Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra informatiky a výpočetní techniky

Diplomová práce

Systém pro extrakci informací z kriminalistických textů

Plzeň 2016 Marek Naggy

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 11. května 2016

Marek Naggy

Poděkování

Rád bych poděkoval prof. Ing. Karlu Ježkovi, CSc. za cenné rady, ochotné vedení práce a čas, který mi věnoval. Dále děkuji plk. Ing. Janu Hořínkovi za jeho rady a poskytnutí možnosti otestovat program v reálném prostředí. V neposlední řadě patří můj dík mé rodině za podporu během studia.

Abstract

System for extracting information from criminalistic texts

The aim of this diploma thesis is processing of unstructured documents and further data processing of extracted information. The main attention was devoted to extraction of personal names. From obtained personal names was created a social (criminal) network. An effective destabilization of this network is shown. Also the detection of communities, which occur frequently together is demonstrated and the example of spatial and the temporal analysis is presented. Our system could facilitate the work of investigative reporters or police, which has an available large set of unstructured documents. Manual processing of these documents may be difficult. Mainly, if they look for clues between multiple documents. However, these documents may contain some personal data. Therefore the anonymizator was created similarly as the deanonymizer. The function of this system is demonstrated. Used test data was created from articles on terrorist attacks in Paris and Brussels.

Keywords: unstructured documents processing, social networks, criminal network destabilization, personal data anonymization, named entity recognition.

Abstrakt

Systém pro extrakci informací z kriminalistických textů

Práce se zabývá zpracováním nestrukturovaných dokumentů a následným zpracováním extrahovaných dat. Největší pozornost je věnována extrakci jmen osob, ze kterých je následně vytvářena sociální (kriminální) síť. Dále je ukázán způsob, jakým tuto síť efektivně destabilizovat. V práci jsou ukázány možnosti detekce komunit, které se spolu často v textech vyskytují, prostorové a časové analýzy. Systém by mohl usnadnit práci např. investigativním reportérům nebo policii, která má k dispozici velké množství textových dokumentů. Jejich ruční zpracování, zejména pokud jsou hledána vodítka napříč několika dokumenty, může být obtížné. Jelikož tyto záznamy mohou obsahovat osobní údaje, je zde představen anonymizátor, který tyto údaje dokáže anonymizovat a následně deanonymizovat. Funkčnost systému byla ověřena na testovací sérii článků, které se věnují teroristickým útokům v Paříži a Bruselu.

Klíčová slova: zpracování nestrukturovaných dokumentů, sociální sítě, destabilizace kriminální sítě, anonymizace osobních údajů, rozpoznávání pojmenovaných entit.

Obsah

1 Úvod

Dle oficiálních statistik se v České republice ročně odehraje průměrně 300 tisíc kriminálních činů. Ročně se škody s nimi spojené odhadují na 20 miliard korun. S rostoucím výpočetním výkonem se objevují postupy, jak rozkrývat zločinecké sítě nebo zločin předpovídat. Je zaznamenáno několik případů (Hruška a kol., 2015), kdy na základě výstupů systémů pro podporu rozhodování, policie dopadla pachatele zločinu ihned po jeho spáchání nebo dokonce během páchání trestného činu. Ve městech kde tyto systémy používají, údajně klesla kriminalita o několik, někde dokonce o desítky procent. Podobné experimentální systémy se začínají objevovat i v České republice. Většina těchto systémů je však napojena na databázi policie, nebo používá semistrukturovaná data. Velká část dokumentů, které však policie vytváří, nebo byly v minulosti vytvořeny, mají podobu nestrukturovaných dokumentů. Jejich manuální zpracování je časově náročné a některé informace, které je nutné vyvodit z několika dokumentů, nemusí být na první pohled patrné. Tato skutečnost může práci vyšetřovatelů ztěžovat a zpomalovat. Diplomová práce se zabývá návrhem a vytvořením části systému, který by tuto práci mohl usnadnit.

Jelikož vstupem systému budou nestrukturovaná data, jsou v druhé kapitole popsány prostředky, jak nestrukturované dokumenty zpracovávat a jak rozpoznávat hlavní informace v těchto dokumentech. Následuje kapitola, která popisuje základy analýzy sociálních sítí, která je následně na zpracované dokumenty aplikována. Ve čtvrté kapitole jsou uvedeny některé metody, pro rozkrývání zločineckých sítí a predikci kriminality, které podobné systémy používají.

Předpokládá se, že v budoucnu program může pracovat s reálnými daty, ve kterých se mohou objevovat osobní údaje. Je proto nutné, aby tyto informace mohl program anonymizovat. Anonymizací osobních dat se zabývá šestá kapitola. Návrhem a popisem architektury vytvořeného systému pak kapitola pátá.

Vybrané implementované metody ze čtvrté kapitoly a výsledky pomocí nich získané jsou ukázány v kapitole sedm. Jedná se zejména o rozkrývání a destabilizaci kriminálních sítí. Jako testovací soubor dat byly vybrány novinové články pojednávající o teroristických útocích z Paříže a Bruselu. Na těchto datech je předvedena funkčnost programu. Kapitola zároveň ukazuje, že tento nástroj může být vhodný např. i pro investigativní reportéry. Při dokončování se podařilo program otestovat na reálných datech v infrastruktuře PČR. V závěru práce jsou tak mj. popsány známé problémy, se kterými se bude potřeba, zejména pokud by měl být program v praxi použit, vypořádat.

2 Vyhledávání informací v textu

Tato kapitola se zabývá přehledem pojmů a metod, které jsou v práci použity ve vztahu s extrakcí informací z přirozených textů. Dále popisuje způsob, jakým je úspěšnost extrakce informací hodnocena a jakých výsledků se v této oblasti dosahuje.

2.1 Extrakce informací

S rozvojem digitalizace dokumentů, dostupností internetu, popularitou on-line sociálních sítí a s rostoucí výpočetní kapacitou, se v posledních dekádách extrakce informací (IE) setkává se stále větším zájmem. Tato část tuto úlohu a její podúlohy popisuje, přičemž vychází z Piskorski a Yangarber (2013).

Extrakce informací se zabývá nalezením relevantních, předem definovaných faktů v přirozeném textu. Dle ACM ontologie[1] je IE řazena do oblastí:

- Computing methodologies – Artificial intelligence – Natural language processing – Information extraction
- Information systems – Information Retrieval – Retrieval tasks and goals – Information extraction

Faktem budeme označovat strukturované záznamy, které mohou zachycovat entity a události v textu zmíněné nebo vztahy mezi těmito entitami a událostmi. Dále stavy těchto událostí, účastníků nebo entit. Fakta jsou v přirozeném textu hledána za konkrétním účelem, který bývá pro zkoumanou doménu specifický. Fakta relevantní v jedné doméně mohou být v jiné doméně irelevantní. V našem případě se jedná o nalezení faktů ze záznamů kriminálních činů, zápisů a dalších policejních zpráv. Extrahovaná fakta jsou dále použita pro tvorbu kriminální sítě, prostorové analýzy aj.

Extrahovaná fakta jsou ukládána do předem definovaných struktur, které se mohou skládat z různého počtu atributů. Tyto atributy jsou zpravidla tvořeny řetězcem, jednou nebo více předdefinovanými hodnotami či referencí na jinou strukturu. IE tak vytváří z nestrukturovaných přirozených textů *strukturované záznamy.*

Extrakce informací se dále rozděluje na několik podúloh. Některé z nich, kterými se budeme zabývat jsou:

- *Rozpoznávání pojmenovaných entit* (Named Entity Recognition, NER) – zabývá se nalezením a klasifikací entit, které v textu nesou nejdůležitější infor-

[1] Viz The 2012 ACM Computing Classification System – `http://www.acm.org/about/class`.

mace. Jsou jimi například jména osob, geografické jména nebo jména organizací (viz část 2.4).

- *Identifikace koreferencí* (Co-reference Resolution, CO) sjednocuje shodné entity, které mohou být vyjádřeny různě nebo mohou vyplývat z kontextu (viz část 2.5).
- *Extrakce relací mezi entitami* (Relation Extraction, RE) se zabývá detekcí a klasifikací relací mezi entitami (viz část 2.6).

Vzhledem k následnému použití dat a jejich povaze, která nám dovoluje pouze jejich anonymizované zpracování, se v práci budeme zabývat zejména podúlohou NER a částečně úlohami CO a RE. Vzhledem ke komplexitě přirozeného jazyka, zejména pak českého jazyka, IE představuje netriviální a vyzývavou úlohou. Z tohoto důvodu se v práci objeví několik zjednodušujících předpokladů, které budou v práci průběžně popisovány.

2.2 Základní pojmy ve zpracování textů

Tato část přibližuje pojmy, které budou v části věnující se extrakci informací používány. Není-li uvedeno jinak, definice vycházejí z Cvrček (2015).

Lemma, lemmatizace a stematizace

V lingvistickém smyslu *lemma* označuje základní slovníkovou podobu hesla tj. jeho základní tvar. Lemmatem podstatných jmen je první pád jednotného čísla. Například tvary „hrady“, „hradu“ a „hradem“ mají společné lemma „hrad“. U přídavných jmen je lemmatem přídavné jméno v prvním pádu jednotného čísla mužského rodu. U sloves pak infinitiv slovesa. *Lemmatizací* označujeme proces, kterým je slovo převedeno na lemma. Lemmatizace bere v úvahu kontext, ve kterém se slovo nachází.

Stematizace je proces podobný lemmatizaci, při němž je však hledán základ slova, který se nemusí shodovat s lemmatem. Na rozdíl od lemmatizace bývá stematizace rychlejší, ale méně přesným procesem. Často je založena na odstraňování předpon a přípon, což vede k tomu, že např. slovo „je“ nebude stematizováno na „být“ apod. Stematizace nebere v úvahu kontext v jakém se slovo nachází (Chmelař a kol., 2011).

Token a tokenizace

Jako *token* označujeme nejmenší smysluplnou jednotku textu. Většinou se jedná o slovo, zkratku nebo číslo, které je v textu odděleno mezerami či interpunkcí. Existují však výjimky, kdy může být více slov považováno za jeden token jako např. „mohu-li“, „Česko-slovenský“ apod. Na rozdíl od lemmatu je token konkrétní rea-

lizací slova a nemusí být v základním tvaru. Proces, který rozděluje text na tokeny, se nazývá *tokenizace*.

Segmentace

Segmentace je proces rozdělení textu na menší celky. Nejčastěji se provádí segmentace větná. Na tokenizaci se můžeme dívat jako na segmentaci dle slov. Segmentaci nelze provádět triviálním přístupem, který rozdělí text dle teček. Ty se mohou vyskytovat ve zkratkách, po řadových číslovkách apod.

Morfologie, morfologická analýza a značka

Morfologie se zabývá zkoumáním podob slov, které vznikají ohýbáním slov a jejich základní formou. Jako *morfologická analýza* se označuje přiřazení všech možných lemmat a morfologických údajů ve formě morfologické značky k tokenu. Pokud není z výrazu jednoznačně určitelné o jaké lemma se jedná, je mu přiřazeno více údajů než jeden. Příklad morfologické analýzy uvádí tabulka 2.1. Program, který morfologickou analýzu provádí, se nazývá *morfologický analyzátor*.

Morfologická značka (nebo morfologický tag) označuje gramatickou informaci, kterou slovo v konkrétním kontextu nese (viz morf. značka v tabulce 2.1). *Úloha tag* vybírá nejpravděpodobnější morfologickou značku a lemma v daném kontextu z výstupu morfologické analýzy. Morfologické značky nesou informace o slovním druhu, jmenném rodu, číslu, mluvnickém pádu, času apod. Pro příklad v tabulce 2.1 morfologická značka lemmatu „být“ značí, že u výrazu „je“ se jedná o sloveso (V) v přítomném čase (p), jednotného čísla (S), ve třetí osobě (3) a vidu nedokonavém (n). Pro popis dalších značek z tabulky 2.1 viz Cvrček (2015).

Tab. 2.1: *Příklad morfologické analýzy pro slovo „je“ z věty „Sním je místo něho.“. Převzato z Cvrček (2015).*

Výraz	Lemma	Morfologická značka
je	být	VpS–3n
	ono	PPSN4–
	oni	PPPM4–
	ony	PPPI4–
	ony	PPPF4–
	ona	PPPN4–

Morfologické značky mohou mít různý formát a jejich počet se může lišit. České morfologické analyzátory často používají značky a pozice, které používá The Prague Dependency Treebank[2]. Tento formát používá patnáct pozic. Pro naši práci jsou nejdůležitější slovní druhy na první pozici, jmenný rod na třetí pozici, číslo na čtvrté

[2]Bližší popis jednotlivých pozic a významů značek naleznete na adrese `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s02s01.html`.

pozici a mluvnický pád na páté pozici. Jako příklad tohoto formátu uveďme morfologickou značku `P5ZS2--3------1` pro slovo „něho“ ze shodné věty, kterou používá tabulka 2.1.

2.3 Vyhodnocování úspěšnosti extrakce informací

Úspěšnost rozpoznávání faktů se měří pomocí hodnot *přesnost*, *pokrytí* a jejich kombinací *F-mírou* (Piskorski a Yangarber, 2013; Ševčíková a kol., 2007b).

Chybová matice

Při vyhodnocování úspěšnosti rozpoznávání se mohou vyskytnout dva druhy chyb. Tyto chyby lze vizualizovat pomocí chybové matice, kterou zobrazuje tab. 2.2.

Tab. 2.2: *Chybová matice zobrazující různé možnosti výstupu systému v porovnání se skutečnou hodnotou.*

		Výstup systému	
		p	**n**
Skutečná hodnota	**p′**	Skutečně pozitivní (TP)	Falešně negativní (FN)
	n′	Falešně pozitivní (FP)	Skutečně negativní (TN)

Pokud systém rozpozná fakt jako pozitivní a pokud pozitivní skutečně je (v matici p a p'), jedná se o pozitivní rozpoznání (true positive, TP). Totéž platí pro negativní případ (v matici n a n', true negative, TN). Pokud však systém označí fakt jako pozitivní, přičemž není (v matici p a n'), jedná se o falešně pozitivní rozpoznání (false positive, FP). Naopak, pokud systém fakt neoznačí, přičemž měl (v matici n a p'), jedná se o falešně negativní chybu (false negative, FN).

V některých systémech může být jeden druh chyby závažnější nežli druhý. V našem případě lze předpokládat, že je závažnější chyba FN. To zejména z důvodu, že se v analyzovaných textech mohou vyskytovat osobní údaje, které musí být z co největší části anonymizovány.

Přesnost

Přesnost (precision) vyjadřuje poměr jakou část faktů se systému podařilo rozpoznat. Vypočte se pomocí vzorce (1).

$$precision = \frac{TP}{TP + FN} \tag{1}$$

Pokrytí

Pokrytí (recall) vyjadřuje poměr jaká část rozpoznaných faktů je skutečně relevantních a správných. Vypočte se podle vzorce (2).

$$recall = \frac{TP}{TP + FP} \tag{2}$$

F-míra

F-míra (F_1-míra) je harmonický průměr přesnosti a pokrytí a vypočte se podle vzorce (3).

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

Obecnou verzi F-míry, ve které je možné upravit váhy přesnosti a pokrytí uvádí vzorec (4). Význam přesnosti lze zvýšit snížením hodnoty β, přičemž β je nezáporná hodnota. Pokud $\beta = 1$ pak jsou váhy hodnot přesnosti a pokrytí nastaveny shodně a vzorec (4) odpovídá vzorci (3). Často používané jsou rovněž $F_{0,5}$ a F_2 míry. F_2-míra upřednostňuje pokrytí, zatímco $F_{0,5}$-míra přesnost.

$$F_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{4}$$

Jazykový korpus

Pro výpočet přesnosti, pokrytí a F-míry je potřeba *jazykový korpus* s anotovanými daty. Oproti jazykovému korpusu se poté výstup systému porovnává.

Jazykové korpusy obsahují reálné texty, které zobrazují jazykové jevy, slova a slovní spojení v jejich přirozeném kontextu (Cvrček, 2015). V tomto korpusu jsou fakta, označeny pomocí anotací. Ty se zpravidla vytváří manuálně pomocí speciálních programů. Pro ověření shody mezi anotátory mohou být stejná data zpracovávána několika osobami a následně mezi sebou tyto anotace porovnávány.

2.4 Rozpoznávání pojmenovaných entit

Termín *pojmenovaná entita* (Named Entity, NE) byl zaveden roku 1995 v souvislosti s konferencí MUC-6 (Message Understanding Conference). Cílem konferencí

MUC byla podpora výzkumu v oblasti IE a vývoj systému, který v textech identifikuje jména lidí, organizací, geografické názvy, časové údaje nebo peněžní částky. Na těchto konferencích se nejčastěji vyhodnocovaly finanční a vojenské zprávy nebo zprávy informačních služeb o teroristických útocích (Ševčíková a kol., 2007b; Nadeau a Sekine, 2007).

Definice pojmenované entity dle MUC-6 a navazujících projektů

Termín NE byl na MUC-6 definován velmi úzce. Jednalo se o sedm druhů pojmenovaných entit:

- ENAMEX s atributem ORGANIZATION pro jména organizací, PERSON pro jména osob a LOCATION pro geografická místa.
- TIMEX s atributem DATE pro data nebo roky a TIME pro časové údaje.
- NUMEX s atributem MONEY pro peněžní částky a PERCENT pro procentuální hodnoty.

Dle této definice nepovažujeme za NE např. název artefaktu nebo jméno významné události, které mají jedinečné jméno a intuitivně by měly jako NE vystupovat.

Projekty IREX (Information Retrieval and Extraction Exercise) a CoNLL (the Conference on Natural Language Learning) navazující na MUC-6 tuto definici dále rozšířily o typ ARTIFACT. S rozšiřováním úlohy NER na další domény vznikla potřeba opustit tuto úzkou definici a zobecnit ji.

Pojmenovaná entita dle TEI

Podle směrnic TEI[3] (Text Encoding Initiative), ve kterých jsou definovány prostředky pro značkování textů v přirozeném jazyce, jsou NE[4] rozšířeny o adresy, čísla, časové úseky, zkratky aj. Zároveň mohou být NE dle konkrétní domény a úkolu dodefinovány (Ševčíková a kol., 2007b). Nadeau a Sekine (2007) uvádějí, že tímto rozšířením vznikají kategorie NE jako jsou názvy filmů, jména chemických prvků, drog, emailových adres apod.

Mějme následující větu:

Velká francouzská revoluce je označení pro období dějin Francie mezi lety 1789 a 1799.

---

[3] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html`

[4] V této směrnici se explicitně jako pojmenované entity neoznačují, jejich význam je ale stejný (Ševčíková a kol., 2007b).

Vyznačené slova a slovní spojení můžeme dle směrnic TEI považovat za NE. Dle definice MUC se však slovní spojení „Velká francouzská revoluce“ za NE nepovažuje i když se jedná o jméno významné události.

Jak vyplývá z předchozího textu na formální definici pojmenované entity nepanuje shoda a je spíše pragmatická v závislosti na doméně a řešeném problému. Vzhledem k tomu, že pro další zpracování textů bude nutné anonymizovat osobní údaje a budeme s těmito NE dále pracovat, bude v této práci použita definice NE dle směrnice TEI.

Rozpoznání pojmenovaných entit

Rozpoznáváním pojmenovaných entit (Named Entity Recognition, NER nebo též Named Entity Recognition and Classification, NERC) označujeme identifikaci a klasifikaci NE. Ty v textu často nesou nejvýznamnější informace a jejich rozpoznávání má pro zpracování přirozeného jazyka zásadní význam. NER se často objevuje i v jiných úlohách. NER se např. používá v systémech pro zodpovídání dotazů nebo u automatického překladu. V těchto systémech se používá jako první krok pro předzpracování textů (Nadeau a Sekine, 2007). V našem systému tomu bude podobně. Systém nejprve pomocí NER rozpozná NE, určí jejich koreference, anonymizuje je a následně je budeme dále zpracovávat.

Techniky rozpoznávání pojmenovaných entit

Při tvorbě NER systémů se využívají dva odlišné přístupy. První z nich se snaží o vytvoření obecného systému, který bude schopen detekovat NE bez ohledu na doménu a jazyk, ve kterém analyzované texty jsou. Druhým přístupem je tvorba specifických systémů, které se zabývají pouze konkrétním jazykem a doménou. Tyto systémy pak zpravidla dosahují lepších výsledků (Nadeau a Sekine, 2007).

Pro nalezení NE, které mají předem jasný tvar je možné používat pravidlový přístup. Ten dokáže pomocí regulárních výrazů takové entity nalézt. U nás to zajišťuje tzv. třetí průchod, viz část 6.3. Dalším způsobem jak rozpoznávat vlastní jména může být použití seznamů (v rámci NER označovaných jako „gazetteer“, lexikon nebo slovník) jmen, obcí apod. U nás to zajišťuje tzv. druhý průchod, který na základě seznamu osobních jmen zvyšuje jejich pokrytí.

Moderní NER systémy nejčastěji používají strojové učení s učitelem (Nadeau a Sekine, 2007). To využívá trénovacího korpusu, jenž obsahuje označené NE, na kterých se klasifikátor trénuje. Poté je klasifikátor schopný označit i NE, které v trénovacích datech nebyly. Strojové učení je založeno na pravděpodobnostním přístupu. Ke zpracování jazyka je vhodné, neboť přirozený jazyk má velmi komplexní strukturu a je jí pomocí pravidel obtížné popsat. Tyto klasifikátory jsou nejčastěji založené na:

- Rozhodovacích stromech (Ševčíková a kol., 2007a).

- Modelu maximální entropie (Konkol a Konopík, 2011), (Straková a kol., 2013).
- Support Vector Machines (Kravalová a Žabokrtský, 2009).
- Conditional Random Fields (Konkol a Konopík, 2013).

Pro trénování klasifikátoru je potřeba určit atributy, podle kterých budou NE detekovány. Na základě těchto atributů klasifikátor určí, zda a o jaký typ NE se jedná.

Těmito atributy v češtině např. mohou být:

- velká písmena na počátku slova,
- celé slovo je velkými písmeny,
- okolí slova – před jménem osoby se často budou vyskytovat slova pan, paní, prezident apod.,
- četnost výskytu slova,
- délka slova.

Další typy příznaků naleznete v (Ševčíková a kol., 2007b) a (Král, 2011).

Jiným přístupem, vhodným zejména pokud není dostupný rozsáhlý trénovací korpus, je metoda využívající slabého učení s učitelem. Využívá se metody *bootstrappingu*, která umožňuje systému z několika počátečních anotovaných výrazů odvodit další. Například z výrazu „New York“, který bude označen jako město, systém odvodí, že se nejčastěji vyskytuje v kontextu „město New York“. Z toho se následně vyvodí, že v kontextu „město Praha“ je „Praha“ též NE. Při další iteraci jsou tyto výrazy přidány k původním a postup se opakuje (Nadeau a Sekine, 2007).

Metody vyhodnocování rozpoznání pojmenovaných entit

Úspěšnost NER se vyhodnocuje pomocí hodnot přesnost, pokrytí a F-mírou (viz část 2.2). Pro vyhodnocení, zda je rozpoznaná entita skutečně správná, existuje několik metod (Nadeau a Sekine, 2007). Jsou jimi :

- *Metoda vyhodnocování dle MUC* hodnotí výstupy systému dvěma hodnotami, které určují správnost určení hranice NE a zda se jedná o správný typ entity. Výhodou této metody je, že bere v potaz všechny možné typy chyb a částečně je ohodnocuje.
- *Metoda přesné shody* považuje za správné pouze přesné rozpoznání NE. Systém tedy musí správně určit její hranice i typ, jinak je entita považována za chybně rozpoznanou. Tuto metodu hodnocení používají IREX a CoNLL. Nevýhodou je,

že částečně rozpoznané entity mohou být započteny jako dvě chyby. Jednak jako falešně pozitivní a dále jako falešně negativní[5].

- *Metoda vyhodnocování dle ACE* (Automatic Content Extraction) používá mechanismus, který umožňuje bodování částečných chyb, podobně jako metoda vyhodnocování dle MUC. Různé typy NE mohou mít odlišné váhy, které určují jejich významnost. Tato metoda má dva zásadní problémy. Pro porovnání systémů mezi sebou je nutné, aby byly váhy jednotlivých typů NE a další parametry zvoleny shodně. Dalším problémem je, že jejich nevhodnou volbou je velmi snadné výsledky výrazně zkreslit.

Pro měření úspěšnosti rozpoznávání NE je v práci použita metoda přesné shody. Tato metoda je použita i v pracích (Ševčíková a kol., 2007b) a (Straková a kol., 2013), což umožní snadné porovnání výsledků systémů. Tato metoda je též vhodná vzhledem k povaze dat. Je žádoucí, aby za správně rozpoznané byly brány pouze entity, u kterých se podaří rozpoznat a anonymizovat jejich celé jméno apod.

Dosahované výsledky v rozpoznávání pojmenovaných entit

Na úspěšnost vyhodnocování má významný vliv jak jazyk, ve kterém je přirozený text uveden, tak i zkoumaná doména. Nadeau a Sekine (2007) uvádí, že některé systémy mohou při přesunu na jinou doménu vykazovat snížení výkonosti systému o 20% – 40%. Jelikož na cílových datech nelze provést trénování, pokusíme se zjistit, zda k tomuto jevu dojde použitím jiného korpusu, než na kterém bylo provedeno trénování (blíže viz část 6.7).

Mezi nejčastěji zpracovávané jazyky patří angličtina. U té se dosahuje jedné z nejlepších úspěšností rozpoznávání NE. To je jednak dáno počtem prací, které se tomuto jazyku věnují a také nepříliš bohatou morfologií tohoto jazyka, což zpracování textů v tomto jazyku usnadňuje. Zpracování některých druhů jazyků může být komplikovanější. V čínštině nebo japonštině komplikuje určení hranic slov absence mezer mezi slovy. V češtině a dalších slovanských jazycích zpracování komplikuje bohatá morfologie, kdy lze ohýbáním slovo upravit do mnoha tvarů. Tato vlastnost znesnadňuje jak rozpoznávání pojmenovaných entit, tak identifikaci koreferencí (viz část 2.5). Dalším problémem, který se vyskytuje v českém jazyce, zejména pro zpracování koreferencí, představuje nevyjádřený podmět. Ten například angličtina nevyužívá.

Systémy pro NER, které dosahují nejlepších výsledků, jsou založeny na strojovém učení. Jak ukazuje tabulka 2.3, pro český jazyk dosahují tyto systémy hodnot F-míry v rozmezí 62-79%, pokud se berou v úvahu detailně rozdělené kategorie. Pokud jsou brány v úvahu pouze „nadtypy“ těchto kategorií potom 68-83%. Pro anglický jazyk se udává úspěšnost rozpoznávání 86-90% (Straková a kol., 2013).

[5]Jako příklad uveďme „Spojené království Velké Británie a Severního Irska“ oproti „Spojené království“. Z druhého názvu je též jasné, o který stát se jedná, ale i tak se chyba započte dvakrát.

Tab. 2.3: *F_1-míra rozpoznávání pojmenovaných entit pro český jazyk. Data byla převzata z (Straková a kol., 2013).*

Práce	**Typy**	**Nadtypy**
Straková a kol. (2013)	79,2%	82,8%
Konkol a Konopík (2011)	–	72,9%
Kravalová a Žabokrtský (2009)	68,0%	71,0%
Ševčíková a kol. (2007)	62,0%	68,0%

Obecně lze říci, že některé typy pojmenovaných entit lze rozpoznat snadněji nežli jiné. Nejsnadněji rozpoznávanými typy entit jsou často jména osob, data, či entity, které lze rozpoznat regulárním výrazem. Mezi problematické entity patří jména institucí a adresy (Konkol a Konopík, 2011; Král, 2011). Dalším problematickým jevem jsou víceslovné entity. U těch je hlavním problémem určení přesných hranic pojmenované entity. Straková a kol. (2013) uvádějí rozdíl F_1-míry jednoslovných a víceslovných entit v jejich práci 6%. Ševčíková a kol. (2007b) uvádí tento rozdíl 40%.

