VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

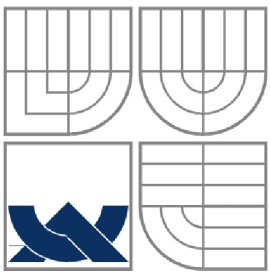
METODY ANALÝZY SOCIÁLNÍ SÍTĚ PRO DOLOVÁNÍ ZNALOSTÍ

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

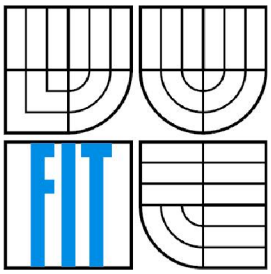
AUTOR PRÁCE
AUTHOR

TOMÁŠ MACHULKA

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

METODY ANALÝZY SOCIÁLNÍ SÍTĚ PRO DOLOVÁNÍ ZNALOSTÍ

METHODS OF SOCIAL NETWORK ANALYSIS FOR DATA MINING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

TOMÁŠ MACHULKA

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JAN SAMEK

BRNO 2011

Abstrakt

V této bakalářské práci jsou popsány některé z mnoha metod analýzy sociálních sítí, spolu s popisem a vizualizací těchto dat. Součástí práce je popis implementace aplikace provádějící tuto analýzu na základě několika implementovaných metod. Je zde popsán výstup analýzy a jeho srovnání s jiným softwarem pro analýzu sociálních sítí.

Abstract

This bachelor thesis describes some of many methods for social network analysis. There is also a description of the data visualization. The thesis contains description of implementation of aplication for social network analysis using several methods. Output of analysis is confronted with output of other software for social network analysis.

Klíčová slova

Sociální síť, Analýza, Teorie grafů, Dolování dat, JUNG, Java, Facebook.

Keywords

Social network, Analysis, Graph theory, Data mining, JUNG, Java, Facebook.

Citace

Tomáš Machulka: Metody analýzy sociální sítě pro dolování znalostí, bakalářská práce, Brno, FIT VUT v Brně, 2011

Metody analýzy sociální sítě pro dolování znalostí

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Jana Samka. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Tomáš Machulka
18. května 2011

Poděkování

Tímto bych chtěl poděkovat svému vedoucímu Ing. Janu Samkovi za cenné připomínky, ochotu a trpělivost.

Obsah

1 Úvod

Každý člověk se v průběhu svého života stává součástí rozmanitých sociálních skupin. Ať už se jedná o obyvatele města, spolužáky ve třídě, akademickou obec, firmu nebo v dnešní době stále se rozšiřující online komunity jako Twitter a Facebook, jedinec vytváří spletité vazby se svým okolím. Tyto vazby vytvářejí spojení, které nazýváme sociální síť. Sociální sítě a struktury davájí nepřeberné možnosti ke studiu chování jednotlivých komunit, jejich vazeb, síly vztahů a dalších aspektů na první pohled málo patrných nebo dokonce v mnoha případech skrytých. Analyzování sociální sítě je vědou sahající desítky let do minulosti ve formě sociologie. Současná doba výpočetní technologie nám dává nové možnosti ve studiu, rozboru a vizualizaci vybraného sociálního uskupení.

1.1 Historie analýzy sociálních sítí

Vývoj analýzy sociální sítě do současné doby utvářela řada různorodých prvků, které se navzájem prolínaly v historii. Z této složité historie nicméně můžeme vyčíst základní linii tří hlavních kořenů sociální analýzy. Práce Radlcifa–Browna čerpá z těchto tří základních zdrojů:

- Moreova sociometrie 30. let.
- Harvardská škola 30. a 40. let.
- Manchesterská škola 50. a 60. let.

Zakladatelem sociometrie byl psychiatr Jacob L. Moreno. Sociometrie se vyvíjela jako teorie lidského chování. Za nejmenší analytickou jednotku považoval Moreno sociální atom, což je systém socioemocionálních vztahů, které si vytváří individum vůči jiným a jiná individua vůči němu. Pro experimentální výzkum malých neformálních skupin vyvinul sociometrický test založený na rozkrytí struktury výběru a vzájemné přitažlivosti osob. Zajímaly ho toky vlivu mezi osobami nebo to, kdo komu může předávat informace. Vytvořil analytický diagram (sociogram, blíže specifikovaný v kapitole 2.3), v němž zobrazil individua jako body a jejich sociální vztahy jako čáry. Němeli bychom pozorovat individum, ale celek, který vytváří s prostředím. S touto teorií sociálního pole přišel Kurt Lewin.

Ve 30. a 40. letech se na Harvardské univerzitě ustavila skupina vědců se zájmem o to, jak lze ve větších sociálních uskupeních nalézat kliky, klastry a bloky. W. L. Warner a E. Mayo vedli v té době výzkum v továrnách na výrobu elektrických spotřebičů. Zkoumali vliv pracovních podmínek a vliv neformálních vztáhů na produktivitu práce a vytvořili první studii, kde byly použity sociogramy k popisu mezilidských vztahů v reálné pracovní situaci. Na jejich základě a na základě pozorování pak byli dělníci rozřazeni do souborů označených jako kliky (z anglického cliques). Členství v překrývajících se klikách a vztahy mezi nimi vytváří sítě vztahů, které strukturují celou

komunitu. Studie prokázala, že hned po rodině jsou kliky důležitým referenčním rámcem sebezařazení v komunitě. Ve stejném období kolem roku 1930 skupina německých emigrantů ovlivněná teorií *"gestalt"* Wolfganga Köhlera, pracovala ve Spojených státech amerických na kognitivní a sociální psychologii. Tato práce vedla ke značnému množství objasnění problematiky sociometrie a dynamiky skupiny. Použitím laboratorních metod a případových studií pozorovali strukturu skupiny a její tok informací a myšlenek prostřednictvím těchto skupin.

V roce 1960 dochází ke konečnému průlomu ve vývoji metodiky analýzy sociálních sítí. Na Harvardské univerzitě začal Harrison White rozšiřovat vyšetřování matematického základu sociální struktury. Díky některým klíčovým postřehům jeho severoamerických předchůdců vytváří unikatní syntézu, který byla později rozvinuta a rozšířena jeho studenty.

V současné době je analýza sociálních sítí značně matematizovaná, data jsou zpracovávána počítačovými programy. Nejsložitější analýzy se zabývají formalizovanými dynamyckými modely procesu vzniku sociálních sítí, a to i na bázi simulovaných dat. V kapitole 1.1 bylo čerpáno ze zdrojů [1], [2] a [3].

1.2 Struktura práce

Práce je koncipována jako ucelený náhled na problematiku dolování znalostí ze sociální sítě a analýzu sociální sítě jako takovou. Jsou zde uvedeny pojmy užívané v této oblasti spolu s rozdělením sociální sítě do specifických struktur pomocí používaných metod analýzy. Součástí práce je popis implementace programu pro analyzování dat sociální sítě a praktické použití analyzačních metod v reálném životě a světě kolem nás. V závěrečných kapitolách je popsán výstup těchto metod a jejich srovnání s jiným analytickým softwarem.

2 Sociální sítě

Sociální síť je struktura, kterou tvoří individua (nebo organizace) nazývané uzly, které jsou vázány (spojeny) prostřednictvím jednoho nebo více typů vzájemné závislosti jako je přátelství, příbuzenství, společné zájmy, finanční výměny, odpor, víra, znalosti apod. [4]

Podle Jamese C. Mitchella [1] konstituují sociální síť dva ideální typy jednání. Vedle instrumentálního jednání (transfer služeb a statků) je to komunikace jako zprostředkovatelka informací, která ustavuje sociální uskupení vztahů zahrnující jak komunity, tak organizace. Pro výzkum je nutno abstrahovat část totální sítě, parciální síť. Tu lze vyčlenit buď na základě obecně definovaných znaků vazeb jedince (politické kontakty, pracovní vztahy, příbuzenské závazky), nebo je parciální síť výsekem egocentrovaných sítí sociálních vztahů jakéhokoliv druhu. [1]

2.1 Pojmy v sociální síti

Mezi základní pojmy v sociálních sítích řadíme tyto: aktér, relační vazba, dyáda, triáda, podskupina, skupina, vztah a síť. V následujících podkapitolách, které čerpají z [5], si definujeme některé z těchto klíčových pojmů.

2.1.1 Aktér

Analýza sociální sítě se zabývá vazbami mezi subjekty a sociálními důsledky těchto vazeb. Tyto sociální entity jsou označovány jako aktéři. Aktéři jsou diskrétní individua, korporace nebo kolektivní sociální jednotka. Příkladem aktérů mohou být lidé ve skupině, oddělení v rámci společnosti nebo národy světa. Naše použití termínu aktér neznamená, že tyto subjekty mají nutně vůli nebo schopnost jednat. Kolekce subjektů, které jsou všechny stejného typu, nazýváme metoda jedné sítě. [5]

2.1.2 Relační vazba

Aktéři jsou spojeni navzájem sociální vazbou, jejíž rozlehlost a typ může být velmi rozsáhlý. Určujícím rysem vazby je, že vytváří spojení mezi dvěma aktéry. Běžným příkladem vazeb používaných v sítích může být hodnocení jedné osoby druhou (kamarádství, sympatie, respekt), přidružení nebo příslušnost (ke stejnému klubu), behavioralní interakce, pohyb mezi místy, formální vztahy a biologické vztahy. [5]

2.1.3 Dyáda

Na nejzákladnější úrovni je pouto nebo vazba ustanovena mezi dvěma aktéry. Toto spojení svou podstatou patří páru. Nevztahuje se tedy pouze na aktéra jako jednotlivce. Mnoho druhů analýzy sítí se zabývá pochopením vazeb mezi páry. Všechny přístupy používají dyádu jako jednotku analýzy. Dyadické analýzy se zaměřují na vlastnosti párových vztahů, například zda je vazba reciprocitní či nikoliv nebo zda mají určité specifické, vícenásobné vazby tendenci vyskytovat se společně. [5]

2.1.4 Triáda

Studovány jsou také vztahy mezi vetší podmnožinou aktérů. Mnoho analýz triád se zabýva takzvanou bilanční teorií, zejména zda jsou trojice tranzitivní (jestliže má aktér i v oblibě aktéra j a zároveň aktér j zase aktéra k, potom bude mít aktér i v oblibě aktéra k) a zda je trojice vyrovnaná (jestliže se aktér i a j mají v oblibě navzájem, potom budou mít i a j podobný vztah vůči třetímu aktérovi k a naopak, pokud se nemají v oblibě, může být jejich vztah vůči k odlišný). [5]

2.1.5 Podskupina

Dyády jsou dvojice aktérů a jejich související vazby, stejně tak triády jsou trojice aktérů a jejich vazby. Z toho vyplývá, že můžeme definovat podskupinu subjektů jako nějakou podmnožinu aktérů spolu se všemi vazbami mezi nimi. Lokalizace a studium podskupiny pomocí specifických kritérií je důležitým zájmem v oblasti analýzy. [5]

2.1.6 Skupina

Síťová analýza se nezabývá pouze kolekcí dyád, triád a podskupin. Do značné míry síla síťové analýzy spočívá ve schopnosti modelovat vztah mezi systémy aktérů. Tyto systémy se skládají z vazeb mezi členy některé (více či méně ohraničené) skupiny. Sociální vědci dali široký rozsah definic pojmu skupina. Pro naše účely budeme skupinu chápat jako kolekci všech aktérů, jejichž vazby jsou měřitelné. Musíme být schopni získat teoretické, empirické nebo koncepční kritéria o aktérech, kteří ve skupině patří k sobě a tvoří méně či více ohraničené soubory. Skupina se pak skládá z konečného počtu aktérů, kteří jsou považováni za konečný soubor jednotlivců. Omezení na konečný soubor aktérů je analytický požadavek. Modelování konečných skupin představuje problematickou otázku síťové analýzy zahrnující například určení a specifikaci hranice sítě, vzorkování a definice (vymezení) skupiny. [5]

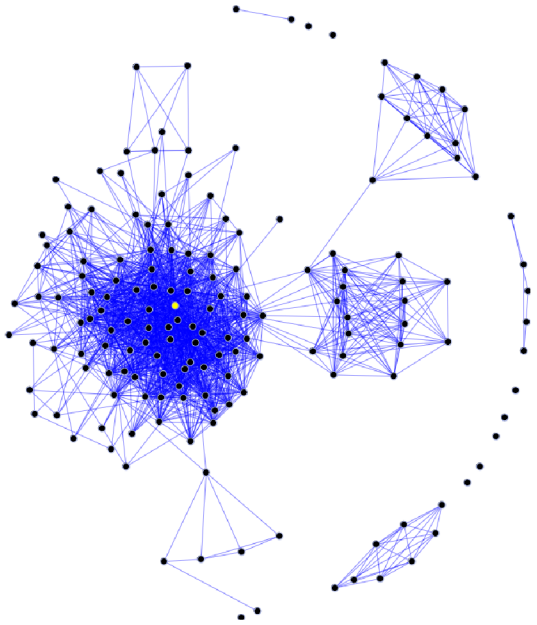
2.1.7 Relace

Jedná se o vazby specifického druhu mezi členy skupiny. Pro příklad soubor formálních diplomatických styků udržovaných mezi páry světových národů. Přítomnost relačního údaje je rozhodující a určující rys sociální sítě. [5]

2.2 Analýza sociální sítě

Analýza sociální sítě se zabývá studiem vztahů mezi jednotlivci, včetně analýzy sociálních struktur, sociálního postavení a mnoha dalšími. Vztahy mezi příbuznými, přáteli, sousedy atd. jsou prezentovány jako síť. Tradiční sociální vědy vycházejí z výzkumu dotazníků, ve kterých respondetni poskytují detaily o interakcích se svým okolím. Poté může být výsledná síť reprezentována jako uzly (respondenti) a vazby (interakce) mezi nimi. [6]

K analýze sociálních sítí slouží relační data (tj. konstanty, vazby a spojení dvou jedinců) uspořádaná do sociogramů a matic. V prvních sociometrických výzkumech výběru přátel používal J. Moreno [1] dotazníky a překročil tak omezení výzkumu skupin pozorováním. Základní sociometrická matice v sobě obsahuje zjištění, s kým je (nebo se považuje být) každý jednotlivý z n aktérů spojen nebo nespojen vazbou. Teorie grafů, která používá algoritmy, umožňuje z výpočtů údajů v maticích stanovit obecné charakteristiky sociální sítě (konexi, vzdálenost), soubory homologických pozic (kliky, třídy ve stejné strukturální úrovni) nebo odhalit zaznamenáníhodné body v síti (centralitu, mosty). Teorie grafů sestává z matematických axiómů a formulí, které popisují vlastnosti modelu utvořeného spojnicemi mezi body. Spojnice mohou indikovat, zda je vztah pozitivní či ne (např. + pro „má rád“), mohou mít směr (A považuje za dobrého přítele B, ale B osobu A ne). Komplexní sociální struktura je složena ze vzájemně se překrývajících triád, které kondenzují poznatky o vztazích. Jednoduché struktury triád tvoří bloky. Sociální síť ve stavu rovnováhy je taková, kde vztahy v podskupinách sítě jsou pozitivní, ale vztahy mezi jednotlivými podskupinami negativní; anebo taková, která sestává ze dvou vzájemně protichůdných soudržných podskupin [1].



