FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Hodnocení smluv zveřejňovaných v Registru smluv s ohledem na podezření z korupce
Student:	Bc. Jan Staněk
Vedoucí:	Ing. Daniel Vašata, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	Do konce letního semestru 2018/19

Pokyny pro vypracování

Cílem práce je navrhnout metriky pro hodnocení smluv zveřejňovaných v Registru smluv (smlouvy.gov.cz) tak, aby bylo možné snadněji nalézt smlouvy podezřelé z využití korupce. K tomu je kromě samotných smluv třeba využít data získaná z Hlídače smluv (www.hlidacsmluv.cz) a dalších veřejných zdrojů (například Věstníku veřejných zakázek, Veřejného rejstříku a Sbírky listin, a Insolvenčního rejstříku).

Detailní pokyny:

1. Získejte data z výše uvedených zdrojů a uložte je ve formátu vhodném pro další zpracování.
2. Proveďte analýzu získaných dat. Zaměřte se především na provázanost jednotlivých entit, které se v souvislosti se smlouvami objevují (firmy a jejich vlastníci).
3. Na základě předchozí analýzy navrhněte vhodné metriky pro hodnocení zveřejňovaných smluv.
4. Ověřte užitečnost navržených metrik na reálných datech.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Karel Klouda, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 5. února 2018

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Hodnocení smluv zveřejňovaných v Registru smluv s ohledem na podezření z korupce

Bc. Jan Staněk

Katedra aplikované matematiky
Vedoucí práce: Ing. Daniel Vašata, Ph.D.

1. května 2018

Poděkování

Rád bych poděkoval Ing. Danielu Vašatovi, Ph.D. za čas věnovaný vedení mé práce, cenné rady a připomínky. Společnosti CRIF – Czech Credit Bureau, Marku Sušickému a Michalu Bláhovi za poskytnutí dat. V neposlední řadě děkuji své rodině za podporu během celého studia.

Prohlášení

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

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

V Praze dne 1. května 2018 .....

Odkaz na tuto práci

Staněk, Jan. *Hodnocení smluv zveřejňovaných v Registru smluv s ohledem na podezření z korupce*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2018.

Abstrakt

Tato diplomová práce se zabývá návrhem metrik sloužících pro nalezení podezřelých smluv zveřejňovaných v registru smluv. Popsány jsou dostupné datové zdroje, kterými je možné data z registru smluv doplnit, integrace dat a výběr příznaků pro detekci anomálií. Vytvořené metriky usnadňují výběr smluv vhodných pro ruční kontrolu.

Klíčová slova analýza dat, detekce anomálií, detekce korupce, Hlídač Státu, integrace dat, předzpracování dat, registr smluv, veřejné datové zdroje

Abstract

This master's thesis describes the design of metrics for identification suspicious contracts published in the register of contracts. It describes public data sources suitable for supplement data from the register of contracts, data integration and feature selection for anomaly detection. Designed metrics simplifies selection of contracts suitable for manual review.

Keywords anomaly detection, data analysis, data integration, data preprocessing, detection of corruption, Hlídač Státu, public data sources, register of contracts

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Zákon o registru smluv, účinný od července 2016, nařizuje státu, obcím, jimi zřizovaným organizacím a dalším subjektům zveřejňovat všechny své smlouvy s hodnotou nad 50 000 Kč a zvyšuje tak transparentnost nakládání s veřejnými financemi. Veřejnost díky němu získala možnost kontrolovat hospodaření a také porovnávat ceny zboží a služeb pro různé úřady a firmy.

V současnosti je v registru smluv zveřejněno přes 1,2 milionu smluv a denně přibývají tisíce nových. Není tedy možné provádět ruční kontrolu každé zveřejněné smlouvy.

Tento problém jsem se rozhodl řešit pomocí data miningu, který se úspěšně používá při detekci podvodů v oblasti finančnictví nebo pojišťovnictví, a navržení metrik tak, aby bylo možné se při kontrole zaměřit na mnohem menší množství smluv.