2.5 Identifikace koreferencí

Pojmenované entity se v přirozených textech mohou vyskytovat v různých tvarech a na různých místech v textu, přičemž se však může jednat o stejnou entitu. Tyto výrazy mohou být ve formě zkratek, zájmen, mohou vystupovat jako nevyjádřený podmět nebo být v jiném morfologickém tvaru (Piskorski a Yangarber, 2013). Aby bylo možné s NE dále pracovat, je nutné aby shodné pojmenované entity byly spolu náležitě propojeny.

Následující fiktivní příklad ukazuje, jak mohou v textech osoby vystupovat:

Dne 10.2.2016 bylo v Berouně Kamile Novákové, nar. 15.3.1990, odcizeno horské kolo. Nováková jeho cenu odhaduje na 20 tisíc Kč. S Novákovou byl na stanici sepsán protokol. Podezřelého Vojtka Ondřeje, nar 1.2.1986, **poškozená** při identifikaci rozpoznala. Kolo bylo dne 25.4.2016 předáno matce poškozené *Kamile Novákové*, narozena nar. 3.9.1965.

Podtržením je vyznačena hlavní entita v dokumentu „Kamila Nováková“. Lze předpokládat, že ve zkoumaných dokumentech budou osoby nejprve identifikovány celým jménem, dále však už mohou být označovány pouze jménem nebo příjmením (např. „Kamila“, „Nováková“, „Novákovou“ apod.). Tučně označené slovo je též vázáno na entitu „Kamila Nováková“, nejedná se však o pojmenovanou entitu a v našem případě se je nebudeme snažit propojit. Zájmena ani nevyjádřené podměty není nutné anonymizovat a ve vytvořené sociální síti se entity, které na ně odkazují stejně objeví. Jediným omezením, které tímto vznikne je, že se nebude brát v úvahu jejich několikanásobný výskyt v analyzovaném dokumentu (viz část 6.5).

Dalším problémem, ten je vyznačen kurzívou, může být výskyt jmenovců v textech. Nejčastějším mužským jménem v ČR[6] je „Jiří“ s výskytem cca 300 tisíc a nejčastějším mužským příjmením je „Novák“ s výskytem cca 34 tisíc. Budeme-li přibližně uvažovat 5,26 milionů mužů dává tento nejhorší případ cca 1950[7] jmenovců „Jiřích Nováků“. Pro omezení tohoto jevu bude využito data narození, které se za jménem osoby může objevit.

2.6 Extrakce relací mezi entitami

Jednotlivé entity mohou mít mezi sebou různé relace. Může se například jednat o relaci „pracuje“ mezi entitami „Petr“ a „IBM“. Smysl těchto relací může být kladný, záporný nebo neutrální. Relace v předchozím případě je kladná, naopak relace „okradl“, „zabil“, „napadl“ apod. je relace záporná.

V této práci se omezíme pouze na identifikaci relace. Pokud se spolu entity vyskytnou v textovém dokumentu, budeme předpokládat, že spolu v nějakém smyslu souvisí a mají mezi sebou relaci. Toto zjednodušení spolu však nese jistá rizika, na která bude nutné při vyhodnocování výsledků pamatovat. Mezi osobami, které budeme dále zkoumat a určovat u nich významnost v kriminální síti, se může například jako nejvlivnější osoba objevit kriminalista, pokud bude v textech podepsán nebo uveden jako vyšetřující. Řešením tohoto problému je odebrání této entity nebo její odfiltrování z výsledků. Druhým problémem je, že nerozpoznáme kladné a záporné relace. Např. relaci kradl a byl okraden. Pokud bychom se chtěli tomuto problému vyhnout, bylo by nutné provádět sémantickou analýzu.

[6] Dostupné např. na `http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx`.

[7] Vypočteno jako $P(Jmeno) \cdot P(Prijmeni)$, přičemž zanedbáme možnou závislost jména na příjmení. Pak $0,057 \cdot 0,0065 \cdot 5,26M = 1948$.

3 Analýza sítí

Tato kapitola popisuje pojmy, které jsou dále v práci používány ve spojitosti s analýzu sociálních, respektive kriminálních sítí. Praktickou aplikací analýzy sociálních sítí na dostupná data se pak zabývá část 4.4 a kapitola 7. Úvodní část této kapitoly, která vychází z (Čada a kol., 2004), popisuje základní pojmy z teorie grafů, ze kterých analýza sociálních sítí vychází. Dále v této kapitole naleznete několik metod, pomocí kterých je možné v síti určit významné aktéry, popis předpokládaného modelu sítě a postup, jak ji efektivně destabilizovat.

3.1 Pojmy z teorie grafů

Graf, jeho ohodnocení a orientace

Jako *graf* označujeme dvojici množin V a E, kde V je konečnou množinou a $E \subset \binom{V}{2}$, přičemž $\binom{V}{2}$ je množina všech neuspořádaných rozdílných dvojic prvků z množiny V. Prvky z množiny E nazýváme *hrany* grafu, prvky z V pak *vrcholy* nebo uzly grafu. Jelikož jsou v tomto případě hrany neuspořádané dvojice, jedná se o *neorientovaný graf*. Neuspořádanou dvojicí $\{u, v\}$ značíme *neorientovanou hranu*. U neorientovaného grafu neuvažujeme směr hran a platí $\{u, v\} = \{v, u\}$. U *orientovaného grafu* naopak směr hran uvažujeme. U orientovaného grafu jsou hrany uspořádané dvojice vrcholů, které značíme (u, v) – jedná se o *orientovanou hranu*, přičemž platí $(u, v) \neq (v, u)$. Jako (u, v) značíme počátek hrany ve vrcholu u a její konec ve vrcholu v. Vrcholy u a v jsou *sousední*, pokud mezi nimi existuje hrana. Počet vrcholů značíme jako n a počet hran jako m.

Graf G označíme jako hranově *ohodnocený*, pokud je současně s množinami V a E definován pomocí reálné funkce w, kde $w : E(G) \to (0, \infty)$. Hodnotu hrany e, která je přiřazená funkcí $w(e)$, nazveme *ohodnocením* nebo *váhou* hrany e. Pokud graf pomocí funkce w definován není, nebo tato funkce přiřadí všem jeho hranám shodnou hodnotu, označíme graf jako *neohodnocený*.

Nejkratší cesta

Cestou z vrcholu u do vrcholu v označujeme libovolný přechod z $u = v_0$ do $v = v_k$, kde v_i je vrchol grafu. Tento přechod je možný pouze mezi sousedními vrcholy a je realizován pomocí hran $\{v_0, v_1\}$,$\{v_1, v_2\}$ ... $\{v_{k-1}, v_k\}$. Pro orientovaný graf obdobně pomocí (v_0, v_1),(v_1, v_2) ... (v_{k-1}, v_k), přičemž respektujeme směr hrany. V tomto přechodu se může každý vrchol v_i objevit právě jednou. Číslo k udává *délku* této cesty. *Nejkratší cestou* mezi vrcholy u a v rozumíme cestu, jejíž ohodnocení (tj. součet ohodnocení hran, které byly pro cestu použity), je ze všech existujících cest mezi

vrcholy u a v minimální. U ohodnoceného grafu nemusí být nejkratší cesta shodná s cestou přes nejméně vrcholů.

Souvislost grafu a jeho komponenty

Graf označujeme jako *souvislý*, pokud pro každé vrcholy u a v existuje v grafu alespoň jedna cesta. V opačném případě je graf *nesouvislý*. U orientovaného grafu rozlišujeme slabou a silnou souvislost. Jako *silně souvislý* graf označíme orientovaný graf, ve kterém pro každé vrcholy u a v existuje jak cesta z u do v, tak i z v do u. U *slabě souvislého* grafu platí pouze jeden z těchto případů. Pokud je graf silně souvislý, je zároveň i slabě souvislý.

Jako *komponenty* $K_1, K_2, ..., K_i$ označujeme maximální souvislé podgrafy grafu G, přičemž pokud $i \neq j$ platí $K_i \cap K_j = \emptyset$. Pokud mezi u a v neexistuje cesta, leží každý v jiné komponentě a nejkratší cesta mezi nimi je nekonečná. Pokud je graf souvislý, má právě jednu komponentu.

Stupeň vrcholu

Jako *stupeň* vrcholu v označujeme počet hran, pro které platí $\{u, v\}$, (u, v) nebo (v, u) a značíme ho jako $d_G(v)$. U orientovaného grafu můžeme rozlišovat výstupní stupeň, který se značí jako $d_G^-(v)$ a udává počet výstupních hran, tj. hran (v, u). Pro počet vstupních hran, tj. hran (u, v), používáme značení $d_G^+(v)$. V případě váženého grafu můžeme uvažovat nejen počet hran, ale i jejich váhu.

3.2 Analýza sociálních sítí

Sociologové definují *sociální síť*[1] jako množinu aktérů, kteří mezi sebou mohou mít sociální vztah (Buštíková, 1999). Jako *aktéři* mohou vystupovat entity, které je možné na základě jejich sociálních vztahů propojit. Může se tedy jednat např. o osoby, organizace, státy nebo komunity. Samotné *sociální vztahy* mohou představovat přátelství, citové vztahy, obchodní partnerství, příslušnost do stejného klubu, komunikační kanály, příbuznost, vztahy moci nebo autority atd. Tyto vztahy mohou být ohodnocené a mít kladný, neutrální, nebo záporný význam – ve smyslu „mám rád“, „považuji za nepřítele“ apod. (Everton, 2008; Buštíková, 1999).

Sociální síť často bývá vizualizována pomocí grafu ve smyslu, jakým jej chápe teorie grafů (viz část 3.1). Tento diagram budeme nazývat *sociogram* (Hanneman a Riddle,

[1] Je třeba poznamenat, že v kontextu této práce pojem síť neodpovídá pojmu jak ho definuje Ryjáček (2014). Ten popisuje síť ve smyslu analýzy toků v sítích (flow network) jako: „*Síť je orientovaný graf* $\vec{G}$ *s ohodnocením hran* $r : H(\vec{G}) \longrightarrow (0; \infty)$ *a ohodnocením uzlů* $a : U(\vec{G}) \longrightarrow R$“. Nicméně toto při analýze sociálních sítí nemusí být splněno. Analýzu sítí je možné provádět i na neválených neorientovaných grafech (Hanneman a Riddle, 2005; Everton, 2008; Réka a Barabási, 2002). Pokud tedy dále budeme mluvit o sítích, budeme mít na mysli libovolný graf. Bylo by možné použít spojení „analýza sociálních grafů“, nicméně tento pojem se nepoužívá.

2005; Buštíková, 1999). Aktéři jsou v sociogramu zobrazeni jako vrcholy a jejich sociální vazby jako hrany grafu[2].

Analýza sociálních sítí se zabývá zkoumáním sociálních vazeb mezi aktéry a jejich vzory. Největší pozornost je směřována k vazbám, mocenské nadřazenosti aktérů, zkoumání slabých a silných vazeb a dynamiky sociálních sítí. (Buštíková, 1999; Hanneman a Riddle, 2005).

Makroskopické údaje o síti

Makroskopickým údajem sítě, který udává celkovou propojenost, je *hustota sítě*. Ta vyjadřuje poměr mezi existujícími a všemi možnými hranami v síti. Pro orientovanou síť se vypočte jako $\frac{m}{n(n-1)}$, pro neorientovanou pak jako $\frac{2m}{n(n-1)}$, kde m je počet hran a n je počet vrcholů v síti. Hustotu sítě značíme jako Δ. Její hodnota se pohybuje v intervalu $\langle 0, 1\rangle$. Dalším makroskopickým údajem sítě je *průměrný stupeň vrcholu*[3]. Značíme ho jako $\langle k\rangle$ a vypočteme ho jako $\frac{1}{n}\sum d_G(v_i)$. *Průměrná délka nejkratší cesty* (average path length) ℓ je definována jako průměrná délka nejkratší cesty, která se vypočte jako $\frac{1}{n(n-1)}\sum d(v, v_i)$, kde $d(v, v_i)$ je délka nejkratší cesty mezi vrcholy v a v_i, přičemž $v \neq v_i$.

Koeficient shlukování udává schopnost sítě tvořit shluky. Jako *shluky* označujeme skupinu vrcholů, která je mezi sebou intenzivně propojena. Shluky si můžeme představit jako skupinu spolužáků nebo přátel, kteří se mezi sebou všichni navzájem znají. Koeficient shlukování v neorientované síti pro vrchol v vypočteme jako $C_v = \frac{2E_v}{k_v(k_v-1)}$, kde E_v je počet vzájemně propojených sousedů a k_v je počet sousedních vrcholů vrcholu v. Pro orientovanou síť se vypočte jako $\frac{E_v}{k_v(k_v-1)}$. Koeficient shlukování pro celou síť se vypočte jako průměrný koeficient shlukování všech vrcholů a značíme ho jako C (Réka a Barabási, 2002).

3.3 Určení významných aktérů v sociální síti

Tato část popisuje několik metod, pomocí kterých je možné v síti odhalit zaznamenání hodné aktéry. Buštíková (1999) označuje tyto aktéry za mocensky nadřazené, Hanneman a Riddle (2005), jako *mocné*. V analýze sociálních sítí vychází koncept moci z množství vztahů v síti nebo důležitých vztahů s dalšími mocnými aktéry. Mocní aktéři, kteří mají v síti výhodné postavení, mají lepší přístup k informacím nebo ke zboží. Totéž platí naopak. Mohou informace nebo zboží efektivně šířit nebo je blokovat. Mají na méně mocné aktéry větší vliv a ti k mocným aktérům mohou vzhlížet (Hanneman a Riddle, 2005). Mocenská nadřazenost aktérů se v sociálních

[2]Dále tedy budeme pojmy „aktér“ a „vrchol“ považovat za synonyma, totéž bude platit pro pojmy „hrana“ a „sociální vazba“.

[3]V orientované síti můžeme obdobně definovat průměrný vstupní a výstupní stupeň vrcholu jako $\langle k_{in}\rangle$ a $\langle k_{out}\rangle$.

sítích často zkoumá pomocí hodnot centralit aktérů (Hanneman a Riddle, 2005; Buštíková, 1999). Centralitu lze měřit pomocí několika metod. Nejčastěji používanými jsou míry centralit, které představuje Freeman (1978). Patří mezi ně: „Degree“, „Closeness“ a „Betweenness“ centrality, které budou dále popsány. Další míry centralit, které na tyto metody navazují naleznete v Hanneman a Riddle (2005).

Degree centrality

Centralita měřená stupněm uzlu (degree centrality, C_D) určuje cenatralitu, jak název napovídá, na základě stupně vrcholu (viz část 3.1). Její hodnota pro vrchol v je prosté přiřazení $C_D(v) = d_G(v)$, kde $d_G(v)$ je stupěň vrcholu v. V případě orientované sítě můžeme rozlišovat vstupní a výstupní stupeň vrcholu (vypočtou se obdobně jako C_D, značíme je C_D^+ a C_D^-). Pokud bychom mezi sebou dvě sítě chtěli porovnávat, můžeme provést normalizaci hodnot pomocí podílu C_D s hodnotou $n-1$, kde n je počet vrcholů (Freeman, 1978).

Aktéři s vyšší hodnotou této centrality[4] jsou zvýhodněni možností volby jak v síti komunikovat či předávat zboží. Tím se zvyšuje jejich nezávislost na ostatních aktérech v síti. Nevýhoda této metody spočívá v tom, že nedokáže odhalit globálně významné aktéry, kteří nutně nemusí mít vysoký stupeň vrcholu, ale mohou např. propojovat dvě významné komunity (Hanneman a Riddle, 2005). Asymptotická složitost této metody je pro všechny vrcholy $O(n^2)$.

Closeness centrality

Centralita měřená blízkostí polohy (closeness centrality, C_C) určuje jak blízko se vrchol nachází k ostatním vrcholům. Aktér, který je k dalším aktérům blíže, může při komunikaci, či předání zboží využít méně prostředníků než aktér, který je vzhledem k dalším aktérům dále. Využívání prostředníků je nevýhodné kvůli možnému zdržení nebo zablokování informace či zásilky. Aktér s vyšší hodnotou této centrality tak sám může být vyhledávaným prostředníkem, což zvyšuje jeho význam a moc (Hanneman a Riddle, 2005). Výpočet hodnoty C_C pro vrchol v se provede pomocí vzorce

$$C_C(v) = \left[\sum_{v \neq v_i}^{n} d(v, v_i) \right]^{-1}, \tag{5}$$

kde $d(v, v_i)$ je délka nejkratší cesty z vrcholu v do vrcholu v_i. Z tohoto vzorce plyne, že pokud by se v síti vyskytovalo více komponent, její hodnota by byla u všech vrcholů nulová. Z tohoto důvodu počítáme její hodnotu pro každou komponentu zvlášť. Hodnotu C_C můžeme normalizovat vynásobením $n-1$ (Freeman, 1978). Oproti metodě C_D dokážeme pomocí C_C odhalit globálně významné aktéry. U ohodnocené sítě může být potřeba dle významu váhy hrany výpočet patřičně upravit, např. převráce-

[4]Označení vyšší hodnota centrality je relativní pojem, který závisí na pravděpodobnostním rozdělení stupně vrcholů v síti.

ním hodnot hran. Asymptotická složitost výpočtu hodnoty C_C pro všechny vrcholy je $O(n^3)$, pokud využijeme n krát Dijkstrův algoritmus (Čada a kol., 2004).

Betweenness centrality

Centralita měřená mezilehlostí (betweenness centrality, C_B) udává, jak často vrchol leží na nejkratší cestě mezi dalšími vrcholy. Efektivní komunikace v síti, či předávka zboží je zajišťována právě pomocí aktérů s vyšší hodnotou C_B. To platí zejména mezi téměř oddělenými komunitami, kde jsou právě tito aktéři klíčoví. Pokud se rozhodnou informace, či zboží blokovat, mohou jejich předání značně zdržet, nebo mu zcela zabránit. To jim dává moc jak dění v síti kontrolovat. Metoda C_B odhalí globálně významné aktéry obdobně jako metoda C_C (Hanneman a Riddle, 2005). Výpočet hodnoty C_B pro vrchol v vypočteme pomocí vzorce

$$C_B(v) = \sum_{v_i \neq v \neq v_j} \frac{g_{v_i v_j}(v)}{g_{ij}} , \tag{6}$$

kde $g_{v_i v_j}(v)$ je počet nejkratších cest[5] mezi vrcholy v_i a v_j, které procházejí vrcholem v a kde g_{ij} je celkový počet nejkratších cest mezi vrcholy v_i a v_j. Normalizace se provede vynásobením hodnoty C_B s $\frac{2}{(n-1)(n-2)}$ pro neorientovanou síť[6]. Pro orientovanou síť pak vydělením hodnoty C_B s $(n-1)(n-2)$ (Freeman, 1978). U vážené sítě je nutné brát ohled na význam ohodnocení hran, podobně jako u C_C a případně vypočítávat nejkratší cestu z převrácené hodnoty váhy hrany. Asymptotická složitost je opět $O(n^3)$. Bader a Madduri (2006) ukazují paralelní algoritmy pro výpočet C_C a C_B. Využívají vlastností řídkých sítí, pomocí kterých je možné výpočet urychlit. V našem případě však nepočítáme s tím, že by síť byla natolik rozsáhlá, aby bylo nutné k paralelizaci přistoupit.

3.4 Bezškálové sítě a jejich destabilizace

Réka a Barabási (2002) ukazují, že reálné komplexní sítě jsou často založeny na *bezškálovém modelu*. Ten je charakteristický tím, že několik vrcholů má několikanásobný stupeň vrcholu, než je průměrný stupeň vrcholu. Pravděpodobnostní rozdělení stupně vrcholů se řídí podle *mocninného zákona* (power-law). Toto pravděpodobnostní rozdělení je předepsáno jako $P(k) \sim k^{-\gamma}$, kde $P(k)$ značí četnost výskytu vrcholů stupně k a jako γ je označen exponent konektivity. Ten má v reálných sítích nejčastěji hodnotu $2 \leq \gamma \leq 3$ (Réka a Barabási, 2002). U orientovaných sítích můžeme zavést obdobně $P_{out}(k) \sim k^{-\gamma_{out}}$ a $P_{in}(k) \sim k^{-\gamma_{in}}$. Model bezškálových sítí dostatečně modeluje vlastnosti, které se v reálných sítích vyskytují, jako je malý poloměr sítě a existence shluků v síti (viz část 3.2).

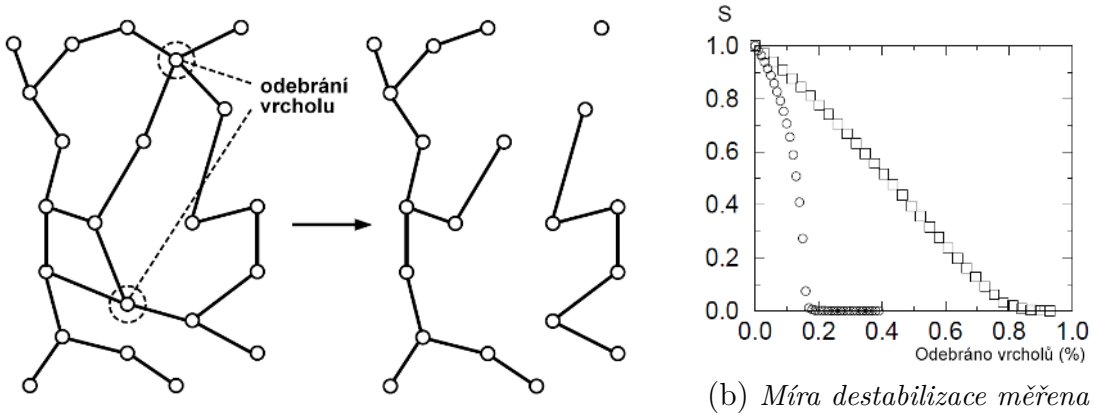
[5]Nejkratších cest mezi vrcholy u a v může být několik, přičemž musí mít stejné ohodnocení.
[6]Všechny nejkratší cesty jsou započteny dvakrát.

Jedna z důležitých vlastností těchto sítí je jejich odolnost vůči náhodným výpadkům. Bezškálové sítě jsou velmi robustní a i proto mají dle Evertona (2008) kriminální a teroristické organizace podobnou topologii. Pokud bude zkoumaná síť tvořena dle tohoto modelu, je neefektivní odebírat vrcholy bez předběžné analýzy. Na druhou stranu je tento model velmi náchylný na cílené útoky. Pokud by nám byla známa topologie bezškálové sítě, můžeme ji snadno destabilizovat.

Míru destabilizace můžeme měřit pomocí relativní velikosti největší komponenty, která zůstává po odebrání vrcholu propojená – v obrázku 3.1b značeno jako S. Druhým způsobem jak míru destabilizace měřit, je průměrná délka nejkratší cesty v síti ℓ. Po odebírání vrcholů v síti bude hodnota S klesat, zatímco hodnota ℓ bude, do doby než se síť začne rozpadat na menší komponenty, stoupat.

Pro vyhodnocování destabilizace budeme používat hodnotu S (viz obr. 3.1b), která je snadněji interpretovatelná a je výpočetně méně náročná. S vypočteme jako $|K_{max}| \,/\, |K_{max\,start}|$, kde $|K_{max}|$ je aktuální počet vrcholů v největší komponentě a $|K_{max\,start}|$ je počáteční počet vrcholů v největší komponentě. Hodnota S se tak může pohybovat v intervalu $\langle 0; 1 \rangle$ a začíná na hodnotě jedna.

Obrázek 3.1a ukazuje myšlenku cíleného odebrání vrcholů v síti tak, aby její topologii byla způsobena co největší škoda. Obrázek 3.1b ukazuje významný rozdíl mezi cíleným (∘) a náhodným (□) odebráním vrcholů z bezškálové sítě. Réka a Barabási (2002) při cíleném odebírání míří na vrcholy s největším stupněm, což odpovídá metodě C_D (viz část 3.3). V části 7.3 bude ukázána destabilizace dle C_D a C_B.



(a) *Cílené odebírání vrcholů ze sítě.*

(b) *Míra destabilizace měřena pomocí hodnoty S.*

Obr. 3.1: *Ukázka destabilizace bezškálové sítě. Převzato z Réka a Barabási (2002).*

4 Šablony zločinného chování a kriminální sítě

Tato kapitola popisuje metody používané pro analýzu a predikci kriminality, které umožňují určení nejpravděpodobnějších míst a časů, na nichž se trestný čin odehraje. Tyto metody je možné uplatnit jak na extrahované informace z přirozených textů (viz kapitola 2 a 6), tak i na strukturovaná data. Praktické provedení některých metod a návrh pro další zpracování naleznete v kapitole 7. Kapitola vychází z (Clarke a Eck, 2010) a (Hruška a kol., 2015).

Část 4.4 popisuje aplikaci analýzy sociálních sítí na kriminální sítě. Pomocí té je možné analyzovat osoby vystupující v případech organizovaného zločinu[1] nebo teroristické skupiny.

4.1 Analýza dat při zajišťování bezpečnosti

Trestný čin není náhodný, je buď plánovaný nebo příležitostný. Ke spáchání zločinu dojde, pokud se potencionální pachatel a vhodný cíl ocitne ve stejné době na stejném místě za absence ochránce – např. policisty, bezpečnostního zámku apod.

V moderním pojetí se kriminologie nezaměřuje na pochopení psychologických a společenských vlivů kvůli kterým se z lidí stávají pachatelé, ale na samotný trestný čin a *šablony trestných činů* (crime patterns). Šablonou trestného činu mohou být častá místa, kde k trestným činům dochází, typy obětí, jenž jsou napadány, produkty, které jsou odcizovány apod. Hledají se způsoby jak minimalizovat příležitosti k páchání zločinu, jak zvýšit riziko dopadení pachatele a jak zločin předpovídat.

V srpnu 2015 byla ukončena veřejná zakázka ministerstva vnitra „Mapy Budnoucnosti“[2] s plněním 3 miliony korun. Cílem této zakázky byl průzkum metod, postupů a softwarových řešení, které jsou v zahraniční využívány pro analýzu a predikci kriminality. Zkušenosti byly čerpány ze Spojených států, Velké Británie, Německa a dalších zemí z EU. Výstupem zakázky jsou dva sborníky[3] a (Hruška a kol., 2015) shrnující získané poznatky.

[1] Organizovaný zločin či zločinné spolčení můžeme definovat jako: „*strukturovaná a hierarchizovaná skupina (organizace) vzniklá za účelem ziskové trestné činnosti, která funguje na základě dělby funkcí a úkolů.*“, viz `www.mvcr.cz/soubor/zlocinne-spolceni.aspx`.

[2] Celý název veřejné zakázky je „Mapy budoucnosti – moderní nástroj ke zvýšení efektivity a kvality výkonu veřejné správy v oblasti prevence kriminality založený na analýze a predikci kriminality“. Registrační číslo projektu: CZ.1.04/4.1.00/B6.00041. Bližší informace naleznete na `http://esfcr.cz/projekty/mapy-budoucnosti-moderni-nastroj-ke-zvyseni-efektivity-a`.

[3] Dostupné na: `http://prevencekriminality.cz/projekty/mapy-budoucnosti`.

V policejní praxi má *analýza kriminalistických dat* nezanedbatelné místo. Pomocí těchto analýz jsou identifikovány a interpretovány vzory trestných činů a na jejich základě mohou být například plánovány hlídky, navrhovány legislativní změny, zavedeny preventivní opatření nebo může být vyšetřovatelům doporučeno na jaké zájmové osoby se mají zaměřit.

Tyto analýzy jsou prováděny většinou manuálně, ale vzhledem k velkému objemu relevantních dat, které na lidskou paměť působí, může snadno dojít ke zkreslení výsledků. Dalšími problémy manuálních analýz je dostupnost lidských zdrojů, administrativní zátěž a fluktuace policistů s místní znalostí. S růstem dostupných dat a rozvojem pokročilých algoritmů se nabízí použití automatické tvorby těchto analýz a predikcí. Je nutno předeslat, že tyto systémy nemohou a nemají nahradit úsudek zkušených policistů, ale urychlit a usnadnit jim práci.