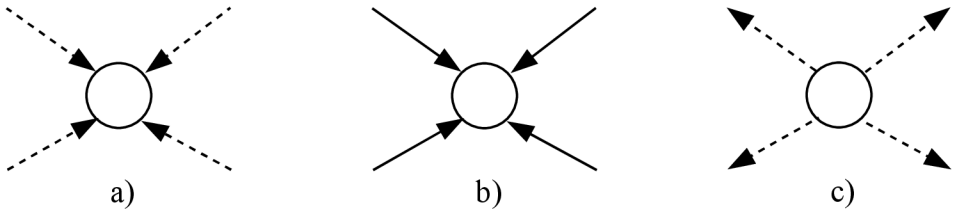
Obrázek 2.1: Ukázka diagramu sociální sítě [4].

2.3 Sociogram

Základem pro grafické znázornění vazeb a vztahů vycházejících z odpovědí získaných sociologickým testem je tzv. sociogram (mapa neformálních vztahů ve skupině). Sociogram vyjadřuje sociometrickou strukturu skupiny podle daného kritéria. Jaroslav Čihovký ve své publikaci [14] rozlišuje několik typů sociogramů. Individuální, kruhový, osový, terčový, topologický a strukturální.

Některé užívané typy individuální pozice jednotlivce ve skupině jsou uvedeny ve [14]. Jako příklad uveďme:

- Černá ovce: Jedinec získal především negativní volby (Obrázek 2.2–a).
- Hvězda: Jedinec získal především pozitivní volby (Obrázek 2.2–b).
- Ježek: Vyjadřuje zejména negativní volby k ostatním (Obrázek 2.2–c).



Obrázek 2.2: Individuální pozice jednotlivce [14].

Individuální sociogram neznázorňuje strukturu skupiny, ale jen sociogram pro každého jednotlivce.

Nejčastěji používaným typem sociogramu je kruhový sociogram. Členové jsou rozmísněni po obvodu kružnice.

Terčový sociogram vyjadřuje míru neformální autority vzdáleností od středu. Čím má jedinec více pozitivních voleb, tím blíže je ke středu kružnice (terče). Izoláti, kteří nezískali žádnou volbu, jsou mimo kružnici, případně na nulové kružnici. Analogicky lze sestavit terčový sociogram negativních voleb. Nejsložitějším sociogramem je strukturální. Jeho konstrukce je více popsána v [14].

2.4 Teorie grafů

Mnoho charakteristik sociálních sítí vychází právě z matematického odvětví teorie grafů. Pojem síť zde budeme chápat jako neformální označení pro objekt složený z prvků a interakcí, nebo spojení mezi těmito prvky. Například Internet jako síť složenou z uzlů (routery, hosté) a propojení mezi těmito uzly (kabely). V kapitole čerpáme ze zdoje [7].

Graf $G=(V,E)$ je abstraktní objekt tvořený množinou vrcholů (uzlů) V a množinou hran E, které propojují jednotlivé dvojice vrcholů. Množinu všech vrcholů grafu G označujeme $G(V)$, množinu všech hran grafu G jako $G(E)$. Kardinalita uzlů V je obvykle označována n, kardinalita hrany jako m. Jestliže jsou dva uzly spojeny hranou, jsou si přilehlé a nazýváme je sousední uzly. Graf může být neorientovaný nebo orientovaný. U neorientovaného grafu není pořadí koncových vrcholů hrany uvažováno. Neorientovaná hrana spojující vrcholy $u, v \in V$ je označována $\{u,v\}$. U orientovaného grafu má každá orientovaná hrana svůj počátek a cíl. Hranu s počátkem $u \in V$ a cílem $v \in V$ reprezentujeme jako pár (u,v). Zkrácená notace hran $\{u,v\}$ nebo (u,v) může být vyjádřena jako uv. V orientovaném grafu je zkrácený zápis uv pouze pro (u,v), zatímco u neorientovaného grafu oba zápisy uv, nebo vu odpovídají $\{u,v\}$. Graf, který obsahuje orientované a neorientované hrany zároveň, označujeme jako smíšený. V následujících podkapitolách jsou popsány některé speciální podoby grafů.

2.4.1 Multigrafy

V obou grafech, orientovaných a neorientovaných, je umožněno, aby soubor hran E obsahoval totožnou hranu několikanásobně. Jestliže se hrana vyskytuje v E, pak její kopie nazýváme paralelní hrany. Grafy s paralelními hranami nazýváme multigrafy. Pokud se naopak hrana v E vyskytuje právě jedenkrát, nemá tedy paralelní hrany, nazýváme tento graf prostý.

Hrana spojující vrchol se sebou samým (hrana, jejíž koncové uzly jsou identické) se nazývá smyčka (z anglického *loop*).

2.4.2 Ohodnocený graf

Často je velmi užitečné asociovat numerickou hodnotu (váhu) s hranou nebo uzlem grafu $G=(V,E)$. My budeme uvažovat pouze váhové ohodnocení hrany. Váha hrany může být reprezentována jako funkce $\omega: E \to \mathbb{R}$, která určuje každou hranu $e \in E$ a váhu $\omega(e)$. Podle kontextu může váha hrany reprezentovat různé vlastnosti jako je cena, vzdálenost, kapacita, síla interakce nebo podobnost. V mnoha případech je použit neohodnocený graf. Ten je ekvivalentem ohodnoceného grafu, kde jsou hrany $\omega(e)=1$ pro všechny $e \in E$.

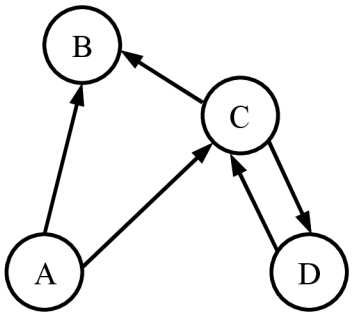
V analýze sociálních sítí je možnost ohodnocení hran velmi důležitou vlastností. Můžeme tak vyjádřit například sílu vztahu mezi dvěma aktéry. A to jak kladnou, tak zápornou hodnotou. Zde si jako příklad uveďme sociometrický výzkum (formou dotazníku), kdy respondent hodnotí vztah k okolním aktérům na stupnici od 1–5, kdy stupeň 5 znamená velmi kladný vztah. Tento údaj tedy bereme jako váhu hrany.

2.4.3 Stupeň uzlu

Stupeň (z anglického degree) uzlu v v neorientovaném grafu $G=(V,E)$ označovaný jako $d(v)$ je počet hran v E, které mají v jako koncový uzel. Jestliže G je multigraf, paralelní hrany se počítají podle jejich množství v E. Soubor hran, které mají v jako koncový uzel, označujeme jako $\Gamma(v)$. Soubor sousedů v je označován jako $N(v)$. V orientovaném grafu $G=(V,E)$ je pojem *out-degree* $v \in V$ označovaný jako $d^+(v)$ a určuje počet hran v E které, mají počátek v. *In-degree* $v \in V$ označovaný jako $d^-(v)$ určuje počet hran s destinací v. Tyto pojmy můžeme do češtiny přeložit jako vstupní stupeň pro in-degree a výstupní stupěň pro out-degree. Pro ohodnocené grafy jsou všechny tyto notace zobecněny součtem vah hran spíše než počtem hran. Soubor hran s počátkem v značíme $\Gamma^+(v)$, soubor hran s destinací v jako $\Gamma^-(v)$. Soubor destinací hran v $\Gamma^+(v)$ je značen $N^+(v)$, soubor počátků hran v $\Gamma^-(v)$ jako $N^-(v)$.

Průměrný stupeň značíme jako $\bar{d}(G)=\frac{1}{|V|}\sum_{(v \in V)} d(v)$. Neorientovaný graf, který má všechny vrcholy stejného stupně, nazýváme regulární a r–regulární, pokud je stupeň roven r. [8] Příklad demonstrující stupeň uzlu je na znázorněn na nasledujícím obrázku 2.3.

Vrchol	OUT (d^+)	IN (d^-)
A	2	0
B	0	2
C	2	2
D	1	1



Obrázek 2.3: Ukázka stupně uzlu.

2.4.4 Podgraf

Graf $G'=(V',E')$ je podgrafem grafu $G=(V,E)$ pokud $V'\subseteq V$ a $E'\subseteq E$. O uzlem indukovaný podgraf se jedná, pokud E' obsahuje všechny uzly $e\in E$, které spojují uzly v V'. Indukovaný podgraf se souborem vrcholů $V'\subseteq V$ označíme jako $G[V']$. Hranou indukovaný podgraf se souborem hran $E'\subseteq E$ značený $G[E']$ je podgraf $G'=(V',E')$ grafu G, kde V', je soubor všech vrcholů v V,které jsou koncovými uzly poslední hrany v E'. Jestliže C je podmnožina V, potom $G-C$ označuje graf získaný z G odstraněním všech vrcholů v C a jejich incidenčních hran. Pokud F je podmnožinou E, potom $G-F$ označuje graf získaný z G odstraněním všech hran v F. Jinými slovy můžeme říct, že soubor vrcholů podgrafu G' z G je podmnožinou souboru vrcholů G. Zároveň platí, že dva vrcholy jsou v G' přilehlé vždy, když jsou přilehlé v G. [10]

2.5 Data v sociální síti

Konvenční sociologická data obsahují pole naměřených (získaných) hodnot. Řádky tohoto pole vyjadřují případy, subjekty nebo náhledy. Sloupce obsahují záznamy (kvantitativní nebo kvalitativní) atributů, proměnných nebo míry. Každá buňka tohoto pole poté popisuje hodnocení aktéra. Základní datová struktura nás vede k porovnání vzájemné (ne)podobnosti jednotlivých aktérů a jejich atributů napříč celým polem a to po řádcích. Porovnáním jednotlivých sloupců naopak získáme (ne)podobnost proměnných mezi sebou.

Jméno	Pohlaví	Věk	Stupeň
Bob	♀	32	2
Carol	♀	27	1
Ted	♂	29	1
Alice	♀	28	3

Tabulka 2.1: Data s atributy.

Data sítě v jejich nejzákladnější podobě jsou reprezentována čtvercovým polem, kde řádky představují subjekty a sloupce vztahy mezi aktéry. Zjednodušeně kdo „má rád“ koho, což zobrazuje následující tabulka.

	Volba:			
Volič:	Bob	Carol	Ted	Alice
Bob	X	0	1	1
Carol	1	X	0	1
Ted	0	1	X	1
Alice	1	0	0	X

Tabulka 2.2: Vztahy mezi aktéry.

Na tyto data můžeme nahlížet podobně jako na data s atributy v předchozím případě. Porovnáním jednotlivých řádků pole vidíme, kteří aktéři jsou si podobní s jinými a koho zvolili. Pohledem na sloupce naopak vidíme podobnost z hlediska volby jiného aktéra. Získáváme tak užitečný náhled na data–pozici aktéra v sociální síti. Toto je jeden z cílů analýzy sociální sítě. Lokalizovat aktéra v celé síti. Dalším může být pohled na strukturu dat jako na celek či analytické porovnání dat nad a pod diagonálou. Zjistíme tak reciprocitu vztahů (Bob zvolil Teda, zvolil také Ted Boba?) a podobně. Základní rozdíl mezi konvenčními a síťovými daty je, že konvenční data se zaměřují na aktéry a atributy a síťová data na aktéry a vztahy. Kapitola čerpá ze zdoje [8].

3 Struktura sociální sítě

Tato kapitola popisuje význačné struktury a uskupení uzlů, které díky analýze sociální sítě můžeme naleznout.

Diagramy sociálních sítí, které znázorňují vzory spojení mezi body, mohou být popsány těmito základními pojmy:

- Orientovaný graf znázorňuje šipkou orientovaná spojení od jednoho aktéra (zde bodu v grafu) k druhému a směr šipky indikuje oriantaci vztahu. Dva body spojené vazbou jsou považovány za přilehlé a tvoří sobě navzájem rozličně velké sousedství.
- Body mohou být spojeny různým způsobem, i zprostředkovaně. Sekvence vazeb–spojnic, které vedou a zprostředkovávají vztah jednoho bodu s druhým, se nazývá cesta a délka cesty se měří počtem vazeb, které tvoří spoj.
- K celkovému popisu studovaného grafu slouží hustota a centralita nebo také centralizace. Dalším pojmem může být hvězda. Tento pojem pochází ze sociometrie. Je to osoba ve skupině nejvíce oblíbená, jež je středem pozornosti. Centrální bod může mít lokální nebo globální charakter. Bod je centrální lokálně, jestliže je spojen větším počtem vztahů s jinými body v jeho nejbližším okolí. Bod centrální globálně je takový, který má strategickou pozici v celkové struktuře sítě a zároveň leží v krátké vzdálenosti od ostatních bodů. Body s více centrální povahou ve srovnání s těmi, se kterými jsou spojeny, jsou vrcholy. Mosty jsou pak centrální body, které spojují dva nebo více vrcholů [1].

