 *Shlukovací koeficient uzlu v, kde C_v je počet hran mezi všemi sousedy uzlu v je dán vzorcem $\frac{2C_v}{k_v(k_v-1)}$, tedy počet hran mezi všemi sousedy uzlu v normalizováno počtem všech možných hran mezi nimi. [12]*

Shlukovací koeficient udává, jak moc jsou nazvájem propojeni sousedi sledovaného aktéra. Pokud je roven 1, uzel se sousedy tvoří společně kliku. Pokud je roven 0, sousedi aktéra nejsou propojeni vůbec.

Shlukovací koeficient pro graf je průměrem shlukovacích koeficientů všech jeho uzlů. [10] Ukazuje se, že přirozené sítě, jako jsou například sítě komunikace, mývají poměrně vysoký shlukovací koeficient, narozdíl od náhodně generovaných sítí. [12]

5.9 Vážený shlukovací koeficient

Přehled různých způsobů, jak měřit shlukovací koeficient na vážených grafech, poskytuje [13]. Způsobů je relativně mnoho a výrazně se liší v tom, co vlastně měří. Vyskytují se zde metody, které uvažují jen váhy vazeb mezi sousedy a naopak takové, které uvažují jen váhy vazeb zkoumaného uzlu se sousedy. Vybrána byla metoda, která uvažuje váhy všech vazeb a je poměrně jednoduchá. Jedinou její nevýhodou je, že váhy musí být z intervalu $\langle 0, 1 \rangle$, což ale lze jednoduše zařídit vydělením vah jejich maximem.

Definice 6 *Vážený shlukovací koeficient podle [14] uzlu i se definuje jako $\frac{2\sum_{j,k \in N}(w_{i,j}w_{j,k}w_{i,k})^{1/3}}{k_i(k_i-1)}$, kde k_i je stupeň uzlu i, N_i je množina všech sousedů uzlu i a $w_{a,b}$ je váha hrany z uzlu a do b, a 0 pokud mezi nimi hrana neexistuje.*

5.10 Centralita

Centralita měří něco, co se dá obvykle nazvat relativní významností aktéra v síti [15]. Podle [7] je primární využití SNA právě pro vyhledávání významých aktérů, tedy zjišťování centrality aktéra.

Centralitu je možné měřit na neorientovaných grafech. V [7] se analogická veličina pro orientované grafy označuje jako prestiž. Prestiž lze obvykle měřit analogicky jako všechny dále uvedené druhy centrality, a proto zde nebude popsána. Navíc i proto, že komunikační síť je vlastně neorientovaný graf. Obě vlastnosti jsou obvykle normalizovány, tak, aby nabývaly hodnot v intervalu $\langle 0, 1\rangle$ a bylo tak snadné okamžitě zhodnotit velikost centrality relativně k ostatním aktérům.

Název „prestiž“ vychází z jedné z možných interpretací této vlastnosti a může být tedy, jak uvádí autoři [7], zavádějící. Pokud by se uvažovaly vazby, které znamenají like, štouchnutí či tweet, což jsou fenomény známe ze sociálních sítí (ve smyslu sociálních aplikací jako je Facebook), pak by tato vlastnost měřila skutečnou prestiž. Pokud by ale vazba byla negativní, například „cílový jedinec je nejvíce lhostejný tomu zdrojovému“, pak by vlastnost „prestiž“ měřila úplný opak reálné prestiže. Z toho lze usoudit, že význam centrality a prestiže závisí na vazbách, na kterých se měří. Přesný význam závisí také na konkrétním druhu centrality či prestiže. Podle [16] obvykle měří reputaci, vliv či prostřednictví.

Seznam zde uvedených druhů centralit není kompletní, už jen proto, že stále přibývají (nejnovější nalezená perkolační centralita je z roku 2013 [17]).

5.10.1 Degree centralita

Nejjednoduší způsob měření centrality je na základě stupně uzlu. V kontextu komunikační sítě odpovídá taková centralita aktéra celkovému počtu aktérů, se kterými komunikoval. Tato metoda bohužel nezohledňuje, s kým aktér komunikuje ani nebere v potaz jeho případnou strategickou pozici v síti. Hodí se zejména pro menší sociální sítě. U větších lze počítat s tím, že aktér „velitel“ začne svou moc delegovat, jelikož nemůže zvládat rozkazovat všem přímo, v takovém případě může být tato centralita nevhodná.

Definice 7 *Degree centralita podle [18] v neorientovaném grafu uzlu $u \in U$, označena jako $C_D(u)$, se definuje jako*

$$\frac{k_u}{|U| - 1}$$

.

Centralita, zde přímo úměrně odpovídá stupni uzlu, který je navíc normalizován maximálním stupněm uzlu $|U| - 1$.

5.10.2 Closeness centralita

Další koncept centrality je založen na vzdálenosti uzlů, tedy minimálním počtu hran mezi nimi.

Closeness centralita měří, jak blízký je uzel ke všem ostatním po nejkratších cestách. V komunikační síti je tento koncept užitečný, jelikož aktéři mezi sebou nekomunikují jen přímo, ale také zprostředkovaně nebo na sebe přenáší své názory, a tím nepřímo šíří svůj vliv.

Definice 8 *Closeness centralita podle [18] v neorientovaném souvislém grafu uzlu u, označena jako $C_C(u)$ se definuje jako*

$$\frac{1}{\sum_{v \in U} d(u,v)}$$

.

Centralita zde nepřímo úměrně odpovídá součtu vzdálenosti uzlu k ostatním uzlům.

5.10.3 Betweenness centralita

Betweenness centralita uzlu se zvyšuje tím, že uzel leží v nejkratších cestách mezi jinými uzly. V komunikační síti se tedy tato centralita zvyšuje tím, jak je aktér potřebný pro nepřímou komunikaci různých dvou vzdálených aktérů v síti. Tento druh centrality zohledňuje případnou strategickou pozici aktéra. Pokud má například aktér jen málo vazeb v rámci své skupiny, ale má vazbu s někým z jiné skupiny, pak je významný tím, že přináší do skupiny nové a tím i cenné informace/názory. Tento fenomén se podle stejnojmenného článku [19] nazývá „síla slabých vazeb“.

Tato centralita ovšem neuvažuje případy, kdy nepřímá komunikace probíhá jinudy, než přes nejkratší cestu, například přes „skoro nejkratší cestu“, protože některý z aktérů v nejkratší cestě komunikaci blokuje.

Definice 9 *Betweenness centralita podle [18] v neorientovaném grafu uzlu u, označena jako $C_B(u)$ se definuje jako*

$$\sum_{\forall j,k \in \langle 1,|U| \rangle : j<k} \frac{p_{n_j,n_k}(u)}{p_{n_j,n_k}}$$

, kde U je množina všech uzlů kde $p_{n_j,n_k}(u)$ je nejkratší cesta z uzlu n_j do uzlu n_k, která zároveň vede přes uzel u, a kde p_{n_j,n_k} je počet nejkratších cest z n_j do n_k.

5.10.4 Eigenvector centralita

Eigenvector centralita podle [20] se podobá degree centralitě, ovšem zavádí myšlenku, která je v oblasti komunikační sítě velice intuitivní a samozřejmá, a sice, že nezáleží jen na počtu vazeb, ale také na tom, jakou centralitu (významnost) mají protějšky vazeb.

Definice 10 *Eigenvector centralita v neorientovaném grafu uzlu u, označená jako $C_E(u)$, se definuje jako $\frac{1}{\lambda}\sum_{v\in U} A_{uv}C_E(v)$, kde A_{uv} jsou prvky matice sousednosti a λ je normalizační konstanta, tak aby $\forall v \ \in U \ \max C_E(v) = 1$.*

Aby pro eigenvector centralitu byly splněny podmínky dané definicí, počítá se tak, že se C_E každého uzlu nastaví na 1. Poté se iterativně vždy pro všechny uzly vypočítá nová hodnota C_E, tím, že se sečtou C_E jeho sousedů a na konci iterace se C_E všech uzlů vydělí maximální hodnotou C_E. Maximální hodnota C_E na konci iterace je tudíž 1. Algoritmus iteruje až do doby, kdy se žádná hodnota C_E nezmění.

5.11 Centralita pro sítě s váženými vazbami

Jelikož pro komunikační síť bylo vhodné zavést vážené vazby, aby byla nějakým způsobem rozlišena intenzita různých vazeb, je nutné zavést centralitu, která váhu vazeb bere v úvahu, narozdíl od přechozích zmíněných.

Newman v [21] ukazuje, že mnohé vlastnosti, které se meří pro neválené grafy lze jednoduše měřit i pro válené tak, že se válený graf převede na multigraf. Pokud mezi uzly u a v existuje hrana o váze n, pak se mezi nimi vytvoří n neválených hran. Po tomto převodu lze měřit degree a eigenvector centralitu bez změny jejich definice.

5.11.1 Válená degree centralita

Barrat a spol v [22] zavádí pojem síla uzlu, který by se dal nazvat i degree centralitou pro válené grafy.

Definice 11 *Síla uzlu u, označena jako $C_D^w(u)$, se definuje jako $\sum_{v\in N_u} w_{uv}$, kde N_u je množina uzlů sousedících s u.*

Jedná se tedy o jednoduchý součet vah všech hran incidujících s uzlem. Autoři [23] mají k této centralitě výhrady. Říkají, že stejnou centralitu pak má například uzel se čtyřmi vazbami o váze 1 a uzel s jednou vazbou o váze 4, což může být špatně u sítí, kde je obzvláště výhodné mít mnoho slabých vazeb, nebo naopak mít málo silných. Zavádí tak vlastní degree centralitu s parametrem α

Definice 12 *Degree centralita v neorientovaném grafu s váženými hranami uzlu u s parametrem $\alpha : \alpha \geq 0 \wedge \alpha \in R$, označena jako $C_D^\alpha(u)$, se definuje jako $k_u(\frac{C_D^w(u)}{k_u})^\alpha$, kde k_u je stupeň uzlu u.*

Pokud parametr $\alpha = 0$, pak C_D^α měří klasickou degree centralitu C_D bez ohledu na váhy. Pokud $\alpha = 1$, pak $C_D^\alpha(u) = C_D^w(u)$. Nastavení $\alpha < 1$ je vhodné pokud má centralitu zvyšovat především počet vazeb, tedy je výhodné mít mnoho slabých vazeb. Nastavení $\alpha > 1$ je vhodné pokud má být výhodné mít především silné vazby.

5.11.2 Vážená betweenness a closeness centralita

Autoři [23] podotýkají, že zobecnění betweenness a closeness cetrality pro vážené grafy je poměrně jednoduché. Stačí místo jednoduché nejkratší cesty, tak jak byla uvedena v definici v 5.10.3 a 5.10.2, uvažovat váženou nejkratší cestu s novými váhami $x_{ij} = 1/w_{ij}$, kde w_{ij}, jsou původní váhy, které byly brány jako intenzita vazby a ne jako cena cesty a je proto potřeba uvažovat jejich inverzní hodnotu. Tato upravená definice closeness a betweenness cetrality v komunikační síti odráží skutečnost, že aktér může snadněji komunikovat delší cestou, pokud vede přes silné vazby. Tyto centrality budou označeny jako C_B^w a C_C^w.