Hruška a kol. (2015), Clarke a Eck (2010) ukazují, že je možné z informací, které jsou získány z předešlých zločinů, *predikovat* budoucí zločiny. Výstup této predikce pak může sloužit jako podklad při rozhodování, jak efektivně využít dostupné zdroje pro řešení trestných činů, prevenci kriminality a přispívat tím ke zlepšení stavu bezpečnosti.

4.2 Prostorové analýzy

Ukazuje se, že pořekadlo „blesk nikdy neudeří dvakrát do stejného stromu“, v případě trestných činů, neplatí. U míst, nebo jejich okolí, která se jednou stala obětí trestného činu, je zvýšené riziko výskytu dalšího trestného činu. Tento jev je možné zkoumat pomocí metod *prostorové analýzy* (spatial analysis). Tato část některé z metod prostorové analýzy představuje.

Opakovaná viktimizace

Metoda *opakovaná viktimizace* (repeat victimization[4]) vychází z předpokladu, že pokud se vyskytne na některém místě trestný čin, má toto místo a okolí tohoto místa v blízké budoucnosti zvýšenou pravděpodobnost výskytu dalšího trestného činu. K opakování dochází v rychlém sledu – často do týdne. Platí to zejména pro vloupání, k jehož predikci se tato metoda nejčastěji používá. Uvádí se, že na 4% populace dojde ke spáchání zhruba 40% vloupání. V komerčních prostorách dochází dokonce v 9% podnicích zhruba k 90% trestných činů.

Je to odůvodňováno následujícím:

- Znalostí prostředí, kterou po prvním spáchaném činu pachatel o místu získá. Pachatel získá další informace o dané oblasti a může je při opakovaném trest-

[4]V angličtině se objevují pojmy „repeat victimzation“, „near-repeat victimisation“ a další mezi kterými pro zjednodušení nebudeme rozlišovat.

ném činu využít. Případně o nich může vypovědět dalším lidem, kteří tuto informaci mohou též využít.

- Silná přitažlivost daného místa, či jejich zranitelnost. Může se jednat o objekt vykazující špatné zabezpečení nebo se v něm nachází žádané zboží.

Pachatele mohou rovněž lákat předměty, které na místě při prvním vloupání v minulosti zanechal. Což je faktická kombinace obou výše uvedených případů. Uvádí se, že 76% pachatelů, kteří se dopouštějí vloupání, se k opakované viktimizaci přiznává. Tito pachatelé často mají delší trestní záznamy. Je tak pravděpodobné, že se podaří objasnit sérii trestných činů a odhalit recidivisty páchající tyto činy.

Opakované činy se často týkají obětí s podobnými rysy, jaké měl jejich prvotní cíl. Při vloupání v jedné čtvrti je tedy vhodné, věnovat této čtvrti zvýšenou pozornost. Pro predikci opakované viktimizace existuje volně dostupný program Near Repeat Calculator[5], který některé policejní složky v zahraničí používají.

Opakovaná viktimizace poukazuje na skutečnost, která se zprvu jeví jako nedostatek naší reprezentace kriminální sítě, ve které nebudou rozpoznávány pachatelé a oběti (viz část 2.6). Pokud se bude některý aktér často vyskytovat v záznamech trestných činů a tím se stane v kriminální síti významným, je vhodné mu věnovat zvýšenou pozornost i když nepůjde o pachatele, ale bude se jednat o oběť. Je totiž pravděpodobné, že se v budoucnu může stát obětí trestného činu znovu. Je na místě zjistit, z jakého důvodu je tato oběť častým cílem a pokusit se zjednat nápravu.

Místa se zvýšenou kriminalitou

Jako *místa se zvýšenou intenzitou jevu* (hot spots), se označují statisticky významné oblasti, ve kterých se zkoumaný jev intenzivní. V našem případě je to oblast, ve které se zločin často opakoval a kde v budoucnosti hrozí zvýšené riziko, že k němu dojde znovu. Výstupem této analýzy jsou statisticky významná místa, která jsou prezentována formou snadno čitelné mapy. Nevýhodou této metody je, že pro její správnou funkčnost je potřeba mít dostatečné množství dat. V opačném případě může dojít ke zkreslení výsledků.

Clarke a Eck (2010) uvádějí, že hot spoty často vznikají na místech kde se shromažďuje velký počet lidí jako nákupní centra, dopravní uzly apod. Dále na místech, kde je slabý dohled nad dodržováním pravidel (např. nemonitorované parkoviště). Do těchto míst pak policisté mohou soustředit hlídky a pokusit se tím kriminální jevy omezit. Pokud se ukáže, že některá místa jsou stálými hot spoty, je vhodné provést další analýzu a určit z jakého důvodu a pokusit se tento jev odstranit, například přidáním bezpečnostních kamer, přídavných zámků apod.

[5] `http://www.cla.temple.edu/cj/center-for-security-and-crime-science/projects/nearrepeatcalculator`

K určení statisticky významných míst existuje několik metod. Tou nejjednodušší může být porovnání s podobným, ale *kriminálně neutrálním*, tj. místem, které kriminalitu nepřitahuje. V kriminálně neutrálních místech je počet trestných činů relativně nízký, jsou izolované a nedochází k nim na základě vzorce. Oproti neutrálnímu místu můžeme provést porovnání a určit, zda je zkoumané místo z pohledu hot spot analýzy významné. To můžeme učinit vydělením počtu trestných činů s potencionálním počtem cílů. Pokud je výsledný poměr větší než poměr u kriminálně neutrálního místa, zobrazíme toto místo na mapě[6].

Dalšími metodami jak významné oblasti určit mohou být:

- Určení směrodatných odchylek – nejprve se vyhledají místa s největší koncentrací bodů, ze kterých se vytvoří shluky. Pro ty se následně vypočítají směrodatné odchylky, které polohou, protáhlostí a sklonem určují rozmístění a intenzitu sledovaného jevu.
- Mapování na administrativní území – trestné činy jsou agregovány podle administrativních území, následně normalizovány na relativní hodnoty dle počtu obyvatel a dle výsledné hodnoty jsou administrativní území v mapě obarveny.
- Kvadrátová metoda – mapa je rozdělena na části dle mřížky, v té jsou trestné činy agregovány a následně obarvením buňky v mřížce zobrazeny.
- Jádrové odhady (kernel density estimation) – nad mapou se vytvoří mřížka, v každé buňce mřížky se provede součet příspěvků jednotlivých trestných činů v závislosti na vzdálenosti a váze této události. Buňky mřížky jsou dle tohoto výpočtu obarveny a následně zobrazeny obdobnou formou jako tepelné mapy.

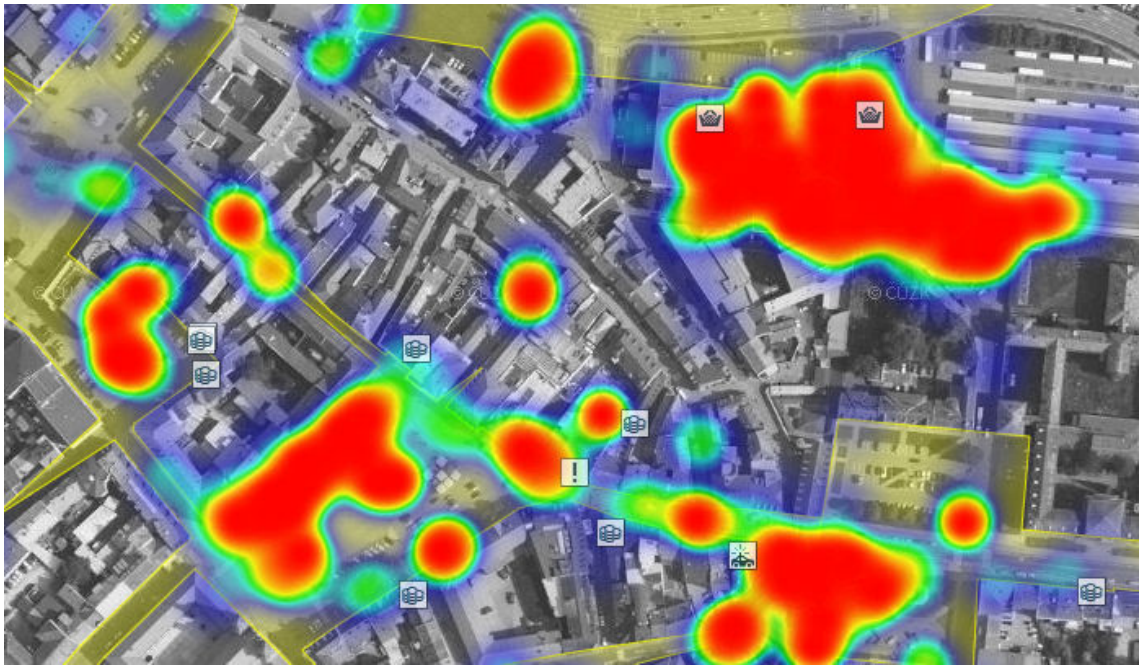
Tepelné mapy

Jinou prostorovou metodou jsou *tepelné mapy*. Jejich výstup je podobný jádrovým odhadům hot spotům, nicméně nelze je interpretovat stejně. Při metodě tepelných map se vizualizuje hustota výskytu jevu na daném území, přičemž nemusí být statisticky významná, což snižuje objektivitu této metody. Ač je interpretace tepelných map subjektivní záležitostí, je významným prvkem pro rozeznání míst se zvýšenou kriminalitou. Zvláště pokud pro metodu hot spotů nemáme dostatek dat. Na základě této metody lze plánovat rozmístění hlídek policistů.

Ukázku tepelné mapy můžete vidět na obrázku 4.1, kde jsou teplými barvami zobrazeny místa se zvýšeným počtem přestupků[7]. V této mapě jsou dále šedivým čtvercem vyznačeny body, které by mohly být využity pro analýzu rizikových oblastí (viz dále). Všimněte si zvýšené koncentrace vpravo nahoře. Jedná se o dvě nákupní centra, což může ukazovat na vysoké počty pokusů o krádež zboží.

[6] Hruška a kol. (2015) ukazují zanesení všech míst s trestnými činy do mapy, lze však namítnout, že tato vizualizace nezobrazuje statisticky významné místa, což je jednou z výhod hot spot analýzy.

[7] Jedná se o systém používaný městskou policií, proto zde nalezneme pouze přestupky.



Obr. 4.1: *Ukázka tepelné mapy zaznamenaných přestupků z portálu „e-Analýza bezpečnosti“*[8].

Analýza rizikových oblastí

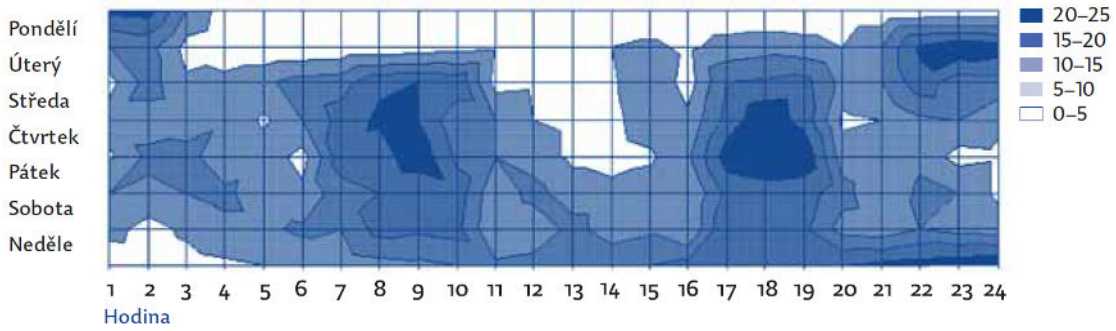
Metoda *rizikových oblastí* (risk terrain modeling, RTM) se zaměřuje na identifikaci geografických prvků, které ke zvýšenému riziku kriminality vedou. Tato metoda tak odpovídá nejen na otázku kde, ale i proč. Forma výstupu je shodná s metodami hot spotů a tepelných map formou spojitého zvýraznění intenzity jevu. Model analýzy rizikových oblastí pracuje s prvky, které mají silnou asociaci ke kriminálním činům, jako jsou bary, zastavárny, obchody, přítomnost podmínečně propuštěných jedinců apod. Oblasti jsou rozděleny podle mřížky, ve které jsou tyto prvky následně sčítány a na základě tohoto výpočtu se poté provede obarvení.

Hlavní odlišností od metody hot spotů je možnost predikce zločinu na místech, která jsou podobné místům, na kterých se trestné činy odehrávají. Přičemž se na těchto místech nemusejí objevovat, ale mohou se zde vyskytnou až v budoucnu. Není nutné mít přesné rozsáhlé data historických trestných činů. Z předešlých dat je však potřeba určit asociace, které se ke konkrétním trestným činům vážou. V obrázku 4.1 jsou některé významné body, ač nejsou pro tvorbu této tepelné mapy použity, vyznačeny.

[8] `http://analyza-bezpecnosti.tmapserver.cz/analyza-bezpecnosti`

4.3 Časová analýza

Významným faktorem, který s výskytem trestných činů souvisí jsou denní a týdenní rytmy. Ať už se jedná o špičky v dopravě, nebo páteční či sobotní noci, kdy lidé vyrážejí za zábavou. Tyto události mají společné to, že se v těchto časech zvyšuje počet potencionálních obětí. Určení těchto rytmů a dalších informací, na základě času nebo data, budeme dále nazývat jako *časovou analýzu*. Základním nástrojem pro určení důležitých rytmů je sestavení tabulky dnů, hodin a počtu trestných činů. Poté na základě této tabulky můžeme provést výpočet aritmetických průměrů a výsledky graficky znázornit. Dalším nástrojem je grafické zobrazení denního a týdenního rytmu do jednoho grafu. To je vhodné pokud se tyto rytmy v jednotlivých dnech mohou lišit. Jak tento graf sestrojit můžete vidět na obrázku 4.2. Buňky v grafu jsou obarveny podle počtu trestných činů v danou dobu.



Obr. 4.2: *Ukázka sestrojení grafu denních a týdenních rytmů na základě dne a hodiny. Převzato z Clarke a Eck (2010).*

Časovou analýzu je vhodné provádět zejména u často se opakujících jevů. Například u přestupků nebo dopravních nehod. Při aplikování na méně časté jevy, jako jsou např. vraždy, je nutné použít delší časovou řadu, aby nedocházelo ke zkreslení výsledků. V tomto případě je však nutné mít na paměti, že se tento rytmus může časem měnit. Při časové analýze je nutné znát dobu, kdy ke spáchání trestného činu došlo.

Obdobně je možné analyzovat dlouhodobé trendy. Například měsíční. Touto analýzou je např. možné pozorovat, zda zavedené preventivní opatření má dopad na zmírnění problému, kvůli kterému bylo zavedeno. Pokud bychom se snažili rozpoznat měsíční cykly, můžeme tak učinit porovnáním stejných měsíců, či týdnů v jednotlivých letech. Jelikož data mohou být zatížena náhodnými výkyvy, může se pro jejich odstranění použít metoda klouzavého průměru. Tato metoda vypočte aktuální hodnotu jako průměrnou hodnotu z aktuální hodnoty a z x předchozích hodnot.

Samotné časové analýzy mohou vést k odhalení možných příčin problému, který k trestnému činu vede. Dále mohou prozradit, kdy je potřeba nasadit více policistů a kdy je naopak jejich nasazení do ulic neefektivní. Časová analýza se často

kombinuje s prostorovou analýzou, například ve formě hot spotů, do časoprostorové analýzy. Pomocí té získáme další informace o tom, kdy a kde se zločiny nejčastěji vyskytují a podle toho můžeme cílit policejní hlídky nebo zavést další preventivní opatření.

4.4 Kriminální sítě

Odlišným přístupem oproti dříve uvedeným je *analýza kriminálních sítí*[9]. Tuto metodu popisují Everton (2008), Xu a Chen (2005), Krebs (2002) a Al-zaidy a kol. (2012). Jako *kriminální síť* budeme označovat sociální síť (viz část 3.2), ve které jsou aktéři účastníci kriminálních činů, podezřelí, vyslýchané osoby či jiné osoby, které jsou s trestným činem spojeni. Povšimněte si, že se jedná o účastníky, nikoli jen o pachatele. Jednak je to z důvodu, že při rozpoznávání pojmenovaných entit v textech momentálně nedokážeme rozpoznat oběti či jiné účastníky od pachatelů (viz část 2.6). Dále je to proto, že i odhalení několikanásobné oběti, která se tak stane v síti významnou, může přinést užitečné informace (viz část 4.2).

Jako první navrhuje Sparrow (1991) použití analýzy sociálních sítí na kriminální sítě. Popisuje přínos vytvoření sítě z dat jako je spatření podezřelých pohromadě, jejich telefonních hovorů, finančních transakcí apod., dále přínosy vizualizace této sítě a určení centrálních aktérů, na které se zaměřit. Hlavními otázkami, na které by analýza kriminální sítě měla pomoci dát odpověď jsou:

- Který aktér je v síti významný?
- Odstraněním kterých aktérů síť co nejefektivněji destabilizovat?
- Pomocí kterých aktérů síť infiltrovat?
- Která spojení jsou často využívána a měla by být monitorována ?

Jak bude ukázáno v kapitole 6, na tyto otázky je možné odpovědět aplikováním metod, které byly představeny ve druhé kapitole. Sparrow (1991) zároveň popisuje hlavní problémy, s nimiž se analýza kriminálních sítí setkává:

- Nekompletnost – některé vazby nebo aktéři nejsou při analýze známy.
- Ohraničení kriminální sítě – koho do kriminální sítě zařadit a koho již ne.
- Dynamičnost – kriminální sítě se neustále mění. Aktéři se k síti připojují a jiní naopak odpojují. Dalším problémem je reakce na vnější podnět policie. Při vyslýchání nebo zatčení členů sítě se může struktura kriminální sítě záměrně změnit, aby probíhající vyšetřování bylo ztíženo (Everton, 2008).

[9]Kriminální sítě jsou v angličtině označovány jako: „dark networks“, „criminal networks“ nebo „covert networks“.

Analýza kriminálních sítí nenalezne uplatnění, pokud pachatel jedná samostatně, nebo pokud s dalšími kriminálníky nekomunikuje. V tomto případě ho nelze s nikým propojit. Tato analýza je tak vhodná především pro organizovaný druh zločinu jako je např. distribuce drog či rozkrývání teroristických skupin.

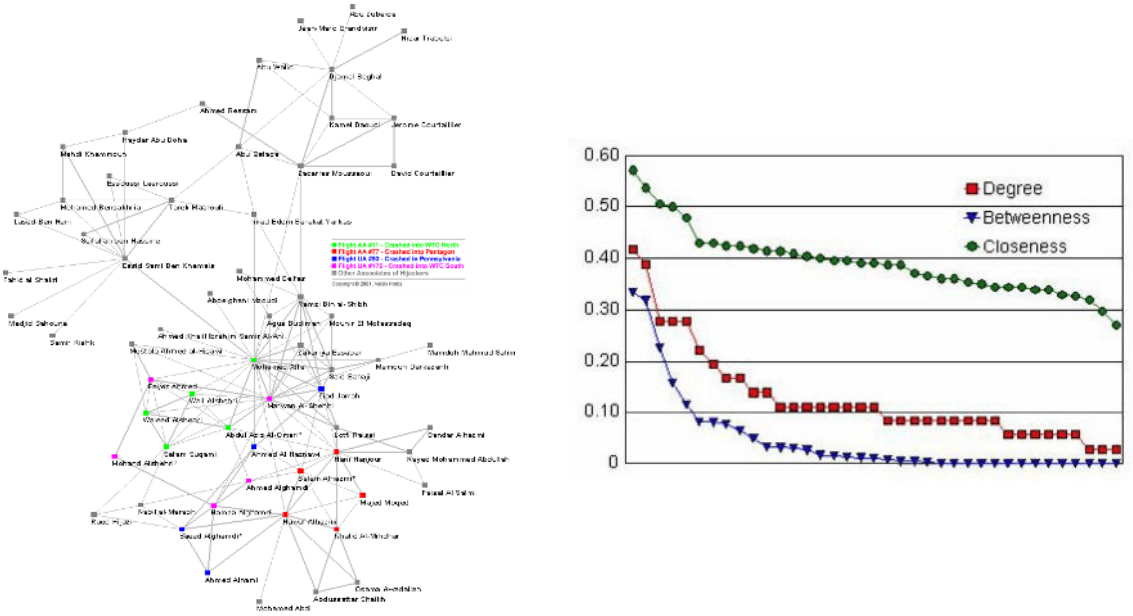
O analýze kriminálních sítí se ve větším měřítku začalo mluvit po událostech 11. září 2001. Tímto útokem se zabývá Krebs (2002). Ukazuje některé vlastnosti, které se speciálně v kriminálních sítí objevují. Kriminální sítě jsou velmi řídké. Krebs (2002) a Everton (2008) to odůvodňují snahou o minimalizaci škody, pokud bude nějaký její člen dopaden policií. V kriminálních sítích jsou často aktéři rozděleni do komunit, přičemž každá komunita je zodpovědná za specifický úkol. Tyto dvě vlastnosti však zároveň snižují efektivitu komunikace v síti a zvyšují její náchylnost na cílené útoky (viz část 3.4). Pokud by z této sítě byli odstraněni aktéři, kteří propojují oddělené komunity, je pro tyto komunity velmi těžké spolu komunikovat a synchronizovat další společný postup. Toto platí i opačně. Pokud odstraníme jednu z komunit, zasáhneme tím i další komunity, které mohou na odebrané komunitě být závislé[10] (Xu a Chen, 2005).

Pokud bychom k aktérům dokázali přiřadit jejich vlastnosti, můžeme se též snažit odstranit aktéry, kteří mají unikátní dovednosti. V případě událostí 11. září 2001 by to byli ti, kteří dokázali pilotovat letadlo (Krebs, 2002). Touto variantou se však dále nebudeme zabývat.

Krebs (2002) ukazuje použití metod C_D, C_C, a C_B (viz část 3.3) pro určení nejvýznamějších aktérů v síti. Tuto síť můžeme vidět na obrázku 4.3. Jedná se o teroristickou síť zodpovědnou za útoky z 11. září 2001. Obrázek 4.3 zároveň ukazuje distribuci centralit v této síti. Již z distribuce hodnot C_D se zdá, že tato síť odpovídá bezškálové topologii, která byla popsána v části 3.4. Nejvýznamnějším aktérem dle hodnot centralit byl Mohamed Atta. Toho za vůdce této sítě označil i bin-Ládin. Obecně se však může jednat o významného, nikoli nejvýznamnějšího, aktéra. Tyto metody jsou náchylné ke změně počtu vrcholů, přičemž některé, jak upozorňuje Sparrow (1991) a Krebs (2002), budou nepochybně chybět.

V našem případě budeme pracovat pouze s aktéry a hranami, které máme k dispozici a jsou nám známé. Ukážeme, jak síť efektivně destabilizovat (viz části 3.4 a 7.3) a jak nalézt aktéry, kteří spolu pravděpodobně spolupracují nebo jsou např. členy společného gangu. K tomu použijeme algoritmus pro odhalování prominentních komunit (viz části 4.5 a 7.4). Pro destabilizaci sítě použijeme cílené útoky. Kromě hodnoty C_D, kterou používají pro destabilizaci sítě Réka a Barabási (2002) použijeme hodnotu C_B. Ta identifikují globálně významné aktéry v síti. Což jsou

[10]Můžeme si to představit na příkladu drogového kartelu. Pokud by se podařilo zadržet aktéry, kteří mají na starost propojení varny a distribuce, nemohou tyto dvě komunity spolupracovat, dokud nenajdou náhradu. Totéž platí i pokud odstraníme pouze aktéry z varny nebo distribuční sítě, tj. jednu komunitu.



Obr. 4.3: *Ukázka analýzy kriminální sítě z útoků 11. září 2001 a aplikace metod pro výpočet centralit. Převzato z Krebs (2002).*

mocensky nadřazení aktéři, kteří by intuitivně měli být kandidáty na odstranění z kriminální sítě. Everton (2008) udává, že samotné odebrání nejvýznamnějšího aktéra nestačí, ba naopak může vést ke zhoršení problému. Odebraný aktér může být rychle nahrazen podobným. Kriminální síť následně zareaguje na vnější podnět a stane se ještě více skrytou a decentralizovanou. Odebírání aktérů navrhuje řešit i na základě topologie sítě. Pokud je v síti několik slabých vazeb, je vhodné je přerušit. Pokud je v síti pouze několik komunit, je vhodné je odstranit. Dále navrhuje infiltraci této sítě z více bodů a další. Některé z těchto postupů budou ukázány v kapitole 7.

4.5 Prominentní komunity a jejich detekce

Al-zaidy a kol. (2012) představují pojem prominentní komunity a způsob jakým je detekovat. Detekci prominentních komunit používají pro analýzu dat získaných ze zařízení (např. notebook, telefon apod.), které byly zabaveny policií pro další vyšetřování.

Jako *prominentní komunitu* označují uskupení dvou nebo více osob, které se společně vyskytují alejspoň v n‘ rozdílných dokumentech. Kde n‘ je uživatelem zadaný minimální společný výskyt těchto osob ve zkoumaných dokumentech. Dále budeme značit tento minimální společný výskyt jako *minSup* (z anglického minimal support). Tabulka 4.1 zobrazuje příklad dokumentů, na kterých bude ukázána detekce prominentních komunit. Komunitu označujeme jako *n-komunitu*, pokud je komunita tvořena n členy.

Tab. 4.1: *Příklad dokumentů, ve kterých budou zjišťovány prominentní komunity.*

Dokument	**Osoby**
1	{Petr, Vojta, Jan}
2	{Petr, Klára, Jiří}
3	{Jan, Petr, Klára}
4	{Jan, Petr, Vojta}

Pokud budeme určovat prominentní komunity s $minSup = 2$, tj. za prominentní komunitu považujeme osoby, které se společně nachází minimálně ve dvou dokumentech, za prominentní komunity budeme považovat {Petr, Jan, Vojta} a {Petr, Klára}. Pokud budeme uvažovat $minSup = 3$ pak pouze dvojici {Petr, Jan}.

Volba vhodného *minSup* je závislá na počtu osob a zpracovávaných dokumentů (je možné nastavit jako parametr, viz příloha A.4). Pokud bude zvoleno malé, bude uživatel zahlcen výsledky, pokud příliš velké, mohou uniknout zajímavá spojení, nebo nebude nalezena žádá prominentní komunita.

Detekce prominentních komunit

Algoritmus vychází z algoritmu Apriori, který se používá pro dolování frekventovaných množin z databází. Nejprve jsou průchodem dokumentů nalezeny všechny prominentní 1-komunity (tj. osoby, které se vyskytují alespoň v *minSup* dokumentech). Pokud budeme uvažovat $minSup = 2$ v tabulce 4.1 jimi budou {Petr}, {Vojta}, {Jan} a {Klára}. Pomocí kombinací z těchto 1-komunit jsou vygenerováni kandidáti na 2-komunity. V našem příkladu tedy {Petr, Vojta}, {Petr, Jan}, {Petr, Klára}, {Vojta, Jan}, {Vojta, Klára} a {Jan, Klára}. U těchto kandidátů je následně ověřeno, zda se společně v dokumentech vyskytují alespoň v *minSup* dokumentech. Pokud ano, je kandidát označen za 2-komunitu. Pokud ne, jsou tito kandidáti odstraněni. V příkladu jsou odstraněny komunity {Vojta, Klára} a {Jan, Klára}. Následně se z 2-komunit vygenerují 3-komunity. Ty je možné vygenerovat ze dvou 2-komunit, ve kterých se nachází shodná osoba. Tedy: {Petr, Vojta, Jan}, {Petr, Vojta, Klára} a {Petr, Jan, Klára}. Opět je provedeno ověření na výskyt v *minSup* dokumentech. Tento postup se opakuje, dokud je možné vytvářet další kandidáty. V našem případě zbude pouze 3-komunita {Petr, Vojta, Jan}, ze které již 4-komunity nelze generovat. Druhou největší komunitou, která není obsažena v 3-komunitě je {Petr, Klára}.