Strukturovat sociální síť lze dle klik nebo kohézních podskupin (např. klastrů, komponent, kruhů) [1]. Výchozím bodem pro všechny tyto prvky, které budou popsány v následujících kapitolách, je idea podgrafu.

3.1 Komponenta

Komponenta (z anglického *component*) je izolovaná skupina spojených bodů v síti. Tyto jednotlivé body mohou být spojeny i zprostředkovaně. Jde o nejjednodušší variantu konceptu podgrafu. Podgraf je propojen, pokud jsou všechny jeho body propojeny prostřednictvím cesty a mohou dosáhnout jiného bodu průchodem jedné nebo více cest. Body zároveň nemají žádná jiná propojení mimo podgraf. Izolované body nemohou být připojeny k existující komponentě, pokud nemají žádné spojení s žádným členem této komponenty. Identifikace komponent se provádí například pomocí metody *Sněhová koule* popsané v kapitole 5.

Komponenta může být hledána jak v grafech orientovaných, tak neorientovaných. Jsou zde však podstatné rozdíly. V případě orientovaného grafu mohou být identifikovány dva rozdílné typy

komponent. Silná (z anglického *strong*) a slabá (z anglického *weak*). Silná komponenta je taková, kde jsou všechny cesty tvořící propojení zarovnány v plynulém řetězci beze změny směru. Všechny cesty nesplňující toto kritérium nebereme v úvahu. Komponenty v orientovaném grafu, které zohledňují pouze výskyt či absenci spojení, nazýváme slabé. Rozdíly mezi silnou a slabou komponentou, stejně tak, jako jejich detekce, samozřejmě odpadají v případě neorientovaného grafu. Zde používáme termín prostá komponenta. Výsledkem analýzy komponent je pohled na graf jako kompozici jedné nebo více komponent (prostých, silých, slabých) a počet izolovaných bodů. Hustota grafů ukazuje na dominanci jedné velké komponenty, speciálně tam, kde je analýza soustředěna na slabé nebo prosté komponenty [3].

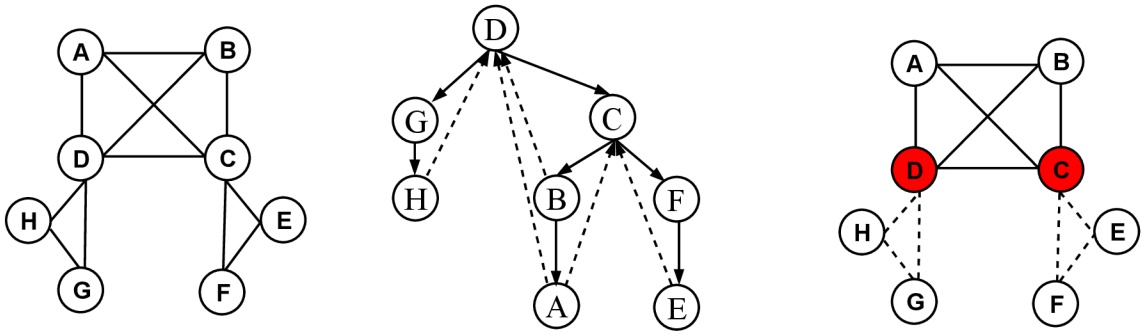
Problematiku, které uzly sítě zůstavají vždy spojeny v případě vypadnutí libovolného uzlu ze sítě, řeší výpočet obousměrně spojených, nebo také neoddělitelných komponent grafu, takzvaných bloků. Uvážíme–li algoritmus prohledávání do hloubky (z anglického *depth first search*, zkratka *DFS*) v neorientovaném spojeném grafu $G=(V,E)$, kde označíme navštívené uzly po sobě jdoucími čísly 1 až $n=|V|$ použitím předem seřazeného číslování *num,* můžeme pozorovat nalezení dvou typů hran. Jedny, které vedou k neoznačeným uzlům, nazýváme *stromové hrany* (*tree edges)*, druhé, které vedou k již objeveným a označeným uzlům, nazýváme *zpětné hrany* (*backward edges)*.

Pro každý uzel v ponecháváme nejmenší hodnotu označení z takového uzlu, který je dostupný prostřednictvím libovolné stromové hrany následované maximálně jednou zpětnou hranou, to jest nejmenší počet všech uzlů, které leží na některém cyklu s v. Kdykoliv je pomocí metody prohledávání do hloubky objeven nový uzel, záznam *low* tohoto uzlu je inicializován svým vlastním číslem. Pokud se navrátíme od předka k potomkovi w, to jest ze stromové hrany (v,w), aktualizujeme $low[v]$ při zachování minima potomkova záznamu $low[w]$ a stávající hodnoty $low[v]$. V případě nalezení zpětné hrany (v,w) aktualizujeme hodnotu $low[v]$. Uzel v je takzvaným *cut–vertex* v případě, že platí právě jedna z následujících podmínek:

1) Uzel v je kořenem stromu a je spojen s nejméně dvěma stromovými hranami.

2) Uzel v není kořenem, ale existuje potomek w tohoto uzlu takový, že platí $low[w]\geq num[v]$.

Důkaz platnosti těchto podmínek spolu s podrobnějším vysvětlením nalezneme v [7]. K nalezení obousměrně spojených komponent, bloků, vložíme každou novou hranu na zásobník. Kdykoliv je splněna podmínka $low[w]\geq num[v]$ po návratu z rekurze pro potomka w ulzu v, hrany na vrcholu zásobníku včetně hran (v,w) utvářejí další blok (a proto jsou ze zásobníku odtraněny). Příklad nalezení bloku můžeme vidět na následujícím obrázku.



Obrázek 3.1: Nalezení bloku v grafu.

Obrázek nalevo je vstupní neorientovaný graf. Uprostřed je rozklad na strom se znázorněním zpětných hran (čárkované hrany). Napravo je poté červeně vyznačen nalezený blok v grafu.

3.2 Cyklická komponenta

Koncept cyklické komponenty závisí na cyklu. Cyklus chápeme jako cestu, která se vrací do vlastního, počátečního bodu. Cyklus v grafu může být popsán délkou jako *3–cyklus*, *4–cyklus* až *k–cyklus*. Užitečným krokem k analýze cyklů je určení si maximální délky. Každý cyklus větší než námi určený, nebude uvažován [3].

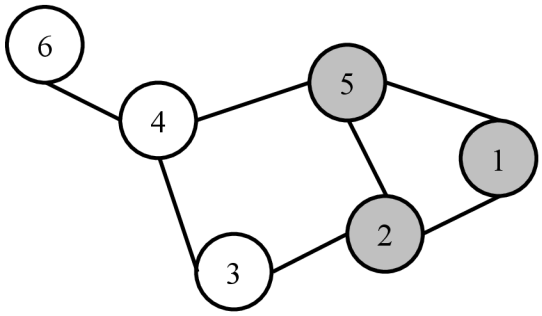
Zajímavým prvkem při rozboru cyklických komponent je takzvaný *most*. Most chápeme jako propojení, které samo o sobě neleží v cyklu, ale spojuje dva různé cykly. Cyklické komponenty lze definovat jako soubor protínajících se kruhů propojených cestami nebo body pro ně společnými. Separované cyklické komponenty se tedy nepřekrývají, ale mohou být spojeny jedním nebo více mosty. Identifikace takových komponent se provádí odstraněním cest označených jako mosty z grafu. Soubor bodů, které nám po odstranění všech přemostění zůstanou, jsou právě hledané cyklické komponenty. Kromě mostu můžeme u těchto komponent rozeznat další body k nim připojené, avšak přímo nepatřící do této komponenty. Rozdělujeme je do následujících tří kategorií:

- Věšáky (z anglického *hangers*) jsou prvky, které jsou spojeny s jedním členem cyklické komponenty, ale samy v cyklu neleží.
- Izolované stromy (z anglického *isolated trees*) jsou řetězce bodů (například dyád) spojených s jakoukoliv cyklickou komponentou. Členové tohoto stromu jsou však mezi sebou propojeny acyklicky.
- Kmeny (z anglického *isolates*) jsou ty body, které nemají žádné propojení. Mají stupeň uzlu 0.

Kapitola čerpá ze zdoje [3].

3.3 Klika

Klika (z anglického *clique*) je inkluzivní skupina entit, které sdílí zájmy, názory, cíle a vzory chování. Klika jako referenční skupina může být nominativní nebo komparativní. V teorii sociálních sítí, stejně jako v teorii grafů, je klika chápána jako podmnožina bodů (nebo uzlů), kde každé dva body jsou přímo spojeny–jsou tedy přilehlé. Vzájemně se překrývající kliky mohou být agregovány do kruhů [9].



Obrázek 3.2: Ukázka kliky.

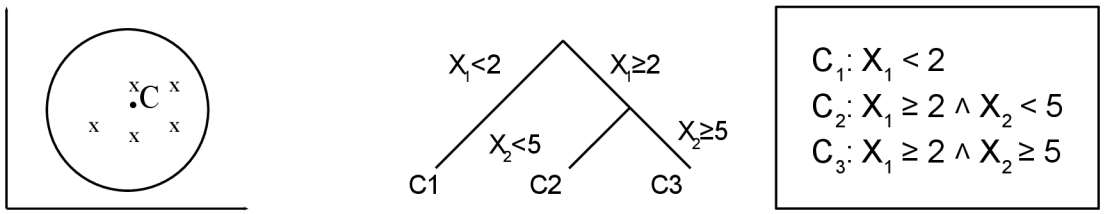
3.4 Klastr

Analýza klastrů nebo také shluků, je soubor metod pro automatickou klasifikaci vzorků. Tento vzorek jsme schopni rozdělit do skupin na základě podobnosti mezi těmito vzorky, v našem případě aktéry. Poté můžeme říct, že aktéři patřící do jedné skupiny jsou si podobní, a naopak aktéři patřící do skupiny jiné si podobní nejsou. Vstupem pro metodu analýzy klastrů je soubor vzorků a atributů. Porovnáváme podobnost atributů mezi dvěma vzorky. Výstupem je množina skupin, klastrů, které tvoří oddíl nebo strukturu oddílů. Vstup může být popsán jako uspořádaná dvojice (X, s) nebo (X, d), kde X je soubor vrozků a s, respektive d jsou hodnoty podobnosti, respektive nepodobnosti (vzdálenosti) mezi vzorky. Výstupem je rozdělení $\wedge = \{G_1, G_2, \ldots, G_N\}$, kde $G_k, k = 1, \ldots, N$ je konkrétní podmnožina taková, že $G_1 \cup G_2 \cup \ldots \cup G_N = X$, a zároveň $G_i \cap G_j = \phi, i \neq j$. Každý klastr může být popsán určitou charakteristikou.

Existuje několik schémat pro formální popis objevených klastrů:

- Reprezentace shluku bodů v n–rozměrném prostoru podle jejich těžiště nebo podle vzdáleností bodu v klastru (Obrázek 3.3–a).
- Grafická reprezentace shluků pomocí uzlů ve stromu klastrů (Obrázek 3.3–b).

- Reprezentace klastrů pomocí logických výrazů (Obrázek 3.3–c).



a) b) c)

Obrázek 3.3: Reprezentace shluků.

Následující tabulka uvádí pro lepší představu jako příklad vzorek devíti zákazníků, kde každému z nich náleží dva atributy. Počet položek, které zákazník nakoupil a celkovou cenu nákupu. [11]

Zákazník	Počet položek	Celková cena	
5	2	1700	klasr 1
7	3	2000	
8	4	2300	
6	10	1800	klasr 2
4	12	2100	
1	11	2500	
3	2	100	klastr 3
2	3	200	
9	3	350	

Tabulka 3.1: Příklad rozdělení vzorku do klastrů.

Výstupem po analýze tohoto příkladu může být rozdělení do celkem tří klastrů, kde v každém nalezneme mezi atributy určitou podobnost. Zákazníci patřící do klastru 1 nakoupí méně věcí za vysokou cenu. Za vysokou cenu nakupují i zákazníci v klastru 2, ale s rozdílem většího počtu nakoupených věcí. Do posledního klastru 3 patří zákazníci kupující méně věcí za nižší cenu.

Pro tuto metodu analýzy demonstrovanou v příkladu výše je tedy nejtěžší a nejdůležitější fází výběr takových atributů zkoumané sociální sítě, které nám po výstupu metody jasně definují hranice jednotlivých klastrů a podobnost mezi aktéry. Na principu podobnosti lze pak v *dendrogramech* seskupovat klastry do stejných hladin. Body (aktéři) se třídí do aglomerací podle toho, s kým sousedí, s přihlédnutím k tomu, zda jsou nositeli předem definovaných atributů. Vedle popisu vzorů uspořádání přímých a nepřímých vazeb lze i detailněji analyzovat vztahy mezi aktéry ve stejných

pozicích. Problém srovnatelnosti sociálních pozic těch, kteří jsou seskupeni do klik nebo klastrů, se řeší konceptem strukturální ekvivalence (dále popsané v kapitole 3.6.1). [1]

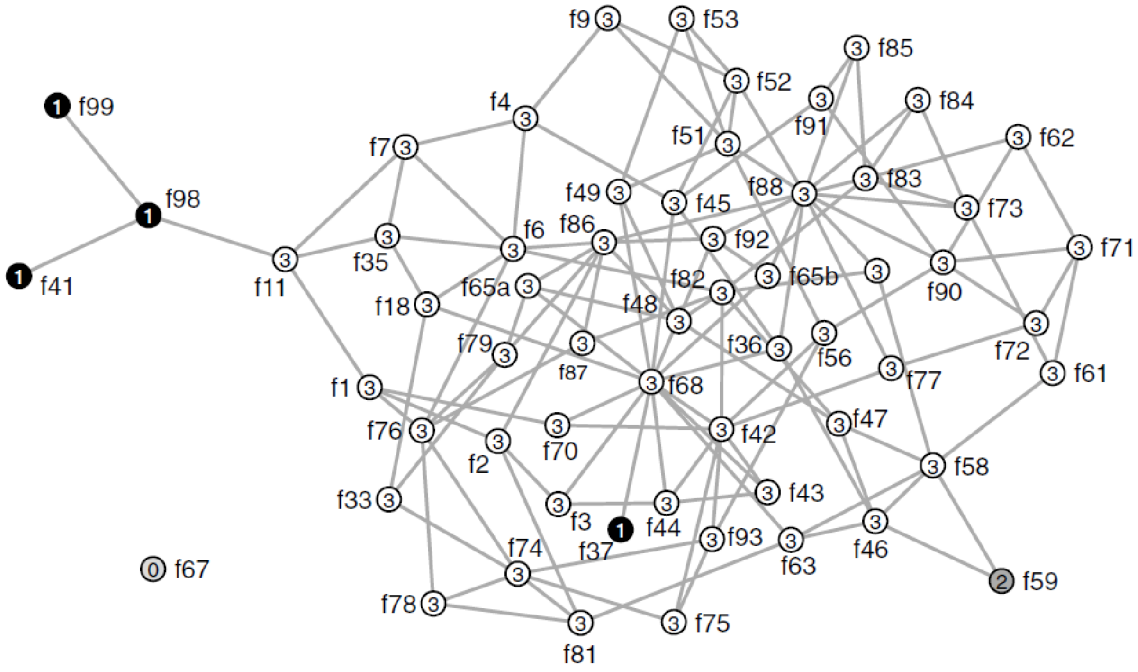
O celkové úrovni organizace sítě a o způsobech, jak jsou utvářeny sociální vazby, vypovídají triády. Triáda (graficky znázorněná jako trojúhelník) B–A–C sestává ze zřetězení dvojic vztahů mezi A–B a A–C, kde A je buď zprostředkovatelem (alespoň v počátku vztahu mezi B–C) vazby zbývajících dvou, nebo se B a C mohou znát nezávisle na A. Vazby, vliv a kontrola jsou zřetězeny [1].