Autoři [23] mají ale k výše zmíněnému zobecnění betweenness a closeness cetrality výhrady stejně jako u degree centrality. Říkají, že vážená nejkratší cesta nebere v potaz, přes kolik uzlů cesta vede, ale pouze kolik je celkový součet vah na hranách. To nemusí vadit například u modelování síťových komponent, kde router vytváří zanedbatelnou latenci, ale typicky to vadí právě v komunikační síti, kde každý další aktér, přes kterého zprostředkovaná komunikace vede vytváří zpoždění a disponuje mocí, že se informace dál vůbec nedostane. Z toho důvodu upravují váhy hran pro algoritmus nejkratší cesty po svém: $x_{ij} = 1/(w_{ij})^\alpha$, kde $\alpha : \alpha \geq 0 \wedge \alpha \in R$, je parametr stejný jako v 5.11.1. $\alpha < 1$ je tedy opět nastavení vhodné pro komunikační síť, jelikož uvažuje celkový počet hran a nejen jejich váhy. Tyto centrality budou označeny jako C_B^α a C_C^α.

5.12 Detekce komunit

Podle Newmana v [24] se hledáním skupin vzájemně hustě propejných uzlů s pouze řídkým propojením skupin zabývá na poli sociálních sítí věda pojmenovaná mimo jiné jako „detekce komunit“. Pojem podskupina je zde zaměňován s pojmem komunita. Podle něj úspěšnost algoritmu v rozdělování sítě na komunity nezávisí jen na algoritmu, ale na síti samotné, konkrétně na tom, jestli se z dobře separovatelných skupin skládá.

5.12.1 Detekce klik

V [10] je zmíněn jeden z nejjednoduších algoritmů založený na vyhledávání klik, tedy maximálních úplných podgrafů. Klika je ovšem celkem silné kritérium, které množina uzlů, která by byla intuitivně nazvána podskupinou, typicky nebude splňovat. Nalezené kliky se navíc mohou překrývat, s čímž je potřeba se nějakým způsobem vypořádávat.

5.12.2 Minimální hranový řez

Jeden z velmi jednoduchých způsobů dělení sítě na podskupiny uvádí Newman v [24] jako algortimus hledání minimálního hranového řezu, který lze v nezměněné podobě převzít z teorie grafů. Jeho nevýhodou je, že preferuje rozdělování například na skupinu jednoho uzlu a skupinu zbývajících uzlů. To lze řešit tím, že se vytvoří omezení, kolik každá skupina musí mít minimálně uzlů. V takovém případě je ale algoritmus nucen graf rozdělit, přestože žádné dělení nemusí být ve skutečnosti vhodné.

5.12.3 Metoda k-jader

Zmíněnými nevýhodami netrpí metoda k-jader. Podle [25] vyhledává takzvaná k-jádra, což jsou maximální podgrafy indukované uzly s minimálním stupněm větším nebo rovným k. 1-jádro je tedy původní graf. Se zvyšujícím se k zůstává v grafu stále méně uzlů a potenciálně se graf může rozdělit na více komponent, které pak představují jednotlivé podskupiny.

5.12.4 Girvan-Newman

Předchozí metody se snažili v grafu hledat útvary, podle předem definovaných kritérií. Girvan-Newmanův algoritmus zveřejněný v [26] přichází, jak uvádí [10], s odlišným přístupem. Jedná se o iterativní algoritmus, který v každém kroku spočítá betweenness všech hran, což je analogická veličina jako betweenness centralita pro uzel, a vymaže hrany s maximální hodnotou. Algoritmus tímto může kdykoliv odstraněním hrany graf rozdělit a zvyšovat tak počet komponent, kde komponenty reprezentují podskupiny, tedy zvyšuje počet podskupin. Jeho výhoda je v tom, že je možné určit, na kolik podskupin má síť rozdělit.

5.12.5 Metoda optimální modularity podle Newmana

Jako další způsob, jak dělit graf, uvádí Newman v [24] metodu optimální modularity. Modularita měří, jak dobře byl graf rozdělen s tím, že v rámci skupiny by mělo být výrazně více hran než v grafu s náhodně rozmístěnými hranami, zatímco mezi skupinami by tomu mělo být naopak. Úkol rozdělení grafu na skupiny je tedy úkolem hledání maximální modularity přes všechna možná rozdělení.

Newman navrhuje vlastní algoritmus založený na modularitě, který je podle něj výpočetně méně náročný a přitom úspěšnější, než předchozí algoritmy založené na tomto principu. Modularitu při rozdělení grafu na dvě skupiny α a β definuje jako

$$Q = \frac{1}{4|H|} \sum_{ij \in \langle 1,|U| \rangle} (A_{ij} - \frac{k_i k_j}{2|H|}) s_i s_j$$

, kde $s_a = 1$, pokud uzel a patří do skupiny α nebo $s_a = -1$, pokud uzel a patří do skupiny β.

Po úpravě

$$Q = \frac{1}{4|H|}\mathbf{s}^T\mathbf{B}\mathbf{s}$$

, kde $\mathbf{s}$ je vektor prvků s_i a $\mathbf{B}$ je takvaná matice modularity s prvky $B_{ij} = A_{ij} - \frac{k_i k_j}{2|H|}$.

Po další úpravě

$$Q = \sum_{i=1}^{n}(\mathbf{u}_i^T\mathbf{s})^2\beta_i$$

, kde β_i je jedno vlastní číslo k matici B a $\mathbf{u}_i^T$ je jemu odpovídající vlastní vektor. Nyní je potřeba najít maximální Q, což je podle autora NP-těžký problém, jelikož $\mathbf{s}$ je vektor 1 a -1. Autor ale dochází k tomu, že dobré řešení dostaneme i pokud zvolíme nejvyšší β_i a k němu $\mathbf{u}_i^T$. Pak je potřeba určit jen $\mathbf{s}$, který bude maximalizovat výsledné Q. Určí se tak, že znaménka musí odpovídat znaménkům na stejných pozicích v $\mathbf{u}_i^T$.

Pro zjištění aproximace rozdělení na dvě skupiny s maximální modularitou, je tedy potřeba:

1. Určit modulární matici.
2. Vypočítat její vlastní čísla a vlastní vektory.
3. Najít maximální vlastní číslo a odpovídající vlastní vektor.
4. Daný vlastní vektor určuje na základě znaménka na pozici i, do jaké ze dvou skupin uzel u_i patří.

Síť lze rozdělit na více částí tím, že se na již rozdělených částech spouští tento algoritmus rekurzivně. Jak ale Newman podotýká, není možné vymazat hrany mezi dvěma částmi, kde bylo rozdělení provedeno, jelikož pak by bylo další dělení nekorektní. Správným způsobem je vypočítání matice modularity o velikosti $n_g \times n_g$ pro již získanou část sítě $\mathbf{B^{(g)}}$ následovně

$$B_{ij}^{(g)} = A_{ij} - \frac{k_i k_j}{2|H|} - [k_i^{(g)} - k_i\frac{d_g}{2|H|}]$$

, kde n_g je počet uzlů části sítě, $k_i^{(g)}$ je stupeň uzlu i, pokud se uvažuje pouze část sítě a d_g je součtem všech takových stupňů.

Dělení části na další části se provede vždy pouze v případě, pokud tím modularita celé sítě stoupne. Je ale nutné uvažovat zobecněnou modularitu pro více než dvě skupiny

$$Q = \frac{1}{4|H|}\sum_{ij\in\langle 1,|U|\rangle}(A_{ij} - \frac{k_i k_j}{2|H|})\rho_{ij}$$

, kde ρ_{ij} je 1, pokud uzly patří do stejné skupiny, jinak -1.

Hlavní výhoda zmíněného algortimu je, že sám volí velikosti skupin a potenciálně může být velikost jedné ze skupin 0, což odpovídá tomu, že graf nebyl rozdělen, jelikož žádné rozumné dělení neexistuje.

Kapitola 6

Existující nástroje

V této kapitole budou popsány nalezené nástroje, které mají potenciál zjišťovat strukturu skupiny účastníků elektronické komunikace. Většina z nich také podporuje pokročilé vizualizační techniky, které zde ovšem nebudou příliš zmíněny.

Výčet nástrojů aktuálních v době psaní práce není vyčerpávající. Jedná se jen o jejich reprezentativní vzorek. Informace o nástrojích, u kterých není vyznačen zdroj, byly získány na základě vlastní instalace a vyzkoušení.

6.1 NodeXL

NodeXL je open-source software (šablona pro Microsoft Excel), který se snaží analýzu sociálních sítí udělat dostupnou jednoduše i těm, kteří neumí programovat, tím, že poskytuje grafické uživatelské rozhraní v tabulkovém editoru Excel. Umožňuje vizualizaci sítí v podobě konfigurvatelného grafu, analýzu různých vlastností grafu a automatickou podporu shlukování uzlů do skupin. Každému uzlu je možné nastavit fotografii. [27]

NodeXL využívá pro veškerý výpočet nad grafy framework Stanford Network Analysis Platform.

6.2 Nástroje od CASOS

CASOS (Computational Analysis of Social and Organizational Systems) Center se na svých stránkách [28] popisuje následovně:

Within CASOS we attempt to understand and formally model two distinct, but complementary types of phenomena: Human groups, organizations, institutions or society, which are universally informatted and continually acquire, manipulate, and produce information (and possibly other material goods) through the joint, and interlocked activities of people and automated information technologies. The artificial computational system, which is generally comprised of

multiple distributed agents who can mutually influence, constrain and support each other as they try to manage and manipulate the knowledge, communication and interaction networks in which they are embedded.

Z dalších informací na těchto stránkách vyplývá, že se skupina zaměřuje mimo jiné na boj s terorismem a drogami. Mezi nástroje, které vyvinula, a které se mohou použít ke komplexní analýze elektronické komunikace patří ORA, Automap, Construct, Biowar a další. Vše implementováno v Javě. Tyto nástroje je možné volně použít pro studijní účely.

6.2.1 ORA

ORA je robustní nástroj pro analýzu a vizualizaci sociálních sítí. Dokáže ze sociální sítě extrahovat klíčové osoby, místa a témata. Nepracuje pouze s jedinou množinou uzlů, ale uzly reprezentují například osoby, zdroje, dovednosti, úkoly atd. [10]

Unikátní schopností ORA je podpora takzvaných dynamických sítí, což umožňuje sledování změny sítě v prostoru (členové sítě se přestěhovali) a změny sítě v čase. ORA umožnuje sledovat síť ve dvou časových okamžicích a porovnávat, jak se změnili její vlastnosti. Umožňuje také spektrální analýzu, což je detekce neobvyklých událostí nebo detekci změny, což je detekce významných strukturálních změn v síti a toho, co je způsobilo. [10]

6.2.2 AutoMap

AutoMap je text miningový nástroj, který získává data z jednoho či více nestrukturovaných dokumentů. Výstup z něho pak může sloužit jako vstup pro nástroj ORA. Informace, jenž může získávat z textu, jsou jmenné entity, souřadnice zeměpisných lokací podle jejich názvů, probíraná témata, sentiment a síť slov podle jejich spoluvýskytu.

6.2.3 Construct

Construct zkoumá síť jako multiagentní systém a zjišťuje, jak jednotlivý agenti mění svoje chování na základě znalostí naučených interakcí s ostatními. Zjištěné skutečnosti mohou být využiti k předpovídání vývoje u podobných sítí.