4.6 Další metody dataminingu

Další možnou metodou, která se dá použít pro predikci kriminality je dolování asociačních pravidel. Pomocí těch by bylo možné přidat dodatečné informace k časoprostorové analýze. Může se jednat například o počasí v době, ve které byl trestný čin spáchán, o místo vstupu do objektu, typ kradených aut a nejčastější místa jejich výskytu apod. Jelikož se ale cílová data nacházejí ve formě volného textu, nikoli ve formě strukturované, či semistrukturované, vyžadovalo by toto zpracování hlubší porozumění textu, což je komplexní problém. Samotné zpracování asociačních pravidel umožňuje například software Weka[11] či R project[12]. Analýzou asociačních pravidel z kriminalistických textů se zabývají Sathyadevan a kol. (2014) a Usha a Rameskkumar (2014).

Další technikou, kterou je možné na textové dokumenty použít, je klasifikace jednotlivých dokumentů. Pomocí té je možné klasifikovat dokumenty do kategorií, které mohou například představovat typy trestných činů. Pro klasifikaci textových dokumentů se často používá Bayesovský klasifikátor, který používají ve své práci Sathyadevan a kol. (2014), SVM či shlukování k-nejbližších. V případě použití učení s učitelem by však bylo nutné nejprve vytvořit trénovací množinu a u té označit, do které kategorie patří. To vzhledem k tomu, že cílové dokumenty obsahují osobní data a nejsou předem dostupné, momentálně není možné. Použití klasifikace bez učitele zde naopak nedává příliš smysl, neboť informace, že jsou si dva dokumenty vzájemně podobné, pokud nevíme z jakého důvodu, v našem případe nemá velké využití.

4.7 V praxi používané programy

Hruška a kol. (2015) popisují několik v zahraničí v praxi používaných programů. Jedná se například o program CrimeView Dashboard[13]. Ten policistům umožňuje identifikovat vzorce kriminality, analyzovat její stav a dle toho reagovat. To pomocí metod prostorových analýz, srovnávání dat mezi obdobími, vyhledávání informací o pachatelích a dalších. Obrázek 4.4 ukazuje prostředí tohoto programu. Tento program údajně již využívá několik stovek organizací. Jeho cena je udávána dle počtu obyvatel ve městě, přičemž se pohybuje v rozpětí 900 tisíc až 1,2 mil. Kč za první rok, za další roky pak 50 tisíc až 100 tisíc Kč (ceny jsou bez DPH). Dalším výdajem je přeložení programu do českého jazyka, které se pohybuje okolo 300 tisíc Kč. Cena za proškolení organizace, která by se dalšímu školení mohla věnovat, je udávána jako 250 tisíc Kč.

[11] `http://www.cs.waikato.ac.nz/ml/weka`

[12] `http://www.r-project.org`

[13] `http://www.theomegagroup.com/police/omega_dashboard_police.html`

Obr. 4.4: *Ukázka programu* CrimeView *ze stránek poskytovatele řešení.*

Dalším zajímavým programem je PredPol[14], který byl vyvinut pro odhalování zločinů proti majetku, drogové činnosti, činnosti gangů a ozbrojené trestné činnosti. Na vývoji se podílelo několik amerických univerzit. Program analyzuje nejpravděpodobnější místa, na kterých k budoucímu zločinu dojde metodami, které byly popsány dříve. Z procesu zpracování jsou vyloučena osobní data. Ta jsou brána přímo z policejní databáze. Oficiální ceny nejsou zveřejněny. Některé organizace využívají program zdarma, výměnou za data. Cena pro hrabství Kent (zahrnuje 1,4 mil. lidí) je údajně cca 4,75 mil. Kč za rok. Policejní oddělení v Seattlu (zahrnuje cca 640 tisíc lidí) má údajně roční licenci za cca 1,13 mil. Kč.

Mezi českými městy predikaci a preventivní opatření na základě softwarových produktů používají města Kolín, Uherské Hradiště a Pardubice. Ve všech případech se jedná o časoprostorovou analýzu, nejčastěji pomocí metody hot spotů. V Kolíně po nasazení tohoto systému klesla kriminalita oproti předchozím rokům o 40%. Pardubice a Uherské Hradiště nesledují dopady po zavedení systémů, chválí však pomoc při plánování tras hlídek. Tyto systémy byly nasazovány mezi roky 2013 - 2015. Data často pořizují policisté přímo na ulici pomocí chytrých telefonů či tabletů, což přináší výhodu přesných souřadnic trestných činů formou GPS souřadnic.

Většina zkoumaných míst ve světě, které podobné systémy využívají zaznamenává pokles kriminality často až o desítky procent. Velmi kladně jsou dále hodnoceny odpadající administrativní činnosti a usnadnění plánování hlídek (Hruška a kol., 2015). Vzhledem k cenám, legislativním povinnostem, českému jazyku, počtu obyvatel v ČR, a skutečnosti, že programy pracují se strukturovanými daty[15] může být pro pilotní řešení prediktivního systému kriminality v ČR vhodné řešení vlastní implementace, která bude spolupracovat s GIS systémem, který používá PČR.

[14] `http://www.predpol.com`

[15] Což vzhledem k tomu, že v ČR dochází k digitalizaci dokumentů a vytváření centrálních policejních databází, s kvalitními daty až v posledních letech, může představovat problém.

5 Architektura a použité technologie

Tato kapitola popisuje architekturu vytvořeného systému, použité technologie, knihovny, databázovou strukturu, proces zpracování a vizualizace dat. Bližší popis jednotlivých částí systémů, včetně použitých algoritmů a testování těchto částí naleznete v kapitolách 6 a 7.

Systém řeší rozpoznávání pojmenovaných entit, jejich anonymizaci a deanonymizaci, identifikaci koreferencí, extrakci relací (formou společného výskytu osob v dokumentu), simulaci destabilizace kriminální sítě pomocí metod centralit a detekci často se spolu vyskytujících osob. Na tu můžeme nahlížet jako na dolování frekventovaných množin v dokumentech. Systém dále generuje několik statistik a ukazuje možnosti prostorové a časové analýzy.

5.1 Použité technologie a knihovny

Aplikace je implementována v jazyce Java. Ten byl zvolen na základě výběru knihoven, skutečnosti, že obsahuje kolekce a stávajících zkušeností s tímto jazykem. Vizualizace výsledků je prováděna pomocí HTML stránek s použitím JavaScriptu. Data jsou ukládána do SQLite databáze (viz část 5.2) a XML souborů (viz část 6.4).

Program používá následující knihovny:

- juniversalchardet[1] – knihovna pro detekci kódování, ve kterém je zpracovávaný soubor uložen. Cílová data se mohou nacházet v kódování UTF-8 nebo Windows-1250.
- MorphoDiTa[2] – morfologický analyzátor pro české texty. Jeho úspěšnost je v (Straková a kol., 2014) pro úlohu tag uváděna jako jedna z nejvyšších – 95,75%. Oproti jiným morfologickým analyzátorům (např. Czech Morphological Analyzer and Tagger, Free morphological analyzer Majka nebo Morče - Czech Morphological Tagger) je až 10x rychlejší a jelikož jsou ohýbaná slova reprezentována jako trie, přičemž její části mohou být použity několika slovy se stejným základem, je výrazně snížena jak paměťová náročnost programu, tak velikost natrénovaného modelu.
- NameTag[3] – knihovna pro rozpoznávání pojmenovaných entit (NE). Obstarává prvotní rozpoznání NE, na které je dále navázáno (viz část 6.3) a tokenizaci.

[1] `https://code.google.com/archive/p/juniversalchardet`
[2] `http://ufal.mff.cuni.cz/morphodita`
[3] `http://ufal.mff.cuni.cz/nametag`

- GraphStream[4] – grafová knihovna, vhodná pro dynamické grafy. Umožňuje výpočet „Betweenness centrality“ (C_B) a dalších algoritmů, pomocí kterých simulujeme cílenou destabilizaci kriminální sítě (viz část 7.3). Jelikož se nepředpokládá síť větší než v řádech tisíců, maximálně deseti tisíců vrcholů, je možné použít tuto knihovnu. V opačném případě je při výpočtu C_B třeba přistoupit k paralelizaci. K tomu lze použít např. knihovnu SNAP[5]. Případně výsledky aproximovat. V části 7.3 je však ukázáno, že dostatečně přesně slouží i „Degree centrality“, která není výpočetně náročná.
- D3[6] – JavaScriptová knihovna používaná pro vizualizaci dat formou grafů.

5.2 Architektura systému a databázová struktura

Aplikace je rozdělena do následujících balíků:

- `anonymizer` – anonymiace NE,
- `db` – komunikace s databází,
- `model.ner` – reprezentace NE a tokenů,
- `model.output` – reprezentace výstupních dat,
- `model.sna` – reprezentace kriminální sítě,
- `output` – generování výstupů systému,
- `reader` – načítání textových souborů, XML souborů a konfigurací,
- `sna` – analýza kriminální sítě a její destabilizace,
- `taggers` – rozpoznávání NE a tokenizace.

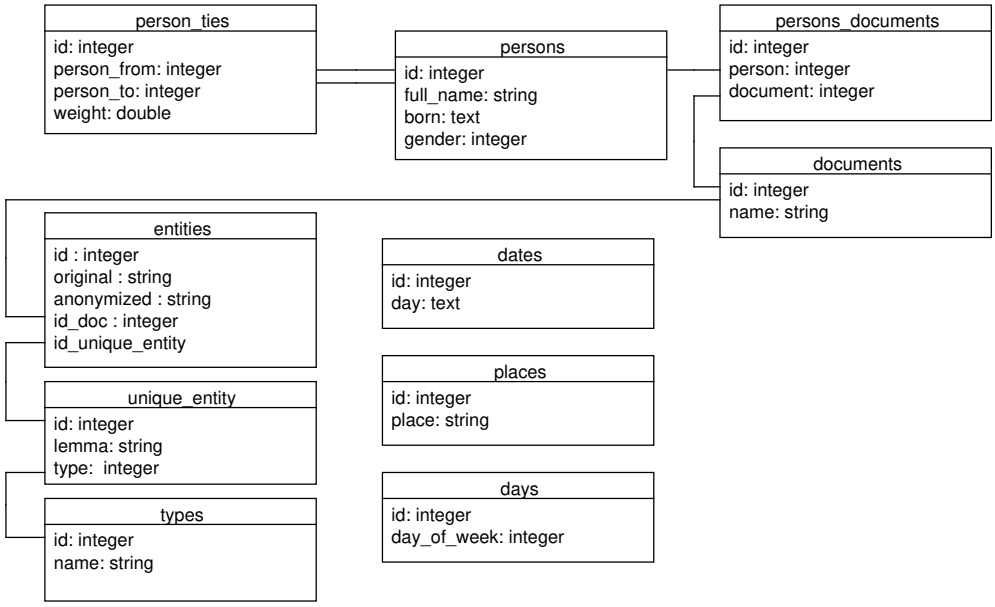
Detailní popis tříd a jejich metod naleznete v přiloženém javadocu či ve zdrojových souborech.

Ukládání dat

Data jsou ukládána do relační databáze SQLite[7] (dále jen DB). Ta byla zvolena z důvodu, že na cílovém počítači není třeba provádět instalaci. Předpokládá se, že program může v budoucnu běžet v infrastruktuře Policie ČR. Takto není třeba řešit přístupová práva a zjednodušuje se tím počáteční používání programu. DB

[4] `http://graphstream-project.org`
[5] `http://snap-graph.sourceforge.net`
[6] `http://d3js.org`
[7] `https://www.sqlite.org`

je ukládána do souboru a lze ji tak snadno přenášet na jiné počítače. Pro připojení k DB je použit ovladač sqlite-jdbc-3.8.11.2[8]. Obsah databáze lze prohlížet nebo upravovat např. prostřednictvím nástroje SqliteBrowser[9]. Soubor s DB se nachází v kořenové složce programu. DB je sdílená pro všechny zpracované soubory. Obr. 5.1 zobrazuje strukturu databáze. Primárními klíči jsou ve všech tabulkách sloupce id. Primární klíče záznamů jsou vytvářeny automaticky. Používá se vlastnosti SQLite, která umožňuje jeho automatickou inkrementaci. V obrázku 5.1 je znázorněno použití cizích klíčů mezi tabulkami relací.



Obr. 5.1: *Struktura databáze. Primárními klíči jsou ve všech tabulkách sloupce* id*. Relace znázorňují použití cizích klíčů.*

Popis jednotlivých tabulek z obr. 5.1:

- `dates` – data, která se v dokumentech nacházejí,
- `days` – dny v týdnu,
- `documents` – zpracované dokumenty,
- `entities` – rozpoznané entity (včetně koreferencí),
- `person_ties` – hrany mezi osobami,
- `persons` – nalezené osoby,
- `persons_documents` – výskyt osob v dokumentech,

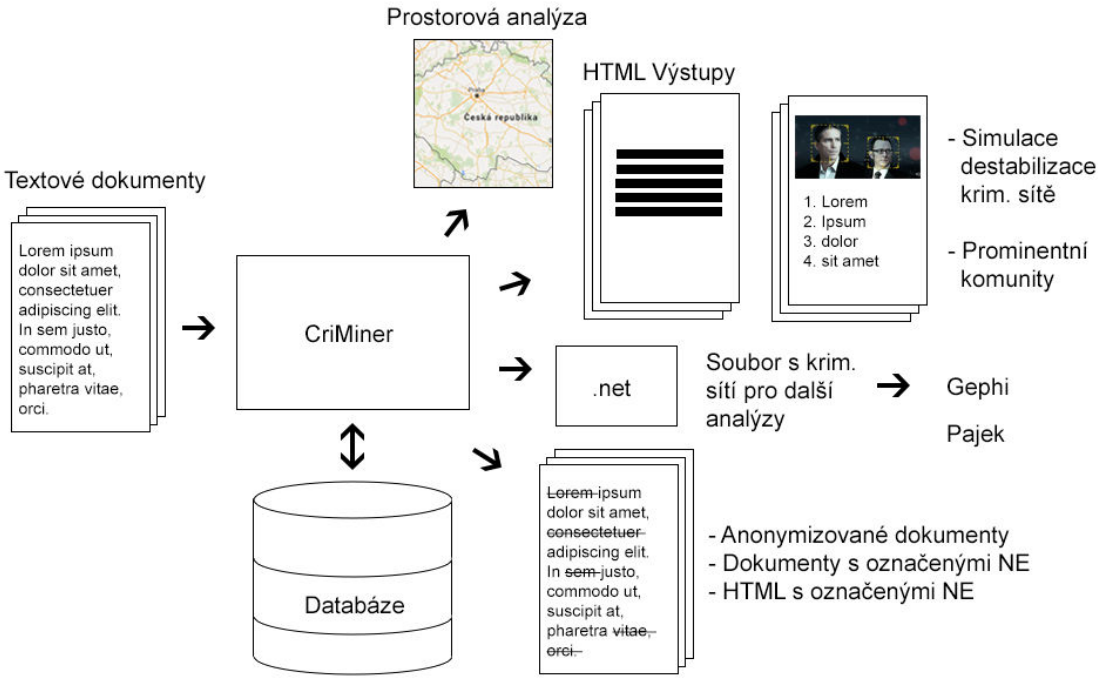
[8] `https://bitbucket.org/xerial/sqlite-jdbc/downloads`
[9] `http://sqlitebrowser.org`

- `places` – nalezená místa,
- `types` –kategorie NE,
- `unique_entity` – unikátní NE napříč dokumenty.

Proces zpracování a vizualizace dat

Na obr. 5.2 je znázorněn proces zpracování dat a generovaných výstupů. Textové dokumenty jsou nejprve zpracovány částí, která zajišťuje rozpoznávání NE a jejich anonymizaci (viz kapitola 6). NE a další informace jsou postupně ukládány do DB. Po zpracování dokumentů jsou na konci programu vytvořeny HTML stránky s postupem, jak kriminální síť destabilizovat, seznam prominentních komunit, soubor, který tuto síť obsahuje, anonymizované verze původních souborů, HTML stránky se statistickými výstupy a ukázky prostorové a časové analýzy. Veškeré výstupy až na prostorovou analýzu jsou anonymizovány.

Pro další analýzy kriminální sítě, kterou program vytváří, doporučuji program Gephi[10] nebo Pajek[11] (resp. PajekXXL). Vygenerovaný soubor s kriminální sítí lze dále analyzovat v obou programech. Základní analýzy kriminální sítě, jako je simulace destabilizace sítě a určení nejvýznamnějších aktérů provádí vytvořený program automaticky.



Obr. 5.2: *Schéma zpracování dokumentů a výstupů programu.*

[10] `http://gephi.org`
[11] `http://mrvar.fdv.uni-lj.si/pajek`

6 Anonymizátor osobních údajů

Tato kapitola popisuje část vytvořeného systému, která zajišťuje rozpoznávání pojmenovaných entit (NE) a anonymizaci osobních údajů, které se v textech vyskytují. V kapitole naleznete popis použitých algoritmů této části systému a způsob, jakým je prováděna deanonymizace osobních údajů. Pomocí té mohou pověřené osoby uplatnit získané poznatky v praxi. V závěru kapitoly je uveden způsob testování úspěšnosti rozpoznávání pojmenovaných entit (NER), jakých výsledků bylo při rozpoznávání dosaženo a jaké problémy se při NER vyskytly.

6.1 Reálná data a důvody anonymizace

Jelikož se předpokládá, že program může pracovat s reálnými daty, jež má k dispozici Policie ČR, je pro další zpracování textů nutné odstranit z těchto dat osobní údaje. Cílová data jsou uložena jako prosté textové soubory, které obsahují policejní zprávy, záznamy trestných činů nebo jiné úřední záznamy. Soubory mohou být uloženy v kódování UTF-8 nebo Windows-1250. Jednotlivé zprávy nebo záznamy jsou uloženy v separátních textových souborech s příponou .txt. Pro představu o jaké soubory se jedná uveďme příklad.

Dokument priklad.txt obsahuje následující text:

Dne 15.10.2008 byl u Franty Nováka, nar. 10.2.1986, při silniční kontrole zjištěn pozitivní nález na omamné látky. Spolujezdcem byl Petr Novák, nar. 8.5.1988. SPZ vozidla 4A1 487.

Podtržené výrazy jsou NE a až na výraz „15.10.2008“ tyto NE obsahují osobní informace, které je dle zákona o ochraně osobních údajů (viz dále) nutné anonymizovat. Jak bylo popsáno v závěru části 2.4, je při NER pro české texty dosahováno $\sim$ 70-80% F-míry. Tento fakt nepochybně způsobí, že se některé osobní údaje nepodaří z textů odstranit. V práci se zaměříme zejména na anonymizaci osobních jmen a identifikátorů (např. RČ), pomocí kterých lze osoby jednoznačně identifikovat. Tyto identifikátory a jména osob mohou představovat největší problém při budoucím získání dat.

Zákon o ochraně osobních údajů, anonymní údaje

Zákon o ochraně osobních údajů (Zákon č. 101/2000 Sb., ZOOÚ) zajišťuje naplnění práva každého na ochranu před neoprávněným zasahováním do soukromí a upravuje

práva při zpracování osobních údajů. *Osobním údajem* se rozumí jakákoliv informace týkající se určeného nebo určitelného subjektu údajů. *Subjektem údajů* rozumíme *fyzickou osobu*, k níž se osobní údaje vztahují[1]. Subjekt údajů se považuje za určený nebo určitelný, jestliže lze subjekt údajů přímo či nepřímo identifikovat zejména na základě čísla, kódu nebo jednoho či více prvků specifických pro jeho identitu. Správce osobních údajů je povinen *zpracovávat osobní údaje pouze v souladu s účelem*, k němuž byly shromážděny.

Je zjevné, že v policejních hlášeních nebo úředních záznamech se osobní údaje mohou vyskytovat a dotčené osoby pravděpodobně nedaly souhlas s jejich zpracováním k jiným účelům, nežli pro úkony orgánů veřejné moci. Z tohoto důvodu, pokud bychom chtěli tyto dokumenty získat a dále s nimi pracovat, je nutné tyto osobní data anonymizovat. *Anonymním údajem* rozumíme takový údaj, který buď v původním tvaru nebo po zpracování nelze vztáhnout k určenému nebo určitelnému subjektu údajů. Nelze tak už hovořit o osobním údaji a nevyvstává tak problém s jejich dalším zpracováním.

6.2 Osobní údaje a pojmenované entity

Dle doménového experta PČR jsou osobními údaji, které se mohou v cílových datech vyskytovat a u kterých je nutné provést anonymizaci, následující:

- jména a příjmení,
- adresy lokalit,
- data narození,
- telefonní a faxová čísla,
- rodná čísla,
- čísla bankovních účtů,
- čísla občanských a řidičských průkazů,
- státní poznávací značky,
- identifikační čísla přístrojů (např. IMEI telefonů),
- webové adresy a e-mailové účty.

[1]Dle této definice a rozsudku Nejvyššího správního soudu (čj. 6 A 83/2001) – Ochrana „osobních údajů“ právnické osoby, se ZOOÚ skutečně vztahuje pouze k fyzickým, nikoli právnickým osobám. Není však vyloučeno, že právnické osoby jsou chráněny jiným zákonem. Nebudeme proto mezi právnickou a fyzickou osobou dále rozlišovat.

Všechny tyto údaje můžeme dle definice TEI (viz. část 2.4) považovat za NE. Anonymizaci osobních údajů je tedy možné provádět anonymizací NE. Jelikož však nedokážeme určit, v jakém případě se jedná o osobní údaj a v jakém ne, budeme anonymizovat veškeré NE. Jako příklad uveďme datum. Pokud je vztaženo jako den narození, jedná se o osobní údaj. Pokud je datum uvedeno např. v kontextu „12. 4. 2016 svítilo slunce“ nelze hovořit o osobním údaji. Jak je vidět, neplatí tedy, že každá pojmenovaná entita je zároveň osobním údajem. Můžeme však říci, že každý osobní údaj je dle definice TEI pojmenovanou entitou. Jak uvidíme dále, je pro další analýzy vhodné zpracovávat i NE, které osobními údaji nejsou.

6.3 Popis NER systému

Tato část popisuje použité technologie a algoritmy, které byly použity k tvorbě části systému, která se stará o rozpoznávání NE.

Výběr výchozí knihovny pro NER a návrh dalších průchodů

Jak bylo popsáno v části 2.4 pojmenované entity lze rozpoznávat pomocí dvou přístupů. Pomocí pravidlového přístupu nebo pomocí strojového učení. Vytvořený systém je kombinací obou přístupů. Pro NER existuje několik volně dostupných knihoven a nástrojů, které jsou založeny na metodách strojového učení. Některé z nich jsou:

- NameTag[2],
- Stanford Named Entity Recognizer[3],
- OpenNLP[4],
- OpeNER[5],
- THD [6],
- LINGVO Named Entity Recognizer[7],
- NERD[8],
- FOX – Federated knOwledge eXtraction Framework[9],

[2] `http://ufal.mff.cuni.cz/nametag`
[3] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[4] `http://opennlp.apache.org`
[5] `http://www.opener-project.eu`
[6] `http://ner.vse.cz/thd`
[7] `http://www.kiv.zcu.cz/cz/vyzkum/software`
[8] `http://nerd.eurecom.fr`
[9] `http://aksw.org/Projects/FOX.html`

- Detektor pojmenovaných entit[10].

Nástroje NameTag, Detektor pojmenovaných entit, THD a FOX nabízejí online dema, která byla testována na části dat z korpusu ČTK (viz část 6.7) a na novinových článcích ze serveru iDnes[11]. Z testování nástrojů FOX a Stanford Named Entity Recognizer vyplynulo, že pro zpracování textů v češtině nejsou nástroje s anglickým či jiným modelem, nežli českým vhodné. Některé entity sice dokáží rozpoznat, nicméně pokrytí není příliš velké.

Jako výchozí knihovnu pro část systému, která zajišťuje NER jsem zvolil NameTag. Tato knihovna obsahuje model natrénovaný na českých datech a dle Strakové a kol. (2013) dosahuje pro český jazyk jednoho z nejlepších výsledků. Dalšími výhodami jsou: velké množství kategorií NE, zařazení detailních kategorií do hlavních kategorií, existence on-line dema a dobrá dokumentace. Knihovna je napsána v jazyce C++ přičemž obsahuje binding na jazyky Java, Perl a Python. NameTag je též možné spustit jako samostatný program, který předané soubory zpracuje a vygeneruje XML soubory, ve kterých jsou NE označeny.

Knihovna ve vytvořeném systému obstarává tzv. první průchod (viz obr. 6.1). Cílem tohoto průchodu je nalézt co největší počet NE při zachování vysoké přesnosti a provést tokenizaci dokumentů.

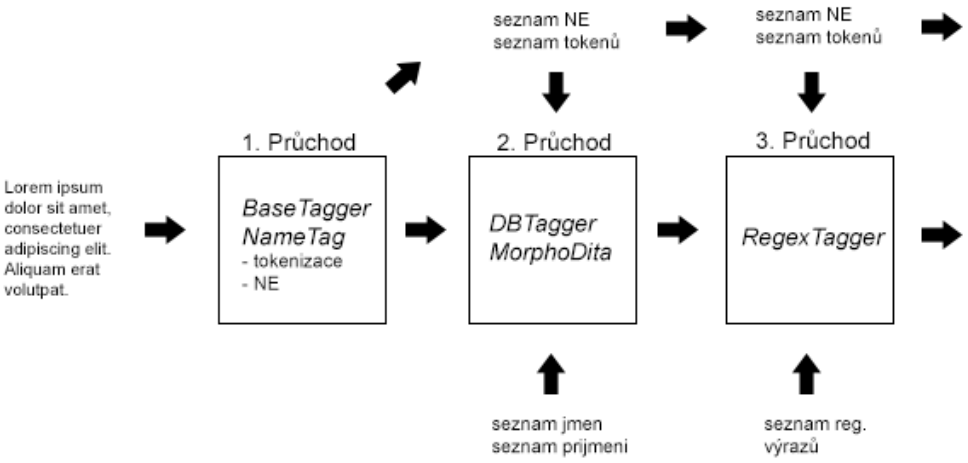
Při testování knihovny se ukázalo, že nedokáže rozpoznat, nebo nepovažuje za NE, všechny osobní údaje, které je třeba rozpoznat a anonymizovat. Jedná se například o rodné čísla, čísla bankovních karet, IMEI čísla apod. To je patrně způsobeno neexistencí těchto dat v trénovacích datech. Možným řešením by bylo natrénovat vlastní statistický model na cílová data. To však vzhledem k tomu, že cílová data nejsou dostupná, časové náročnosti simulace reálných dat, následnému anotování a nedostatečné znalosti cílové domény, není možné. Tento problémový typ NE lze však snadno rozpoznávat regulárními výrazy (viz třetí průchod).

Dalším problémovým typem entit jsou vlastní jména. NameTag dokázal na testovacím korpusu odhalit ∼80% osobních jmen s přesností ∼82%. Jelikož jsou právě osobní jména, zejména jejich co největší pokrytí, pro další práci významné, je přidán další průchod, který je na ně zaměřen. Postup NER ve vytvořeném systému můžete vidět na obr. 6.1.

Druhý průchod – rozpoznávání osobních jmen

Po zpracování dokumentu prvním průchodem, který provede tokenizaci a první označení NE, následuje porovnání některých tokenů (kandidátů na jména, viz dále) se seznamy jmen a příjmení. Ty se nacházejí v souborech jmena.csv a prijmeni.csv. Jedná

[10] `http://nlp.fi.muni.cz/projekty/ner/v2`
[11] `http://www.idnes.cz`



Obr. 6.1: *Schéma rozpoznávání pojmenovaných entit všemi průchody.*

se o upravené soubory zveřejněné Ministerstvem vnitra[12], které obsahují jména a příjmení obyvatel ČR. Původní soubory obsahují i jejich četnost. V souborech se objevují i málo používaná jména nebo příjmení, včetně cizích jmen. Soubory na každé řádce obsahují jedno jméno nebo příjmení, které jsou v prvním pádu. V tomto průchodu mohou být označeny jako NE pouze tokeny, které zatím jako NE označeny nejsou a zároveň se nacházejí v některém z těchto seznamů. Výstupem tohoto procesu je disjunktní množina NE prvního a tohoto průchodu, označení příslušných tokenů za část NE a jejich propojení s NE[13]. Samotné rozpoznávání NE pro tento průchod zajišťuje třída `DatabaseTagger` metodou `tag`. Průběh, jakým je druhý průchod realizován ukazuje algoritmus 1. V následujícím textu budou v závorkách uváděny příslušné řádky algoritmu.