3.5 Jádro

Jádra (z anglického *cores*) odhalují podle rozdělení stupně uzlu lokální koncetraci vazeb kolem těchto uzlů. V této sekci budeme používat stupeň uzlů k identifikaci shluku vrcholů, které jsou těsně propojeny. Každý vrchol má v rámci shluku konkrétní minimální stupeň. Nezajímáme se však o stupeň jednotlivých vrcholů, ale o stupeň všech vrcholů v rámci shluků. Těmto shlukům poté říkáme *k–jádra,* kde *k* značí minimální stupeň každého vrcholu v rámci jednoho shluku. Jak jsme si uvedli dříve, stupně uzlů rozdělujeme na in–degree a out–degree. U dekompozice jádra můžeme navíc uvažovat stupeň uzlu (id–degree) + (out–degree).

Příklad pro *k = 2.* Pak *k–járdo = 2–jádro* a obsahuje všechny vrcholy, které jsou v rámci shluku propojeny a mají stupeň 2 a více.

Jádra dokáží identifikovat relativně husté podsítě a pomáhají naleznout soudržné podskupiny. Jádro s maximálním stupněm nazýváme hlavní (*main core)*. Neorientovaná síť vztahů na obrázku 3.4 má jako hlavní jádro skupinu bílých uzlů se stupněm 3. Jedná se tedy o *3–jádro*. Každý uzel je spojen nejméně se třemi dalšími uzly. Dále v této síti můžeme naleznout *2–jádro* označené tmavě šedě, *1–jádro* označeno černě a *0–jádro* označeno světle šedě. Platí, že *k–jádra* jsou vnořené, to znamená, že uzel s *3–jádrem* je součástí *2–jádra*, avšak ne všechny uzly s *2–jádro* jsou součástí *3–jádra*.



Obrázek 3.4: Příklad jádra [13].

Algoritmus pro nalezení hierarchie jader může být následující: Pokud z daného grafu $G=(V, E)$ rekurzivně vymažeme všechny vrcholy a hrany k nim náležící, se stupněm vrcholu menším než k, poté je výsledný graf *k–járo*.

Pseudokód pro nalezení jádra (vstupem je graf $G=(V, E)$):

```
// compute the degree of vertices
// order the set of vertices V in increasing order of their degrees
for each v in V in order do begin
     core[v] := degree[v];
     for each u in Neighbors(v) do
          if degree[u] > degree[v] then begin
               degree[u] := degree[u] - 1;
               update the ordering of V
          end
end;
```

Označíme–li $n=|V|$ a $m=|E|$. Spočítání všech uzlů můžeme vyjádřit časovou náročností $O(m)$. Použitím jednoho seznamu speciálně pro uložení uzlů se stejným stupněm v kombinaci s tabulkou začátků takovéto části seznamu, můžeme řadit uzly V s časovou náročností $O(n)$.
Tato kapitola čerpá ze zdrojů [12] a [13].

3.6 Ekvivalence a podobnost v sociálních sítích

V analýze sociálních sítí se často rozlišují dva typy podobnosti. Jsou to strukturální ekvivalence a regulární ekvivalence. Oba tyto typy rozděluje hlavně stupeň jejich abstrakce. Strukturální je velmi konkrétní, oproti velmi abstrahované regulární ekvivalenci.

3.6.1 Strukturální ekvivalence

U strukturální ekvivalence říkáme, že dva aktéři jsou přesně strukturálně ekvivaletní, pokud mají stejný vztah ke všem ostatním aktérům. Tedy zjednodušeně řečeno, pokud má aktér A vazbu s B a aktér C má také vazbu s B, poté A a C jsou strukturálně ekvivaletní, nahraditelní jeden druhým. Neuvažuje se zde tedy skutečnost vazby mezi A a C, pokud vůbec nějaká existuje. Výskyt přesných strukturálně ekvivaletních aktérů je v sociálních sítích spíše ojedinělý, proto nás více zajímá zkoumání míry ekvivalence, než výskyt těchto přesných vzorů. [8]

3.6.2 Regulární ekvivalence

U regulární ekvivalence jsou dva uzly regulárně ekvivalentní, pokud mají stejný profil, neboli stejný obraz vazeb s množinou jiných aktérů, kteří jsou také regulárně ekvivaletní. Nejlépe lze tuto problematiku pochopit na příkladu „dvou matek“. Dvě matky jsou ekvivaletní, jelikož každá má podobný vzor vazeb s manželem, dětmi a příbuznými. Ve většině případů nemají dvě matky stejnou vazbu s jedním manželem nebo dětmi, nejsou tedy strukturálně ekvivalentní. Množiny regulárních ekvivalencí popisují sociální role nebo typy, které jsou základními stavebními bloky všech sociálních struktur. Platí, že aktéři strukturálně ekvivaletní jsou nutně regulárně ekvivaletní. Aktéři regulárně ekvivaletní ale nutně strukturálně ekvivaletní nejsou. [8]

4 Metody analýzy sociální sítě

Existuje několik stategií pro volbu jak postupovat ve sběru vzorků a vztahů mezi nimi. *Kompletní metoda*, popsaná dále, poskytuje maximum informací, ale nemůže být z důvodu náročnosti prováděna. Další přístupy poskytují sice méně informací o struktuře sítě, ale často umožňují snazší náhled na vzorek rozsáhlejších dat. V této kapitole jsou uvedeny některé používané metody pro analýzu sítí, které utvářejí z prvotní množiny sociálních dat ucelenější skupiny a podkupiny.

4.1 Kompletní metoda

Kompletní metoda (z anglického *Full Network Method*) vyžaduje sběr informací o každém vazbě aktéra se všemi ostatními. V podstatě jde o sčítání vazeb v populaci aktérů, spíše než vzorku. Příkladem může být tok e–mailů mezi všemi zaměstnanci společnosti. Jelikož shromažďujeme informace o vazbách mezi všemi páry nebo dyádami, dostáváme kompletní obraz vztahů v populaci. Kompletní metoda poskytuje velmi silný nástroj pro analýzu sociální struktury. Avšak je velmi drahá a náročná na sběr dat. Většina osob, skupin a organizací má spíše omezené množství vazeb, nebo alespoň omezený počet silných vazeb. Je to pravděpodobně proto, že sociální aktéři mají omezené zdroje a čas, proto nemohou udržet velké množství silných vazeb. [6]

4.2 Metoda Sněhová koule

Metoda Sněhová koule (z anglického *Snowball method*) probíhá dotazováním aktéra na jeho vazby s ostatními aktéry. Dále jsou všichni jmenovaní (vystopovaní) aktéři opět rekurzivně dotazováni. Celý proces ukončujeme po identifikaci všech potřebných entit, nebo v případě, že se rozhodneme skončit (často z časových důvodů nebo když vystopovaní aktéři přesahují hranice zkoumané skupiny). Metoda Sněhové koule je zejména vhodná pro stopování „speciální“ populace (často numericky malé podmnožiny, která je součástí větší skupiny) jako například spojenci, příbuzní a mnohé další skupiny. Omezení počtu silných vazeb a tendence akterů vazby opětovat umožňují snadno naleznout hranice.
Exitují dvě hlavní omezení a nedostatky této metody:

- Izolované aktéry, kteří nemají žádná propojení nelze touto metodou nalézt. Izolované entity jsou však velmi důležitým rysem pro analytické účely.
- Neexistuje garantovaný způsob nalezení všech propojení.

Problém také nastává na začátku metody při rozhodování, kde začít. V případě začátku v nevhodném uzlu můžeme přeskočit celou podmnožinu aktérů.

V mnoha případech je nemožné použití metody Sněhové koule pro vystopování všech uzlů. Alternativním přístupem je začít se sekcí lokálních uzlů a identifikací s kým jsou propojeny. Poté určíme, které z uzlů zkoumaných v první fázi jsou spojeny s někým jiným. Tento způsob může být velmi efektivní při sběru dat velkých sociálních skupin. [6]

4.3 Detekce komunit

Aktéři v síti mají sklon formovat skupiny s blízkými vazbami. Skupiny jsou také nazývány komunity, klastry, soudržné podskupiny nebo moduly v různých kontextech. Zjednodušeně řečeno, interakce mezi individui probíhá častěji v rámci skupiny než mezi skupinami. Detekce kohezivních skupin a sociálních sítí zůstává jádrem problému analýzy sociálních sítí.

Kritéria skupin se dělí do několika kategorií. My si uvedeme detekci skupiny spadající do kategorie metod založených na uzlu (z anglického *node–centric*). Podrobnosti o dalších skupinách a podskupinách nalezneme v [6].

Detekce komunity založená na kritériu uzlu vyžaduje, aby každý uzel ve skupině splňoval určité vlastnosti jako je vzájemnost (*mutality*), dosažitelnost (*reachability*) nebo stupeň (*degree*).

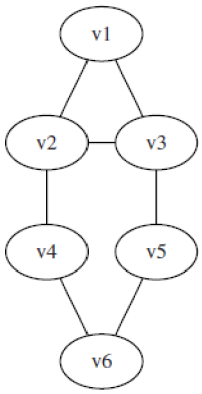
Skupiny založené na kompletní vzájemnosti. Ideálem soudržené skupiny je klika. Klikou rozumíme maximální úplný podgraf ze tří nebo více uzlů, z nichž všechny leží vedle sebe. Pro orientovaný graf se ukazuje velmi vysoká pravděpodobnost, že se zde budou vyskytovat úplné bipartity v komunitě. Tyto úplné bipartity fungují jako jádro komunity. Bipartitu označíme jako *(i, j)*, kde je v grafu každý *i* uzel spojen s jiným *j* uzlem.

Nelezení úplných klik v síti je takzvaně NP–těžké (z anglického *Non–deterministic Polynomial–time hard*). Dokonce nalezení přibližného řešení je velmi obtížné. Jeden z přístupů pro nalezení klik je použití hrubé síly (z anglického *brute–force*), tedy průchod přes všechny uzly v síti. Pro každý uzel se zkontroluje, zda existuje nějaká klika určité velikosti, která obsahuje uzel. Klika je poté uložena a uzel je odebrán. Tento přístup funguje pro malé sítě, avšak pro větší se stává nepraktický. Hlavní strategií tohoto problému je efektivně prořezávat tyto uzly a hrany, které pravděpodobně nejsou obsaženy v úplných klikách nebo kompletní bipartitě. K identifikaci úplých klik ve velkých sociálních sítích slouží algoritmus, kdy se po každém vzorkování nalezne podmnožina sítě. Na základě této podmnožiny může být klika nalezena pomocí metody nenasytného vyhledávání (z anglického *greedy–search*). Úplná klika nalezená v podmnožině (řekněme, že obsahuje *q* uzlů) slouží jako dolní mez pro prořezávání. To znamená, že úplná klika obsahuje alespoň *q* členů, takže uzly se stupněm nižším než *q* mohou být odstraněny. Tento proces prořezávání se opakuje, dokud není síť redukována na rozumnou velikost. Podobnou strategii lze využít pro nalezení bipartit zaměřené na nalezení úplných bipartit fixní velikosti–*(i, j)* bipartity. Iterativní prořezávání je

použito na odstranění uzlů s out–degree menším než *j* a in–degree menším než *i*. Po tomto počátečním omezení je prořezávací strategie použita na eliminaci seskupení nebo nalezení *(i, j)* bipartity.

Skupiny založené na dosažitelnosti. Tento typ komunity uvažuje dosažitelnost mezi aktéry. V krajním případě lze i dva uzly považovat za komunitu, pokud mezi nimi existuje cesta. Poté je každý z nich komunitou. Více úsilí musíme vynaložit pro nalezení komunity ve vetších sítich. Koncepčně by zde měla být krátká cesta mezi libovolnými dvěma uzly ve skupině. Uveďmě si tedy několik dobře známých struktur ze sociálních věd:

- *k–klika:* maximální podgraf, kde nejdelší vzdálenost mezi dvěma libovolnými uzly není větší než *k*. To znamená $d(i, j) \leq k \; \forall \, v_i, v_j \in V_s$
- *k–klan:* je k–klika, ve které není vzdálenot $d(i, j)$ mezi všemi uzly v podgrafu větší než k pro všechny cesty uvnitř podgrafu. K–klan musí být k–klika, ne však naopak.



cliques: $\{v_1, v_2, v_3\}$
2-cliques: $\{v_1, v_2, v_3, v_4, v_5\}, \{v_2, v_3, v_4, v_5, v_6\}$
2-clans: $\{v_2, v_3, v_4, v_5, v_6\}$
2-clubs: $\{v_1, v_2, v_3, v_4\}, \{v_1, v_2, v_3, v_5\}, \{v_2, v_3, v_4, v_5, v_6\}$

Obrázek 4.1: Klika, klan, klub [6].