6.3 NetMiner

NetMiner je komerční produkt od firmy Cyram. Umožňuje zjišťovat velkou škálu informací o síti jako stupěň, ego sítě, strukturální díry, homophily, assortavity, equicentrality, dyad census, triad census, triad combination, nejkratší cesty, minimální řez, maximální tok, topologické uspořádání, vliv, komponenty, kliky, k-jádra, nejrůznější druhy centrality atd.

Obsahuje data miningové metody: klasifikaci, shlukování, regresi, redukci, detekci odlehlých hodnot a text miningovou metodu LSA, která je určená k extrakci témat z textu, který může být součástí vstupní sítě.

Výhoda NetMineru je poměrně dobrá vizualizace. Verze 4 umožňuje vykreslovat 3D sítě a přiřazovat uzlům různé popisky a zobrazení. [29]

6.4 NetworkKit

NetworkKit je open-source nástroj pro analýzu rozsáhlých sítí čítající až miliardy hran. Z toho důvodu je implementován v C++ s podporou vícevláknového zpracování. [30]

Výsledkem analýzy jsou statistické rozdělení stupně pro uzly, průměr, shlukovací koeficient, komponenty, k-jádra, centrality (betweenness, eigenvector), PageRank, rozdělení na komunity. Podporuje také grafové algoritmy jako BFS, DFS, Dijkstra a generování náhodných grafů. Všechny algoritmy jsou implementovány ve snaze o co nejlepší složitost. [31]

6.5 Commetrix

Podle Dr. Matthiase Triera, což je vedoucí projektu, je nástroj zamýšlen primárně pro analýzu elektronické komunikace jako jsou emaily, diskuzní skupiny a instant messaging. Lze ho ale použít i na jiné typy sítí, než komunikační. [32]

Jeho hlavní výhodou je možnost animovat vývoj sítě v čase, pokud data obsahují časové známky. Například je možné zobrazit, jak se mezi jednotlivými aktéry v síti šíří konkrétní téma. Jinak tento nástroj provádí klasickou SNA, tedy například umí zjišťovat centralitu, statistické rozdělení stupně, hustotu či ego sítě, obohacenou o analýzu témat a klíčových slov.

6.6 Gephi

Gephi je open-source nástroj pro SNA s přívětivím uživatelským rozhraním, tudíž je použitelný pro analýzu i méně zkušenými uživateli. Jeho vývoj je řízen neziskovou organizací Gephi Consortium.

Funkcionalitou je velice podobný předchozímu Commetrixu, umožnuje SNA (centralita, průměr, shlukovací koeficient, průměrná nejkratší cesta, PageRank, HITS, detekce komunit, výpočet modularity), vizualizaci na základě zjištěných hodnot či dynamickou analýzu sítě, tedy sledování změn jejích vlastostí v čase [33]. Stejně jako ostatní SNA nástroje umožňuje volit algoritmus na vykreslení grafu, zde je ovšem navíc možné algoritmus ve vhodné chvíli zastavit. Umožňuje filtrování uzlů na základě vlastností, typicky vymazání uzlů s malým stupněm.

Narozdíl od ostatních nástrojů podporuje širokou škálu vstupních formátů.

6.7 Socilyzer

Socilyzer je placená webová aplikace, která nabízí uživatelsky přívětivé prostředí pro jednoduchou SNA, která zahrnuje rozdělení aktérů do skupin, do komponent a zjišťování jejich stupně a centrality.

6.8 Shrnutí

Nalezené nástroje dovolují měření množství různých vlastností ze SNA. Některé se zabývají i dynamickou analýzou sítí či text miningem a data miningem. Programy jako NodeXL, Gephi či Socilyzer jsou uživatelsky velmi přívětivé a umožňují analýzu i méně technicky zdatným uživatelům. Problémem je, že data z elektronické komunikace je před analýzou nutné převést na síť, s čímž zmíněné programy nijak nepomáhají. Navíc uživatel neznalý v SNA se jen těžko vyzná ve vlastnostech sítě a jedinců, které je možné měřit, natož v tom, co přesně měří.

Kapitola 7

Návrh

V této kapitole budou navrženy metody, které na základě dat z elektronické komunikace v jednotném formátu, který zde bude také navržen, budou schopny určit strukturu skupiny komunikujících jedinců, tedy zjistit hierarchii a rozdělení na podskupiny.

Jelikož byly v kapitole 5 nalezeny pokročilé algoritmy z oboru SNA, které umí zjišťovat to co bylo vytyčeno nad komunikační sítí, budou zvlášť navrženy metody převodu z komunikující skupiny na komunikační síť a zvlášť metody samotné analýzy komunikační sítě.

Navržených algoritmů pro analýzu komunikační sítě bude více než jeden pro zjišťování hierarchie a jeden pro rozdělení na podskupiny, jelikož nebylo rozhodnuto, který z algoritmů by měl být na reálných datech nejlepší.

7.1 Vstupní formát

Vstupem analýzy nebude surová databáze například ze zabaveného zařízení, ale textový soubor v následujícím formátu:

autor;příjemce;obsah zprávy
autor;příjemce;příjemce;obsah zprávy
...

Každá zpráva je na samostatné řádce. Položky v rámci zprávy jsou odděleny symbolem „;“. První položkou je autor. Další až předposlední položka jsou příjemci. Poslední položkou ve zprávě je obsah zprávy. Položky nesmí obsahovat symbol „;“ ani symbol odřádkování. V obsahu zprávy je tedy nutné nahradit tyto symboly například mezerou, což nijak nevadí. Není uvažována velikost písmen, tedy „A“ je stejný jedinec jako „a“. Množina jedinců není v souboru explicitně uvedena a je dána jedinci, kteří se vyskytují jako autoři nebo příjemci ve zprávách.

Zde je praktický příklad toho jak může vstupní soubor vypadat:

A;B;Co děláš B
B;A;Krátím daně jako vždy
C;A;B;Chlapy, odposlouchávají nás

7.2 Převod komunikující skupiny na komunikační síť

V této části bude navržen převod z komunikující skupiny $K = (J, Z)$ na komunikační síť $N = (A, V, w)$, kde A je množina aktérů, V je množina neorientovaných vazeb, říkajících, že koncoví aktéři spolu komunikovali a $w : V \rightarrow R$ je ohodnocení vazeb, vypovídající o jejich intenzitě. Je evidentní, že aktéři i jedinci jsou obrazem skutečně komunikujích lidí, a proto $A = J$. Převod z množiny zpráv Z na množinu vazeb V již tak jednoznačný není, a proto bude rozebírán dále.

7.2.1 Vazba pokud byla poslána zpráva

Nejjednoduší způsob je mezi dvěma aktéry vytvořit vazbu, pokud jeden druhému alespoň jednou poslal zprávu. Otázkou je, zda by měli mezi sebou dva aktéři mít vazbu, pokud jeden poslal druhému pouze hromadnou zprávu, kterou zároveň poslal i všem ostatním. Při takovém způsobu převodu navíc není uvažován rozdíl mezi jedinci, kteří si vyměnili stovky zpráv a mezi takovými, kteří si vyměnili pouze jednu.

7.2.2 Vazba pokud byl poslán dostatečný počet zpráv

Lepším způsobem je vytvořit vazbu jen pokud počet vyměněných zpráv přesáhne určitý limit. Takový způsob je použit například v [34]. Nevýhodou takového přístupu ovšem je, že slabší vazby, které nepřesáhnout daný limit, budou ignorovány. Navíc v síle uvažovaných vazeb můžou být řádové rozdíly, které se nadále ignorují.

7.2.3 Vážené vazby

Třetím způsobem je zavést váhy vazeb, které vyjadřují jejich sílu. Tento přístup předchozími nevýhodami netrpí.

Váhou by mohl být počet poslaných zpráv mezi konkrétními dvěma jedinci. V takovém případě by ovšem nijak nebylo využito cenného obsahu zpráv, který je v rámci komunikující skupiny k dispozici.

Bude se předpokládat, že délka zprávy je dobrou aproximací toho, kolik nese informace a jak posiluje vazbu mezi jedinci. Váha vazby mezi dvěma jedinci by tedy mohla být součtem délek všech zpráv poslaných mezi nimi. V tom případě by mohly být váhy velmi rozdílné a případně obrovské. Na jedné straně můžou být v síti konverzace obsahující jednotky krátkých zpráv, tedy váha by byla například 100, na druhé straně zde mohou být konverzace obsahující stovky dlouhých zpráv, tedy vznikly by vazby s váhami například 1 000 000. Jelikož ale důležité jsou hlavně řádové rozdíly v objemu komunikace, bude váha vazby logaritmem celkového součtu délek zpráv v konverzaci. Konkrétně to bude logaritmus o základu 10, jelikož je poměrně názorný. Váha vazby mezi dvěma aktéry a, b s konverzací K_{ab} bude tedy

$$\log_{10} \sum_{m \in K_{ab}} l_m$$

pro $\sum_{m \in K_{ab}} l_m >= 10$ a 1 pro $\sum_{m \in K_{ab}} l_m < 10$, kde l_m je délka zprávy m. Hodnota pro $\sum_{m \in K_{ab}} l_m < 10$ musí být vždy 1, aby nevznikla vazba se zápornou váhou nebo s váhou 0, což by bylo ekvivalentní tomu, že vazba neexistuje.

7.2.4 Vybraný způsob

Třetí způsob bude použit, jelikož je považován za nejpokročilejší a byly nalezeny algoritmy, které s váženou komunikační sítí umí pracovat. Na druhou stranu některé algoritmy s váženými hranami pracovat neumí. V tom případě by mohl být použit druhý způsob. Tím by ovšem mohla vzniknout komunikační sít rozdělená do více komponent, s kterou všechny algoritmy pracovat neumí. Proto v případě, že algoritmus neumí pracovat s váhami, budou váhy jednoduše ignorovány a bude v podstatě použit první nejjednoduší přístup.

7.3 Hledání hierarchie nad komunikační sítí

Jak bylo vysvětleno v 5.10, centralita je přesně to, co by mělo být na síti měřeno za účelem zjištění významnosti aktéra vzhledem k dané síti. Jelikož je snaha o využití vah, které dávají cenou informaci o intenzitě komunikace mezi aktéry, musí být použity algoritmy pro váženou centralitu. Budou to Degree, betweenness a closeness centralita s α parametrem. Konrétní použité α bude $\alpha = 0$, $\alpha = 0,5$ a $\alpha = 1$. α větší než 1 použito nebude, jelikož jak bylo uvedeno, v komunikační síti je typicky důležitější mít více slabých vazeb, než málo silných. Pro $\alpha = 0$ budou vlastně váhy zcela ignorovány a bude se tedy jednat o jednoduchou centralitu zmíněnou v 5.10. Pro $\alpha = 1$ budou váhy uvažovány stejně jako v jednoduché vážené centralitě. Pro $\alpha = 0,5$ budou váhy uvažovány také, ovšem bude více hleděno na celkový počet vazeb, tedy bude méně výhodné mít jednu silnou vazbu, než více slabších vazeb v součtu o stejné

hodnotě jako ona silná vazba. Hodnota 0,5 nebyla zvolena nijak exaktně, ale jen proto, že je právě v půli cesty mezi předchozími dvěma případy.

Eigenvector centralita bude měřena také ovšem pouze ve variantě bez uvažování vazeb a s uvažováním vazeb pomocí převodu na multigraf. Varianta s parametrem α nebyla nalezena.