Jména a příjmení uvedená v seznamech se nacházejí v prvním pádě. Zpracovávané tokeny je tedy třeba lemmatizovat nebo provést jejich stemming. Jak bylo uvedeno v části 2.2 stemmatizace je méně přesný proces. Proto je v tomto kroku prováděna lemmatizace pomocí morfologického analyzátoru. Jako morfologický analyzátor je použita knihovna MorpohoDiTa (viz část 5.1). Aby bylo zajištěno co největší pokrytí, je u kandidátů na osobní jméno (3) provedena kompletní morfologická analýza, která u tokenu určí všechny jeho možná lemmata (4). Pro zvýšení přesnosti lemmatizace, je používán i předchozí token. Tyto lemmata jsou následně porovnána se seznamy osobních jmen (6). Pokud se alespoň jedno lemma v seznamu objevuje, je token označen jako část NE (9) a z tokenu je NE vytvořena (10).

[12]Původní, neupravené, soubory jsou dostupné na `http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx`.

[13]NE se může skládat z více tokenů. Token může být označen jako část NE a přiřazen k ní. Obdobně může mít token přiřazenou NE.

Algoritmus 1: Rozpoznání osobních jmen dodatečným průchodem

Data: T_{okeny}, NE_s, J_{mena}, $P_{rijmeni}$
Výstup: T_{okeny}, NE_s

```
1  foreach token z T_okeny do
2      osobní jméno = false
3      if token ≠ NE, začíná velkým písmenem, délka > 3, není počátek věty
       then
4          L_emmas = lemmatizuj(token, předchozí token)
5          foreach lemma z L_emmas do
6              if lemma ∈ J_mena OR lemma ∈ P_rijmeni then
7                  osobní jméno = true
8          if osobní jméno then
9              označ token jako NE
10             vytvoř z tokenu NE_t a přidej jí do NE_s
11     if osobní jméno, další token začíná velkým písmenem, další token ≠
       NE then
12         označ další token jako NE a přidej ho k NE_t
```

Dále se pro urychlení a zvýšení pokrytí vychází z následujícího předpokladu: za jménem se často budou objevovat příjmení. Jelikož je rozpoznání jména snadnější, je na základě tohoto předpokladu přistupováno k rozpoznání neznámých příjmením (11 a 12). V seznamu s příjmeními se sice nacházejí i cizí příjmení, ta však v průběhu času mohou přibývat, nebo se u cizích příjmení nemusí podařit korektní lemmatizace. Pokud se za tokenem, který byl rozpoznán jako jméno objevuje token s velkým písmenem a je delší než tři znaky, je považován za příjmení. Chybovost tohoto předpokladu pomáhá snižovat první průchod. Již rozpoznané NE nejsou přeznačovány ani zpracovávány (3 a 11).

Aby token mohl být označen za kandidáta na NE (3), musí být splněno několik podmínek: token je delší než 3 znaky, začíná velkým písmenem a není na začátku věty – tj. jeho předcházející token není některý ze znaků `. > \ - „ ? !`. Pokud bychom tyto podmínky vynechali, došlo by ke značnému zhoršení přesnosti. V souborech s příjmeními se totiž objevuje např. vietnamské příjmení „Do“. U každé věty, která by začínala slovem „Do“ by toto slovo bylo bráno jako jméno, i když jím často nebude. Tím, že se zpracovávají pouze kandidáti, též dojde k urychlení zpracování. Na druhou stranu aplikací těchto podmínek snižujeme potenciální pokrytí.

Třetí průchod – regulární výrazy

Jak bylo řečeno dříve, některé osobní údaje, jako jsou rodná čísla, SPZ, IMEI čísla apod. se nepodaří během prvního průchodu rozpoznat. Je proto přidán další průchod dokumentů, který se právě na tyto entity zaměřuje. Zpracovává celý text a jeho výstup je opět disjunktní množina NE s prvním a druhým průchodem, příslušné propojení NE s tokeny a případné označení zda jsou tokeny součástí NE. O rozpoznávání NE se v tomto průchodu stará třída RegexTagger.

Defaultně zpracovávané regulární výrazy jsou tyto:

- `\d{1,2}[\.\-/] ?\d{1,2}[\.\-\/] ?\d{4}` pro datum ve tvaru 14.5.2001, 14/05/2001 a 14-05-2001 přičemž mohou obsahovat významové nuly.
- `(\+{1}\d{1,4})? ?\d{3}?\?\d{3}\?\d{3}` pro telefonní čísla ve tvarech (+420) 607 123 456 a 607 123 456 s mezerami i bez mezer.
- `\d{6}[/]\d{3,4}` pro rodná čísla ve tvaru 123456/1231, přičemž není ověřována jejich platnost.
- `\d([A-Z]|\d){2} \d{4}` pro SPZ vozidel ve tvaru 4A2 3000.
- `\d+[-]?\d+[/]\d{3,4}` pro bankovní účty ve tvaru 000-111111111/1234 nebo 11111111111/1234.
- `\d{4} \d{4} \d{4} \d{4}` pro čísla kreditních karet ve tvaru 5133 1111 2222 3333.
- `[a-zA-Z0-9]+@[a-zA-Z0-9]+.[a-zA-Z]{2,5}`, pro e-mailové adresy ve tvaru adresa@domena.cz.
- `\d{15}` pro identifikační čísla telefonů (IMEI) a sim karet (IMSI) ve tvaru 230011223344556.

Jelikož se předpokládá, že může být požadováno přidání dalších regulárních výrazů, či úprava stávajících, pokud by se na cílových datech ukázala jejich neúplnost, jsou regulární výrazy uloženy v konfiguračním souboru regex.ini. Soubor je tvořen následujícími dvojicemi řádek:

```
REG = regex
TAG = oznaceni NE
```

Přidáním této dvojice řádek, je možné regulární výrazy přidávat. Stávající regulární výrazy je možné upravovat. V souboru je potřeba při uvedení dopředného lomítka znak escapovat pomocí zpětného lomítka. Pokud bude soubor chybně naformátován, bude na to uživatel upozorněn při spuštění programu.

Typy rozpoznaných entit

Program detekuje a dále používá několik typů NE. Tyto kategorie vycházejí z hlavních kategorií NameTagu[14]. Pro naše potřeby není třeba používat všechny detailní kategorie, které mohou snižovat přesnost klasifikace. Omezení jejich počtu umožní snadnější vyhodnocení úspěšnosti rozpoznávání na korpusu ČTK (viz část 6.7). V tomto korpusu jsou používány odlišné kategorie NE a ne všechny lze řádně namapovat na detailní kategorie. Několik detailních kategorií, které jsou pro další zpracování vhodné, je však použito. Jsou to kategorie pro akademické tituly, jména ulic, náměstí a jména měst. Dalšími používanými kategoriemi, jsou kategorie vytvořené pomocí regulárních výrazů. Tyto kategorie by patrně bylo možné zařadit do kategorie „Specifické užití čísel“, nicméně pro další zpracování je vhodné tyto kategorie zvláště odlišovat.

Pro další analýzu jsou nejdůležitější kategorie „osobní jména“ pro tvorbu a analýzu kriminální sítě (viz část 7.2). Dále „geografické názvy“ pro prostorovou analýzu (část 7.5), „časové údaje“ pro časovou analýzu (část 7.5) a kategorie rozpoznané pomocí regulárních výrazů, které mohou v budoucnu být zařazeny do kriminální sítě.

6.4 Anonymizace, zachování morfologie a deanonymizace

Při anonymizaci NE je vhodné zachovat její morfologii. Pokud bude morfologie zachována, nebude výsledný text anonymizací příliš poznamenaný, bude čitelnější a budou zachovány původní morfologické informace, které mohou pomoci při trénování statistického modelu, jenž může být na datech v budoucnu trénován. Vytvořený anonymizátor tak při anonymizaci bere ohled na slovní druh, rod, pád a číslo entity. Tyto informace jsou při anonymizaci zachovávány. Aby data nebylo možné rozšifrovat, nevychází anonymizovaný řetězec z rozpoznané entity, ale z předem připraveného řetězce.

Reprezentace pojmenované entity ve výstupních souborech

Aby byl text použitelný pro další zpracování, není vhodné anonymizované entity ukládat jako obyčejný řetězec (například ve formátu Novák_001, či Novákovy). Vhodnější je anonymizované data ukládat do souboru XML[15], který je možné v budoucnu snadno zpracovávat. V našem případě je anonymizovaná entita označena značkou:

<ENTITY type='person' id='003' entityId='1'> Petru Novákovi </ENTITY>

[14] `http://ufal.mff.cuni.cz/~strakova/cnec1.0/ne-type-hierarchy.pdf`

[15] Entity jsou zároveň ukládány i do relační databáze, viz 5.2.

Přičemž jméno uvnitř entity již není reálné, ale generické jméno, v tomto případě pro osoby mužského pohlaví se zachovanou morfologií. Pokud bychom použili řetězce typu Novák_003, ztratíme morfologickou informaci, text příliš zdeformujeme a učiníme ho těžko čitelným. S touto reprezentací se však objevuje problém, jak určovat koreference entit. To je řešeno pomocí atributu entityId. Hodnota tohoto atributu je pro koreference entit (i napříč dokumenty) stejná. Atribut id je unikátní identifikátor pro každou NE, který koreference neuvažuje. To umožňuje deanonimizaci celého dokumentu 1:1 (viz dále). Eliminujeme tak možné chyby, které by mohly nastat při převodech pomocí morfologického analyzátoru.

Anonymizace zachovávající morfologii

Anonymizace je prováděna na základě morfologické značky, která je určena morfologickým analyzátorem úlohou tag (viz část 2.2). Ze značky (používá se formát PDT2.0[16]) se vyberou pozice, které odpovídají slovnímu druhu, číslu, pádu a rodu zpracovávaného tokenu (tj. znaky na pozicích 1, 3, 4 a 5). Na základě této značky je vybrán anonymizovaný řetězec ze souboru `ane.ini`. Ten obsahuje mapování částí morfologických značek na připravené anonymizované řetězce. Soubor má tvar:

NMS1::Novák

Každá řádka odpovídá jednomu morfologickému tagu a příslušnému anonymizovanému tokenu. V tomto případě se jedná o podstatné jméno, mužského rodu, jednotného čísla v prvním pádě. Tokeny, které mají být anonymizovány a odpovídají této značce, budou anonymizovány jako „Novák“. Úpravou souboru je možné upravovat jakou anonymizovanou entitou budou původní NE nahrazovány nebo je pro další značky přidávat. Pokud bude výstupem morfologické analýzy značka, která se v tomto souboru nenachází, bude provedena anonymizace jako „X“.

Anonymizace je prováděna pomocí třídy Anonymizer, metodou doAnonymize, která postupně zpracovává jednotlivé tokeny. Pokud se jedná o NE, je provedena anonymizace, podle výše uvedeného postupu. Anonymizovaný text je zapsán do výstupního souboru a pokračuje se dalším tokenem. Pokud se o NE nejedná, je token zapsán do souboru bez úprav a pokračuje se dalším tokenem. Pokud se NE skládá z více tokenů, budou anonymizovány všechny, každý zvlášť. Víceslovné entity jsou obaleny jedním tagem ENTITY.

Deanonymizace

Zpětná projekce původních NE je pro celé soubory prováděna na základě atributu id, který každý element ENTITY obsahuje. Tento identifikátor je sdružen s původním jménem entity v tabulce převodů (viz sekce 5.2, tabulka entities). Díky tomu, že je ukládána entita v původním tvaru (není u ní provedena lemmatizace) je tato deanonymizace provedena 1:1. Tím je zabráněno chybovým převodům, které

[16] `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s02s01.html`

by mohly vzniknout při převodech pomocí morfologického analyzátoru, na úkor zvětšení velikosti databáze. Pokud je třeba deanonymizovat pouze konkrétní NE, lze použít identifikátor entityId, který sdružuje shodné entity (viz obr. 5.1 tabulka `unique_entity`).

Deanonymizace je přístupná pouze osobám, které mají přístup k databázi (viz obr. 5.1 a 5.2). Pokud není databáze poskytnuta, program nedokáže anonymizované entity deanonymizovat. Vzhledem k tomu, že anonymizované entity nevycházejí z původních NE není je možné ani rozšifrovat.

6.5 Koreference osob a jiných entit

V textech se mohou vyskytovat stejné NE na několika místech, v různých tvarech a ve více dokumentech. Pro další analýzy je potřeba tyto NE náležitě propojit a pracovat s nimi jako s jednou entitou. Program rozlišuje dva případy, při kterých je prováděno zpracování koreferencí. Prvním případem je koreference osob. Koreference osob je prováděna při vytváření objektů typu Person. Ty mohou být vytvářeny pouze z NE typu „osobní jména“ dle algoritmu 2.

Osoby jsou ukládány do tabulky persons. Zpracování koreferencí zajišťuje třída PersonManipulator. Z takto vytvořených osob je následně vytvářena kriminální síť (viz část 7.2) a prováděna detekce prominentních komunit (viz část 7.4). Algoritmus 2 ukazuje zpracování koreferencí u entity typu „osobní jména“. V dalším textu čísla v závorkách odkazují na příslušné řádky algoritmu 2.

Při zjišťování koreferencí u osob je nejprve provedena lemmatizace celého jména (1). Aby se zajistila větší přesnost lematizace, je prováděna nejen na základě textu, který má být lemmatizován, ale i na základě předchozího tokenu, což se při experimentech ukázalo jako dostatečně přesné. S ním se například ve výrazu „premiéra Bohuslava Sobotky“ již dosáhneme korektní lemmatizace – „Bohuslav Sobotka“. Pokud by předchozí token nebyl brán v úvahu, byla lemmatizace úlohou *tag* (vybere nejpravděpodobnější lemmu) provedena jako „Bohuslava Sobotka“.

Dále se vychází z následujícího předpokladu: při prvním výskytu osoby v textu musí být osoba jednoznačně identifikovatelná. Předpokládá se tedy, že při prvním výskytu bude osoba zmíněna celým jménem, minimálně jménem a příjmením (3). Pokud se jedná pouze o jméno, či příjmení, předpokládá se, že osoba byla již dříve zmíněna a jedná se o koreferenci (19). Dále program rozlišuje jmenovce (4-13). Ty jsou rozlišováni na základě možné zmínky o datu narození za jménem, která se v kriminalistických textech může vyskytovat. Jsou prohledávány 4 tokeny za jménem (5). Pokud některý z nich bude NE s datem (6), jsou následně prohledávány předchozí 3 tokeny (8). Pokud některý z nich je řetězec „narozen“, „nar“ nebo „n“ je tento datum k osobě přiřazen (10, 11). Počet prohledávaných tokenů za NE a možné ozna-

čení data narození, bylo zvoleno za nákladě možného výskytu data narození např. „Petr Novák, nar. 11.5.1985“, pro volbu přesnější hodnoty a všech typů označení by bylo nutné mít k dispozici dostatečně velký vzorek reálných dat.

Algoritmus 2: Identifikace koreferencí osob

Data: NE, O_{soby}
Výstup: O_{soby}, *osoba*

```
1  normalizované jméno = lematizuj(tokeny NE, předchozí token NE)
2  osoba = new Osoba(normalizované jméno)
3  if osoba má jméno a příjmení then
4      token_d = token následující po NE
5      while počet za > 4 do
6          if token_d je částí NE_x a NE_x je datum then
7              token_p = token_d
8              for i = 0; i < 3; i++ do
9                  token_p = předchozí token token_p
10                 if token_p == ("nar"OR "narozen"OR "n") then
11                     osoba.datum narození = NE_x
12         token_d = token následující po token_d
13         počet za++
14     if osoba ∈ O_soby then
15         koreference = true
16     else
17         přidej osoba do O_soby
18 else
19     koreference = true
```

Pokud se osoba již nachází v seznamu osob (sdílený pro všechny dokumenty), je považováno jméno osoby za koreferenci (14, 15). Pouze u jmen nebo příjmení (např. pouze Novák) je rozsah koreferencí uvažován v rámci zpracovávaného dokumentu. Pokud osoba v seznamu osob není a nejedná se o koreferenci, je do tohoto seznamu přidána (17). Aby osoba byla považována za stejnou, musí mít shodné jméno, příjmení a pokud jsou u obou osob uvedena data narození, musí být shodná. Pokud u jedné osoby uvedeno nebude, nejsou osoby považovány za shodné. Pokud není uvedeno ani u jedné, je testováno na shodu pouze jméno a příjmení osob.

Po zpracování všech dokumentů je z nalezených osob vytvořena kriminální síť. Pomocí identifikátorů v této síti (nemusejí se shodovat s identifikátorem entityId, osoby mají vlastní identifikátor) lze jména osob deanonymizovat (viz příloha A.3).

Druhým případem koreferencí je zpracování koreferencí všech druhů NE. Toto zpracování provede lemmatizaci NE a zkontroluje, zda se lemma nevyskytuje v tabulce unique_entity (viz obr 5.1). Pokud ano, je k entitě přiřazen identifikátor z této tabulky (sloupec unique_entity v tabulce entities a atribut entityId v anonymizovaných souborech). Pokud ne, je do tabulky lemma přidáno. Toto zpracování je prováděno u všech NE, i u osob. Nejsou však již na ně kladeny žádné další požadavky, jako je např. přítomnost jména a příjmení. Na základě identifikátoru entityId je sestavena tabulka nejčastěji se vyskytujících NE v textech, která může být přínosná pro určení šablony zločinného chování – zejména pak např. často odcizovaných předmětů nebo předmětů, které jsou k zločinům často používány[17].

6.6 Výstupy anonymizátoru a NER

Program generuje pro každý zpracovávaný dokument tři výstupní soubory. Tyto soubory jsou ukládány do složky output, která je dále členěna do podsložek anonymized, html a orig. Složka anonymized obsahuje anonymizované soubory ve formátu XML (viz část 6.4), ve kterém jsou rozpoznané NE vyznačeny a anonymizovány. Tyto soubory mohou být, přístupné i nepověřeným osobám. Osobní údaje, které se podařilo rozpoznat, jsou v souborech anonymizovány a bez databáze je není možné deanonymizovat ani rozšifrovat.

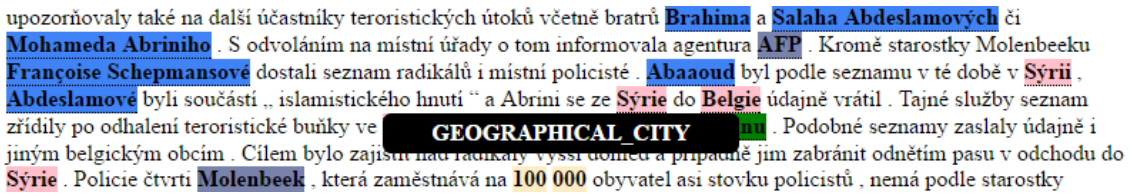
Složka orig obsahuje shodné soubory jako složka anonymized, které však nejsou anonymizovány a neměly by tedy být poskytovány nepověřeným osobám. Rozpoznané NE jsou opět v XML souborech vyznačeny. Tyto výstupy mohou sloužit pro testování a další vývoj programu.

Složka html obsahuje HTML soubory, které nejsou anonymizované a slouží pro další manuální analýzy. Soubory jsou snadno čitelné a umožňují snadnou vizuální analýzu označených entit. Rozpoznané entity jsou zde vyznačeny pomocí HTML značek a různé druhy jsou barevně zvýrazněny pomocí CSS. Na tyto soubory je dále odkazováno v části programu zabývající se detekcí prominentních komunit (viz část 7.4), což může usnadnit rozhodování, v jaké souvislosti se tyto osoby spolu vyskytují. Ukázku souboru s označenými NE zobrazuje obr. 6.2.

6.7 Testování úspěšnosti anonymizace

Úspěšnost NER byla testována vůči korpusu pojmenovaných entit textminingové skupiny na FAV. Korpus je tvořen 9 soubory s články ČTK a nachází se v něm

[17]Jedná se o NE typu „artefakt“. V tabulce jsou zahrnuty i další typy NE, tento typ je však dle mého názoru v této tabulce nejpřínosnější.



Obr. 6.2: *Ukázka označení typů NE v HTML souborech. Různé kategorie jsou barevně odlišeny. Po označení kurzorem se zobrazí typ NE.*

téměř 48,5 tisíc anotovaných NE. Anotace se nacházejí v XML souborech .annotation a každá NE je popsána pomocí trojice značek: pozicí prvního znaku (start), pozicí posledního znaku (end) a typu NE (type). V korpusu se těchto typů vyskytuje celkem 16. Tímto korpusem se pokusíme simulovat přechod na jinou doménu a zjistit zda a případně o kolik klesne výkonnost NameTagu, kterým je obstaráván první průchod. Dále tímto bude ověřena možnost mapování kategorií mezi systémy užívající různé označení pro NE[18]. To může být, zejména pokud by PČR používala nebo požadovala jiné kategorie NE, též vhodné.

Nalezené problémy

Při testování anonymizace proti zvolenému korpusu bylo zjištěno několik problematických skutečností, které se v korpusu objevují a které získané výsledky podhodnocují.

- V korpusu se objevují delší úseky anglických textů. Vzhledem k tomu, že program bude pracovat s českými texty, z čehož bylo při návrhu a následném vývoji vycházeno, nedaří se v anglických částech textu dosahovat vysoké úspěšnosti rozpoznávání.
- Není normalizováno, zda je jméno a příjmení uvažováno jako jedna NE (tento případ je častější), nebo zda je jméno bráno jako jedna NE a příjmení jako druhá NE[19]. To při vyhodnocování dle přesné shody způsobí, že je toto jméno započítáno jako tři různé chyby. Např. první chybou je falešně pozitivní entita „Karel Lang“, druhou a třetí chybou pak falešně negativní entity „Karel“ a „Lang“.
- V korpusu se objevuje velké množství sportovních utkání a (HTML) tabulek. Údaje v těchto tabulkách nejsou zcela správně rozpoznávány. Nepředpokládá se však, že v kriminalistických datech by se podobné případy vyskytovaly. Pro zvýšení pokrytí byl přidán regulární výraz \d+ pro číselné údaje. Aplikuje se pouze pokud číslo není částí jiné NE.

[18] Jelikož je dle části 2.4 možné kategorie NE dodefinovávat, nejsou tyto kategorie normalizovány.
[19] Viz např. „Karel Lang“ v prvním souboru. Je zde označen jako dvě NE – první začíná na pozici 2027 a končící na pozici 2032, druhá pak začíná na pozici 2033 a končící na 2037.

- Některé kategorie z korpusu je problematické, nebo nemožné namapovat na námi používané kategorie. Jedná se především o kategorie „náboženství“ a „artefakt“. Nejčastěji vyskytující se artefakty v korpusu jsou měny. V korpusu jsou označeny jako artefakty celé výrazy např. „500 Kč“. Námi používané kategorie však používají dělení na „500“ jako číselný údaj a „Kč“ jako název artefaktu. Kategorie „artefakt“ proto není do výsledného hodnocení zahrnuta (obsahuje 1,4 tisíc entit, což představuje 1,8% entit). U kategorie se sportovními týmy není zcela jasné, na které kategorie je namapovat (např. „Curych“ ve sportovních tabulkách). Vzhledem k tomu, že se nepočítá s tím, že by se v cílových datech vyskytovaly sportovní týmy, byly namapovány na města pro ověření rozpoznávání měst, které bude nutné též anonymizovat. To však může způsobit chyby ve vyhodnocení u týmů, které nejsou pojmenované podle města jako např. „Sparta“.
- Časové údaje jsou definovány velmi volně. Jako časový údaj jsou uvažovány např. i „dnes“, „v říjnu“ (ne pouze „říjnu“), „od 12 : 00“, „v noci na dnešek“ apod. Platí zde pak podobné jako u jmen osob, chyba bývá započtena dvakrát a celkové pokrytí časových údajů je malé.

V závěru práce se podařilo program otestovat v prostředí PČR na 200 souborech s reálnými daty, ze kterých byla vytvořena anonymizovaná kriminální síť. Vzhledem k některým specifikům reálných dat jako např.: velká část příjmení nebo měst je psána verzálkami; jména jsou často za příjmeními; obsahují velké množství cizích jmen nebo anglických textů atd., by před reálným použitím bylo potřeba provést další předzpracování nebo na reálných datech natrénovat model, který je použit pro první průchod.

Úspěšnost rozpoznávání

Tabulka 6.1 ukazuje úspěšnost rozpoznávání jednotlivých kategorií NE pro první průchod. Vyjma posledního řádku musí být určeny jak přesné hranice, tak typ NE. V opačném případě nebude entita započtena jako skutečně pozitivní – jedná se o metodu vyhodnocování přesné shody. Hodnoty „celkem“ z následujících tabulek jsou vypočteny pomocí součtu všech měřených NE. Nikoli jako aritmetický průměr hodnot uvedených v tabulkách. Častěji zastoupené kategorie tak mají v celkovém výsledku větší váhu. Výrazný pokles proti úspěšnosti 82,82% uváděný Strakovou a kol. (2013) je dán zvoleným testovacím korpusem a dříve zmíněným problémům, které se v něm vyskytují. Zejména velkou částí sportovních reportáží a tabulek, kterou korpus tvoří. Po odfiltrování číselných a časových údajů je F_1-míra 63,32%, F_2 míra pak 58, 75%. Jako překvapivě obtížné se ukázalo mapování kategorií mezi různými typy značení. V budoucnu je vhodné testovat úspěšnost anonymizace spíše pomocí korpusu Czech Named Entity Corpus 2.0[20], který používá shodné kategorie a vyhod-

[20] `http://ufal.mff.cuni.cz/cnec/cnec2.0`

nocování tak nebude zatíženo chybami, které může mapování způsobovat. Pravděpodobně se však na něm nepodaří nasimulovat přechod na jinou doménu.

Tab. 6.1: *Úspěšnost rozpoznávání kategorií prvním průchodem dle metody přesné shody (typ a ohraničení) a dle ohraničení (pouze ohraničení).*

Kategorie	**Přesnost**	**Pokrytí**	F_1**-míra**	F_2**-míra**
Geografické názvy	78,15	54,05	63,90	57,60
Města	81,14	62,45	70,58	65,47
Instituce	56,41	42,23	48,30	44,47
Číselné údaje	56,13	21,92	31,53	24,96
Jména osob	82,54	80,32	81,41	80,75
Časové údaje	54,16	28,07	36,97	31,06
Celkem typ a ohraničení	69,37	44,25	54,03	47,71
Celkem pouze ohraničení	76,24	49,09	59,73	52,85

Tabulka 6.2 ukazuje změny, které se projevily po přidání druhého a třetího průchodu. Přidání regulárních výrazů se projevilo téměř pouze na rozpoznávání číselných údajů. V korpusu se nevyskytuje dostatečné množství dalších entit, které jsou jimi rozpoznávány. Bylo nalezeno pouze ~ 100 dalších e-mailových adres (které však podle tohoto korpusu nejsou považovány za NE), telefonních čísel nebo dat.

Tab. 6.2: *Změny úspěšnosti rozpoznávání po přidání druhého a třetího průchodu. Hodnoty v závorkách uvádějí změnu hodnoty oproti tabulce 6.1.*

Kategorie	**Přesnost**	**Pokrytí**	F_1**-míra**	F_2**-míra**
Číselné údaje	66,29 (+10,16)	71,66 (+49,74)	69,87 (+38,34)	70,52 (+45,56)
Jména osob	75,04 (-7,50)	83,74 (+3,42)	79,15 (-2,26)	81,84 (+1,09)
Celkem typ a ohraničení	68,74 (-0,63)	61,93 (+17,68)	65,16 (+11,13)	63,18 (+15,47)
Celkem pouze ohraničení	74,64 (-1,60)	67,48 (+18,39)	70,88 (+11,15)	68,80 (+15,95)

Přidáním druhého průchodu bylo u kategorie „jména osob“ na úkor přesnosti zvětšeno pokrytí. Při anonymizaci osobních údajů je žádoucí, aby jejich pokrytí bylo co největší. Za významnější tak lze v našem případě považovat F_2-míru (viz vzorec (4), kde $\beta = 2$), která upřednostňuje pokrytí před přesností.