- *k–klub:* omezuje délku cesty v rámci skupiny. Jedná se o maximální průměrnou spodní hranici *k*. Všechny k–klany jsou k–kliky, k–kluby jsou obvykle obsaženy v k–klice.

Skupiny založené na stupni uzlu: Vyžaduje subjekty v rámci skupiny, které přiléhají k relativně velkému počtu členů skupiny. Mezi dvě běžné studie patří:

- *k–ple*x: Jedná se o maximální podgraf obsahující n_s uzlů, v němž je každý uzel přilehlý k méně než $n_s - k$ uzlům v podgrafu. Jinými slovy, každý z uzlů nemusí mít žádné vazby až do *k* členů skupiny. K–plex se stane klikou když $k = 1$.

■ *k–jádro*: každý uzel (v_i) propojuje nejméně k členů obsažených ve skupině. $d_s(i) \geq k \forall v_i \epsilon V_s$

Kapitola čerpá ze zdroje [6].

4.4 Hierarchické shlukování

Při analýze hierarchie shluků jako součást vstupu nespecifikujeme předem očekávaný počet shluků. Vstupem do analýzy je dvojice (X, s), kde X je množina vzorků a s je míra podobnosti. Výstupem je množina shluků na různé úrovni, hierarchii. Mnoho procedur pro hierarchické shlukování není založeno na optimalizaci. Algoritmus této analýzy rozdělujeme do dvou kategorií. Rozkládací algoritmus a aglomerační algoritmus.

Rozkládací algoritmus začíná s množinou vzorků X a dělí je do oddílů podmnožin. Poté opět rozděluje tyto podmnožiny na menší množiny atd. Takto algoritmus generuje sekvenci oddílů, které jsou seřazeny od hrubšího po jemnější dělení.

Aglomerační algoritmus na začátku bere každý objekt jako počáteční shluk. Shluky jsou poté sloučeny do hrubších oddílů. Proces slučování probíhá do doby, něž jsou všechny objekty obsaženy v jednom společném shluku. Tento proces shlukování je typu zdola nahoru, kdy se oddíly s jemnějším dělením shlukují do hrubšího dělení.

Většina hierarchických aglomeračních algoritmů jsou varianty algoritmů prostého nebo úplného propojení. Ty se liší pouze v popisu podobnosti mezi dvojicí shluků. U metody prostého propojení je vzdálenost mezi vzorky minimem vzdáleností mezi všemi dvojicemi vzorků získanými ze dvou shluků. U algoritmu úplného propojení je vzdálenost mezi dvěma shluky maximem všech vzdáleností mezi dvojicemi ze dvou shluků. V obou případech jsou dva shluky sloučeny do shluku většího na základě kritéria minimální vzdálenosti. Pro obě varianty těchto algoritmů platí následující kroky:

– Vlož každý uzel do samostatého shluku. Sestroj seznam vzdáleností mezi shluky a dvojicemi rozdílných uzlů. Seřaď tento seznam vzestupně.

– Průchodem přes seřazený seznam vzdáleností získáme pro každou rozdílnou prahovou hodnotu d_k graf uzlů, kde dvojice uzlů bližší než d_k jsou propojeny do nového shluku hranou. V případě, že jsou všechny uzly členy propojeného grafu, skonči. Jinak opakuj tento krok.

– Výstupem algoritmu je vnořená hierarchie grafů, která může být rozdělena na požadované úrovně dle podobnosti.

Jako příklad můžeme uvést pět bodů $\{x_1, x_2, x_3, x_4, x_5\}$ ve dvoudimenzionálním prostoru. Souřadnice bodů jsou $x_1 = (0;2)$, $x_2 = (0;0)$, $x_3 = (1{,}5;0)$, $x_4 = (5;0)$ a $x_5 = (5;2)$.

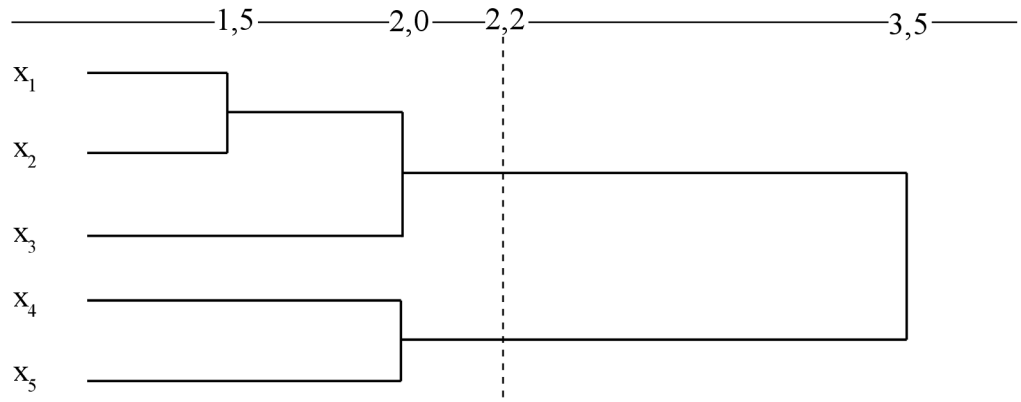
Vzdálenosti mezi těmito body v euklidovské metrice jsou:

Pro x_1 :	Pro x_2 :	Pro x_3 :	Pro x_4 :
$d(x_1, x_2)=2$	$d(x_2, x_3)=1,5$	$d(x_3, x_3)=3,5$	$d(x_4, x_5)=2$
$d(x_1, x_3)=2,5$	$d(x_2, x_4)=5$	$d(x_3, x_4)=4,03$	
$d(x_1, x_4)=5,39$	$d(x_2, x_5)=5,29$		
$d(x_1, x5)=5$			

Po prvním kroku iterace jsou vzdálenosti mezi body stejné pro oba algoritmy prostého i úplného propojení. V dalším kroku se již výpočet rozchází. Použitím aglomeračního algoritmu prostého propojení jsou provedeny následující kroky pro vytvoření shluků pro reprezentaci v dendogramu.

Nejprve jsou x_2 a x_3 sloučeny a shluk $\{x_2, x_3\}$ je vytvořen s minimální vzdáleností rovnou 1,5.

Ve druhém kroku jsou x_4 a x_5 sloučeny do nového shluku $\{x_4, x_5\}$ na vyšší úroveň 2,0. Ve stejnou dobu je minimální vzdálenost mezi shlukem $\{x_2, x_3\}$ a $\{x_1\}$ také rovna 2,0. Tyto dva shluky jsou tedy na stejné úrovni podobnosti jako x_4 a x_5. Zbylé dva shluky $\{x_1, x_2, x_3\}$ a $\{x_4, x_5\}$ jsou nakonec sloučeny na nejvyšší úroveň s minimální prostým propojením vzdáleností 3,5. Výsledný dendogram poté vypadá následovně.



Obrázek 4.2: Ukázka dendogramu pro prosté propojení [11].

4.5 Algoritmus Chameleon

Chameleon je shlukovací algoritmus, který se snaží vylepšit kvalitu shlukování na základě komplikovanějších kritérií při slučování dvou částí. Dvě části budou sloučeny, pokud je vzájemné propojení a těsnost sloučených shluků velmi podobné vzájemnému propojení dvou individuálních shluků ještě před sloučením.

Tento algoritmus nejdříve inicializuje graf $G=(V, E)$, kde každý uzel $v \in V$ reprezentuje vzorek dat. Dále pak váhovou hranu $e(v_i, v_j)$, která představuje spojení mezi uzly v_i a v_j. Váha

každé hrany v grafu reprezentuje přilehlost mezi dvěma uzly. Algoritmus dále využívá metody dělení grafu pomocí minimálních řezů k rekurzivnímu rozdělení grafu do mnoha malých, nespojených podgrafů. Minimální řez probíhá na každém stupni rekurze a vztahuje se k rozdělení grafu do dvou částí s blízkou, stejnou velikostí takovou, že celková váha řezaných hran je minimální. Každý podgraf je poté považován za počáteční podshluk. Algoritmus se opakuje dokud není dosaženo požadovaného kritéria. Druhá fáze algoritmu probíhá zdola nahoru. Algoritmus určí podobnost mezi každým párem z elementárních shluků C_i a C_j dle jejich relativní vzájemné propojenosti $RI(C_i, C_j)$ a jejich relativní přilehlosti $RC(C_i, C_j)$. Vzhledem k tomu, že propojení shluků je definováno jako celková váha hran, které jsou odebrány při metodě minimálího řezu je relativní propojení $RI(C_i, C_j)$ definováno jako poměr mezi propojeností sloučeného shluku C_i a C_j ku průměrné propojitelnosti C_i a C_j. Nápodobně, je relativní přilehlost $RC(C_i, C_j)$ definována jako poměr mezi přilehlostí sloučených shluků C_i a C_j ku průměrné vnitřní přilehlosti C_i a C_j. Zde se těsnost shluku odkazuje na průměrnou váhu hran, které jsou odebrány při min–cutu na shluku. Zde se těsnost shluku odkazuje na průměrnou váhu hran, které jsou odebrány při metodě minimálního řezu prováděné nad shlukem. Funkce podobnosti je poté součinem $RC(C_i, C_j) * RI(C_i, C_j)^{\alpha}$, kde α je parametr mezi 0 a 1. Algoritmus získal své jméno díky adaptaci na vnitřní charakteristiku shluků. Je účinný při detekci libovolně formovaných shluků různé hustoty, nicméně není vhodný pro rozsáhlá data, jelikož je jeho časová náročnost pro n prvků rovna $O(n^2)$.

Kapitoly 5.5 a 5.6 čerpají ze zdroje [11]. Jsou zde uvedeny také další užívané metody hledání shluků jako je rozdělovací a inkrementální algoritmus.

5 Analýza sociálních sítí v praxi

V současné době je pojem analýzy sociálních sítí spojován hlavně s oblastí online komunit pro získání zajímavých informací převážne pro marketingové využití. Odhalení specifických subkomunit, jejich vlastností a chování, napomáhá marketingovým agenturám k lepšímu zaměření reklamy na určitou cílovou skupinu. Analýza sociálních sítí však není omezena pouze na populaci jako takovou. Principy aplikované na sociální struktury lze aplikovat i na sítě jiného charakteru. Naříklad světovou burzu, vztahy států apod. Zajímavou analýzou je práce Jany Bakalové[1], která se věnuje analýze sociální sítě hudebních skladatelů. Cílem bylo zjistit, jaké jsou vztahy mezi skladateli, jejichž díla se v současné době hrají a jak se vzájemně kombinují. Existují také sociální analýzy zaměřené na topologii a provázanost teroristických buněk či jiných potenciálně nebezpečných komunit.

5.1 Software pro analýzu sociálních sítí

V oblasti sociální analýzy existuje široká škála softwaru od komerečních aplikací po volně šířitelné distribuce. Díky této široké škále programů máme k dispozici nástroje zaměřené převážně na vizualizaci dat, jako jsou například *NetDraw,* ale také sofistikovanější aplikace s pokročilými analyzačními nástroji, jako například *NetMiner*. Na aplikace můžeme nahlížet z hlediska následujících vlastností:

- Datový vstup, formát a manipulace s daty.
- Vizualizační techniky.
- Postupy při analýze sociální sítě, které rozdělujeme do těchto kategorií:
    a) Statictické metody (jako je centralita, tranzitivita).
    b) Postupy založené na komplexních algoritmech (shlukování apod.).
    c) Statictické modelování založené na rozložení pravděpodobnosti.

V dalších podkapitolách jsou uvedeny dvě velmi známé aplikace v oblasti analýzy sociálních sítí. Jedná se o software Pajek a UCINET. Informace o těchto aplikacích vycházejí ze zdroje [15].

5.1.1 UCINET

UCINET je jedním z komplexních programů na analýzu dat ze sociální sítě. Jedná se o nejrozšířenější a nejvíce používanou aplikaci pro analýzu síťových dat obsahujících velké množství analytických rutin. Tento program spadá do sféry komerčních produktů, s možností 30denní zkušební verze, která

[1] Dostupné na webu: http://tlampac.webnode.cz/news/jana–bakalova–analyza–socialni–site–skladatelu/

je poskytována zdarma. UCINET poskytuje dva výstupy analýzy. V textovém a grafickém formátu. Vstupem programu je vetšinou matice nebo soustava více matic, na které je program převážně orientován. Mezi hlavní atributy patří plná podpora operací s maticemi, pravděpodobnostmi, slučování datových množin apod. Program zobrazuje dendogramy nebo stromové diagramy příslušné sítě. Jelikož program sám o sobě neobsahuje procedury pro vizualizaci sociální sítě jako takové, což považuji za nedostatek, je zde možnost zobrazení těchto dat pomocí aplikace NetDraw, která přijímá exportovaná data. Program je schopen rozeznat v datech sociální sítě tyto struktury: kliky, klany, komponenty, jádra, centralitu a mnohé další. Ukázka analýzy dat a další specifikace je uvedena v [15].

5.1.2 Pajek

Dalším programem pro analýzu a vizualizaci dat sociální sítě je Pajek, navržený převážně pro analýzu velkého objemu dat (schopnost analyzovat data s více než 1000000 uzly). Hlavní výhodou této aplikace je redukce dat z velké sítě do několika menších sítí, které mohou být lépe analyzovány. Na rozdíl od předchozího je tento software poskytován zdarma. Data rozsáhlých sociálních sítí je velmi obtížné vizualizovat v jednom náhledu, proto je důležité identifikovat nosné konstrukce v síti, které mohou být vizualizovány samostatně. Pajek používá 6 různých datových struktur:

1) Síť (uzly, hrany).
2) Oddíly (klasifikace uzlů, kde je každý uzel zařazen do třídy).
3) Permutace (uspořádání uzlů).
4) Shluky (podmnožiny uzlů).
5) Hierarchie.
6) Vektory (parametry uzlů).

Struktura programu je založena na těchto šesti datových strukturách a přechody mezi nimi. Další parametry této aplikace spolu s ostatními používanými programy, jako je například NetMiner pro analýzu dat, jsou uvedeny v [15].