Cíl práce

Cílem teoretické části práce je získání přehledu o veřejných datových zdrojích, významu dat z nich, jejich formátu, datové kvalitě a seznámení se s řešeními pro integraci dat, uložení dat a algoritmy vhodnými pro detekci problematických smluv.

Cílem praktické části práce je navržení ETL procesů (datových transformací ze zdrojových systémů) pro integraci a předzpracování dat, analyzování dat a navržení příznaků pro popis smluv.

Struktura práce

Práce je členěna do pěti kapitol. Kapitola „Zdroje dat“ se zabývá významem dat z jednotlivých datových zdrojů, jejich formátem a datovou kvalitou. V kapitole „Předzpracování dat“ je popsána volba technologií pro uložení dat, návrh datového modelu a ETL procesů pro integraci dat. Kapitola „Analýza dat“ se

věnuje analyzování dat z různých pohledů a získání informací pro návrh metrik. V kapitole „Detekce podezřelých smluv“ je popsáno hodnocení veřejných zakázek používané projektem DIGIWHIST, návrh příznaků, vizualizace mnohodimenzionálních dat pomocí Self organizing map a algoritmy používané pro detekci anomálií. Poslední kapitola „Vyhodnocení“ se zabývá vyhodnocením a srovnáním jednotlivých algoritmů pomocí ruční kontroly smluv označených jako nejvíce podezřelé.

Kapitola 1

Zdroje dat

Tato kapitola popisuje jednotlivé datové zdroje dostupné v České republice. Zabývá se významem dat, obsaženými informacemi a datovou kvalitou.

1.1 Registr smluv

Od 1. července 2016, kdy vstoupil v účinnost zákon o registru smluv [1], musejí vybrané subjekty uveřejňovat smlouvy s hodnotou vyšší než 50 000 Kč bez DPH do 30 dnů od jejich uzavření. Od 1. července 2017 platí, že smlouva, na niž se vztahuje povinnost uveřejnění, může nabývat účinnosti nejdříve dnem uveřejnění, pokud není smlouva uveřejněna ani do tří měsíců od uzavření, je zrušena.

Povinnost uveřejňovat smlouvy v registru smluv mají například tyto subjekty:

- Česká republika,
- územní samosprávné celky,
- příspěvkové organizace, ústavy a obecně prospěšné organizace založené státem nebo územními samosprávnými celky,
- veřejné výzkumné instituce nebo veřejné vysoké školy,
- státní podniky nebo národní podniky,
- zdravotní pojišťovny,
- Český rozhlas a Česká televize,
- právnické osoby, v nichž mají stát nebo územní samosprávné celky většinový podíl.

Naopak se povinnost nevztahuje na Kancelář prezidenta republiky, Poslaneckou sněmovnu, Senát, Ústavní soud, Nejvyšší kontrolní úřad a Kancelář veřejného ochránce práv, další výjimky jsou popsány v zákoně. [2]

Smlouva musí být uveřejněna v otevřeném a strojově čitelném formátu a obsahovat následující metadata:

- identifikaci smluvních stran,
- vymezení předmětu smlouvy,
- cenu, případně hodnotu předmětu smlouvy, lze-li ji určit,
- datum uzavření smlouvy.

Smlouvy nesplňující tyto podmínky nejsou považovány za zveřejněné a nemohou být platné.

1.1.1 Hlídač Státu

Data z registru smluv jsou od jeho počátku přístupná v XML formátu. Díky tomu mohly vzniknout nástroje jako Hlídač Státu [2]. Soubory jsou zveřejňovány po měsících, soubor s daty z aktuálního měsíce se aktualizuje každý den.

Hlídač Státu si klade za cíl lépe zpřístupnit data z registru smluv, informace z něj obohatit o data z dalších databází, identifikovat plýtvání a zneužití moci v úřadech, analyzovat data a umožnit analýzu i veřejnosti, zvýšit kontrolu veřejných prostředků ze strany občanů a propojit data z více datových zdrojů (v současnosti s informacemi o politicích a údaji z transparentních účtů).