7.4 Rozdělení na podskupiny nad komunikační sítí

Pro řešení problému rozdělení na podskupiny budou použity dva algoritmy, které jsou ze zde zmíněných považovány za nejpokročilejší. Girvan-Newmanův algoritmus je hojně implementován (je ostatně jako jediný pro tyto účely implementovan v knihovně JUNG, kterou tato práce využívá), a měl by tedy být poměrně dobrý na reálných datech. Newmanův algoritmus tolik implementován není, ale výsledky v [24] v tomto případě Girvan-Newmana překonal.

7.4.1 Girvan-Newmanův algoritmus

Girvan-Newmanův algortimus vyžaduje určení počtu výstupních podskupin. To je bohužel pro účely této práce nežádané, jelikož dělící algoritmus by si sám měl určit počet poskupiny do kterých síť rozdělit. Algoritmu se tedy nebude předávat počet výstupních skupin, ale počet hran, které má odebrat. Tento počet bude konstatní část celkového počtu hran, konkrétně 1/3. Bude tedy odebrána třetina hran s největší betweenness centralitou. To by mělo umožnit kompaktním sítím, aby nebyly rozděleny vůbec a naopak dobře separovatelným sítím, aby z nich byly vytvořeny komponenty představující podskupiny. Girvan-Newmanův algoritmus má navíc tu nevýhodu, že neuvažuje váhy vazeb. Bude tedy ze zde uvedených algoritmů tvořit jedinou výjimku a bude ignorovat celkový objem komunikace provedený v rámci konverzace.

7.4.2 Newmanův algoritmus optimální modularity

Newmanův algoritmus optimální modularity se dokáže sám zastavit ve chvíli, kdy již další dělení není vhodné a nepotřebuje tak určovat počet výstupních podskupin. Dokáže pracovat s váženými vazbami, pokud dojde k transformaci sítě na multigraf způsobem uvedeným v 5.11. Váhy se pak promítnou do matice sousednosti a do stupňů a stanou se tak přirozenou součástí algoritmu.

Kapitola 8

Implementace a testování navržených metod

8.1 Testovací dataset

Jako testovací dataset byl vybrán Enron Corpus. Enron Corpus byl poprvé představen Klimtem a Yangem v [35] na CEAS (Collaboration, Electronic messaging, Anti-Abuse and Spam) konferenci 2004, a v [35] na ECML (European Conference on Machine Learning) 2004. Podle autorů se jedná o velký soubor reálných emailových zpráv (dále jen zpráv), které byly zveřejněny v průběhu vyšetřování společnosti Enron. Původně obsahoval 619 446 zpráv patřících 158 pracovníkům společnosti Enron. Autoři ale zredukovali obsah zpráv na 200 399, tím, že odstanili duplikáty a počítačem generované zprávy. Zprávy jsou v datasetu členěny do různých složek, tak, jak si je rozčlenili lidé, kterým patřili. Ty jsou pak ve složkách odpovídajících konkrétnímu člověku.

Zveřejněn byl v témže roce Williamem Cohenem na [36], kde je navíc uvedeno, že původní dataset byl nejdříve zveřejněn na internetu americkou Federal Energy Regulatory Commission v průběhů vyšetřování a patřil především zaměstnancům z vyššího managementu. Později byl odkoupen Leslie Kaelblingem z MIT a po předzpracování dat, kdy byly navíc smazány přílohy a některé zprávy na žádost původních zaměstnanců společnosti Enron, byl upravený zveřejněn na uvedené stránce.

8.1.1 Enron

V [37] se uvádí, že Enron byl založen v roce 1985 Kenethem Layem a spekulací s elektrickou energií se z něj stala sedmá největší americká akciová společnost. Keneth Lay byl až do roku 2002 v čele. V té době společnost zaměstnávala přes 20 000 zaměstnanců. Někteří zaměstnanci společnosti Enron, včetně jejího kontrolora Arthura Andersena, byli vyšetřováni a následně odsouzeni za zastí-

rání finanční krize, která se ve společnosti začala projevovat a vedla k jejímu bankrotu v roce 2002.

Jak uvádí [6], zprávy byly získány v časovém úseku 1998-2002. Jedná se o relativně dlouhý interval, kdy se mohly pozice jednotlivých členů ve společnosti měnit, což se například podle [37] stalo v roce 2000, kdy byl na několik měsíců Jeffrey Skilling místo Kenetha Laye na nejvyšším postu jako CEO.

8.1.2 Instance Enron Corpusu

Jak je uvedeno v [6], v literatuře se může pracovat s různými verzemi datasetu, jelikož mnoho výzkumných skupin opravovalo integritní problémy a nekonzistence po svém. Navíc na původní stránce [36] byl dataset z roku 2004 v roce 2009 dodatečně upraven a ten z roku 2004 již není oficiálně dostupný. Shetty a Adibi vytvořili ze starší verze MySql databázi uvedenou v [38], která obsahuje 151 zaměstnanců a 252 759 zpráv.

8.1.3 Zvolená instance Enron Corpusu

Instance datasetu, která bude v této práci použita na testování navržených metod, byla uvedena v [39]. Jak je zde uvedeno, autoři poskytnou na požádání MongoDB databázi, kterou popisují. Ta obsahuje jak původních 158 zaměstnanců, jejichž data byla získána a jsou zde označováni jako core, tak dalších 1360 noncore zaměstnanců. Noncore zaměstnanci jsou ti, kteří se účastnili komunikace s core zaměstnanci a nebyly předmětem sledování a jimi provedená komunikace v rámci společnosti tedy byla zaznamenána jen z části, jelikož nebyly získány obsahy jejich emailových schránek.

Přestože autoři [39] pracují mimo jiné s oněmi 158 core zaměstnanci, v databázi není zanesena informace, kteří zaměstnanci jsou core a kteří noncore.

Databáze neobsahuje jen zaměstnance a zprávy, ale také informace o tom, že je nějaký zaměstnanec na vyšší pozici než jiný nebo například, že zaměstanci patří v rámci společnosti pod nějakou organizační jednotku například pod HR oddělení. Přestože u zaměstnanců je možné zjistit i názvy pozic, které ve společnosti zastávali, jména organizačních jednotek přítomna nejsou. Autoři tuto část databáze nazývají jako „zlatý standard pro organizační hierarchii společnosti Enron“(dále jen „zlatý standard“).

8.1.4 Data k testování správnosti navržených metod

„Zlatý standard“ bude v této práci použit k validaci zde navržených metod na zjišťování hierarchie a rozdělení na podskupiny. Přestože obsahuje pouze informace o formální hierarchii a v [6] autoři podotýkají, že formální hierarchie nemusí vždy odpovídat hierarchii v každodenním životě, považuje se zde tato formální hierarchie za nejlepší dostupný odhad reálné hierarchie organizace a bude proto použita k validaci zde navržených metod. Rozdělení na podskupiny

je ve „zlatém standardu“ také pouze formální a navíc nepřímé, takže není pevně dán počet skupin, ale použito k validaci bude ze stejného důvodu.

8.2 Implementace

Metody navržené v přechozí kapitole byly implementovány v programovacím jazyce Java s využitím open-source knihoven nabízejících některé algoritmy ze SNA a teorie grafů. Konkrétně se jedná o knihovnu JUNG (Java Universal Network/Graph Framework), která je chráněna BSD licencí a nabízí i pokročilé algoritmy ze SNA, jako různé druhy centralit, shlukovací koeficient či Girvan-Newmanův algoritmus na hledání podskupin a knihovnu JGraphT, která je chráněna LGPL a EPL licencí a nabízí především algoritmy z teorie grafů. Dále je využita knihovna JAMA, která je dostupná jako volné dílo a poskytuje základní algoritmy z lineární algebry. Použita byla pro výpočet vlastních čísel a vlastních vektorů u metody optimální modularity podle Newmana. Pro načítání testovací databáze byl využit driver pro MongoDB.

Implementace se skládá mimo jiné z následujících tříd:

- Třída načítající vstupní soubor ve formátu popsaném v 7.1 a implementující rozhraní komunikující skupiny, tedy poskytující množinu jedinců a množinu zpráv.

- Testovací třída načítající data z databáze uvedené v 8.1, která implementuje rozhraní komunikující skupiny, tedy poskytující množinu zpráv a jedinců. Dále poskytuje skutečnou hierarchii a skutečné rozdělení na podskupiny. Testovací data jsou načítána přímo z databáze a ne ze vstupního souboru především kvůli jejich velikosti.

- Třída převádějící komunikující skupinu na komunikační síť v předchozí kapitole navrženým způsobem.

- Třída provádějící nad komunikační sítí předchozí kapitolou navržené SNA algoritmy na měření centrality a detekci komunit.

- Testovací třída zjišťující na základě výsledků ze SNA, skutečné hierarchie a skutečného rozdělení na podskupiny procentuální úspěšnost metod.

- Třída, která komunikační síť společně s informací o zjištěné hierarchii či zjištěném rozdělení na podskupiny uloží ve formátu popsaném v [40] do dvou souborů. Jeden soubor obsahuje vazby a druhý aktéry společně se zjištěnou informací. Tyto soubory je možné vizualizovat v programu Gephi zmíněném v 6.6.

Tabulka 8.1: Přehled vlastností testovacích komunikačních sítí pro hledání hierarchie

	počet vazeb	součet vah	průměrná váha	medián vah
S_{core}	1424	5357	3,76	4
S_{all}	22494	81580	3,62	4

8.3 Testování hledání hierarchie

V následující části budou otestovány navržené metody na hledání hierarchie ve skupině komunikujících lidí.

8.3.1 Komunikující skupiny použité k testování

Jak je uvedeno v 8.1.3, přestože autoři testovací databáze ve své práci [39] pracují odděleně s core a noncore zaměstnanci, v databázi není zanesená informace o tom, o které se jedná. Z [41] byla získána jména core záměstnanců společnosti Enron. Z toho 109 bylo pomocí jmen identifikováno v databázi. Jména těchto zaměstnanců jsou mimo jiné uvedena v obrázku 8.2. V této části se tedy bude pracovat se 109 zaměstnanci, u nichž je dostupná celá jejich komunikace v rámci společnosti. Bude pracováno s komunikující skupinou $K_{core} = \langle J_{core}, Z_{core} \rangle$, kde J_{core} je oněch 109 zaměstnanců a Z_{core} jsou veškeré dostupné zprávy poslané mezi nimi.

Další komunikující skupinou, se kterou se zde bude pracovat, je $K_{all} = \langle J_{all}, Z_{all} \rangle$, kde J_{all} je všech 1518 zaměstnanců u nichž je informace o formální hierarchii a nemusí být pravda, že je známa veškerá komunikace jedince v rámci společnosti, ze kterých bylo ale navíc odebráno 129 zaměstnanců, kteří nebyli komunikačně propojeni s ostatními z 1518 zaměstnanců. Zbylo tedy 1389 zaměstnanců, kteří jsou komunikačně navzájem propojeni a splňují tak definici komunikující skupiny. Z_{all} jsou veškeré dostupné zprávy poslané mezi nimi.

Komunikační síť, vytvořená z K_{core} navrženým způsobem, bude označena jako S_{core}. Komunikační síť, vytvořená z K_{all} navrženým způsobem, bude označena jako S_{all}. V tabulce 8.1 je přehled několika vlastností těchto sítí. Mimo jiné z ní lze vyčíst, že průměrný objem komunikace mezi dvěma aktéry je zhruba 10 000 znaků.