6 Návrh aplikace

Aplikace pro analýzu dat sociální sítě je vytvořena v programovacím jazyce Java a vývojovém prostředí NetBeans IDE verze 6.9.1. Jsou zde implementovány celkem tři metody analýzy sociální sítě. V následujících kapitolách jsou tyto metody popsány. Dále je popsána množina dat, na kterých je výsledná analýza prováděna.

6.1 Knihovna JUNG

Jung[1] je knihovna, která poskytuje nástroje pro modelování, analýzu a vizualizaci dat, které mohou být primárně reprezentovány jako graf. Knihovna je napsána v programovacím jazyce Java. Architektura této knihovny je navržena pro různorodou reprezentaci entit a jejich vztahů (orientované a neorientované grafy, multimodální grafy apod.). Jung umožnuje vytváření analytických aplikací pro složitější soubory dat, které zkoumají vztahy mezi subjekty. Dále usnadňuje vizualizaci díky mnoha implementovaným rozložením grafu. Tato knihovna spadá do kategorie volně šiřitelných a modifikovatelných (open–source) nástrojů.

6.2 Analyzovaná data

Aplikace je navržena pro zpracování dat v podobě matice dosažitelnosti, kde každý řádek reprezentuje jednoho aktéra v sociální síti a jeho vztah s ostatními. Pokud existuje vazba k určitému aktérovi, je v matici zaznamenána hodnotou 1. V opačném případě hodnotou 0.[2] Tento vztah není bipartitní, nemusí se tedy jednat o opětovanou vazbu. Jinými slovy, pokud má aktér (řádek) X vztah s aktérem (sloupcem) Y, neznamená to nutně, že aktér Y má vazbu s aktérem X. Pro analýzu jsem zvolil celkem dvě množiny rozdílných sociálních sítí.

První sociální sítí jsou data získaná ze sociologického průzkumu doktora Johna Gagnona z Univerzity Stony Brook v New Yorku, který v roce 1950 sesbíral sociometrické volby od 67 trestanců, kteří byli požádáni, ať vyjmenují všechny spoluvězně, které považují za přátele. Každý z dotázaných měl možnost zvolit libovolný počet těchto přátel. Výsledkem je tedy matice 67x67, kde hodnota 1 označuje pozitivní volbu aktéra X na řádku s aktérem Y ve sloupci (ne naopak). Tuto sociální síť bych zařadil do střední, spíše menší kategorie, co se velikosti nebo hustoty týče.

[1] Java Universal Network/Graph Fremework, poslední vydaná verze JUNG 2.0.1, Leden 2010

[2] Prvky na diagonále matice musejí být vždy nulové. Vazba aktéra se sebou samým se nepovažuje za validní.

Rozsáhlejší sociální síť je reprezentována druhým vzorkem dat. Jedná se o data internetové sociální sítě Facebook studentů univerzity Caltech sesbíraná v září 2005. Hranicí sítě je tedy samotná univerzita. Aktéři a propojení mimo univerzitu nejsou uvažovány. Součástí dat byly informace o každém aktérovi (pohlaví, kolej, údaj o střední škole apod.). Já jsem využil pouze části dat, matici, která vyjadřuje vztah mezi aktéry. Zde jsou, na rozdíl od dat z předchozí části, vztahy mezi aktéry bipartitní. To je dáno vlastnostmi sociální sítě Facebook. Neexistuje zde pouze jednostranné přátelství. Akteři jsou uvedeni pod identifikačním číslem. Není zde tedy dohledatelná návaznost na konkrétního studenta univerzity. Jedná se o datový vzorek ceklem 769 studentů, tedy výsledná matice má velikost 769x769.

6.3 Vytvoření grafu sítě

Sociální síť je v aplikaci reprezentována jako graf. Tedy jako množina vrcholů a množina hran. Naše data jsou v podobě matice, proto je potřeba tuto reprezentaci konvertovat na graf.

6.3.1 Načtení a formát dat

Prvotním úkolem je načtení dat konkrétní sociální sítě. Tato data mohou být uložena v souborech typů: `*.dat`, `*.TXT`, ale i například `*.net`. Soubory typu `*.net` jsou využívány i jinými aplikacemi pro zpracování a analýzu sociálních sítí. Ukázka formátu souboru je následující:

```
A1    0    1    1    0
A2    1    0    0    1
A3    1    1    0    0
A4    1    0    0    0
```

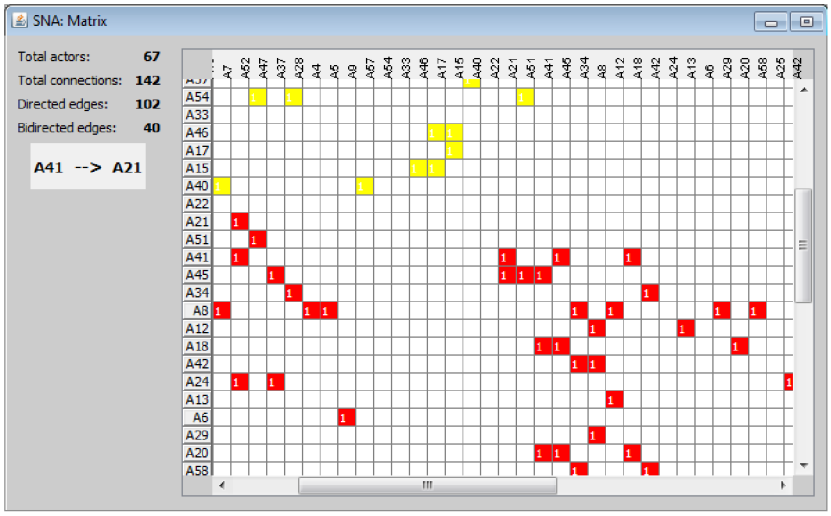
Údaje v prvním sloupci reprezentují jednotlivé aktéry. Může být použito libovolného identifikačního označení, například číslo nebo jméno konkrétního aktéra. Podle tohoto označení budeme schopni identifikovat daný uzel v sociální síti. Další sloupce již reprezentují jednotlivé vztahy mezi aktéry. Například aktér A1 má vazbu s aktérem A2 a A3. Názvy sloupců se v souboru nevyskytují, jsou uvažovány analogicky podle názvů řádků. Načtení souboru je zpracováno ve třídě `LoadFile`. Výstupem instance této třídy je `Object[][]`, kde je uložena matice dat spolu se jmény aktérů. Zpracování těchto dat je provedeno ve třídě `Network`.

6.3.2 Inicializace uzlů, hran a grafu

Po načtení dat ze souboru je potřeba tyto data převést do reprezentace grafu. Jak bylo uvedeno dříve, do množiny vrcholů a hran. Ty reprezentují vztahy mezi aktéry. Celému procesu napomáhá rozhraní `Graph<V,E>` knihovny JUNG. Každý uzel je reprezentován jako samostatný objekt třídy `MyVertex`, kde jsou uloženy informace o tomto uzlu (aktérovi). Například jeho ID, stupeň uzlu, pozice v rámci sociální sítě atd. Podobně jsou vytvořeny i všechny hrany. Každý hrana je objektem třídy `MyEdge`. Při vytváření a reprezentaci hran je jednostranný vztah aktéra k jinému aktérovi reprezentován jako orientovaná hrana. Je–li vztah dvou aktérů bipartitní, příslušná hrana mezi nimi je reprezentována jako neorientovaná a chápeme ji jako obousměrnou. Rozhraní `Graph` udržuje informace o celém grafu spolu s dalšími vlastnostmi uzlů a hran. Pomocí zavolání příslušné metody můžeme tedy jednoduše získat seznamy takzvaných následníků nebo předchůdců zvoleného uzlu, jeho stupěň, seznam vstupních a výstupních hran a další užitečné informace.

6.4 Vizualizace sociální sítě a dat

Pro vizualizaci dat sociální sítě je použito dvou metod. Každá metoda je promítána ve vlastním okně aplikce. První metoda je v podobě tabulky, která reprezentuje matici dosažitelnosti a propojení jednotlivých aktérů. V levé části tohoto okna s tabulkou můžeme vidět také některé informace o sociální síti, jako je celkový počet aktérů, celkový počet vazeb, počet vazeb, které jsou jednostranné a počet vazeb, které jsou oboustranné. Pro lepší orientaci v této tabulce je po označení libovolné buňky vypsán vztah mezi jednotlivými aktéry k ní náležící (zda je jednosměrný nebo bipartitní). Ukázka této grafické reprezentace sítě je na následujícím obrázku.



Obrázek 6.1: Reprezentace matice sociální sítě.

Ve druhém okně je zobrazení vlastní sociální sítě jako grafu. Pro samotné rozmístění prvků je použito algoritmu Fruchterman–Reingold [16], který je impementován v knihovně JUNG a třídě `FRLayout2`. Tento algoritmus rozmisťuje jednotlivé uzly na základě přitažlivé a odpudivé síly jednotlivých uzlů a jejich hran. U hran spojujících dva uzly si můžeme představit jistou elasticitu a energii mezi každým propojením. Algoritmus poté počítá takzvanou atrakci (přitažlivost) v rámci skupiny spojených uzlů. Stejně tak je uvažován opak, tedy odpor jednotlivých uzlů. Jedná se o iterativní algoritmus patřící do skupiny algoritmů založených na síle. Iterace probíhá do doby, než síly mezi uzly v celém grafu dosáhnou rovnováhy a ustálení. Hlavní nevýhodou je vysoká časová náročnost, která dosahuje až $O(V^3)$, kde V je počet uzlů v grafu. Počet iterací je odhadován na $O(V)$ a v každé iteraci je potřeba projít dvojice uzlů spolu s výpočtem jejich odpudivých a přitažlivých sil. Mimo rozmístění celého grafu v prostoru je součástí grafické reprezentace třída pro různé barvy uzlů a hran dle použité analyzační metody. Odlišné barvy uzlů napomáhají k identifikaci význačných struktur. Jejich význam bude vysvětlen v dalších kapitolách. Třída je implementována v `MyVisualViewer` a rozšiřuje třídu `VisualizationViewer` knihovny JUNG. Tato třída poskytuje metody pro náhled na celý graf, posun, rotaci, přiblížení a oddálení celého grafu pomocí kolečka myši. Ve třídě lze nastavovat prakticky celou podobu vykreslovaných objektů. Tedy tvar uzlu, šířku čáry, popisky uzlů apod. Podoba vykresleného grafu sociální sítě je uvedena na obrázcích v následujících kapitolách.

6.5 Použité metody analýzy

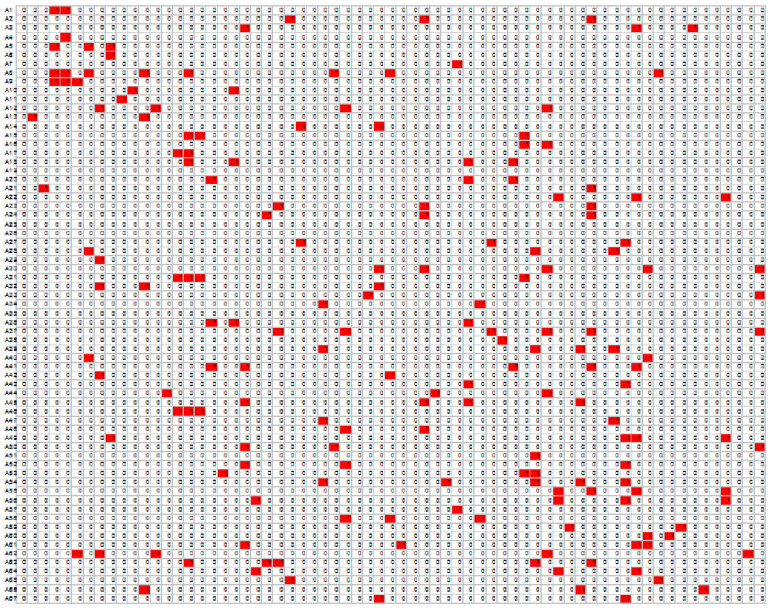
Spolu s inicializací grafu a jeho převodu z maticové reprezentace do grafové jsou ve třídě `Network` zahrnuty i metody pro samotnou analýzu načtených dat. Na data byly aplikovány celkem tři postupy analýzy. Nalezení jader, silných komponent a shluků.

6.5.1 Nalezení jader

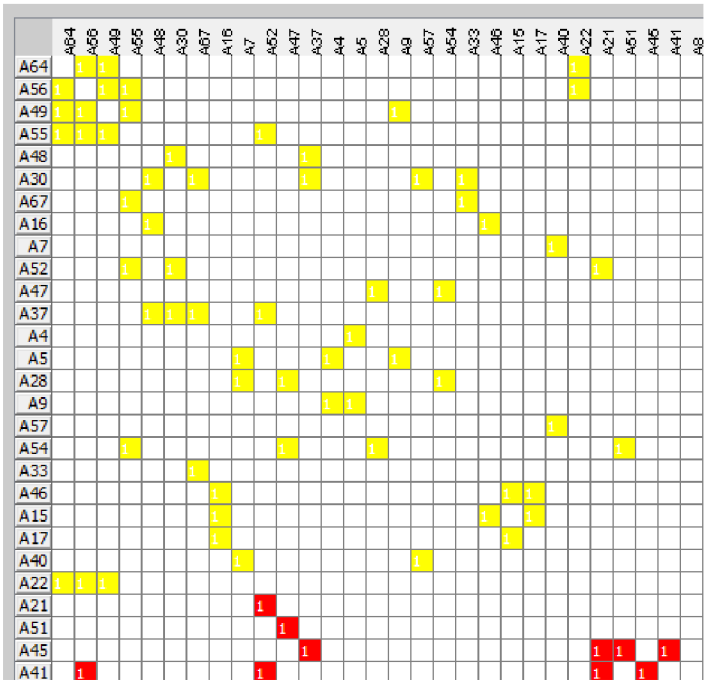
Tato metoda vychází z kapitoly 3.5 o hledání jádra v sociální síti. Konkrétně koncentraci vazeb kolem ulzu podle jeho stupně. Metodě je na vstup předána neseřazená matice dat načtených ze souboru. Výstupem je opět matice, kde jsou již uzly (řádky a sloupce) uspořádány s ohledem na zachování nulové diagonály, podle příslušnosti k jednotlivému jádru. U této metody je grafickým výstupem barevná tabulka reprezentující tuto uspořádanou matici. Jednotlivá jádra vytvářející shluky jsou rozlišena podle barev. Není–li mezi uzly žádná vazba, pole zůstává bílé. Aplikováním metody na sociální síť vězňů získáme přehled o několika zajímavých, na první pohled viditelných shlucích. Je tak identifikována silná koncentrace vazeb v rámci malé skupiny. Podobně je například v ukázce příkladu ze zdoje [13] identifikována skupina hispánského entika v rámci větší sociální sítě. Můžeme

se tedy domnívat, že aktéři takto hustě propojení mají v našem sociálním uskupení jistý společný rys.