Přesnost byla zhoršena zejména kvůli rozpoznávání NE, které mohou být zároveň jak jménem, tak institucí např. „Reuters“, nebo městem např. „Bělehrad“. Tyto entity, zejména geografické názvy, je však též nutné rozpoznávat a anonymizovat. Mohou obsahovat např. bydliště nebo místa trestných činů. Pokud jej tedy nerozpozná první průchod, může nastat situace, kdy budou anonymizovány pomocí druhého průchodu. Budou však označeny chybným typem. Ve většině případů se při chybně rozpoznaných typech z druhého průchodu jedná o jednoslovnou entitu. V části 6.5 byl popsán algoritmus, kterým je zajištěno rozpoznávání osob a jejich koreferencí ze kterých je následně vytvářena sociální síť (viz část 7.2). Aby byla z NE vytvořena osoba, se kterou budeme dále pracovat, je nutné aby bylo uvedeno jak její jméno, tak příjmení. To u entit, které jsou chybně označeny jako osoby není splněno. Nejsou tak z nich vytvářeny osoby, které jsou zahrnuty do sociální sítě a snížení přesnosti se již dále neprojevuje.

Rozpoznávání (včetně ukládání dat do DB, bez tvorby sociální sítě a destabilizací) pro celý korpus pouze prvním průchodem trvalo na stroji AMD Phentom II X4 955, 12 GB RAM 4,5 minuty. Pro všechny průchody pak 5,25 minuty. Z 47 tisíc NE, které byly z korpusu zahrnuty, bylo všemi průchody rozpoznáno celkem 29 tisíc skutečně pozitivních entit. Jako falešně pozitivních bylo označeno 13,2 tisíc entit (nerozlišujeme chybný typ nebo ohraničení). 16 tisíc entit z korpusu je tvořeno číselnými údaji. 8,6 tisíc jmény osob. Skutečně pozitivních jmen se podařilo nalézt 7,2 tisíc. Jelikož se používalo vyhodnocení přesné shody a v korpusu se vyskytl problém s normalizací zápisu jmen, bude patrně tato hodnota vyšší.

7 Analýza extrahovaných dat z textů

Tato kapitola popisuje analýzy, které jsou na základě extrahovaných faktů z textů prováděny a shrnuje získané poznatky z testovacích dat. Naleznete zde analýzu, jak efektivně destabilizovat kriminální síť, detekci prominentních komunit, shrnutí několika statistických údajů a ukázky možného odhalování šablon zločinného chování, jako jsou časové a prostorové analýzy, často se vyskytující artefakty, které můžeme považovat za často odcizované nebo k zločinům používané předměty. Jsou tak zde uplatněny některé metody popsané v kapitolách 3 a 4.

7.1 Testovací data

Jelikož v průběhu práce nebyla dostupná cílová data, na kterých by bylo možné následující analýzy provádět, bylo vytvořeno a pro další analýzy použito několik testovacích souborů dat.

Hlavní soubor s daty

Večer 13. listopadu 2015 bylo v Paříži provedeno šest teroristických útoků, které si vyžádaly 130 obětí. K útokům došlo u fotbalového stadionu Stade de France, u několika barů a restaurací. Nejtragičtější útok se odehrál v koncertním sále Bataclan. Dle dostupných informací byl hlavním strůjcem těchto útoků Abdelhamid Abaaoud, který byl 18. listopadu při zásahu policie zabit. 18. března 2016 byl zatčen další podezřelý z těchto útoků – Salah Abdeslam. Za dva dny po jeho zatčení následovaly teroristické útoky v Bruselu.

Výběr dat je inspirován prací (Krebs, 2002). V té zkoumá události z teroristických útoků z 11. září 2001, ze kterých manuálně vytvořil kriminální síť a ověřil, že hlavním aktérem byl vůdce této skupiny. Cílem výběru těchto dat je otestování tvorby kriminální sítě a ověření, zda v ní budou hlavní pachatelé těchto útoků označeni jako hlavní aktéři, dále otestování prostorové a časové analýzy, kterou, vzhledem k tomu, že je tato událost všeobecně známá, můžeme vyhodnotit.

Data o událostech v Paříži a Belgii byla ručně stažena ze zpravodajského serveru iDnes. K útokům v Paříži byly staženy všechny články, které byly označeny tagem „situace po teroristických útocích v Paříži“[1], od počátku útoků až do dne 29.3.2016. Celkem se jedná o 147 článků. Na přiloženém DVD jsou uloženy jako soubory PX.txt. K belgickým útokům pak bylo staženo 51 článků, které byly označeny tagem „teror

[1] `http://zpravy.idnes.cz/situace-po-teroristickych-utocich-v-parizi-fek-/zahranicni.aspx?klic=64265`

v Bruselu“[2]. Opět byly staženy všechny články ke stejnému datu. Články jsou uloženy jako BX.txt. Obě události byly analyzovány jak zvlášť, tak dohromady.

Data pro testování a ověřování funkčnosti programu

Pro testování a vývoj programu bylo dále staženo 30 pseudonáhodných článků o politickém dění v ČR. Cílem tohoto testovacího souboru je ověřit tvorbu sociální sítě na malé sadě známých jmen, na které může být program testován a dále vyvíjen. Data byla opět stažena ze serveru iDnes.

7.2 Tvorba sociální sítě

Jak bylo uvedeno v části 4.4, principem kriminální sítě je vytvořit sociální síť z účastníků kriminálních činů. V této síti pak můžeme analyzovat, kteří aktéři jsou významní a jak kriminální síť destabilizovat. To můžeme uplatnit v případě, že se budeme snažit „rozbít“ např. drogový nebo pouliční gang či teroristickou buňku. Dále můžeme analyzovat jak síť efektivně infiltrovat nebo které spojení je vhodné monitorovat. Pro zjednodušení, v našem případě nebereme v úvahu možnou nekompletnost této sítě a predikce možných spojení, které nám nejsou známy.

Vytvořená síť může být jak anonymizovaná, tak neanonymizovaná. Anonymizovaná verze je vhodná pro další analýzy. Ty mohou provádět i nepověřené osoby. Žádné osobní údaje v ní nejsou obsaženy. Osoby, které se podařilo rozpoznat jsou zde již anonymizované. Ty, které se rozpoznat a anonymizovat nepodařilo zde obsaženy nejsou. Po určení důležitých aktérů, či spojení mezi aktéry, je možné pomocí programu osoby deanonymizovat a uplatnit tak získané poznatky. Deanonymizaci mohou provést pouze osoby, které mají k dispozici databázi, ve které je mapování anonymizovaných a původních entit uloženo.

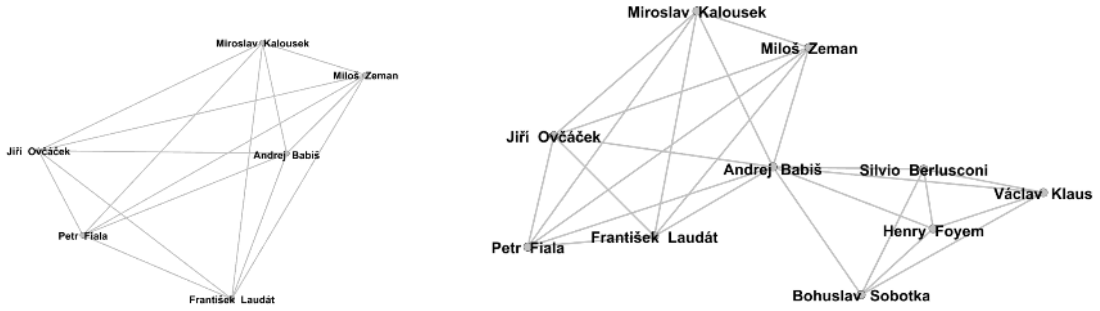
Vygenerovaná síť je ve formátu .net. Soubor je možné otevřít v programech Gephi nebo Pajek, ve kterých je možné provádět další analýzy vygenerované sítě. Simulaci destabilizace sítě a určení významných aktérů provádí program automaticky.

Postup vytváření sociální sítě

Do sítě jsou zahrnovány pouze osoby (objekty typu Person). Osoby jsou vytvářeny z NE typu „osobní jména“ podle algoritmu 2, který je uveden v části 6.5. Aby byla osoba do sítě přidána, je nutné, aby bylo uvedeno jak její jméno, tak její příjmení (volitelně i datum narození). Tím je zaručena identifikace mezi různými osobami napříč dokumenty. Zároveň je tím snížena chybovost vytváření osob, které ve skutečnosti osobami nejsou.

[2] `http://zpravy.idnes.cz/teror-v-bruselu-0i8-/zahranicni.aspx?klic=64275`

Postup tvorby sítě zobrazuje obr. 7.1 (příklad neuvažuje hranové ohodnocení, které je popsáno dále). Mějme dokumenty A a B. V dokumentu A se vyskytují osoby *{J. Ovčáček, P. Fiala, F. Laudát, A. Babiš, M. Kalousek a M. Zeman}*. V dokumentu B se vyskytují osoby *{A. Babiš, S. Berlusconi, H. Foyem, B. Sobotka, V. Klaus}*. Při zpracování dokumentu jsou všechny osoby, které se v něm vyskytují mezi sebou propojeny (obr. 7.1a). Pokud se již některá z těchto osob nachází v seznamu osob, jsou k ní přidány hrany k osobám ze zpracovávaného dokumentu (obr. 7.2b). Tímto způsobem můžeme propojit libovolné množství dokumentů. Pokud se žádná osoba z dokumentu v seznamu osob nenachází, je vytvořena nová komponenta v síti. Jak ukazuje obr. 7.1b, dokumenty A a B z předchozího příkladu tak můžeme propojit osobou *A. Babiše*.



(a) *Síť vytvořená z dokumentu A.* (b) *Síť vytvořená z dokumentů A a B.*

Obr. 7.1: *Ukázka tvorby sociální sítě z dokumentů.*

Osoby, pomocí kterých jsou tato propojení realizována, jsou podle ohodnocení metod centralit významné (viz část 3.3). Jelikož se de facto jedná o propojení komunit, které se v dokumentech nacházejí, budeme dále používat ohodnocení aktérů dle míry „Betweenness centrality“ (C_B). C_B považuje za významné právě ty aktéry, kteří různé komunity propojují.

Jelikož nerozpoznáváme typ událostí nebo relací, kterými jsou osoby spojené, uvažujeme všechny hrany za neorientované. Ve vytvořené síti není možné, aby hrana z uzlu v vedla do uzlu v. Tato hrana by v našem případě byla nesmyslná[3].

Určení váhy hrany

V dokumentech jsou u každé osoby počítány koreference. Osoba je reprezentována vrcholem. Počet výskytů osoby (resp. koreferencí) v dokumentu odpovídá ohodnocení vrcholu, který tuto osobu reprezentuje (viz příklad na obr. 7.2). Váha hrany je vypočtena dle vzorce (7).

$$e_k(v_i, v_j) = \frac{v_{ic} + v_{j_c}}{2} \cdot \frac{|V|}{\sum\limits_{x \in V} v_{xc}} + e_{k-1}(v_i, v_j)\,, \tag{7}$$

[3]Vyslýchaná osoba by např. označila za komplice či za dalšího člena gangu sama sebe.

kde v_i a v_j jsou vrcholy, jejichž váhu hrany určujeme, přičemž $v_i \neq v_j$. Počet koreferencí osoby v dokumentu, kterou reprezentuje vrchol v_i je značen jako v_{ic}, obdobně pak v_{j_c} u vrcholu v_j. Počet osob v dokumentu označujeme jako $|V|$ a $e_{k-1}(v_i, v_j)$ je váha hrany vypočtená z předchozích dokumentů mezi vrcholy v_i a v_j (k udává počet dokumentů, ve kterých se osoby společně vyskytly). Pokud hrana dříve neexistovala tj. $k = 1$ (pro první společný výskyt osob v dokumentech), platí $e_{k-1}(v_i, v_j) = 0$. Jako $\sum v_{xc}$ značíme součet koreferencí všech osob v dokumentu.

Vzorec vychází z následujících úvah:

- Pokud se budou dvě osoby vyskytovat společně ve více dokumentech, je třeba váhu hrany zvýšit. To zajišťuje člen $e_{k-1}(v_i, v_j)$.
- Pokud se některé osoby budou v dokumentu vyskytovat častěji, jsou v dokumentech významné a jejich váha hrany by měla být větší než u osob, které se v dokumentu vyskytují méně často. Vycházíme z myšlenky, že např. v dokumentu : „*Novák udává, že šéfem drogového kartelu je Corleone. Novák pracuje u Corleoneho rok jako dealer. Corleone řídí několik dalších dealerů. Dvěma z nich jsou Mrázek a Josh ...*“ jsou významné osoby Novák a Corleone. Novák a Corleone tak mezi sebou budou mít hranu jejíž ohodnocení je v rámci dokumentu největší, budou následovat hrany Corleone s Mrázkem a Joshem, poté Novák s Mrázkem a Joshem. Hranu s nejmenším ohodnocením v rámci dokumentu mezi sebou budou mít Mrázek a Josh. Ve vzorci (7) toto zajišťuje výraz $(v_{ic} + v_{j_c})$.
- V rámci dokumentu je vhodné váhu hrany normalizovat. Pokud budeme mít dokument pouze se dvěma osobami, je vhodné aby měly při libovolném počtu výskytů shodnou váhu. To zajišťuje výraz $|V| / (2 \cdot \sum_{x \in V} v_{xc})$.

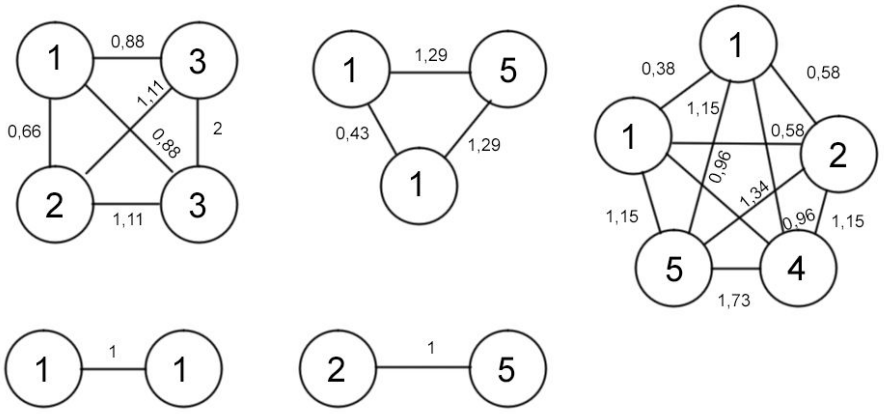
Příklad ohodnocení hran dle vzorce (7) na různých dokumentech ukazuje obr. 7.2 (uvažujeme první společný výskyt, tj. člen $e_{k-1}(v_i, v_j) = 0$). Vrcholové ohodnocení udává počet výskytů osoby v dokumentu. Každá komponenta značí jeden dokument.

Ukázka výpočtu pro dokument se třemi osobami z příkladu na obr. 7.2:

$$
\begin{aligned}
e_k(v_a, v_b) &= \tfrac{1+1}{2} \cdot \tfrac{3}{7} + 0 \doteq 0,43\,, \\
e_k(v_b, v_c) &= \tfrac{1+5}{2} \cdot \tfrac{3}{7} + 0 \doteq 1,29\,, \\
e_k(v_c, v_a) &= \tfrac{5+1}{2} \cdot \tfrac{3}{7} + 0 \doteq 1,29\,.
\end{aligned}
$$

Takto vypočtená síla hrany odpovídá na otázku, kterým spojením je vhodné věnovat zvýšenou pozornost. Jak ukazuje tabulka 7.1, jsou nejvýznamnější hrany ve většině případů shodné s detekovanou prominentní komunitou (viz 4.5 a 7.4), jejíž určení též na tuto otázku odpovídá[4].

[4]Může se v nich však nacházet více osob, tímto se od určení síly hrany odlišují.



Obr. 7.2: *Příklad ohodnocení hran podle vzorce (7).*

Tab. 7.1: *Nejvýznamnější hrany a jejich porovnání s prominentními komunitami. Síť byla vytvořena z 30 článků zabývajících se politickým děním v ČR.*

#	Aktér A	Aktér B	Síla hrany	V prom. komunitě
1.	A. Babiš	M. Kalousek	7,7	x
2.	M. Zeman	V. Havel	6,0	x
3.	M. Kalousek	P. Fiala	5,7	x
4.	A. Babiš	B. Sobotka	5,1	x
5.	M. Zeman	J. Ovčáček	4,9	x
6.	A. Babiš	P. Fiala	4,2	x
7.	M. Zeman	V. Klaus	3,6	–
8.	A. Babiš	P. Gazdík	2,9	–
9.	J. Paroubek	R. John	2,8	–
10.	M. Zeman	Pavel Mareš	2,7	–

Prominentní komunity v tabulce 7.1 byly vytvořeny s parametrem $minSup = 3$, tj. za prominentní komunitu budou považovány osoby, které se společně vyskytují ve 3 dokumentech (tj. 10% dokumentů z 30 testovaných dokumentů). Pro tento parametr bylo nalezeno 5 prominentních komunit s 8 různými osobami. Jediná komunita, která se v takto ohodnocených hranách nevyskytuje, je {M. Kalousek, K. Schwarzenberg}. Vyskytují se spolu totiž pouze v dokumentech, ve kterých je velké množství osob a dominantním je zde pouze jeden z nich. Vzhledem k tomu, že několik článků se

týkalo zavedení EET pokladen a udělení amnestie prezidentem Pavlu Marešovi, lze takto ohodnocené hrany, i s přihlédnutím na prominentní komunity v tabulce 7.1, považovat za správné.

Testování tvorby sociální sítě

Vytvořená sociální síť z testovacích dat o politickém dění v ČR obsahuje 113 osob. Při bližší analýze bylo zjištěno, že se v síti vyskytuje 11 (tj. 9,74%) chybně vytvořených nebo duplicitních osob. U těch se v některém kontextu nepodařila korektní lemmatizace nebo jsou vytvořeny z chybně rozpoznaných pojmenovaných entit. Jedná se např. o osoby „svatý Václav“, „White Media“, „Andrea Vrbovský“, „Libanonec Alí Fajádem“ atd.

Součet stupňů chybně vytvořených aktérů je 6,1% (53) z celkového počtu (868). Též byly ověřeny další metody pro výpočet centralit. Žádný z chybných nebo duplicitních vrcholů není v síti významný. Jejich hodnota C_B je 0 a dle metody C_C jsou řazeny na posledních místech. Při analýze dalších sítí se však ukázalo, že zejména u arabských, nebo jiných cizích jmen bude tato hodnota pravděpodobně vyšší. Morfologickému analyzátoru se u těchto jmen v některém kontextu nepodaří zcela správná lemmatizace a mohou se tak objevit osoby jako Salah Abdeslam, Salahus Abdeslam, Salaha Abeslama apod., které odkazují na stejnou osobu.

V budoucnu je tak vhodné zvážit, zda nepoužít Levenštejnovu vzdálenost a podobná jména neslučovat. Je však nutné brát v úvahu, že tímto způsobem mohou být spojeny osoby, u kterých to není žádoucí (např. Miroslav Novák a Miloslav Novák).

7.3 Destabilizace kriminální sítě

V této části naleznete několik informací o sociálních sítích, které byly podle části 7.2 sestaveny z osob nacházejících se v novinových článcích o Pařížských a Bruselských teroristických útocích (viz část 7.1). Jak bylo zmíněno v části 4.4, jedná se o účastníky[5] trestných činů, nebo vyšetřované osoby. V následujících sítích, jelikož jsou tvořeny z novinových článků, se navíc vyskytují i osoby, které se k těmto činům vyjádřily. Tato část dále uvádí tabulku nejvýznamnějších aktérů, kteří se v síti na-

[5]Nezabývali jsme se rozlišováním pachatelů od dalších účastníků trestných činů. V dalším textu se tak v kriminálních sítích nacházejí i osoby, které kriminálníky jistě nejsou. Lze namítnout, že kriminální sítě nejsou zcela přesné označení. Předpokládá se však, že na ostrých datech se osoby, které nejsou podezřelé či v minulosti trestané, příliš často vyskytovat nebudou. Pokud ano, bude se pravděpodobně jednat pouze o několik výskytů, což by na výsledky nemělo mít vliv. Pokud se budou opakovat vícekrát, dochází na nich k opakované viktimizaci (viz část 4.2), což je též důležitý poznatek. Je však nutné, aby získané výsledky byly správně interpretovány.

cházejí. V závěru této části je ukázáno, jak kriminální sítě efektivně destabilizovat[6] a návrh jak tyto sítě infiltrovat.

Makroskopické údaje o sítích

Tabulka 7.2 zobrazuje makroskopické údaje o vytvořených sociálních sítích, s nimiž budeme dále pracovat.

Xu a Chen (2005) popisují analýzu kriminálních sítí, které tvoří členové drogového a pouličního gangu. Největší komponenty (K_{max}) v jejich práci obsahují 502 aktérů (narkotický gang, $|V| \sim 13$ tisíc, $K_{max} \sim 3{,}9\%$) a 2595 aktérů (pouliční gang, $|V| \sim 4{,}4$ tisíc, $K_{max} \sim 59{,}4\%$). První síť v jejich práci obsahovala 2618 komponent (K) s více než 2 členy, druhá pak 284 komponent.

Tab. 7.2: *Informace o vytvořených sociálních sítích. P pro pařížské útoky, B pro belgické útoky a S pro spojení obou útoků. $|V|$ počet vrcholů, $|E|$ počet hran, $\langle k \rangle$ průměrný stupeň vrcholu, ℓ průměrná nejdelší cesta, D nejdelší nejkratší cesta, Δ hustota sítě, C koeficient shlukování, K počet komponent, K_{max} % vrcholů v největší komponentě.*

	$\|V\|$	$\|E\|$	$\langle k \rangle$	ℓ	D	Δ	C	K	K_{max}
P	337	1368	8,12	3,08	7	0,024	0,87	21	87,2%
B	132	392	5,94	3,08	6	0,045	0,91	16	68,9%
S	438	1736	7,93	3,14	7	0,018	0,88	33	81,7%

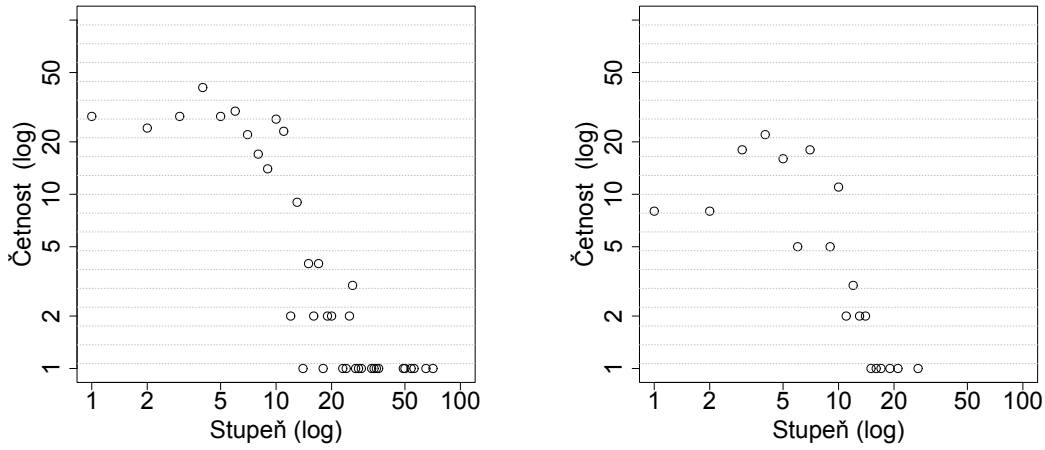
Pokud srovnáme tabulku 7.2 s dříve uvedeným, naše sociální sítě jsou proti reálným kriminálním sítím hustší (Δ), obsahují méně komponent a největší komponenty obsahují více procent vrcholů. Z tohoto a z povahy dat, předpokládám, že v reálných kriminálních sítích budou největší komponenty tvořeny maximálně několika tisíci aktéry (není příliš pravděpodobné, že v záznamech výslechů by se mohlo vyskytovat např. 50 tisíc osob, které by navíc tvořili jednu komponentu). Tato vlastnost nám dovolí provést dostatečně rychlý[7] výpočet i pro výpočetně náročnější destabilizace.

Průměrný stupeň vrcholu ($\langle k \rangle$) a obr. 7.3, vzhledem k rozdělení stupňů vrcholů, napovídá, kolik unikátních osob se průměrně v dokumentech vyskytuje. Jelikož jsou

[6] Tj. znemožnit osobám v těchto sítích vzájemnou spolupráci. Pokud bychom byli např. finančními, časovými nebo lidskými prostředky omezeni a nebylo by možné vyslýchat, vyšetřovat nebo sledovat všechny podezřelé, aktérům, kteří jsou v této simulaci destabilizace označeni na prvních místech, by měla být věnována přednostní pozornost

[7] Výpočet C_B pro tisíc vrcholů trval 5 vteřin. Pro 5 tisících vrcholů pak 3,5 minut. Výpočet C_B přepočítávané (po každém odebrání se výpočet C_B provede znovu) pak 4,2 minuty pro tisíc vrcholů. Test byl proveden na stroji AMD Phentom II X4 955, 12 GB RAM.

osoby u každého dokumentu spolu spojeny, odpovídá tomu i nezvykle vysoký koeficient shlukování (C, viz str. 15). Réka a Barabási (2002) často uvádějí hodnoty C v intervalu $\langle 0,2; 0,7\rangle$. Nejdelší nejkratší cesta (D), stejně jako nízká hodnota průměrné nejkratší (ℓ) cesty ukazuje na dobré propojení všech uzlů v síti. Již z rozložení stupňů v těchto sítích a hodnot ℓ, D a C můžeme říci, že tyto sítě jsou vytvořeny dle bezškálového modelu (viz část 3.4).



(a) *V pařížské síti.* (b) *V bruselské síti.*

Obr. 7.3: *Log-log rozložení stupně vrcholů ve vytvořených sociálních sítích.*

Nejvýznamnější aktéři

Osoby, které jsou označovány jako pachatelé útoků v Paříži a Bruselu a pokud bychom měli k dispozici data o vyšetřování, by se pravděpodobně objevovaly na nejvýznamnějších místech, ukazuje obrázek B.1 (viz příloha B).

Tabulka 7.3 ukazuje nejvýznamnější osoby, kterým dle hodnot C_B (viz vzorec 6) byla po útocích věnována největší mediální pozornost. Tabulka odpovídá nepřepočítávané hodnotě C_B, tj. výpočet je proveden pouze jednou a po odebrání nejvýznamnějšího aktéra nejsou hodnoty přepočítávány. Na předních pozicích jsou osoby, které často propojují jiné osoby z několika dalších článků. V případě reálných dat by se jednalo o propojení ve vyšetřovacích dokumentech. Určení významných aktérů pak může pomoci při vyšetřování trestných činů a pomoci tak odhalit možná spojení, které nemusí být na první pohled patrné.

V tabulce 7.3 jsou tučně označeni pachatelé. Kurzívou osoby, které můžeme považovat za chybné nebo duplicitní[8] (viz závěr části 7.2).

[8]Osoby nebyly úmyslně sloučeny, jelikož na reálných datech může být provedena anonymizace. V tomto případě tak nebude možné rozpoznat, zda osoby sloučit či nikoli.