Data jsou pravidelně stahována z měsíčních XML souborů. Textová vrstva je z příloh uložena do databáze a je tak možné vyhledávat i v textech smluv. Téměř 30 % dokumentů údajně tvoří naskenované dokumenty bez textové vrstvy, u nich je provedena OCR analýza a při úspěšném převodu je použit takto získaný text.

U každé smlouvy je provedeno ověření platnosti formálních údajů, zkontrolován správný výpočet cen s DPH a bez DPH, úplnost údajů, včetně validace pomocí dat z dalších rejstříků. Dále jsou analyzovány protiprávní kroky (například podepsání smlouvy dříve, než vznikla dodavatelská firma), vztahy politiků k dodavatelům, provádějí se statistické výpočty (například dodavatelé s největším podílem na zakázkách) a hledají se podezřelé okolnosti (například dělení zakázek tak, aby jejich hodnota byla nižší než limit specifikovaný v zákoně o zadávání veřejných zakázek, nebo smlouvy s nově založenými firmami).

Díky využití Elasticsearch [3] umožňuje Hlídač Státu pokročilé vyhledávání v metadatech i textech smluv.

Doplněná data jsou dále volně k dispozici prostřednictvím API nebo je možné je stahovat ve formátu JSON, buď jako jeden soubor, nebo soubory rozdělené po dnech.

1.1.2 Popis dat

Použitá data jsou získána stažením kompletního souboru z Hlídače Státu. Soubor obsahuje u každé smlouvy například tyto údaje:

- identifikátory smlouvy a verze smlouvy (unikátní),
- platnost záznamu (pokud má smlouva více verzí, jen jedna může být platná),
- předmět smlouvy,
- odkaz do registru smluv,
- čas zveřejnění,
- datum uzavření,
- číslo smlouvy (interní označení smlouvy),
- jméno osoby, která smlouvu schválila na straně plátce,
- hodnota bez DPH a hodnota s DPH,
- vypočtená cena s DPH,
- kvalita vypočtené ceny (cena může být spočítána z metadat, doplněna automaticky z textu smlouvy, nebo ručně),
- cizí měna (měna a hodnota v cizí měně),
- údaje plátce (název, IČO, adresa, ...),
- seznam příjemců (název, IČO, adresa, ...),
- seznam příloh (odkaz na soubor, metadata souboru, text extrahovaný ze souboru, kvalita extrakce textu, délka textu a počet slov),
- seznam chyb (typ, závažnost, popis),
- důvěryhodnost smlouvy (vypočtená podle chyb a jejich závažnosti),
- seznam provedených vylepšení (například extrahování textu nebo doplnění datové schránky).

Nalezených chyb se v datech vyskytuje 22 typů:

- identifikace smluvní strany (není uvedeno IČO smluvní strany, může se jednat o fyzickou osobu bez IČO nebo zahraniční subjekt, další informace je uvedena v popisu),
- chybí identifikace smluvní strany (není uvedeno IČO smluvní strany, může se jednat o fyzickou osobu bez IČO nebo zahraniční subjekt, další informace je uvedena v popisu),
- chybné strany smlouvy (plátce je i na straně příjemce),
- stejné strany smlouvy (plátce i příjemce je stejný),
- neexistující IČO (je uvedeno neexistující IČO, často chybí jen počáteční nuly, některé organizační složky státu pravděpodobně chybí v databázi, ve které se provádí kontrola, nebo se jedná o subjekt zapsaný ve Slovenské republice),
- vadné IČO (IČO nesplňuje kontrolní součet),
- smlouva uzavřena s nespolehlivým plátcem DPH (smlouva byla uzavřena v době, kdy byl příjemce evidován v registru nespolehlivých plátců DPH),
- firma vznikla až po podpisu smlouvy (často je uveden špatný datum uzavření smlouvy, u některých subjektů neodpovídá datum vzniku datu zápisu, například příspěvkové organizace založené před 1. 1. 2001 se neevidovaly v obchodním rejstříku [4], rejstřík spolků vznikl až 1. 1. 2014 apod.),
- firma vznikla krátce před podpisem smlouvy (smluvní strana vznikla méně než 60 dní před uzavřením smlouvy),
- budoucí datum uzavření (datum uzavření smlouvy je až po uveřejnění v registru smluv),
- neplatný datum uzavření smlouvy (datum uzavření smlouvy je v budoucnosti nebo je vyplněn nesmyslný datum),
- nulová hodnota smlouvy (smlouva nemá uvedenu hodnotu, v některých případech je utajení možné),
- neplatná cena (rozdíl ceny s DPH a bez DPH neodpovídá žádné sazbě DPH),
- cena bez DPH nulová (vyskytuje se pouze u smluv získaných z jiných zdrojů, před účinností zákona o registru smluv),