Na následujících obrázcích je srovnání dat před a po aplikaci algoritmu vyhledávajícího tyto jádra. Druhý obrázek není kompletní, jedná se pouze o část dat.

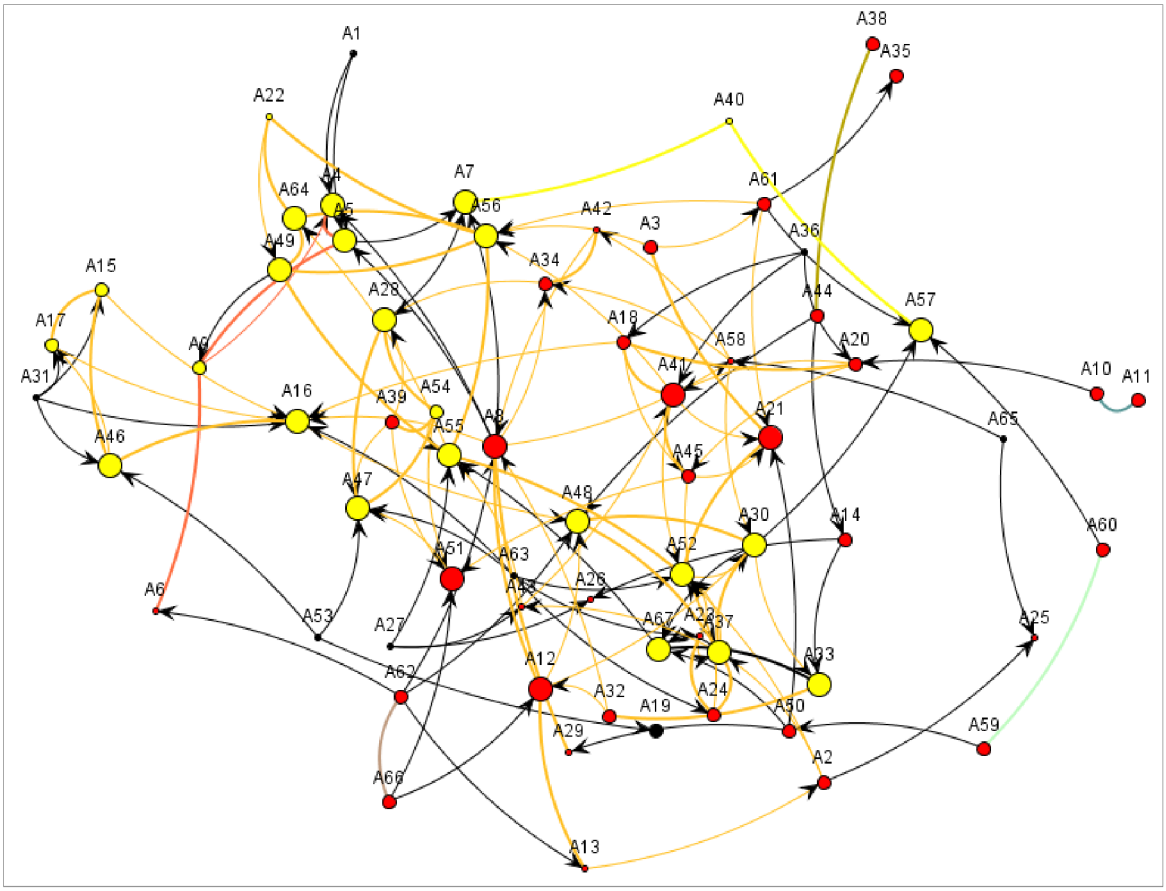


Obrázek 6.2: Matice dosažitelnosti před aplikací metody jader.



Obrázek 6.3: Matice dosažitelnosti po aplikaci metody jader.

Na obrázku 6.3 je tedy patrný shluk aktérů A64, A56, A49 a A55 vlevo nahoře. Tito aktéři jsou spolu úzce provázáni. Z pohledu sociologie se může jednat například o vězeňský gang či jinou propojenou skupinu. Téměř všichni členové v tomto uskupení spolu mají bipartitní vztah. Panuje zde tedy velká důvěra. Další, menší skupinou jsou aktéři A46, A15 a A17. Reprezentace všech jader jako grafu sociální sítě je na následujícím obrázku. Kliknutím myši na zvolenou buňku v tabulce zjistíme minimální hodnotu jádra, která se zobrazí v levé části tabulky. Jedná se o hodnotu k–jádra.

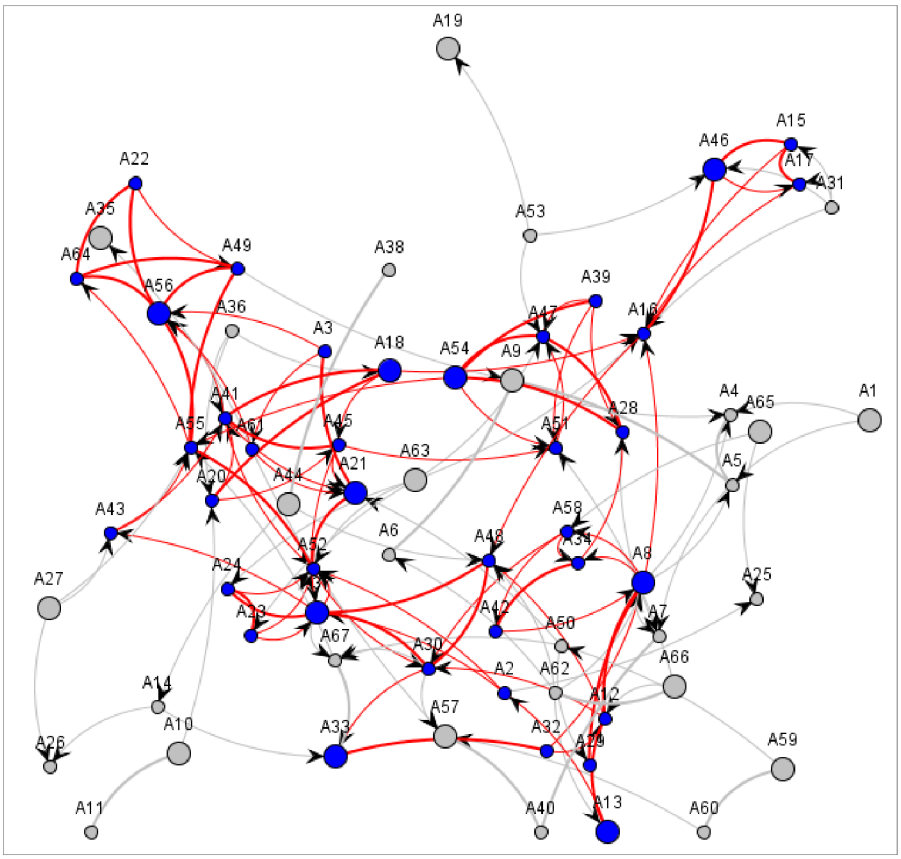


Obrázek 6.4: Reprezentace sociální sítě jako grafu–metoda nalezení jader.

Z obrázku výše je patrná dominance žlutých aktérů. Jedná se tedy o hlavní jádro. Oproti tomu nejmenší stupeň mají aktéři obarvení černě. Z pohledu grafu jsou většinou vytlačeni na jeho okraj. Metoda byla provedena pro porovnání na základě vstupního stupně uzlu (in–degree).

6.5.2 Detekce silných komponent

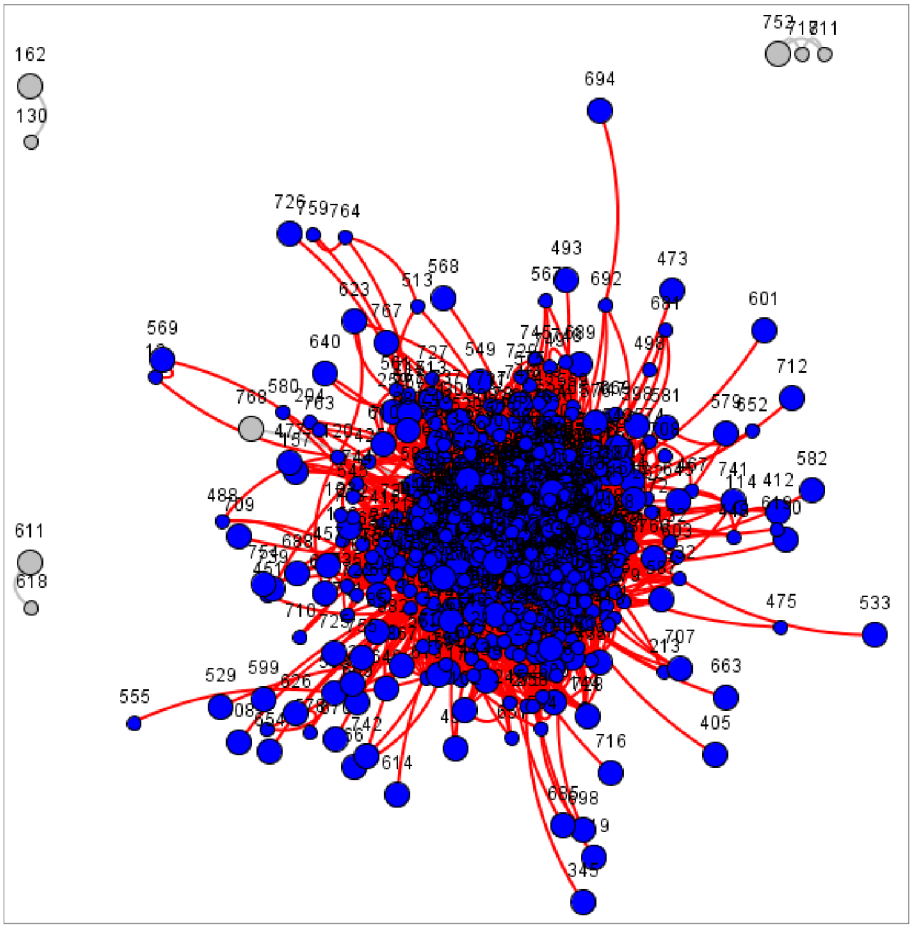
Detekce silných komponent opět vychází z matice dosažitelnosti. Algoritmus je založen na zjišťování dosažitelných vrcholů z výchozího vrcholu a zpět. V případě, že taková cesta existuje, vrchol je zařazen do komponenty. Algoritmus končí, jakmile jsou všechny vrcholy součástí některé komponenty. Jako komponenta je uvažován i samostatný uzel. Jádrem algoritmu je určení trojúhelníkové nerovnosti. Jedná se o variaci Floydova algoritmu popsaného v [7]. Počet komponent a jejich rozložení v rámci sociální sítě si můžeme prohlédnout v okně s grafickou reprezentací sociální sítě v podobě grafu. Ve spodní liště tohoto okna jsou navigátory spolu s počítadlem těchto komponent.



Obrázek 6.5: Dominantní komponenta.

Obrázek 6.5 je vizualizací metody nalezení silných komponent pro první datavou množinu – věznici. Celkem je v této sociální síti 21 komponent, což je z hlediska sociometrie velmi zajímavý jev. Naznačuje to všeobecnou nedůvěru mezi aktéry v rámci tohoto sociálního uskupení. Mezi aktéry převažuje jednostranná volba (102 ku 40 z celkem 142 vazeb). Vztah dotyčných jedinců je ve většině případů neopětován. To znaméná, že aktér A nepovažuje za přítele aktéra B, který jej však jako přítele označil. Velký podíl na vysokém počtu komponent mají i jedinci, kteří jsou tzv. na okraji

skupiny. Nevede k nim prakticky žádná pozitivní volba od ostatních aktérů a jsou tedy osamoceni. Může se jednat například o nováčky, neoblíbené jedince nebo oběti šikany. Pro srovnání je na následujícím obrázku uvedena grafická podoba druhé datové množiny–vazeb studentů univerzity Caltech na síti Facebok.



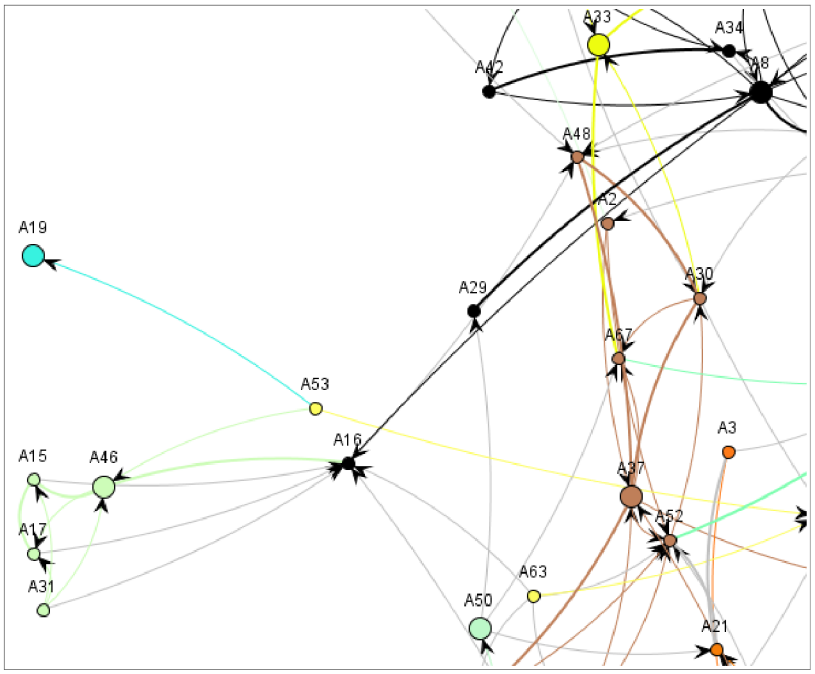
Obrázek 6.6: Dominantní komponenta v síti Facebook.