Tab. 7.3: *Nejvýznamnější aktéři dle C_B. Tučně jsou vyznačeni pachatelé, kurzívou osoby, které lze považovat za chybně vytvořené.*

#	**Paříž**	**Brusel**	**Spojené události**
1	François Hollande	Jan Jambon	*Charlie Hebdo*
2	*Charlie Hebdo*	**Najim Laachraoui**	Bernard Cazeneuve
3	Manuel Valls	Daniel Šabík	François Hollande
4	Bernard Cazeneuve	Milan Chovanec	Manuel Valls
5	**Abdelhamid Abaaoud**	Bohuslav Sobotka	**Salah Abdeslam**
6	Bohuslav Sobotka	Charles Michel	Bohuslav Sobotka
7	**Salah Abdeslam**	*de Standaard*	**Abdelhamid Abaaoud**
8	Barack Obama	*Václav Havel*	Charles Michel
9	*Alláh Akbar*	**Salah Abdeslam**	Barack Obama
10	Angela Merkelová	**Brahim Bakraoui**	*Alláh Akbar*
11	Charles Michel	*Salahe Abdeslam*	Jan Jambon
12	Milan Chovanec	Koe Geens	Milan Chovanec
13	Joachim Herrmann	Bernard Cazeneuve	Angela Merkelová
14	Vladimir Putin	Lubomír Zaorálek	**Najim Laachraoui**
15	Volker Kluwe	Maggia de Blocková	Joachim Herrmann

Charlie Hebdo je v síti nejvýznamnější kvůli častému přirovnávání k útoku na redakci tohoto časopisu z 7. ledna 2015. Tato entita (v tomto kontextu by se spíše mělo jednat o instituci nežli jméno) propojuje velké množství dalších jmen vztažených k těmto útokům, které pomocí jiných osob propojeny nejsou. Další významné osoby jsou státníci, kteří útoky odsuzovali, vyjadřovali Francii podporu nebo soustrast a informovali o průběhu vyšetřování.

Dále následují údajní strůjci útoků – Abdelhamid Abaaoud a Salah Abdeslam[9]. V bruselské sítí jsou navíc významní bratři Bakraouiovi (Khalid je na 16. místě), kteří bruselské útoky spáchali. Osoba Abdelhamida Abaaouda se v bruselských článcích příliš často nevyskytuje, v té době byl již po smrti. Za zajímavé považuji, že

[9]Všimněte si, že i přes to, že jméno Salah Abdeslam je v několika případech chybně lemmatizováno (např. Salahe Abdeslam, Salahus Abdeslam apod.) je zde natolik významný, že ani tato chyba nemá na jeho umístění, zejména při sloučení obou událostí, významný vliv.

v celkové síti je na 14 místě Najim Laachrouri. Ten se přitom v Pařížské síti objevuje pouze v jednom článku. Dle článku v Lidových novinách[10] je to údajně strůjce bomb i pro pařížské atentáty a propojuje pravděpodobně s Abdeslamem pařížskou a bruselskou buňku (viz příloha C). Nejvýznamnějším pachatelem ve spojené síti je Salah Abdeslam, jehož osoba tyto dvě události významně propojuje. Jeho zadržení pravděpodobně zapříčinilo nebo urychlilo události, které se staly v Bruselu.

Destabilizace kriminální sítě

Tato část popisuje, jak kriminální sítě destabilizovat. Jak bylo uvedeno v části 3.4, míru destabilizace měříme dle relativní velikosti největší komponenty pomocí hodnoty S. Dle (Réka a Barabási, 2002) je destabilizace sítí, které jsou vytvořeny podle bezškálového modelu (viz část 3.4), bez předchozí analýzy obtížná. Everton (2008) ukazuje, že odebráním nejvýznamnějšího aktéra kriminální síť nezničíme. Naopak spíše situaci zhoršíme. Můžeme tím posílit aktéra, který se mohl s odebraným aktérem o moc dělit nebo o ni v kriminální síti dokonce soupeřit.

Vytvořený program provádí simulaci destabilizace sítě dle několika metod. Implicitně jsou prováděny čtyři náhodné destabilizace (vrcholy jsou odebírány bez předchozí analýzy náhodně), které slouží jako výchozí body, proti kterým jsou další metody ověřovány (v obr. 7.4 odstíny šedé). Při náhodných destabilizacích se ukazuje významná výhoda bezškálových sítí – jsou velmi robustní (hodnota S klesá pomalu).

Dalším způsob destabilizace, kterou program provádí, je odebírání vrcholů, který ukazují Réka a Barabási (2002). Tím je cílená destabilizace na základě odebírání vrcholů s největším stupněm (v obr 7.4 červeně, po každém odebrání jsou stupně vrcholu přepočítány). Destabilizace podle hodnot C_B je vyznačena zeleně a dle C_B přepočítávané (po každém odebrání vrcholu je hodnota C_B opět přepočítána) modře. Z obrázků je patrné, že C_B přepočítávaná dosahuje nejlepších výsledků. Je však výpočetně náročná. Může být až $O(n^4)$ a jelikož není předem jasné, na jak velké skupině aktérů bude tento výpočet prováděn, je implicitně tento výpočet společně s C_B vypnut (pokaždé se provádí C_D přepočítávaná). Jak bylo uvedeno dříve, nepředpokládá se však, že by kriminální sítě mohly být tvořeny více aktéry než v řádech tisíců a tato vlastnost by tak neměla způsobovat problémy.

Obrázky 7.4 ukazují, že všechna cílená odebrání jsou lepší, nežli náhodné odebírání a přepočítávané varianty jsou lepší, než jejich nepřepočítávané varianty (nepřepočítávanou C_D neuvádíme, proto vždy C_D vyjde lépe nežli nepřepočítávaná C_B). U destabilizace spojených sítí bylo dle C_D nutné oproti náhodnému odebírání (celkem 336, tj. 76,5% všech aktérů) odebrat o 65,8% méně aktérů (celkem 48 tj. 10,7% všech

[10] `http://www.lidovky.cz/kdo-je-najim-laachraoui-mistr-prevleku-schopny-pyrotechnik-a-komplic-teroristu-z-parize-gbm-/zpravy-svet.aspx?c=A160323_122148_ln_zahranici_msl`

aktérů). Oproti náhodnému odebírání je při C_B přepočítávané pro destabilizaci[11] sítě třeba odebrat o 69% procent méně vrcholů (celkem 34, tj. 7,5% všech aktérů). Při přepočítávané C_B je tak proti přepočítávané C_D potřeba odebrat o 3,2% aktérů z celkového počtu aktérů méně. Obrázky 7.4 dále ukazují, že pokud odstraníme pouze několik nejvýznamnějších aktérů, na síť jako celek to nemá téměř vliv.

Program generuje tabulky s návrhy destabilizace sítě do HTML souborů ve složce results (soubory destabilization_typ_destabilizace.html). V těchto souborech je uvedeno, kolik % aktérů se nachází v největší komponentě, kolik % jich již bylo odebráno a kterého aktéra v následujícím kroku ze sítě odebrat. Data jsou opět anonymizovaná a lze je pomocí programu deanonymizovat. Dále jsou generovány CSV soubory obsahující data pro tvorbu grafů, které jsou zobrazeny na obr. 7.4. Ten je možné zobrazit pomocí souboru destabilization.html v internetovém prohlížeči. Jelikož jsou data načítána z lokálních CSV souborů, je třeba v prohlížeči vypnout *Same-origin policiy* (viz příloha A.5). V případě reálných dat jsou osoby, které se v těchto souborech objeví, doporučeny pro bližší prozkoumání v uvedeném pořadí.

Návrh na infiltraci kriminální sítě

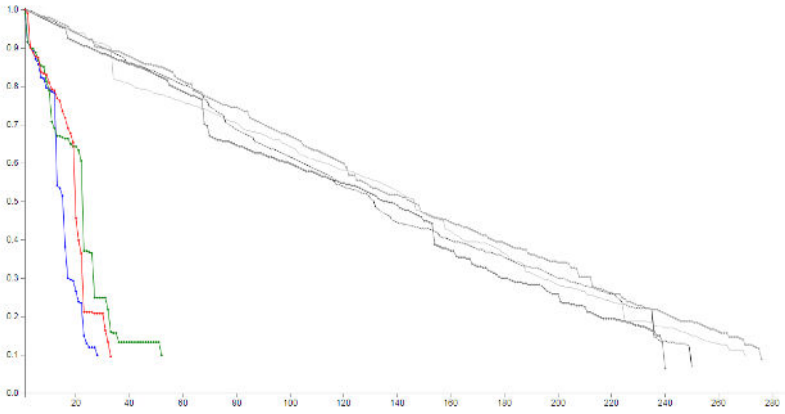
Infiltraci kriminální sítě je vhodné provádět pomocí aktérů, kteří jsou nejlépe ohodnoceni dle metody „Closeness centrality“ (C_C, viz část 3.3). Tito aktéři mají k ostatním aktérům nejblíže. Nemusí se však nutně jednat o nejvýznamnější aktéry, kteří by pravděpodobně nebyli ochotni spolupracovat, nebo by to jejich míra zapojení do kriminálních činů vylučovala. Metoda ohodnocování dle C_C je vhodná pro šíření virů v sítích, za což můžeme infiltraci považovat.

7.4 Detekce prominentních komunit

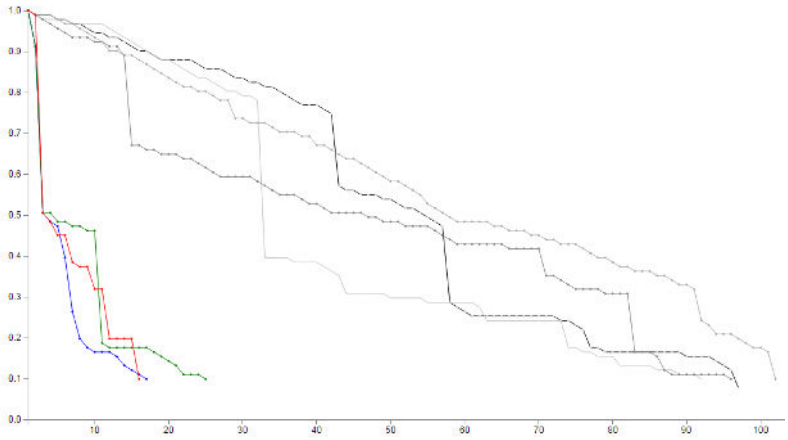
V části 3.4 byl popsán postup jakým detekovat největší možné prominentní komunity. Pomocí něj lze nalézt osoby, které se společně v dokumentech vyskytují alespoň n krát. Lze předpokládat, že tyto osoby jsou v jedné komunitě (v případě trestných činů se může např. jednat o komunitu zodpovědnou za provoz varny drog) a mohou být pro další analýzu významné. Odstraněním těchto komunit, zejména pokud se jich v síti objevuje pouze několik, můžeme efektivnost kriminální sítě (viz část 4.4) snížit[12].

[11] V našem případě považujeme síť za destabilizovanou, pokud je největší komponenta zmenšena na 10% původní velikosti. Tuto hodnotu je možné, dle potřeby, snadno měnit.

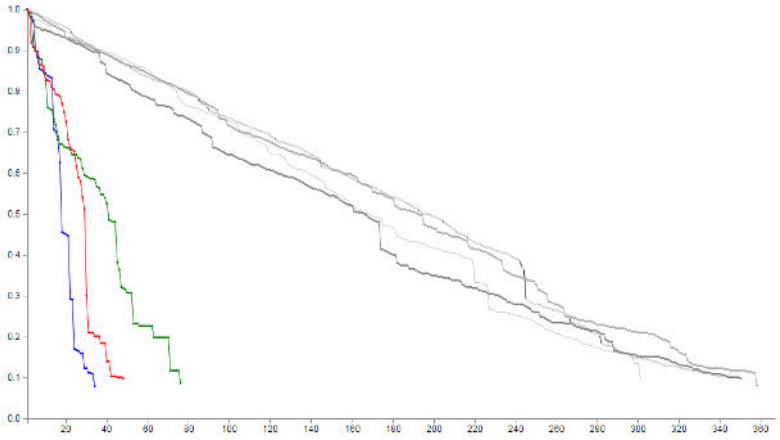
[12] V případě zatčení komunity, zodpovědné za distribuci drog bude kriminální síť do té doby, než zjedná nápravu paralyzovaná. Stejně tak jako při zatčení komunity zodpovědné za provoz varny drog.



(a) *Destabilizace pařížské sítě.*



(b) *Destabilizace bruselské sítě.*



(c) *Destabilizace spojených sítí.*

Obr. 7.4: *Destabilizace sociálních sítí. Osy x uvádějí kolik bylo odebráno vrcholů. Osy y míru destabilizace měřenou pomocí hodnoty S (viz. část 3.4). Šedě jsou označeny náhodné destabilizace, modře destabilizace podle C_B přepočítávané, zeleně dle C_B a červeně dle C_D přepočítávané.*

Zvýraznění zajímavých dokumentů

Program vygeneruje do složky results HTML soubor, kde jsou anonymizované prominentní komunity uvedeny. V ukázkovém adresáři (který není anonymizován) byly do souboru přidány k osobám odkazy na wikipedii, neboť některá jména nemusí být všeobecně známá. Ve vygenerovaném dokumentu jsou pro další manuální analýzy přidány odkazy na HTML verze všech původních dokumentů, ve kterých se tyto osoby nacházejí. Rozpoznané NE jsou v nich barevně zvýrazněny (viz část 6.6). Ve vygenerovaném souboru jsou dále uvedeny odkazy na dokumenty, ve kterých se nacházejí obě osoby. Tyto dokumenty jsou označeny jako „zajímavé dokumenty“. V poslední řadě je zde uveden počet společných výskytů osob v prominentní komunitě. Čím větší je počet společných výskytů, tím významnější může prominentní komunita být.

Získané výsledky z testovacích dat

Prominentní komunity z testovacích dat o politickém dění v ČR byly uvedeny v tabulce 7.1. Tabulka 7.4 ukazuje prominentní komunity v síti, vytvořené z pařížských a bruselských útoků. Jedná se o prominentní komunity s $minSup = 4$, tj. za prominentní komunitu jsou označeny osoby, které se spolu vyskytují alespoň ve čtyřech dokumentech. Při hodnotě $minSup = 2$ se zde vyskytovalo 96 prominentních komunit, při $minSup = 3$ pak 28. Jak je vidět, při nízké hodnotě $minSup$ je uživatel zahlcen výsledky, při vysoké hodnotě mohou uniknout některá významná spojení. Hodnotu $minSup$ může uživatel měnit (viz příloha A.4). Její optimální hodnota je závislá na použitých datech a počtu analyzovaných souborů. V našem případě na testovacích datech $minSup = 2$ odpovídá 1% dokumentů, $minSup = 3$ pak 1,5% dokumentů atd.

Při volbě $minSup = 3$ se v testovacích datech jako největší prominentní komunity, objevovaly 3-komunity (např. B. Sobotka, M. Zeman, M. Stropnický, kteří tyto útoky komentovali). Při volbě $minSup = 2$ pak 4-komunity. Jednou z nich jsou pachatelé útoku – A. Abaaoud, A. A Mohammad, B. Abdeslam a S. Abdeslam.

V budoucnu, by bylo vhodné, automaticky vygenerovat poznámku, z jakého důvodu se jedná o prominentní komunitu (např. stručně shrnout obsah „zajímavých dokumentů“) a do souboru s komunitami tuto poznámku přidat. Pokud by se v kriminální síti nacházelo více prominentních komunit, bylo by též možné vytvářet sociální síť pouze z těchto komunit. Mohly být propojeny na základě výskytu dalších osob nebo osob, které se v prominentních komunitách nacházejí (blíže viz Al-zaidy a kol. (2012)).

Tab. 7.4: *Prominentní komunity v novinových článcích z pařížských a bruselských útoků při minSup = 4. Z 2-komunit zde již nelze vygenerovat 3-komunity, jelikož se každá komunita vyskytuje v jiných souborech, viz příloha D.*

Prominentní komunita	**Společných výskytů**
{M. Chovanec , B. Sobotka}	8
{B. Sobotka , M. Zeman}	7
{F. Hollande , M. Valls}	7
{B. Cazeneuve, F. Hollande}	5
{S. Abdeslam, B. Cazeneuve}	4
{S. Abdeslam, A. Abaaoud}	4
{B. Cazeneuve, A. Abaaoud}	4
{B. Cazeneuve, B. Sobotka}	4
{B. Sobotka , P. Fiala}	4
{B. Sobotka , M. Stropnický}	4
{F. Hollande , A. Hidalgová}	4
{D. Cameron , B. Obama}	4
{B. Obama, V. Putin}	4

7.5 Šablony zločinného chování a statistické údaje

Tato část popisuje ukázku prostorové analýzy, která je tvořena pomocí tepelné mapy (viz část 4.2). Pomocí tepelné mapy je možné vizualizovat často se vyskytující místa, na kterých se mohou zločiny (na testovacích datech místa, které jsou v článcích zmiňována) často odehrávat. Dále je v této části uvedena ukázka časové analýzy (viz část 4.3) a několik statistik, které byly z textů získány. Na reálných datech lze jak prostorovou, tak časovou analýzu, považovat za část šablony zločinu či zločinného chování a pomocí těchto analýz se snažit zločiny predikovat a předcházet jim.

Omezení prostorové analýzy a implementační poznámky

V případě prostorové analýzy se objevil zásadní problém, který detailní prostorovou analýzu, podle které by např. bylo možné plánovat hlídky, znemožňoval. A to sice, jak získané adresy nebo místa, převádět na GPS souřadnice. K tomuto účelu je možné použít proces geokódování, který tento převod provede. Geokódování je

možné provést např. volně dostupnou službou, kterou nabízí Google – *Google Maps Geocoding API*[13]. U těchto volně dostupných služeb se však jedná o on-line řešení, která jistě příchozí požadavky monitorují. Vzhledem k tomu, že tyto informace mohou být osobními údaji, není je možné na tyto služby posílat. Možností, jak tento problém vyřešit, by bylo napojení na geografický informační systém, který by geokódování umožňoval. Potom by nebylo třeba sdílet tyto data s třetími stranami. Tato varianta by však vyžadovala užší spolupráci s PČR. V případě, že bude anonymizace provedena, není již zjevně možné geokódování provést. Dalším problémem, který se vyskytl, je nízká úspěšnost korektně určených hranic adres, ze kterých by bylo možné získat přesné GPS souřadnice. Problematické je též omezení požadavků volně dostupných služeb[14].

Vzhledem k výše uvedeným problémům je provedena pouze ukázka prostorové analýzy. Ta je, vzhledem k tomu aby nedošlo k nechtěnému odeslání osobních dat třetím stranám, spouštěna[15] volitelně a pouze mimo hlavní program. Program vygeneruje soubor places.csv, který tyto geografické názvy obsahuje. Po otevření souboru heatmap.html v internetovém prohlížeči (je třeba vypnout *Same-origin policy*, viz příloha A.5), jsou tyto místa následně odesílána na službu *Google Maps Geocoding API*, která tyto místa vrací jako GPS souřadnice. Ty jsou následně postupně zanášeny do mapy (*Google Maps*) a vizualizovány pomocí tepelné mapy.

Jelikož toto API umožňuje zaslání pouze 10 požadavků za vteřinu, je tím značně omezena velikost dat, na kterých lze tuto analýzu provádět. Ve vytvořeném skriptu se požadavek zašle každou půl-sekundu. Leč by tato hodnota mohla být menší, na základě experimentů byla zvolena tato hodnota jako nejvhodnější. Při nižších hodnotách se občasně vyskytovala chyba se zamítnutím dalších požadavků. Na testovacích datech netrvá odesílání požadavků déle než 30 sekund. Na mapu jsou dále pomocí značek přidány informace s označením, kolikrát se místa v textech vyskytují. Aby se zabránilo posílání požadavků, jako jsou např. státy, které nemůžeme korektně do mapy zanést, jsou do této analýzy zahrnuty pouze města, vesnice, náměstí a ulice (pojmenované entity typu GEOGRAPHICAL_CITY, GEOGRAPHICAL_STREET ale nikoli GEOGRAPHICAL_NAMES).

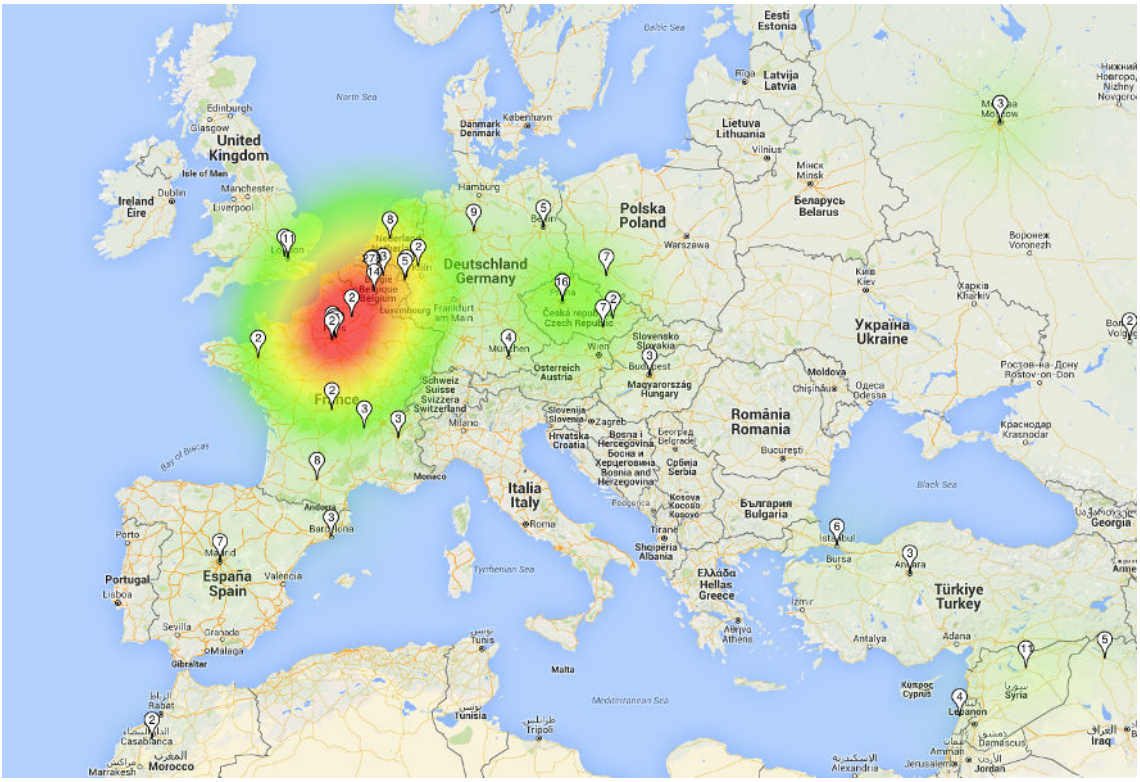
Prostorová analýza pařížských a bruselských útoků

Obr. 7.5 ukazuje prostorovou analýzu pařížských a bruselských útoků. Jako nejvýznamnější místa jsou správně určeny Paříž (492) a Brusel (278), jejich části či blízká města (Saint-Denis 16, Charleroi 14), dále hlavní města států, jejichž představitelé se k událostem vyjadřovali (Praha – 16, Moskva – 13, Londýn – 11 a Madrid – 7). Další je syrské Rakká (11). To je významné vzhledem k tomu, že útočníci dle dostupných údajů byli napojeni na teroristickou organizaci, která v této oblasti působí.

[13] `https://developers.google.com/maps/documentation/geocoding/intro`

[14] Při vytváření detailní prostorové analýzy by tak bylo nutné posílat všechny adresy (např. pouze se změněným číslem popisným) s časovým odstupem, což může být časově náročné.

[15] Formou HTML stránky s JavaScriptem.



Obr. 7.5: *Tepelná mapa útoků v Paříži a Bruselu. Mapa byla vygnerována pomocí Google Maps Geocoding API, Google Maps a novinových článků ze serveru* ldnes.

Časová analýza

Vytvořený program dále generuje statistiky, které je možné použít pro časovou analýzu. Jak bylo uvedeno v části 4.3, může být vhodná pro objevení opakujících se rytmů nebo souvislostí, které nemusí být na první pohled patrné. Rozpoznaná data (NE typu TIME_EXPRESION) jsou uložena do databáze (tabulka dates, viz obr. 5.1). Následně je do složky results z DB vygenerován CSV soubor dates.csv, který agregovaná data obsahuje. Jelikož jsou data agregovaná a není z nich možné určit k jaké osobě, události nebo dokumentu se vztahují, nelze je již považovat za osobní údaje[16]. Otevřením souboru dates.html lze zobrazit graf s daty, které se v textech nacházejí (opět je nutné mít v prohlížeči vypnut *Same-origin policy*, viz příloha A.5).

Obr. 7.6 ukazuje jednoduchou časovou analýzu na našich datech, ke které byl soubor dates.html použit[17]. Z obrázku jsou vidět data, při kterých došlo k významným událostem. Do obrázku byly manuálně přidány intervaly, ve kterých po sobě data bezprostředně následují[18], obrázky událostí a časy počátků událostí. Červeně jsou

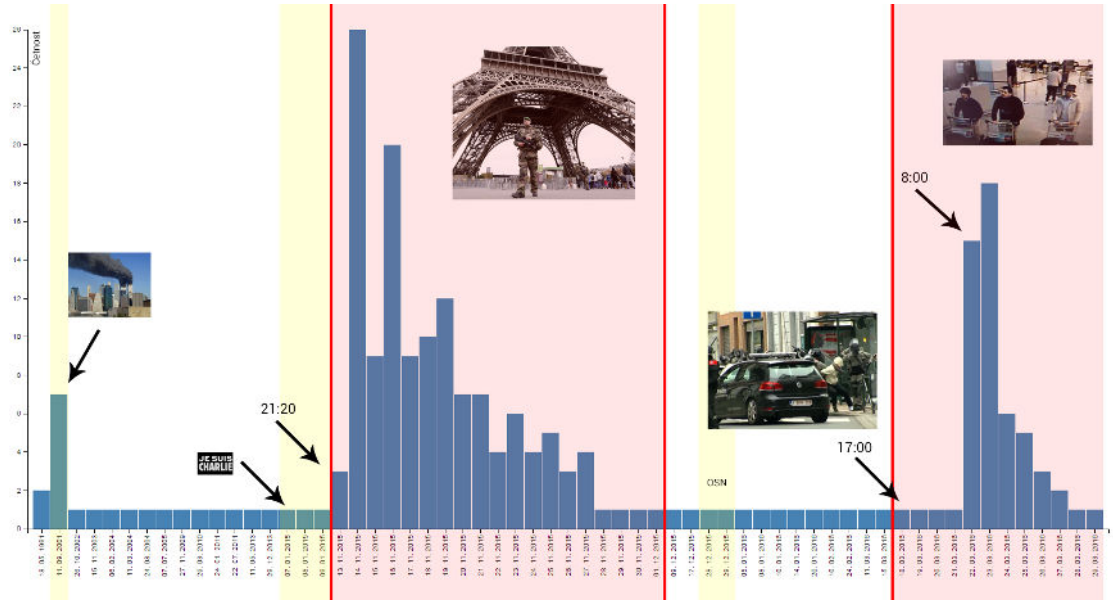
[16]Na rozdíl od místa ze kterého lze např. dohledat kdo zde má trvalé bydliště.

[17]Úpravami JavaScriptu v tomto souboru lze např. docílit vykreslení pouze hodnot až od zvolené prahové hodnoty, vyfiltrování několika měsíců nebo let apod.

[18]V grafu jsou zahrnuty pouze dny, které se v textech objevují, proto je nejmenší hodnota jedna. Vzdálenost na x-ové ose tak neodpovídá „časové vzdálenosti“.

označeny intervaly, které jsou vztaženy k očekávaným událostem. Žlutě pak události, které nemusí být na první pohled patrné a vyplynuly až z této časové analýzy.