- cena s DPH nulová (vyskytuje se pouze u smluv získaných z jiných zdrojů, před účinností zákona o registru smluv),
- záporná cena bez DPH (záporná hodnota je možná pouze u dodatku smlouvy, kterým se snižuje hodnota smlouvy, většinou se jedná o chybně zveřejněné dodatky),
- záporná cena s DPH (záporná hodnota je možná pouze u dodatku smlouvy, kterým se snižuje hodnota smlouvy, většinou se jedná o chybně zveřejněné dodatky),
- bez DPH = s DPH (uvedená hodnota s DPH je stejná jako hodnota bez DPH),
- chybí předmět smlouvy (metadata neobsahují předmět smlouvy),
- nečitelnost smlouvy (text přílohy není strojově čitelný, výjimka platí pro smlouvy uzavřené před účinností zákona o registru smluv, které musejí být uveřejněny společně s uveřejněním dodatku [5]),
- smlouva byla znepřístupněna (vyskytuje se pouze u smluv získaných z jiných zdrojů, před účinností zákona o registru smluv),
- smlouva nespadá pod registr smluv (označuje smlouvy získané z jiných zdrojů, před účinností zákona o registru smluv).

Data ze 14. ledna 2018 obsahují 1 033 263 smluv.

Dále byl poskytnut seznam identifikátorů smluv, jejichž příjemci měly v posledních pěti letech ve svých orgánech politiky nebo sponzorovaly politickou stranu.

1.1.3 Kvalita dat

Data obsahují poměrně velké množství chyb v metadatech vyplňovaných ve formuláři pro zveřejnění záznamu [6].

U některých smluv je u smluvních stran nesprávné označení udávající, zda se jedná o plátce nebo příjemce. Smlouvu do registru smluv může vložit libovolná smluvní strana a tyto smluvní strany jsou označeny jako „publikující smluvní strana“ a „strany smlouvy“, ve formuláři ale není výběr povinný. Pokud tento údaj chybí, označuje Hlídač Státu jako plátce publikující smluvní stranu, pravděpodobně proto, že publikující smluvní strana bývá většinou ta, která má povinnost smlouvy uveřejňovat a s ohledem na typ těchto subjektů, bývá většinou na straně plátce. Problém se projevuje například u nájemních smluv, kde městský úřad pronajímá svůj majetek, nebo smluv, kde mají povinnost smlouvy uveřejňovat obě strany.

Dalším problémem souvisejícím s identifikací smluvních stran je uvedení publikující smluvní strany zároveň mezi stranami smlouvy. Problém nejspíše

způsobuje to, že uživatelé nejprve vyplní všechny smluvní strany smlouvy a následně nechají zaškrtnuté, aby se údaje publikující smluvní strany doplnily automaticky podle datové schránky odesílatele. U automaticky doplněné smluvní strany není ani možné zvolit, zda se jedná o plátce nebo příjemce. Tento problém se týká přibližně 7 600 smluv. Více než 17 500 smluv má obě smluvní strany stejné.