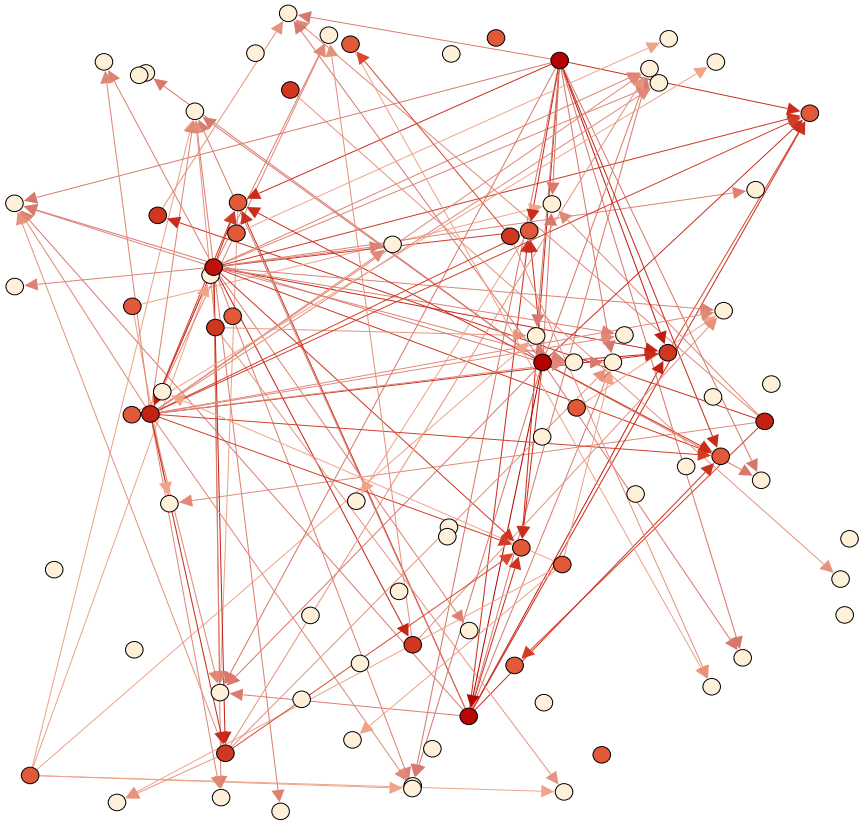
8.3.2 Způsob vizualizace zjištěné hierarchie

Na obrázcích 8.4, 8.5, 8.6, 8.7 je vidět komunikační síť S_{core}. Obrázky se liší pouze v odstínu červené, kterou jsou vykresleny uzly. Čím je červená intenzivnější, tím má aktér vyšší zjištěnou centralitu, tedy je významnější. Nutno podotknout, že hierarchie není stejná pouze na dvou grafech obarvených stejně,

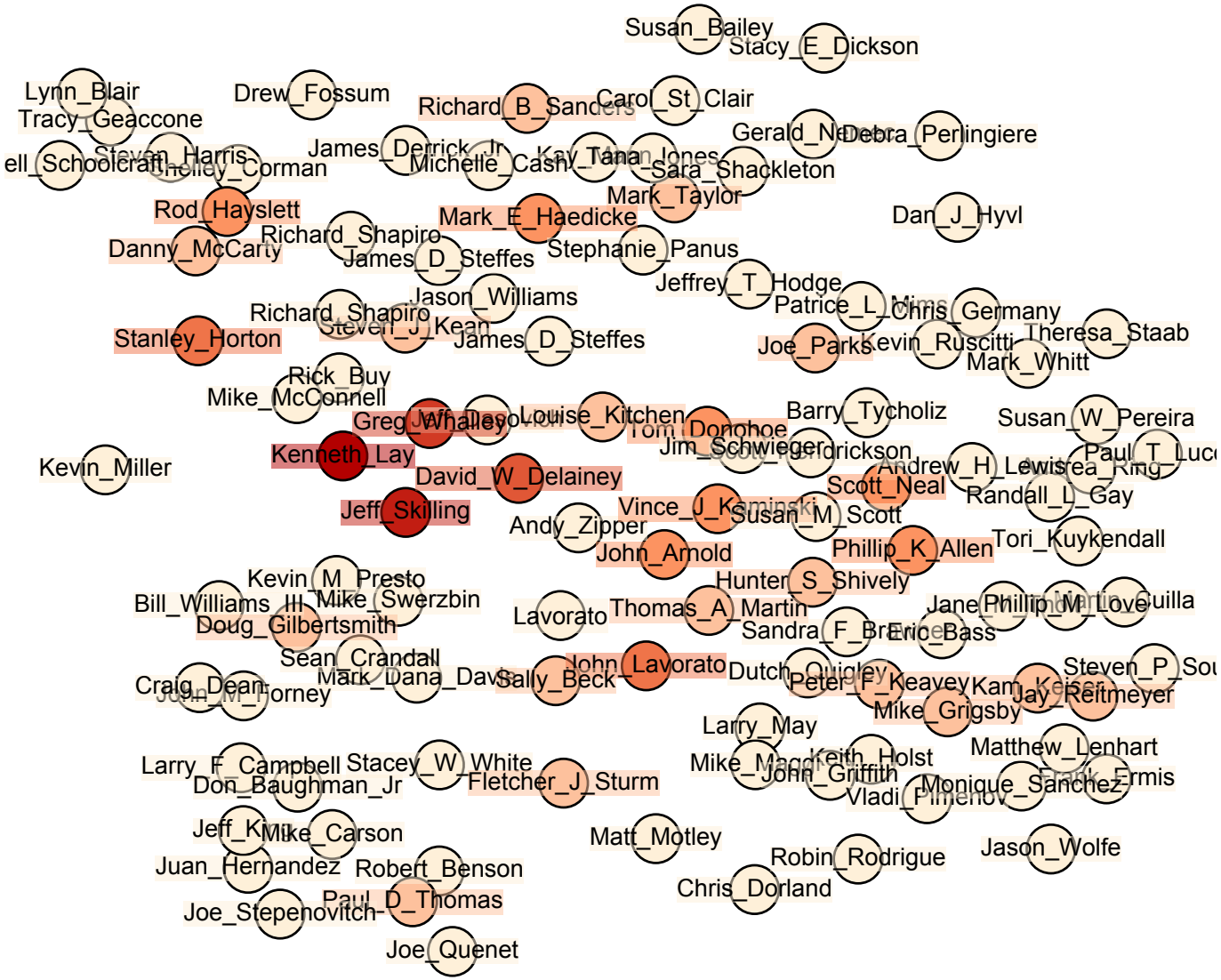
ale také na grafech, kde pro všechny dvojice uzlů platí, že pokud v prvním grafu je jeden uzel světlejší než druhý, je tomu tak i v druhém grafu.

8.3.3 Vizualizace reálné formální hierarchie

Stejným způsobem jako zjištené hierarchie byla vizualizovaná i reálná formální hierarchie získáná z databáze, aby bylo možné pouhým okem odhadnout správnost zjištěných hierarchií. Taková vizualizace ovšem nemůže být správná, jelikož reálná hierarchie není relací lineárního uspořádání, a přesto se zde tak vizualizuje. Z toho důvodu je na obrázku 8.1 ještě korektní vizualizace reálné hierarchie, která ovšem není užitečná k odhadu správnosti zjištěných hierarchií. Znázorněná orientovaná hrana (a, b) je vlastně elementem z relace dominance popsané v 4.1, barva uzlu pak odpovídá barvám v 8.3. Na obrázku 8.2 jsou navíc skutečná jména aktérů v komunikační síti společně s jejich reálnou hierarchií. Jak je vidět Keneth Lay a Jeff Skilling jsou podle reálné hierarchie skutečně na nejvyšších pozicích společnosti, jak se již dalo usoudit z informací v 8.1.1.



Obrázek 8.1: Korektní reálná hierarchie pro K_{core} vyznačena hranami a odvozená „nekorektní“ reálná hierarchie znázorněna barvou



Obrázek 8.2: Jména aktérů v komunikační síti S_{core} společně s reálnou hierarchií

8.3.4 Výsledky

V tabulce 8.2 je vidět úspěšnost zjišťování hierarchie pro S_{all} a S_{core}. Některé hodnoty pro S_{all} jsou nevyplněné, jelikož výpočetní náročnost pro takové množství aktérů je již příliš vysoká.

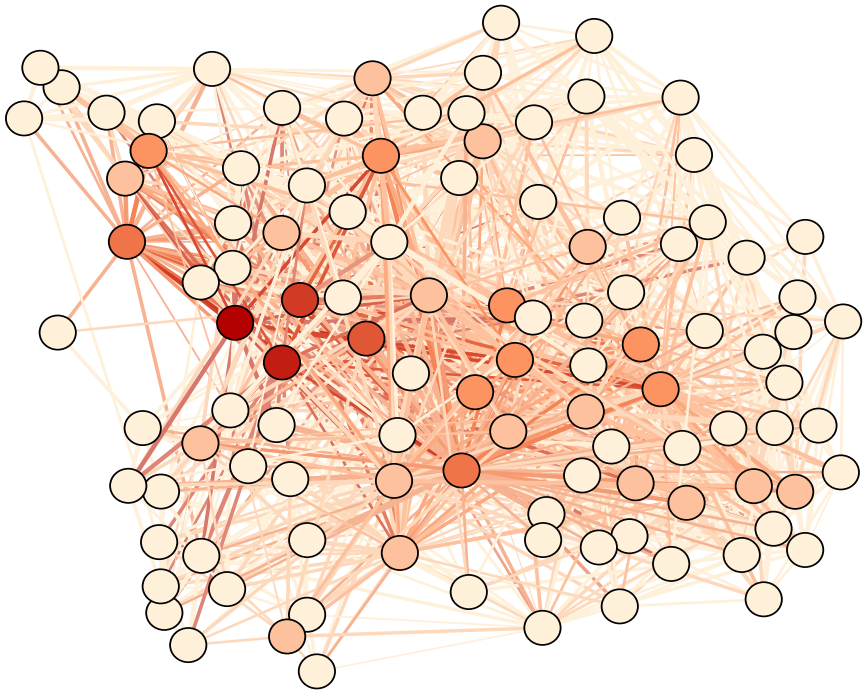
Úspěšnost je celkově vyšší u K_{all}, jelikož jak bylo řečeno v 8.1, vyšetřovaní core zaměstnanci byli z nejvyššího managementu, tedy byli významní, což byl nejspíš i důvod proč byli sledování právě oni. Jelikož u noncore zaměstnanců není dostupná kompletní jimi provedená komunikace, centralita u nich je automaticky nižší a jelikož byli skutečně na nižších pozicích, je úspěšnost hledání hierarchie při uvažování i noncore zaměstnanců vyšší.

Překvapivě, na základě výsledků v tabulce nelze usuzovat, že zavedení vah u vazeb by nějak výrazně zlepšovalo úspěšnost.