Tato síť je díky počtu aktérů výrazně hustější. Ve srovnání z předchozím příkladem však obsahuje pouze 5 silných komponent. Na obrázku je opět zvýrazněna dominantní komponenta v této sociální síti. Pro detailnější a přehlednější pohled na tuto sociální síť a jednotlivé vazby mezi aktéry lze v samotné aplikaci použít přiblížení. Uzavřené skupiny jsou zde patrny na okrajích této sítě. S hlavní dominantní skupinou nemají společnou žádnou vazbu.

6.5.3 Shluky

Poslední aplikovanou metodou analýzy je vyhledání shluků, neboli klastrů. Jedná se o aktéry, kteří mají nějaké společné atributy. V našem případě je shluk utvářen vždy kolem jedince s nejvyšším stupněm ulzu. Tedy kolem autority v rámci tohoto shluku. Algoritmus jedince do shluku zařazuje

podle kritéria, zda má s autoritou vztah a zároveň zda má vztah s jiným členem stejného shluku. Jednotlivé skupiny jsou v grafické reprezentaci rozlišeny společnou barvou uzlu a hrany. Aktér v roli autority je v porovnání s ostatními zobrazen zvětšeně.



Obrázek 6.7: Ukázka shluku.

Na obrázku 6.7 je zobrazen zvětšený náhled na část sociální sítě, ze které můžeme vyčíst několik zajímavých údajů. V levé spodní části se nachází těsně propojený shluk aktérů A31, A13, A15 a autorita A46. Všichni aktéři patřící do tohoto shluku mají vazbu s aktérem A16, který je zařazen mimo tuto skupinu. Odstraněním tohoto aktéra by se shluk odpojil od zbytku sociální sítě. Aktéra A16 tedy z pohledu sociometrie můžeme označit za tzv. most (tento prvek byl diskutován v předchozích kapitolách). Skrze aktéra A16 tedy může probíhat komunikace s ostatními shluky. Například s dalším, větším uskupením, zobrazeným černou barvou na obrázku vpravo nahoře.

7 Závěr

Cílem této bakalářké práce bylo vytvořit aplikaci demonstrující použití metody analýzy sociální sítě pro dolování znalostí. Program představuje převážně grafickou reprezentaci sociální sítě s použitím tří základních metod její analýzy. A to detekce silných komponent, nalezení jádra a shluků. Aplikace dává možnost naleznout v sociální síti specifické struktury aktérů a různé rysy sociálního uskupení. Aplikace není omezena pouze na jednu sociální síť. Díky načítání dat ze souboru lze při zachování formátu tohoto souboru vložit jakoukoliv jinou sociální síť reprezentovanou maticí dosažitelnosti. Zde uvažuji o rozšíření v podobě ohodnocení hrany, tedy váhy propojení mezi aktéry (uvažovány jsou i negativní propojení). To dává další možnosti v pokročilejších metodách analýzy. Z hlediska shlukovacích metod by se jednalo o velmi přínosný údaj. Další vylepšení vidím v optimalizaci algoritmů, aby bylo dosaženo nižší časové náročnosti výpočtů. Pro rozsáhlejší sítě není dosaženo plynulosti vykreslování grafické podoby sociální sítě. Dalším rozšířením může být pokročilejší uživatelské rozhraní, které dovoluje interaktivně měnit parametry sociální sítě (vazbu mezi aktéry, váhu hrany apod.).

Po implementaci všech metod jsem provedl srovnání výstupů s programem Pajek při analýze sociální sítě „vězení“. Konkrétně srovnání s metodou nalezení jádra a silné komponenty. Výstup analýzy programu Pajek, který má možnost zobrazit seznamy aktérů spolu s jejich příslušností k jádru nebo komponentě, je totožný s implementací, kterou jsem provedl. Jednotliví aktéři byli zařazani v porovnání s mou aplikací do stejných jader a komponent. Více o výstupu analýzy programu Pajek je v příloze A.

Tuto práci pro sebe považuji za velmi přínosnou. Rozšířil jsem si znalosti v oblasti metod kreslení grafů a použití teorie grafů obecně. Souvisejícím odvětvím této práce bylo hlubší seznámení se s principy socilometrie, sociologie a chování jedinců v rámci sociální skupiny.

Literatura

[1] BUŠTÍKOVÁ, Lenka. *Analýza sociálních sítí*. Sociologický časopis. 1999, No. 2, s. 193-206.

[2] DOREIAN, Patrick; STOKMAN, Frans N. *Evolution of Social Networks*. Amsterdam: Gordon and Breach Publishers, 1997. 261 s. ISBN 90-5699-538-3.

[3] SCOTT, John. *Social Network Analysis: A Handbook.* 2nd edition. London: SAGE Publications Ltd, 1991. 244 s. ISBN 978-0-7619-6339-4.

[4] Social network. In *Wikipedia: the free encyclopedia* [online]. St. Petersburg (Florida): Wikipedia Foundation, [cit. 5.2.2011]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Social_network>.

[5] WASSERMAN, Stanley; FAUST, Katherine. *Social Network Analysis: Methods and Applications.* USA: Cambridge University Press, 1994. 825 s. ISBN 0-521-38269-6.

[6] AGGARWAL, Charu C.; WANG, Haixun. *Managing and Mining Graph Data.* 1st Edition. New York: Springer, 2010. 610 s. ISBN 978-1-4419-6044-3.

[7] BRANDES, Urlik; ERLEBACH, Thomas. *Network Analysis: Methodological Foundations.* Berlín: Springer, 2005. 471 s. ISBN 0302-9743.

[8] HANNEMAN, Robert A.; RIDDLE, Mark. *Introduction to social network methods* [online]. Riverside, CA: University of California, 2005 [cit. 31.1.2011]. Dostupné z WWW: <http://faculty.ucr.edu/~hanneman/nettext/>.

[9] Clique. In *Wikipedia: the free encyclopedia* [online]. St. Petersburg (Florida): Wikipedia Foundation, [cit. 5.2.2011]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Clique>.

[10] ALBA, Richard D. A GRAPH-THEORETIC DEFINITION OF A SOCIOMETRIC CLIQUE*. Journal of Mathematical Sociology. 1973, 3, s. 113-126.

[11] KANTARDZIC, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. [s.l.]: John Wiley & Sons, 2003. 360 s. ISBN 9780470544341.

[12] KRATOCHVÍL, Jan. *Graph Drawing*. Czech Republic: Springer, 1999. 422 s.

[13] NOOY, Wouter de; MRVAR, Andrej; BATAGELJ, Vladimir. *Stuctural analysis in the social sciences: Exploratory Social Network Analysis with Pajek*. New York, USA: Cambridge University Press, 2005. 334 s.

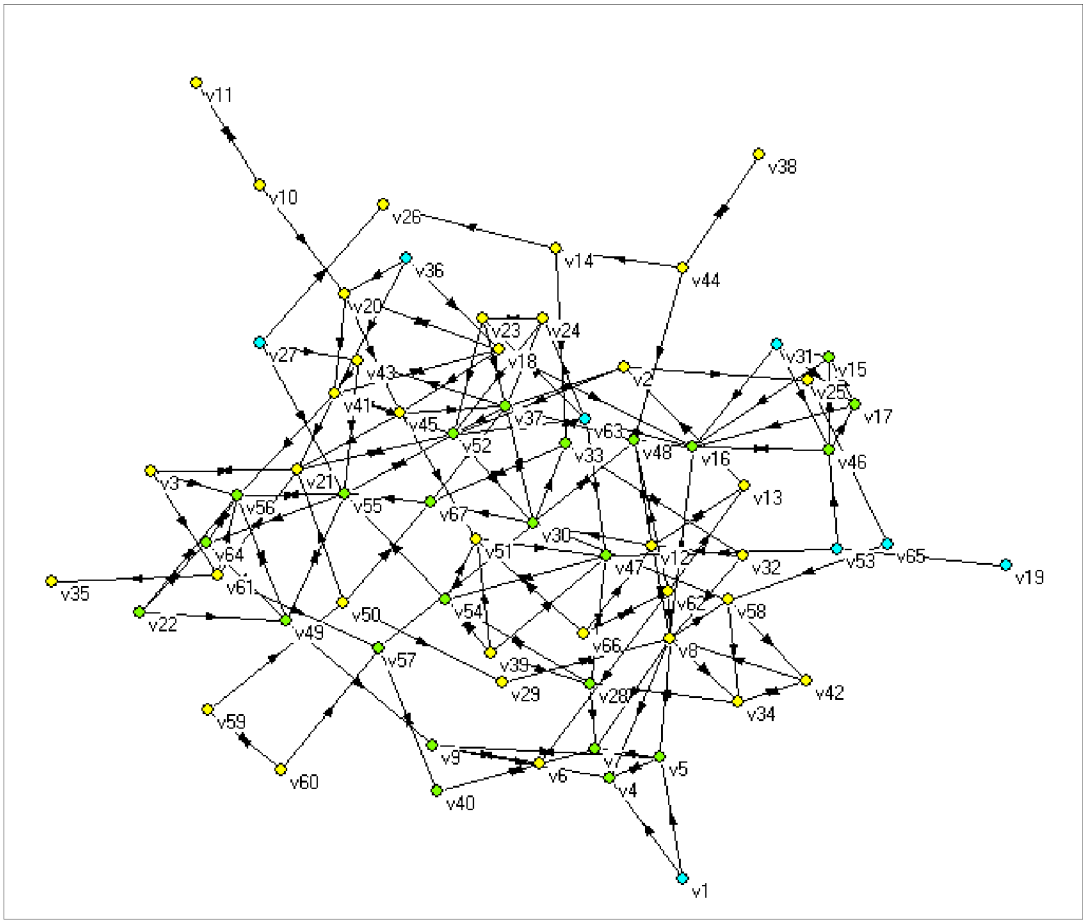
[14] ČIHOVSKÝ, Jaroslav. *Sociologický výzkum*. Olomouc, 2006. 44 s. Studijní text. Univerzita Palackého v Olomouci.

[15] HUISMAN, Mark; VAN DUIJN, Marijtje A.J. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, 2005. Software for Social Network Analysis, s. 270.

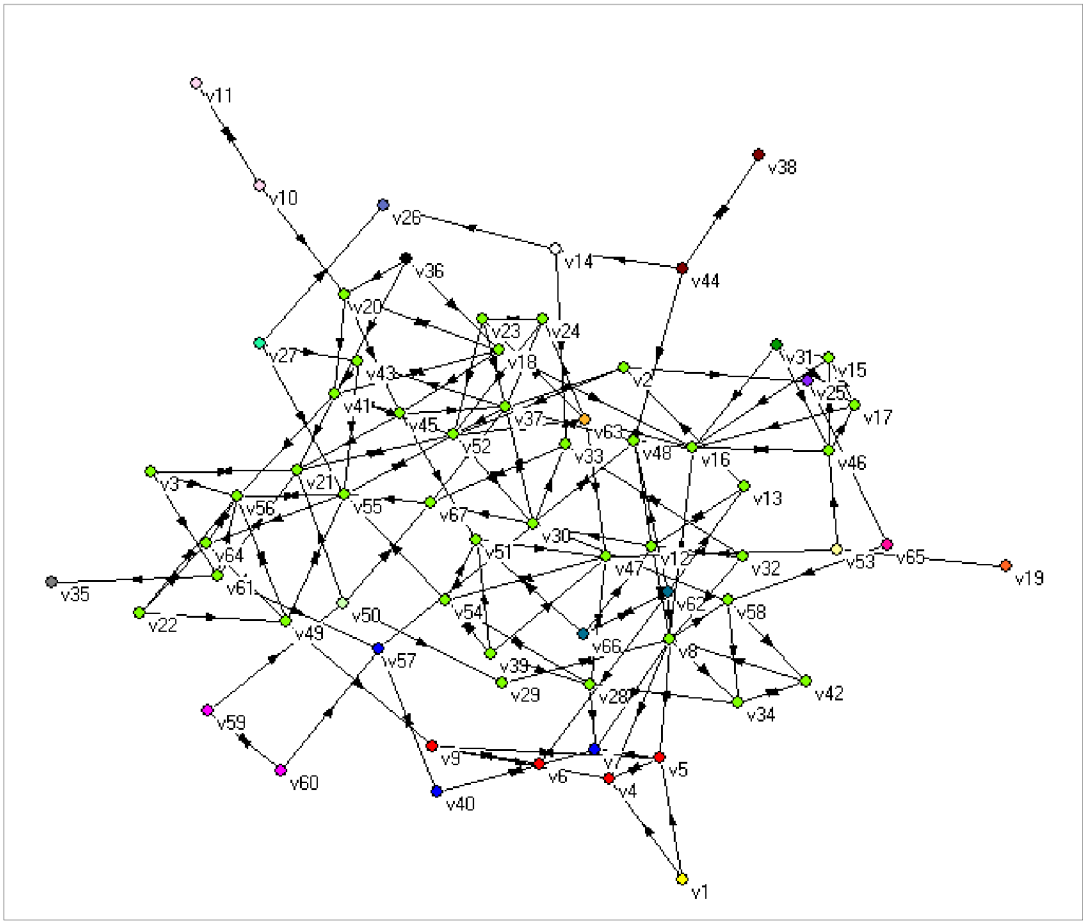
[16] FRUCHTERMAN, Thomas M. J.; REINGOLD, Edward M. *Graph Drawing by Force-directed Placement.* SOFTWARE-PRACTICE AND EXPERIENCE. 1991, 21, s. 1130.

Seznam příloh

Příloha A



Obrázek A.1: Nalezení jádra programem Pajek.



Obrázek A.2: Nalezení silných komponent programem Pajek.