Formulář kontroluje pouze to, zda jsou vyplněna povinná pole, ale nekontroluje platnost údajů. Téměř 300 smluv může proto mít jako datum uzavření uveden nesmyslný datum (1. 1. 0001, 30. 12. 1899, 1. 1. 1900, 1. 1. 1970, ...) a u více než 600 smluv má některá smluvní strana uvedeno IČO nesplňující kontrolní součet (překlepy, 123456).

U části smluv je utajen dodavatel nebo hodnota, což je možné pouze v odůvodněných případech. Největší počet smluv s utajenou hodnotou i dodavatelem má Česká pošta s. p. (tvoří přibližně 80 % jejích smluv). 99,4 % smluv s utajenou hodnotou a 96,3 % smluv s utajeným dodavatelem má Budějovický Budvar, n. p., který podle iROZHLAS [7] proti zákonu dlouhodobě vystupoval a požadoval výjimku i jeho úplné zrušení.

Označení nalezených chyb není konzistentní pro všechny smlouvy, například „Identifikace smluvní strany“ se, mimo jiné, používá v případě, kdy je smluvní stranou fyzická osoba bez IČO, pro stejný případ se u jiných smluv používá „Chybí identifikace smluvní strany“. Mění se i závažnost chyby.

1.2 Veřejný rejstřík a Sbírka listin

1.2.1 Veřejný rejstřík

Veřejný rejstřík je informační systém veřejné správy. Podle zákona o veřejných rejstřících právnických a fyzických osob [8] obsahuje údaje právnických a fyzických osob evidovaných ve spolkovém rejstříku, nadačním rejstříku, rejstříku ústavů, rejstříku společenství vlastníků jednotek, obchodním rejstříku a rejstříku obecně prospěšných společností.

Do veřejného rejstříku se zapisují tyto údaje:

- jméno,
- sídlo nebo adresa místa pobytu, případně také bydliště, pokud se liší od místa pobytu,
- předmět podnikání,
- právní forma právnické osoby,
- den vzniku a zániku právnické osoby,
- u fyzické osoby datum narození a rodné číslo (není zveřejněno ve výpisu),

- identifikační číslo osoby přidělené rejstříkovým soudem,
- členové statutárního orgánu (jméno, adresa, způsob jednání, den vzniku a zániku funkce),
- členové kontrolního orgánu (jméno, adresa, den vzniku a zániku funkce),
- jméno, adresa a způsob jednání prokuristy,
- další skutečnosti stanovené zákonem,
- den k němuž byl zápis proveden.

Jednotlivé rejstříky jsou dále rozšířeny o další údaje, u obchodního rejstříku jsou to například:

- výše základního kapitálu společnosti s ručením omezeným nebo akciové společnosti,
- výše vkladů jednotlivých společníků.

Ne všechny subjekty, které mají přidělené IČO, jsou evidovány ve veřejném rejstříku (obce, školy, politické strany a hnutí, církve a náboženské společnosti, ...). Fyzické osoby podnikající jsou v obchodním rejstříku evidovány buď dobrovolně na vlastní žádost, nebo pokud výše jejich výnosů nebo příjmů snížených o daň z přidané hodnoty, je-li součástí výnosů nebo příjmů, dosáhla nebo přesáhla za dvě po sobě bezprostředně následující účetní období v průměru částku 120 000 000 Kč.

Součástí výpisu jsou také údaje o insolvencích získané z insolvenčího rejstříku.

1.2.2 Sbírka listin

Sbírka listin je součástí veřejného rejstříku. Vkládají se do ní například:

- zakladatelské listiny,
- rozhodnutí o volbě, jmenování nebo odvolání členů statutárního orgánu,
- výroční zprávy,
- účetní závěrky,
- rozhodnutí o zrušení právnické osoby.

Podle technické specifikace [9] musejí být listiny vloženy jako jeden soubor ve formátu PDF, nesmějí obsahovat informace, které nemohou být zveřejněny, velikost jedné stránky je omezena na 150 kB, dokumenty nesmějí být zašifrované a nesmí být omezen jejich tisk. Pokud je to možné, preferuje se, aby byla listina vytvořena převodem z textových dokumentů, a ne jako grafický obraz.