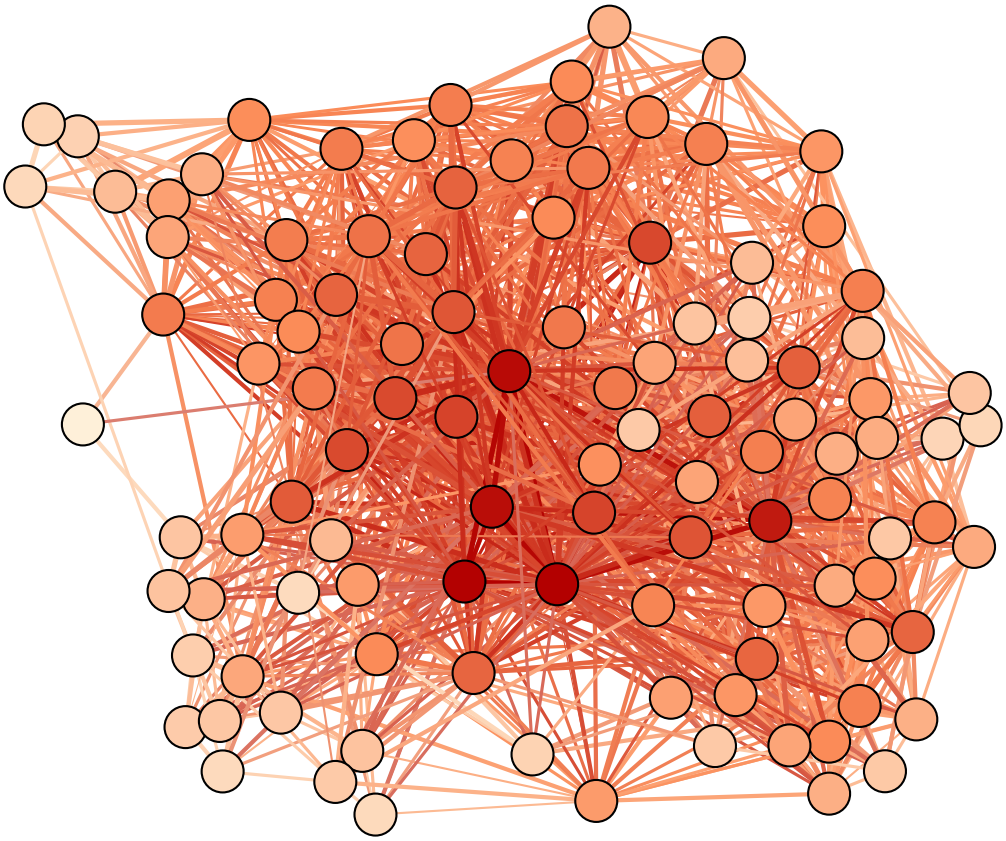
U S_{core} získala nejlepší úspěšnost vážená α betweenness centralita, což je poměrně překvapivé. Očekávaným šampionem u S_{core} byly degree centrality, protože jak zmiňují i autoři [39], síť má poměrně velký počet vazeb, kde aktéři mají vazby téměř každý s každým.

Jednoznačně nejhorší u S_{core} je closenness centralita, která je ale naopak nejlepší u S_{all}. Z toho lze odhadovat, že closenness centralita bude dobrá v případech, kdy je dostupná kompletní komunikace pouze malé části jedinců z komunikující skupiny, což se při vyšetřování typicky bude dít, jelikož získáním komunikace vyšetřovaných jedinců bude získáno i velké množství jedinců, se kterými vyšetřovaní někdy komunikovali, ale nejsou předmětem vyšetřování. Otázkou zůstává, zda je informace o hierarchii nevyšetřovaných jedinců zajímavá.

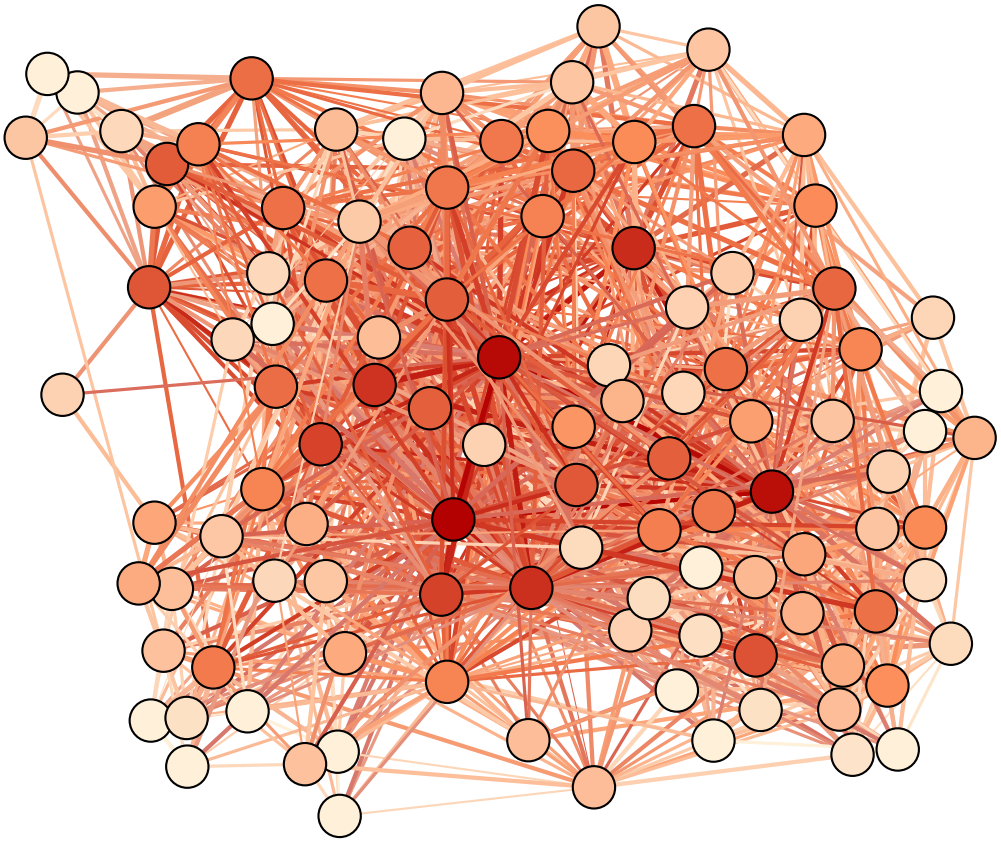
Jak je vidět na obrázcích 8.4, 8.5, 8.6, 8.7, které zobrazují změřenou centralitu na síti pro vybrané druhy centrality, shodly se centrality poměrně dobře na tom, kdo jsou nejvýznámější členové, a to Louise Kitchen, John Lavorato a Phillip Allen. Formálně je Phillip Allen manažer, John Lavorato je CEO Enron America a Louise Kitchen prezident Enron Online. Jeff Skilling a Keneth Lay, kteří jsou na nejvyšších pozicích formální hierarchie se mezi zjištěnými nevýznamnějšími aktéry neumístili. Pouhým neodborným odhadem by se z toho mohlo usoudit, že formálně nejvyšší zaměstnanci delegovali svoji moc na zaměstnance, kteří za ně prováděli většinu manažerské činnosti, nebo jednoduše největší vliv ve společnosti měli oni a nikoliv Jeff Skilling a Keneth Lay. Tento fakt by neměl při reálném použití metod vadit, jelikož například ve zločineckých skupinách ve stylu mafie, kde vládne obvykle tzv. kultura moci [42], vůdce svou moc typicky příliš nedeleguje, a tudíž musí sám provádět velké množství komunikace k ovládání skupiny. I v případě, kdy by se vyšetřovala například podobná společnost jako Enron a vedoucí jedinci by byli špatně označeni, nejedná se o prohru, jelikož označení jedinci budou pravděpodobně znát nejdůležitější informace o skupině a mít úzký kontakt na opravdové vedení. Zároveň budou nést velký podíl viny, protože důležitá komunikace probíhala právě přes ně.



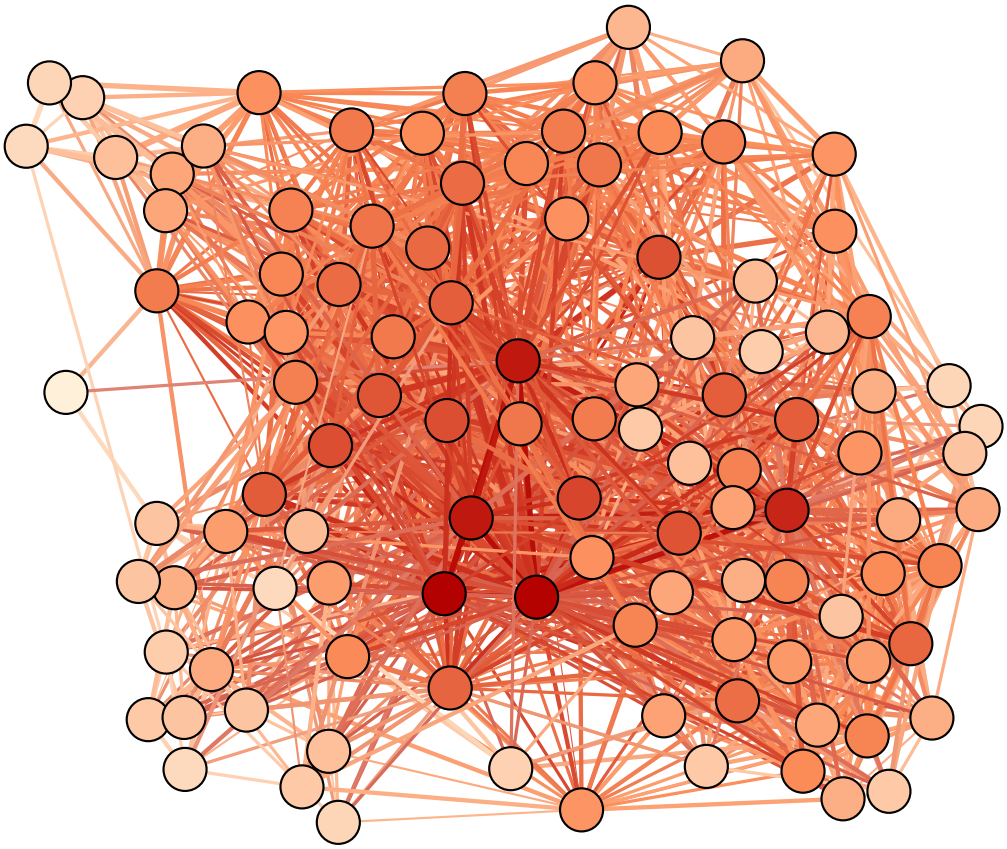
Obrázek 8.3: Komunikační síť S_{core}, reálná hierarchie znázorněna barvou