Označené části, zleva doprava, jsou: události z 11. září 2001, ke kterým jsou pařížské útoky přirovnávány; útok na Charlie Hebdo, který se též odehrál v Paříži; pařížské útoky ze 13. listopadu 2015; shrnutí roku tajemníkem OSN; zatčení Salaha Abdeslama dne 18. března 2016; bruselské útoky ze dne 22. března 2016. Na vyznačených místech je vidět vliv nedělí, což vzhledem k tomu, že se jedná o novinové články, není překvapivé. Program zpracovává i dny v týdnu, se kterými je v budoucnu též možné pracovat (viz obr. 5.1, tabulka days). Nezpracováváme však výrazy jako „dnes“, „včera“, „v pátek“ apod. Pouze ty NE, u kterých je explicitně uvedeno datum.



Obr. 7.6: *Časová analýza provedená na testovacích datech. Červeně jsou označeny intervaly, které jsou vztaženy k očekávaným událostem. Žlutě pak události, které vyplynuly až z časové analýzy. Obrázky byly převzaty ze serveru* ldnes.cz.

Další statistiky

Program generuje do souborů ve složce results několik statistik. Kromě dříve uvedených lokací a dat se jedná o počet mužů a žen vyskytujících se v dokumentech a často se vyskytující NE.

Při ukládání osoby je na základě morfologické značky rozhodnuto o tom, zda se jedná o mužské či ženské jméno a následně je podle toho určeno pohlaví osoby. Pohlaví osoby je ukládáno do tabulky persons (viz obr. 5.1), kde je uloženo jako atribut osoby (pokud není z morfologické značky jasné o jaké jméno se jedná, je pohlaví uloženo jako neznámé).

Další statistikou, která může pomoci odhalit šablony zločinného chování, je generována do souboru importantNE.html. Do tohoto souboru jsou uloženy (anonymizované) nejčastěji se vyskytující NE a počet jejich výskytů. Jejich minimální výskyt je možné nastavit parametrem programu (viz příloha A.4). Jelikož u některých NE může být jejich opakovaný výskyt téměř bezinformativní (např. nejčastější jméno osoby Petr.), je v souboru uváděn rovněž typ entity. V ukázkové složce byly přidány opět odkazy na wikipedii a nebyla provedena anonymizace.

Zajímavými v této tabulce jsou zejména NE typu ARTIFACT_NAMES. U pojmenované entity typu artefakt mohou být odhaleny např. často odcizované předměty nebo předměty k trestným činům často používané. Nalezenými artefakty v našich datech jsou: AK-47 – zbraň, která byla k útokům použita; Facebook – sociální síť, která ve spojitosti s útoky zavedla novou funkci; Renault Clio a Volkswagen Polo – vozidla, která útočníci použili k dopravě na místo činu, nebo po nich bylo vyhlášeno pátrání; Bataclan – místo, kde k jednomu z útoků došlo.

Dalším zajímavým údajem by bylo zjištění četností různých trestných činů. To by bylo možné např. pomocí Bayesovského klasifikátoru. Bylo by však třeba mít k dispozici reálná data, na kterých by byl tento klasifikátor natrénován. Jistě by dále bylo zajímavé sledovat jaké typy zločinů se v dokumentech nejčastěji vyskytují (tj. do které kategorie byl dokument zařazen) a následně se např. pokusit vypočítat průměrnou nebo nejvyšší způsobenou škodu (můžeme k tomu použít NE typu SPECIFIC_NUMBER_USAGES a ARTIFACT_NAMES). Tato statistika by mohla zároveň posloužit jako nástroj pro ověření účinnosti zavedeného preventivního opatření. Např. po zavedení nového dopravního značení by jsme mohli sledovat, zda se škoda způsobená dopravními nehodami snižuje či nikoli.

8 Závěr

V práci byl popsán návrh a vytvořena část systému, který by mohl např. investigativním reportérům nebo policii usnadnit práci při vyšetřování, které může vycházet z analýzy nestrukturovaných dokumentů. Vytvořený program s pomocí několika volně dostupných knihoven nalezne v nestrukturovaných dokumentech pojmenované entity, anonymizuje je a umožní tak jejich další zpracování i nepověřeným osobám. Pověřeným osobám je dále umožněno provést deanonymizaci a mohou tak uplatnit získané poznatky.

Největší pozornost byla věnována analýze osob, ze kterých je vytvořena kriminální síť, jenž může být dále analyzována. V souladu s tímto bylo též nutno řešit rozpoznávání koreferencí entit. Po vytvoření této sociální sítě program formou HTML stránek navrhne, jak tuto síť destabilizovat při zachování nízkého počtu osob, kterým je třeba věnovat pozornost. Bylo ukázáno, že pro destabilizaci vytvořených sítí, je při cíleném odebírání (např. zatknutí či výslechu podezřelého) třeba odstranit o 69% méně aktérů nežli při náhodném odebírání bez předchozí analýzy. Při cíleném odebírání jsou použity metody centralit měřené stupněm uzlu a mezilehlostí uzlu.

Dále program detekuje komunity osob, které se spolu často v dokumentech nacházejí a patrně tak mezi sebou mají důležité spojení. Spolu s ohodnocením hran v síti, které bere v úvahu četnost výskytů osob v jednotlivých dokumentech, poukazují na zajímavá spojení, na které se vyšetřovatelé mohou zaměřit. Též byla ukázána jednoduchá prostorová a časová analýza spolu s několika dalšími statistikami, které je možné ze zkoumaných dat vytěžit.

Program byl testován na datech z novinových článků z teroristických útoků v Paříži a Bruselu. Získané výsledky se shodují s obecně uváděnými informacemi. Program byl také otestován v prostředí PČR na 200 dokumentech s reálnými daty, ze kterých byla vytvořena kriminální síť. Vzhledem k některým specifikům reálných dat (např. velká část příjmení nebo měst je psána verzálkami, jména jsou často za příjmeními atd.) je před možným použitím systému vhodné tyto data předpřipravit či na nich provést trénování. Vzhledem k tomu, že jsou však v těchto dokumentech obsaženy osobní údaje a není tak možné s nimi pracovat přímo, by to vyžadovalo intenzivnější spolupráci s PČR. Za zajímavé pro další práci, kromě zvýšení úspěšnosti rozpoznávání na reálných datech, považuji např. klasifikaci dokumentů, výpočet průměrné nebo nejvyšší škody dle jednotlivých trestných činů, provedení experimentů s infiltrací kriminální sítě, predikci hran mezi aktéry nebo vytvoření vzhledného uživatelského rozhraní. U toho je vhodné zvážit, zda do něj nezakomponovat možnost manuálních zásahů např. ve formě opravy typu pojmenované entity nebo opravy její hranice.

Literatura

ČADA, Roman, RYJÁČEK, Zdeněk a Kaiser, Tomáš, 2004. *Diskrétní matematika*. Západočeská univerzita.

AL-ZAIDY, Rabeah, FUNG, Benjamin CM, YOUSSEF, Amr M a FORTIN, Francis, 2012. Mining criminal networks from unstructured text documents. *Digital Investigation*, **8**(3), 147–160.

BADER, David A a MADDURI, Kamesh, 2006. Parallel algorithms for evaluating centrality indices in real-world networks. In: *Parallel Processing, 2006. ICPP 2006. International Conference on*, s. 539–550. IEEE.

BUŠTÍKOVÁ, Lenka, 1999. Analýza sociálních sítí. *Sociologický časopis*, **35**(2), 193–206.

ČESKO, 2015. Zákon č. 101/2000 sb., o ochraně osobních údajů a o změně některých zákonů, ve znění účinném od 1. ledna 2015. *In: Sbírka zákonů*.

CHMELAŘ, Petr, HELLEBRAND, David, HRUŠECKÝ, Michal a BARTÍK, Vladimír, 2011. Nalezení slovních kořenů v češtině. In: *Znalosti 2011: Sborník příspěvků 10. ročníku konference*, s. 66–77. VŠB-Technical University of Ostrava. ISBN 978-80-248-2369-0.

CLARKE, Ronald V. a ECK, John E., 2010. *Analýza kriminality v 60 krocích*. ProPolice.

CVRČEK, Václav, 2015. *Český národní korpus* [online]. [Cit: 12.3.2016]. Dostupné z: `https://www.korpus.cz`.

EVERTON, Sean S, 2008. Tracking, destabilizing and disrupting dark networks with social networks analysis.

FREEMAN, Linton C, 1978. Centrality in social networks conceptual clarification. *Social networks*, **1**(3), 215–239.

HANNEMAN, Robert A a RIDDLE, Mark, 2005. *Introduction to social network methods*. University of California Riverside.

HRUŠKA, Lubor, FUJAK, Radek a ŠEVČÍK, Jří, 2015. *Mapy budoucnosti*. Vědecovýzkumný ústav ACCENDO – Centrum pro vědu a výzkum, z. ú.

KONKOL, Michal a KONOPÍK, Miloslav, 2011. Maximum entropy named entity recognition for czech language. In: *Text, Speech and Dialogue*, s. 203–210. Springer.

KONKOL, Michal a KONOPÍK, Miloslav, 2013. Crf-based czech named entity recognizer and consolidation of czech ner research. In: *Text, Speech, and Dialogue*, s. 153–160. Springer.

KRÁL, Pavel, 2011. Features for named entity recognition in czech. In: *KEOD*, s. 437–441. SciTePress.

KRAVALOVÁ, Jana a ŽABOKRTSKÝ, Zdeněk, 2009. Czech named entity corpus and svm-based recognizer. In: *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, s. 194–201. Association for Computational Linguistics.

KREBS, Valdis E, 2002. Mapping networks of terrorist cells. *Connections*, **24**(3), 43–52.

NADEAU, David a SEKINE, Satoshi, 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3–26.

PISKORSKI, Jakub a YANGARBER, Roman, 2013. Information extraction: Past, present and future. In: *Multi-source, multilingual information extraction and summarization*, s. 23–49. Springer.

RÉKA, Albert a BARABÁSI, Albert-László, 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, **74**(1), 47.

RYJÁČEK, Zdeněk, 2014. *Teorie grafů, diskrétní optimalizace a výpočetní složitost 1.* Západočeská univerzita.

SATHYADEVAN, Shiju, DEVAN, MS a SURYA Gangadharan, S, 2014. Crime analysis and prediction using data mining. In: *Networks & Soft Computing (ICNSC), 2014 First International Conference on*, s. 406–412. IEEE.

ŠEVČÍKOVÁ, Magda, ŽABOKRTSKÝ, Zdeněk a KRŮZA, Oldřich, 2007a. Named entities in czech: annotating data and developing ne tagger. In: *Text, Speech and Dialogue*, s. 188–195. Springer.

ŠEVČÍKOVÁ, Magda, ŽABOKRTSKÝ, Zdeněk a KRŮZA, Oldřich, 2007b. *Zpracování pojmenovaných entit v českých textech*. Universitas Carolina Pragensis.

SPARROW, Malcolm K, 1991. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, **13**(3), 251–274.

STRAKOVÁ, Jana, STRAKA, Milan a HAJIČ, Jan, 2013. A new state-of-the-art czech named entity recognizer. In: *Text, Speech, and Dialogue*, s. 68–75. Springer.

STRAKOVÁ, Jana, STRAKA, Milan a HAJIČ, Jan, 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, s. 13–18.

USHA, D a RAMESKKUMAR, K, 2014. A complete survey on application of frequent pattern mining and association rule mining on crime pattern mining. *International Journal of Advances in Computer Science and Technology*, **3**(4).

XU, Jennifer J a CHEN, Hsinchun, 2005. Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems (TOIS)*, **23**(2), 201–226.

Obsah DVD

Přiložené DVD obsahuje následující soubory:

- bin – spustitelná verze programu
- data
  - corpus-example – ukázkový soubor použitého korpusu
  - debug-data – testovací data s politickým děním v ČR
  - test-data – články zabývající se teroristickými útoky v Paříži a Bruselu
- javadoc
- libs
  - jars – použité knihovny
  - rars – oficiální distribuce NameTagu a MorpohoDiTy
- poster – poster ve tvaru .pub a .pdf
- results
  - non-anonymized-example – neanonymizovaný příklad spuštěný na složce test-data, obsahuje odkazy na wikipedii
  - pcr – kriminální síť a prominentní komunity z reálných dat (všechny data jsou anonymizována)
- src – zdrojové kódy k programu (pro přeložení je třeba připojit soubory ze složky libs/jars)
- CriMiner.rar – archiv s programem, totožný se složkou bin
- DIP-Marek-Naggy.rar – zdrojové dokumenty elektronické verze BP
- DIP-Marek-Naggy.pdf – elektronická verze DIP
- person-network.net – neanonymizovaná verze sítě vytvořená ze složky test-data

Přehled zkratek

ACE	Automatic Content Extraction
API	Application Programming Interface
CO	Coreference resolution
CoNLL	The Conference on Natural Language Learning
CSS	Cascading Style Sheets
CSV	Comma-separated values
FN	False negative
FP	False positive
GIS	Geographic information system
GPS	Global Positioning System
HTML	HyperText Markup Language
IE	Information extraction
IMEI	International Mobile Equipment Identity
IREX	Information Retrieval and Extraction Exercise
MUC	The Message Understanding Conferences
NE	Named entity
NER	Named entity recognition
RE	Relation extraction
RTM	Risk Terrain Modelling
SVM	Support vector machines
TEI	Text Encoding Initiative
TF	True negative
TP	True positive
XML	Extensible Markup Language

Seznam obrázků

Seznam tabulek

Přílohy

A Uživatelská příručka

A.1 Výstupy programu

Anonymizované soubory jsou ukládány do složky output/anonymized. Originální soubory s označenými entitami do složek output/orig a output/html. Generované výsledky jsou ukládány do složky results. V této složce se žádné neanonymizované výsledky nevyskytují.

Složka results po dokončení programu obsahuje:

- destabilization...procedure.html – destabilizace kriminální sítě - na které osoby se v síti zaměřit. Soubory rand jsou ukázky náhodných destabilizací.
- dots – CSV soubory s daty pro vykreslení grafu destabilizace (pomocí souboru destabilization.html).
- importantNE.html – často se vyskytující pojmenované entity (min. výskyt lze volit parametrem).
- dates.html – ukázka časové analýzy.
- heatmap.html – ukázka prostorové analýzy.
- prominentComunity.html – prominentní komunity s odkazy na příslušné texty (ty nejsou v tomto adresáři obsaženy).
- person-network-anonymized.net – anonymizovaný soubor se sítí osob na základě jejich společného výskytu v dokumentu.

A.2 Spuštění programu

Jelikož program používá knihovny, které byly vytvořeny v jazyce C++ a jsou na Javu bindovány, je v závislosti na používané verzi Javy třeba použít patřičné DDL knihovny. Defaultně jsou v programu používány 64b. verze. Pro použití 32b verze je třeba zkopírovat soubory ze složky DDL32 do hlavní složky s programem. Doporučená verze Javy je 1.8. Program byl vyvíjen a testován na operačním systému Windows 7 a Windows 8.1.

Program lze spustit příkazem:

```
java -jar criMiner.jar data
```

Parametr data udává složku, která obsahuje soubory, které mají být zpracovány.

Po spuštění programu budou z této složky zpracovány všechny textové soubory. Program umožňuje zpracování dokumentů v kódování UTF-8 a Windows-1250. Program funguje dávkově a je možné zpracovat pouze adresář, který je předaný jako parametr. Po opětovném spuštění budou stará data nahrazena novými, nelze tak zpracovávat postupně více adresářů. Parametr se složkou s daty je vždy povinný. Archiv obsahuje testovací data s teroristickými útoky v Paříži a Bruselu (složka test-data), na kterých je program možné vyzkoušet.

A.3 Deanonimizace

Po prvním spuštění programu (které textové soubory zpracuje) je možné data deanonymizovat parametry:

- -DE X – pro deanonymizaci konkrétní entity (např. určené z importantNE.html nebo anonymizovaných XML souborů).
- -DP X – pro deanonymizaci konkrétní osoby (např. určené z prominentComunity.html nebo kriminální sítě).
- -DA – pro deanonymizaci všech souborů ze složky anonymized do složky output/deanonymized

Např. příkazem :

```
java -jar criMiner.jar data -DP 5
```

je možné deanonymizovat osobu s identifikátorem 5.

A.4 Další parametry

- -COM X – udává, kolikrát se osoby spolu musí vyskytnout, aby byly považovány za prominentní komunitu (výchozí hodnota je 3). Odpovídá parametru *minSup*, viz části 4.5 a 7.4.
- -NE X - udává, kolikrát se NE musí v textech vyskytnout, aby byla zahrnuta do seznamu v souboru importantNE.html (výchozí hodnota je 5), viz část 7.5.
- -SB - při zadání bude simulována destabilizace pomocí „Betweennes centrality“. Výpočet však v závislosti na velikosti sítě může být výpočetně náročný, viz části 3.3 a 7.3.
- -SBR - podobně jako předchozí parametr. Při každém odebrání se však „Betweennes centrality“ přepočítává. Je výpočetně náročnější nežli SB, viz část 7.3.

Parametry je možné kombinovat. Př. spuštění programu s dalšími parametry:

```
java -jar criMiner.jar data -COM 4 -NE 10 -SB -SBR
```

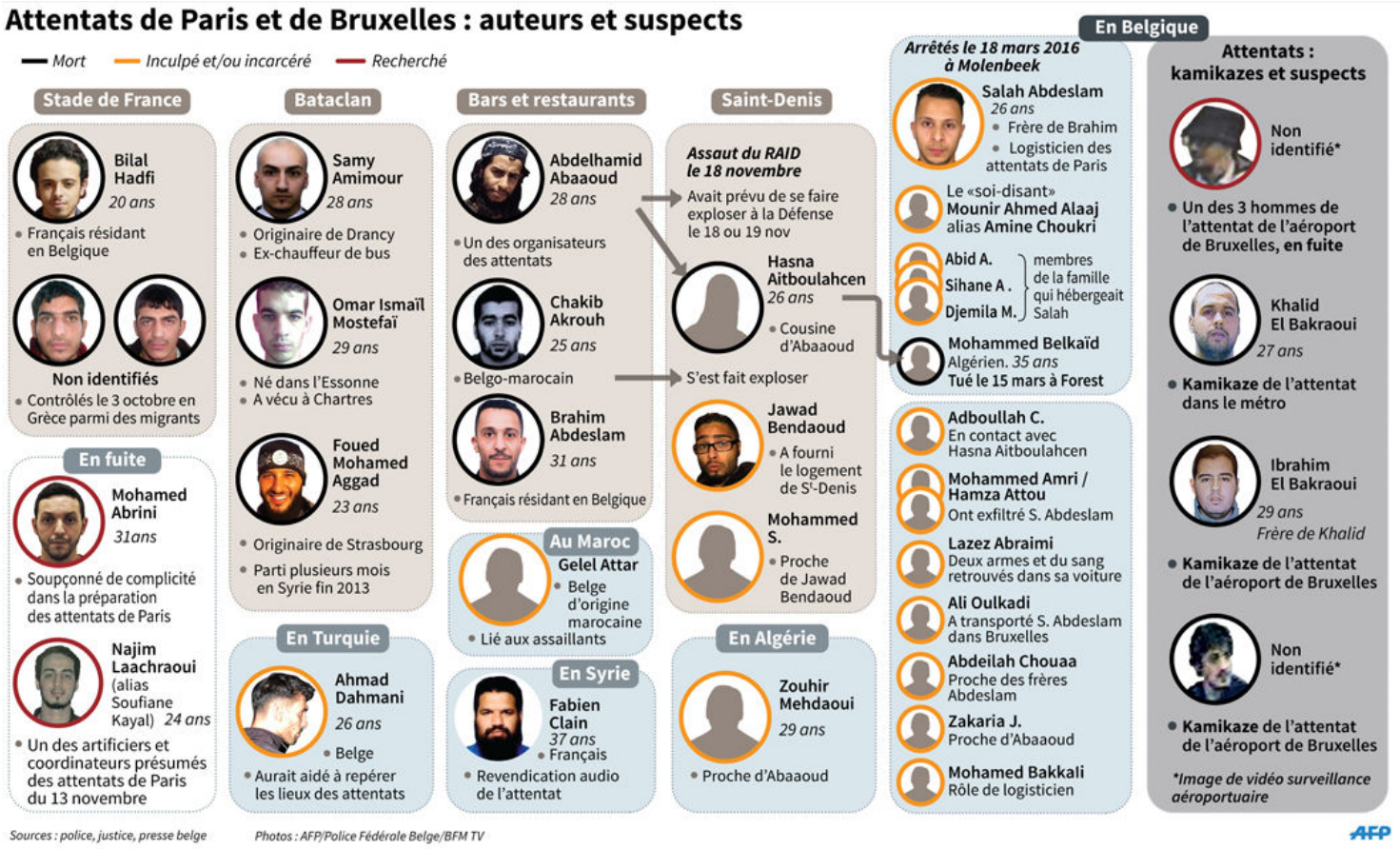
A.5 Same-origin policy a zobrazení grafů

Program vytváří CSV soubory, které jsou následně použity pro zobrazení grafů v prohlížeči (např. graf s destabilizací, daty nebo heatmapu). Jelikož však některé prohlížeče nedovolují načítat podobné soubory ani z lokálního úložiště, je pro zobrazení grafů nutné tuto vlastnost vypnout nebo použít lokální webový server (např. XAMPP).

Pro prohlížeč Chrome toho lze docílit následujícím postupem:

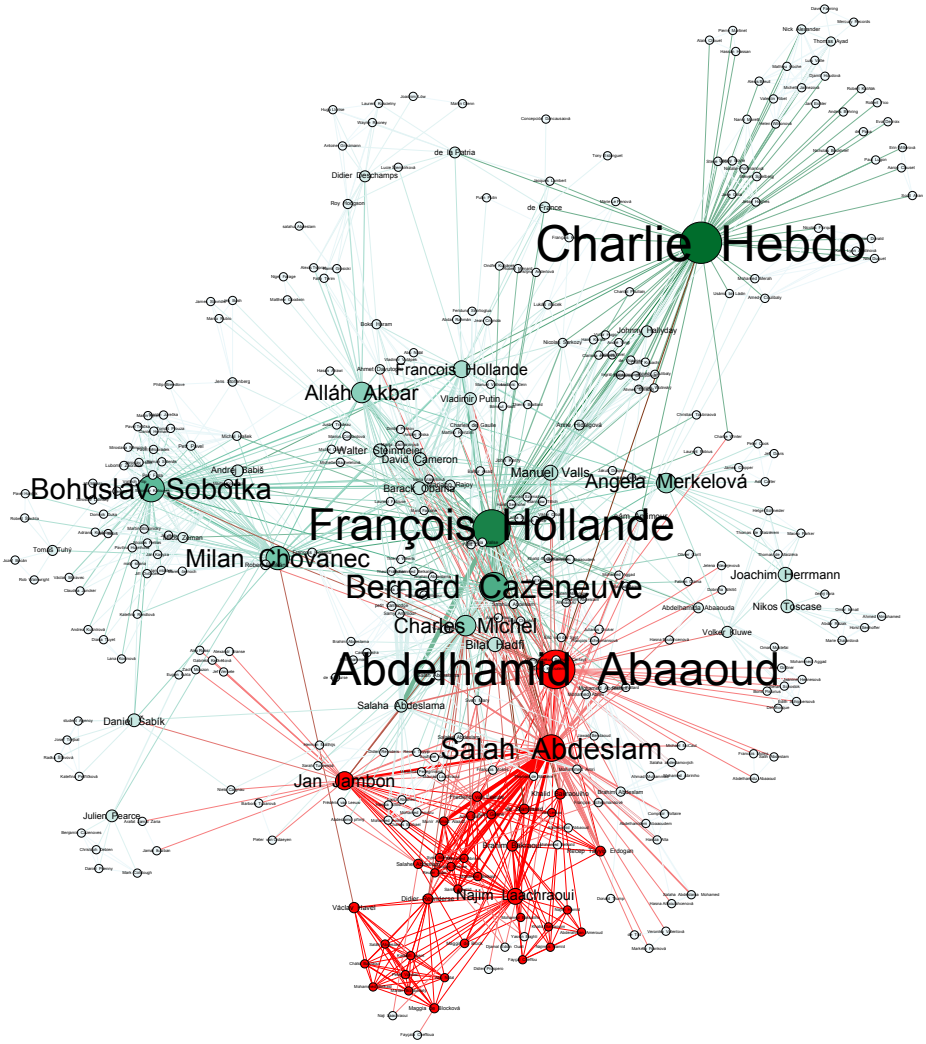
- Zavřít všechny stávající okna prohlížeče.
- Přejít do adresře, kde je prohlížeč nainstalován (nejčastěji C:/Program Files (x86)/Google/Chrome/Application).
- V tomto adresáři spustit příkazovou řádku a příkazem chrome.exe --allow-file-access-from-files spustit prohlížeč.
- Následně je možné grafy zobrazit.

B Pachatelé a podezřelí z útoků v Paříži a Bruselu



Obr. B.1: *Pachatelé a podezřelí z teroristických útoků v Paříži a Bruselu. Převzato z AFP (2016).*

C Ukázka vytvořené sítě



Obr. C.1: *Síť z článků o teroristických útocích v Paříži a Bruselu. Velikost je určena dle metody C_B (pro vizualizaci byl použit program Gephi, nebere tak v potaz váhu hran). Červeně byl obarven Najim Laachouri a všechny jeho sousední uzly. 50% jich tvoří útočnici či podezřelí.*

D Ukázka prominentních komunit

Ukázku výstupu programu – prominentních komunit, zobrazuje obr. D.1. Komunity již nelze dále sloučit do 3-komunit. Každá 2-komunita se totiž vyskytuje v jiných dokumentech. Jména osob nebyla anonymizována. V opačném případě se zobrazuje např. jako „Osoby [16, 96]“. Deanonimizaci lze provést provést pomocí parametru -DP (viz příloha A.3).

Osoby [Milan Chovanec , Bohuslav Sobotka]

16 v dokumentech: 12.txt , 46.txt , 50.txt , P101.txt , P127.txt , P129.txt , P142.txt , P143.txt , P146.txt , P147.txt , P58.txt , P9.txt , P98.txt ,
96 v dokumentech: 45.txt , 46.txt , 48.txt , 50.txt , P116.txt , P122.txt , P127.txt , P139.txt , P14.txt , P141.txt , P143.txt , P146.txt , P147.txt , P19.txt , P58.txt , P62.txt , P98.txt ,
Společných výskytů: 8
Zajímavé dokumenty: P143.txt , P98.txt , 50.txt , P127.txt , P58.txt , 46.txt , P147.txt , P146.txt ,

Osoby [Bohuslav Sobotka , Petr Fiala]

96 v dokumentech: 45.txt , 46.txt , 48.txt , 50.txt , P116.txt , P122.txt , P127.txt , P139.txt , P14.txt , P141.txt , P143.txt , P146.txt , P147.txt , P19.txt , P58.txt , P62.txt , P98.txt ,
99 v dokumentech: 46.txt , 48.txt , P141.txt , P146.txt ,
Společných výskytů: 4
Zajímavé dokumenty: 48.txt , P141.txt , 46.txt , P146.txt ,

Osoby [Bohuslav Sobotka , Miloš Zeman]

96 v dokumentech: 45.txt , 46.txt , 48.txt , 50.txt , P116.txt , P122.txt , P127.txt , P139.txt , P14.txt , P141.txt , P143.txt , P146.txt , P147.txt , P19.txt , P58.txt , P62.txt , P98.txt ,
107 v dokumentech: 48.txt , P116.txt , P122.txt , P127.txt , P139.txt , P14.txt , P141.txt ,
Společných výskytů: 7
Zajímavé dokumenty: 48.txt , P116.txt , P141.txt , P14.txt , P127.txt , P139.txt , P122.txt ,

Osoby [Bohuslav Sobotka , Martin Stropnický]

96 v dokumentech: 45.txt , 46.txt , 48.txt , 50.txt , P116.txt , P122.txt , P127.txt , P139.txt , P14.txt , P141.txt , P143.txt , P146.txt , P147.txt , P19.txt , P58.txt , P62.txt , P98.txt ,
182 v dokumentech: P116.txt , P139.txt , P141.txt , P98.txt ,
Společných výskytů: 4
Zajímavé dokumenty: P98.txt , P116.txt , P141.txt , P139.txt ,

Obr. D.1: *Ukázka prominentních komunit.*