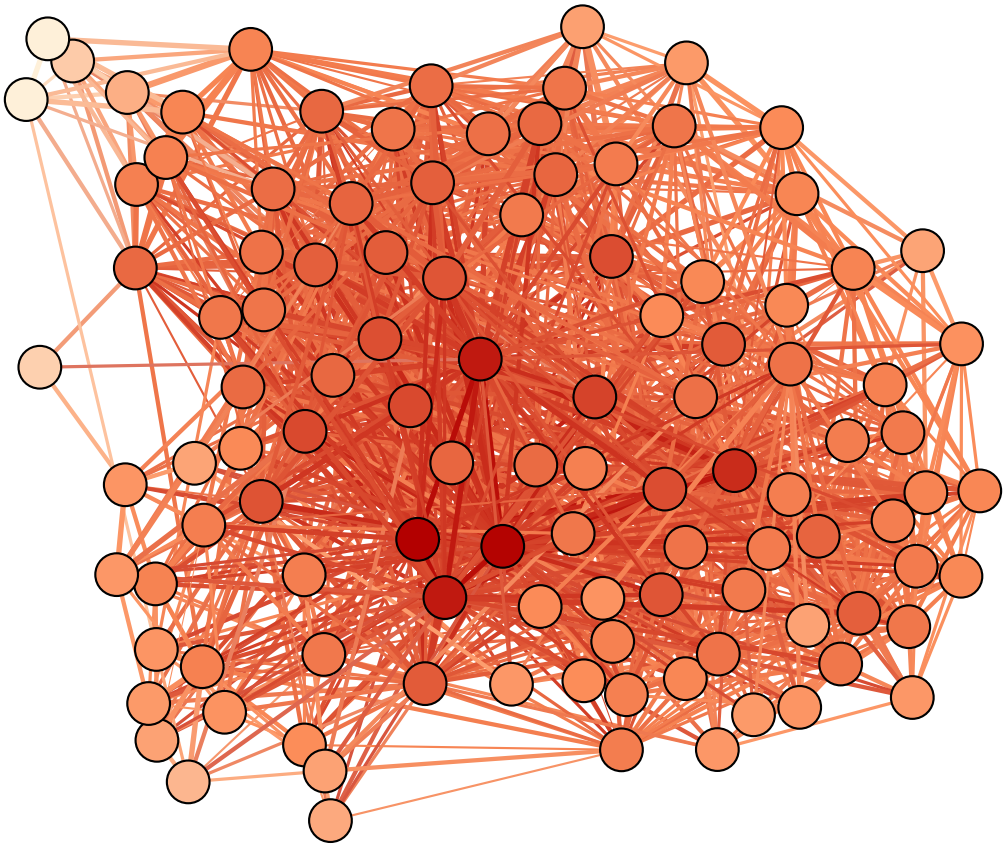
Obrázek 8.4: α degree centralita na komunikační síti S_{core}



Obrázek 8.5: α betweenness centralita na komunikační síti S_{core}



Obrázek 8.6: Eigenvector centralita na komunikační síti S_{core}



Obrázek 8.7: Closeness centralita na komunikační síti S_{core}

Tabulka 8.2: Úspěšnost zjišťování hierarchie pro různé druhy centralit

	S_{all}	S_{core}
C_D	87,69%	79,26%
C_D^w	86,72%	78,95%
$C_D^{\alpha=0,5}$	84,32%	79,88%
C_B	85,50%	79,88%
C_B^w	–	80,80%
$C_B^{\alpha=0,5}$	–	81,11%
C_C	90,02%	73,07%
C_C^w	–	75,23%
$C_C^{\alpha=0,5}$	–	74,92%
C_E	88,08%	79,26%
C_E^w	–	78,95%

8.4 Testování rozdělení na podskupiny

V následující části budou otestovány navržené metody pro rozdělení komunikující skupiny na podskupiny.

8.4.1 Komunikující skupiny použité k testování

Reálné rozdělení na podskupiny, tedy funkce c_{real}, potřebné k zjištění úspěšnosti rozdělení na podskupiny, bylo získáno ze stejné databáze jako hierarchie. Přestože k tomuto použití nejpíše nebyla databáze určena. Obsahuje údaje o tom, že zaměstnanec patří do jakési organizační jednotky, neboli skupiny. Skupin, které tvoří oněch 109 zaměstnanců z K_{core}, bylo extrahováno 32, přičemž 16 z nich byly pouze jednočlenné skupiny. Na tyto jednočlenné skupiny se dá nahlížet také jako na jedince bez skupiny, proto budou z testování vyjmuti a bude použita komunikující skupina K_g, která obsahuje jen jedince patřící v reálném rozdělení na podskupiny do větší než jednočlenné skupiny a jejich komunikaci.

K_g obsahuje komunikaci určitých zaměstnanců společnosti Enron z průběhu celého fungování společnosti, tedy několika let. V průběhu této doby došlo, jak bylo zmíněno i v 8.1.1, ke změnám v celkové struktuře. Z toho důvodu se zde použijí i komunikující skupiny vytvořené z kratších úseků fungování společnosti, v rámci nichž se událo méně změn. Budou to komunikující skupiny z 1. polotetí 2000 a 1. pololetí 2001 označené popořadě K_{20}, K_{21}. Tabulka 8.3 ukazuje přesné počty jedinců, zaslaných zpráv a reálných skupin v jednotlivých komunikujících skupinách.

Komunikující skupina obsahující noncore zaměstnance zde použita nebude, jelikož je evidentní, že jako první by ze sítě byli odděleni právě noncore za-

Tabulka 8.3: Přehled testovacích komunikujících skupin pro rozdělení na podskupiny

	počet jedinců	počet zpráv	počet skupin
K_g	92	11497	16
K_{20}	65	188	7
K_{21}	92	2765	16

Tabulka 8.4: Přehled testovacích komunikačních sítí pro rozdělení na podskupiny

	počet vazeb	součet vah	průměrná váha	medián vah
S_g	1054	3938	3,73	4
S_{20}	179	635	3,37	3
S_{21}	584	2027	3,5	4

městnanci, kteří by tvořili mnoho jednočlenných skupin.

Komunikační sítě vytvořené z K_g, K_{20} a K_{21} navrženým způsobem budou označeny popořadě jako S_g, S_{20} a S_{21}. V tabulce 8.4 je přehled několika vlastností těchto sítí.

8.4.2 Komentář k vyhodnocování úspěšnosti

Úspěšnost rozdělení na podskupiny se pohybuje v intervalu $\langle 0, 1\rangle$. Stejně jako u hierarchie platí, že 50% je typická úspěšnost, pokud problém bude řešen náhodně, tedy pokud bude každý jedinec náhodně vložen do některé z m skupin (m je počet reálných skupin, tudíž náhodný algoritmus má ještě navíc malou výhodu, jelikož ví kolik má skupin být). Narozdíl od hierarchie je ale velice těžké dosáhnout čísel blízkých 1. Na ukázku byl proveden experiment. Pro hledání hierarchie bylo provedeno 10000 pokusů, kdy byla jedincům přiřazena různá čísla (žádná dvě nebyla stejná) a byla měřena úspěšnost při odhadu reálné hierarchie. Průměr byl 0,50013 a průměrná odchylka 0,0223906. Poté bylo provedeno 10000 pokusů pro rozdělení na podskupiny, kdy jedinci byli náhodně vloženi do 16-ti skupin (což je počet reálných skupin u K_g a K_{21}) a byla opět měřena úspěšnost. Průměr byl 0,500042 a průměrná odchylka pouze 0,003951. Tento pokus má demonstrovat, že i úspěšnost rozdělení na podskupiny blízká 50% může být poměrně dobrá.

8.4.3 Způsob vizualizace výsledků

Obrázky v této části ukazují rozdělení jedinců do podskupin podle reálného formálního rozdělení a podle toho, jak bylo určeno algoritmy pro detekci komunit. Pozice uzlů jsou dány použitým algoritmem pro rozložení uzlů v grafu Force Atlas nebo Force Atlas 2. Podle [43] zachází algoritmus s hranami jako s pružinami mezi uzly, které je táhnout k sobě. Z toho lze usoudit, že nazvájem více propojené uzly budou po vizualizaci blíže u sebe. Samotná blízkost na obrázku je tedy poměrně dobrým odhadem toho, že by jedinci měli být ve stejné skupině, jak se ostatně i ukáže. Každá barva vyjadřuje jednu skupinu. Nutno připomenout, že na konkrétních barvách nezáleží. Pokud by tedy došlo k záměně dvou barev, rozdělení na podskupiny by zůstalo stejné.

8.4.4 Výsledky

Na obrázku 8.10 je vidět rozdělení do podskupin provedené pro S_g náhodně, způsobem popsaným v 8.4.2. Evidentně se odlišuje od rozdělení provedených algoritmy na detekci komunit. Záměrem bylo demonstrovat, že algoritmy negenerují náhodná řešení.

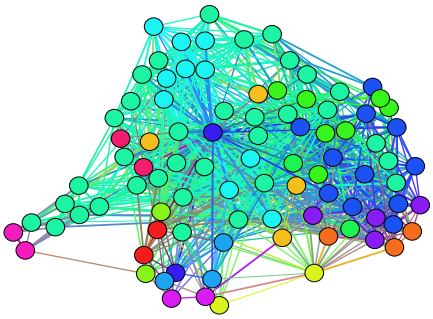
Jak je vidět na obrázku 8.8, reálné rozdělení do podskupiny se opravdu nějakým způsobem projevuje v komunikaci, jelikož Force Atlas vykreslil jedince ze stejné podskupiny blízko sebe. Problémem je, že i v rámci podskupin je komunikace velice intezivní, a tak je pro algoritmy na detekci komunit velice těžké podskupiny rozeznat.

Pro S_g dosáhl algoritmus Girvan-Newman svojí nejvyšší úspěšnosti. Newmanův algoritmus již na začátku odmítl dělení skupiny, jelikož podle něj pro dělení nebyl vhodný, a proto dosáhl pouze padesátiprocentní úspěšnosti. Jedná se o známku toho, že pro S_g neexistuje žádné příliš dobré dělení, jak je ostatně patrné z komunikační sítě na obrázcích 8.8, 8.9, 8.10.

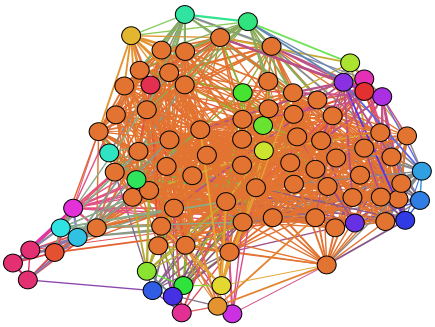
S_{20} již byla lépe separovatelná, pravděpodobně tím, že byla vytvořena z komunikace za kratší časové období. Newmanův algoritmus zde dosáhl nejlepší úspěšnosti, zatímco Girvan-Newmanův algoritmus, zde jen podle úspěšnosti poměrně selhal. Posouzením obrázků 8.13 a 8.12 lze ale dojít k závěru, že velký rozdíl v úspěšnosti by jen na základě obrázků nebyl odhadován a i Girvan-Newmanův algoritmus vygeneroval rozdělení, které se trochu podobá tomu reálnému. Newmanův algoritmus zřejmě byl tak úspěšný hlavně proto, že, jak je vidět na obrázku 8.13, vytvořil zelenou skupinu, která odpovídá v reálném rozdělení 8.11 růžové skupině, která je v rozdělení největší.

8.4.5 Shrnutí

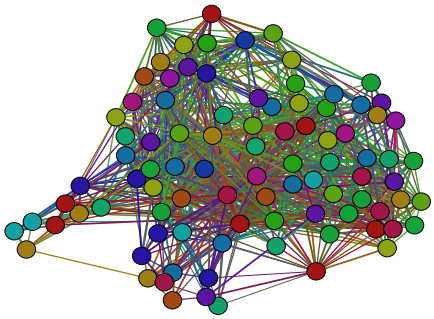
Výsledná úspešnost použitých algoritmů není nijak oslnivá. To je dáno zaprvé důvodem zmíněným v 8.4.2. Za druhé tím, že reálná podoba formálních podskupin ve společnosti byla extrahována ad-hoc z databáze, která k tomu nebyla původně určena, a proto nemusí být příliš přesná. Za třetí, formální



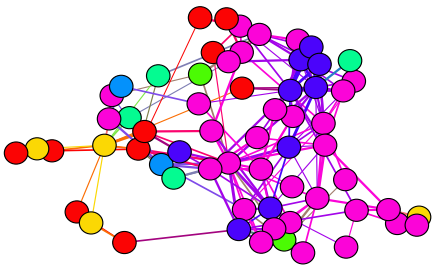
Obrázek 8.8: Reálné rozdělení na podskupiny v S_g



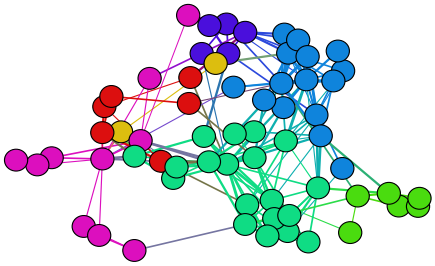
Obrázek 8.9: Rozdělení S_g na podskupin pomocí Girvan-Newman algoritmu



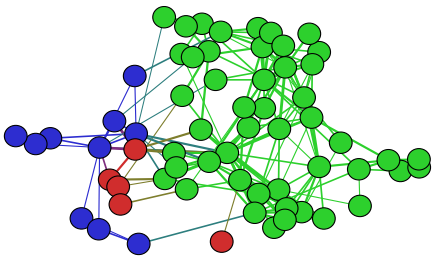
Obrázek 8.10: Náhodné rozdělení S_g na podskupiny



Obrázek 8.11: Reálné rozdělení na podskupiny v S_{20}



Obrázek 8.12: Rozdělení S_{20} na podskupiny pomocí Girvan-Newman algoritmu



Obrázek 8.13: Rozdělení S_{20} na podskupiny pomocí Newmanovi optimální modularity

Tabulka 8.5: Úspěšnost a počet zjištěných skupiny rozdělení na podskupiny různými algoritmy pro různé komunikující skupiny

	Girvan-Newman	Newmanova opt. modularita
K_g	60,89%, 32	50%, 1
K_{20}	54,42%, 7	63,22%, 3
K_{21}	58,7%, 24	56,25%, 3

podskupiny ve společnosti téměř stejně intenzivně komunikují v rámci podskupin jako mezi podskupinami, a tudíž může být nemožné je nějak ostře oddělit pouze na základě jejich komunikace.

Závěr

Absence forenzních nástrojů, které by umožňovaly analýzu elektronických dat od začátku do konce, o které mluví Alzaidy, zdá se, alespoň na poli elektronické komunikace, stále trvá, jak ukázala rešerše existujících nástrojů. Program v programovacím jazyce Java, který je výstupem této práce umožňuje načíst elektronickou komunikaci ve formátu, který obsahuje již pouze zprávy, které mohou pocházet z více fyzických zdrojů. Jeho výstupem je pak soubor, který lze vizualizovat open-source nástrojem Gephi.

Testování implementovaných metod na reálné elektronické komunikaci společnosti Enron ukázalo, že implementované metody na zjištění hierarchie ve skupině označily jako nejvýznamnější jiné jedince, než ty kteří byli opravdu na formálně nejvyšších postech, na závěrech se ale shodly se na nich všechny metody společně. Z toho lze usuzovat, že algoritmy opravdu našly nejvýznamnější, přirozeně nejdominantnější jedince z pohledu proběhlé elektronické komunikace, kteří se ale liší od formálního vedení. Testování metod pro rozdělení skupiny na podskupiny ukázalo, že tyto metody byly poměrně úspěšné, vzhledem k tomu, že zaměstnanci společnosti Enron komunikovala téměř stejně intenzivně mezi podskupinami jako v uvnitř podskupin. Z testování tedy lze usoudit, že metody budou na reálných datech relativně úspěšné i přesto, že reálná data mohou být všelijaká a usnadní tak vyšetřovatelům práci s odkrýváním struktury organizované skupiny.

Předmětem další práce může být návrh systému, využívajících zde implemetovaných metod, který bude data sám rozpoznávat ve fyzických databázích, agregovat je a výsledky analýzy rovnou sám i vizualizovat.

Literatura

[1] Martin Mulazzani, E. W., Markus Huber. Social Network Forensics: Tapping the Data Pool of Social Networks. In *Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics, 2012*, 2012.

[2] Al-Zaidy, R. *Criminal Network Mining and Analysis for Forensic Investigations*. Bachelor thesis, Concordia University, Montreál, 2010.

[3] Al-Zaidy, R.; Fung, B. C. M.; Youssef, A. M.; et al. Mining criminal networks from unstructured text documents. *Digital Investigation*, volume 8, no. 3-4, 2012: pp. 147–160.

[4] Liu, B. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[5] ČESKO. Předpis č. 40/2009 Sb. Trestní zákoník ze dne 8. ledna 2009. In *Sbírka zákonů České republiky*, 2009, částka 11/2009, § 361.

[6] Rowe, R.; Creamer, G.; Hershkop, S.; et al. Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, New York, NY, USA: ACM, 2007, pp. 109–117.

[7] Wasserman, S.; Faust, K. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[8] Kolář, J. *Grafové algoritmy a základy teorie složitosti - lekce 1 (nepublikovaná přednáška)*. Praha, Fakulta informačních technologií ČVUT, 2014.

[9] Kolář, J. *Grafové algoritmy a základy teorie složitosti - souvislost, orientované grafy (nepublikovaná přednáška)*. Praha, Fakulta informačních technologií ČVUT, 2014.

[10] *Encyclopedia of Social Network Analysis and Mining*. New York: Springer Science+Business Media, 2014, ISBN 978-1-4614-6169-2.

[11] Kolář, J. *Grafové algoritmy a základy teorie složitosti - vlastnosti grafů (nepublikovaná přednáška)*. Praha, Fakulta informačních technologií ČVUT, 2014.

[12] Watts, D. J.; Strogatz, S. H. Collective dynamics of'small-world'networks. *Nature*, volume 393, no. 6684, 1998: pp. 409–10.

[13] Kalna, G.; Higham, D. J. A Clustering Coefficient for Weighted Networks, with Application to Gene Expression Data. *AI Commun.*, volume 20, no. 4, Dec. 2007: pp. 263–271, ISSN 0921-7126.

[14] Onnela, J. P.; Saramäki, J.; Kertész, J.; et al. Intensity and Coherence of Motifs in Weighted Complex Networks. *Phys. Rev. E*, volume 71, no. 6, June 2005.

[15] Spizzirri, L. Justifcation and Application of Eigenvector Centrality. 2011, [Cited 2015-03-23]. Available from: `https://www.math.washington.edu/~morrow/336_11/papers/leo.pdf`

[16] Brandes, U. Centrality Concepts and Methods. 2006, [Cited 2015-02-21]. Available from: `http://vw.indiana.edu/netsci06/ws-slides/ulrik_brandes.pdf`

[17] Piraveenan, M.; Prokopenko, M.; Hossain, L. Percolation Centrality: Quantifying Graph-Theoretic Impact of Nodes during Percolation in Networks. *PLoS ONE*, volume 8, no. 1, 01 2013: p. e53095.

[18] Freeman, L. C. Centrality in social networks conceptual clarification. *Social Networks*, 1978: p. 215.

[19] Granovetter, M. The Strength of Weak Ties. *The American Journal of Sociology*, volume 78, no. 6, 1973: pp. 1360–1380.

[20] Mascolo, C. Social and Technological Network Analysis Lecture 3: Centrality Measures. 2011, [Cited 2015-02-19]. Available from: `https://www.cl.cam.ac.uk/~cm542/teaching/2011/stna-pdfs/stna-lecture3.pdf`

[21] Newman, M. E. J. Analysis of weighted networks. *Phys. Rev. E*, volume 70, Nov 2004: p. 056131.

[22] Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; et al. The architecture of complex weighted networks. 2004, pp. 3747–3752.

[23] Opsahl, T.; Agneessens, F.; Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, volume 32, no. 3, 2010: pp. 245–251.

[24] Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, volume 103, no. 23, 2006: pp. 8577–8582.

[25] Alvarez-Hamelin, J. I.; Barrat, A.; Vespignani, A.; et al. k-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases. *NETWORKS AND HETEROGENEOUS MEDIA*, volume 3, no. 2, 2008: p. 371.

[26] Girvan, M.; Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, volume 99, no. 12, June 2002: pp. 7821–7826, ISSN 1091-6490.

[27] Hansen, D.; Shneiderman, B.; Smith, M. A. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2010, ISBN 0123822297, 9780123822291.

[28] Mission [online]. [cit. 2015-03-20]. Available from: `http://www.casos.cs.cmu.edu/mission/`

[29] CYRAM. *Unleashing the Hidden Power*. [cit. 2015-03-12]. Available from: `http://www.netminer.com/images/NetMiner4_E.pdf`

[30] NetworKit - Documentation [online]. [cit. 2015-02-05]. Available from: `https://networkit.iti.kit.edu/documentation/`

[31] NetworKit - Features [online]. [cit. 2015-02-05]. Available from: `https://networkit.iti.kit.edu/features/`

[32] Trier, M. Commetrix Trailer [online]. [cit. 2015-01-04]. Available from: `https://www.youtube.com/watch/?v=df9l8jL9ISs`

[33] Features [online]. [cit. 2015-03-26]. Available from: `http://gephi.github.io/features/`

[34] Diesner, J.; Frantz, T. L.; Carley, K. M. Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is No Different". *Comput. Math. Organ. Theory*, volume 11, no. 3, Oct. 2005: pp. 201–228, ISSN 1381-298X.

[35] Klimt, B.; Yang, Y. The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, 2004: pp. 217–226.

[36] Cohen, W. Enron Email Dataset [online]. [cit. 2015-02-27]. Available from: https://www.cs.cmu.edu/~./enron/

[37] Diesner, J.; Carley, K. M. Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, 2005: pp. 3–14. Available from: http://www.andrew.cmu.edu/user/jdiesner/publications/diesner_carley_siam_enron_03_05.pdf

[38] Shetty, J.; Adibi, J. The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report, University of Southern California*, 2004.

[39] Agarwal, A.; Omuya, A.; Harnly, A.; et al. A Comprehensive Gold Standard for the Enron Organizational Hierarchy. In *ACL (2)*, The Association for Computer Linguistics, 2012, ISBN 978-1-937284-25-1, pp. 161–165.

[40] CSV Format [online]. [cit. 2015-05-01]. Available from: http://gephi.github.io/users/supported-graph-formats/csv-format/

[41] Park, Y. [cit. 2015-03-21]. Available from: http://cis.jhu.edu/~parky/Enron/employees

[42] Spejchal, P. *Manažerská psychologie - podniková kultura (nepublikovaná přednáška)*. Praha, Fakulta elektrotechnická ČVUT, 2014.

[43] Jacomy, M.; Venturini, T.; Heymann, S.; et al. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, volume 9, no. 6, 06 2014.

Příloha A

Seznam použitých zkratek

SNA Social network analysis (analýza sociálních sítí)

HITS Hyperlink-induced topic search

CEO Chief executive officer

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `thesis`
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `src` ..... zdrojová forma práce ve formátu LaTeX
- `impl`
  - `example` ..... ukázkový vstup pro implementované metody a jim odpovídající výstup
  - `project` ..... implementované metody jako projekt v NetBeans
  - `jar` ..... implementované metody jako spustitelný jar soubor