Účetní závěrky musejí podle zákona o účetnictví [10] od 1. 1. 2014 zveřejňovat všechny osoby zapsané ve veřejném rejstříku. Účetní závěrka může být součástí výroční zprávy.

1.2.3 Popis dat

Veřejný rejstřík neposkytuje API a podle podmínek provozu [11] je možné odeslat maximálně 3 000 požadavků denně a 50 požadavků během jedné minuty. Kvůli těmto omezením není možné získat údaje o všech evidovaných subjektech, ale bylo možné alespoň stáhnout údaje příjemců z registru smluv (přibližně 130 000 subjektů). Ukládá se pouze základní kapitál a informace o probíhajícím insolvenčním řízení.

Pro stahování dat byl vytvořen nástroj, který pro zadaný seznam IČO vyhledává subjekty v rejstříku a vybrané údaje získává parsováním HTML kódu výpisu aktuálních údajů. Nástroj umožňuje i stažení dokumentů požadovaného typu ze sbírky listin, z důvodu omezeného počtu požadavků a složitého strojového zpracování PDF dokumentů, se tato možnost nevyužívá. Původně bylo plánováno využít i údaje uvedené v účetních závěrkách, bohužel velká část listin byla do PDF převedena naskenováním vytištěných dokumentů (některé dokumenty jsou dokonce vyplněny ručně), ale ani ostatní listiny není možné jednoduše zpracovat kvůli velké rozmanitosti použitých formulářů nebo zveřejnění účetní závěrky jako součásti výroční zprávy.

1.2.4 Kvalita dat

Údaje jsou aktuální ke dni stažení (průběh ledna a února 2018), nemusejí tedy vypovídat o subjektu v době uzavření smlouvy.

Základní kapitál je v rejstříku zapsán jako text bez definovaného formátu a při jeho zpracování mohlo dojít ke ztrátě informace o tom, že je částka uvedena v jiné měně.

1.3 ARES

Účelem ARES [12] je souhrnné zpřístupnění údajů z informačních systémů pro evidenci ekonomických subjektů. Prohledávané zdrojové systémy jsou:

- veřejný rejstřík,
- živnostenský rejstřík,

- registr ekonomických subjektů,
- rejstřík registrovaných církví a náboženských společností,
- národní registr poskytovatelů zdravotních služeb,
- evidence zemědělských podnikatelů,
- seznam politických stran,
- rejstřík škol a školských zařízení.

Od 1. 1. 2017 je účinné Nařízení vlády o seznamu informací zveřejňovaných jako otevřená data, které definuje, jaká data musejí být zpřístupněna. Mezi ně patří i data ve veřejných rejstřících. Přístup přes API je ale omezen na 10 000 dotazů v době od 8:00 do 18:00 a 50 000 dotazů v době od 18:00 do 8:00. Přistoupit lze pouze k datům z veřejného rejstříku. Od prosince 2018 zveřejňuje ARES tato data také jako archiv XML souborů, jeden soubor pro každé IČO. Archiv je zveřejňován jedenkrát za měsíc, dále je poskytován seznam IČO, u kterých došlo ke změně od zveřejnění archivu.

1.3.1 Popis dat

XML soubory v archivu obsahují, mimo jiné, tyto údaje:

- údaje o výpisu:
  - nadpis,
  - datum a čas výpisu,
  - typ výpisu,
- základní údaje:
  - typ rejstříku,
  - IČO,
  - název,
  - sídlo (kód RÚIAN, stát, PSČ, okres, obec, část obce, městská část, ulice, číslo popisné, číslo orientační, číslo evidenční, ...),
  - datum zápisu,
  - datum výmazu,
  - předmět podnikání,
- statutární orgán:
  - název,

  - způsob jednání,
  - členové (vznik členství, zánik členství, název funkce, vznik funkce, zánik funkce, funkce, adresa, jméno a příjmení),

- jiné orgány:
  - název,
  - členové (vznik členství, zánik členství, název funkce, vznik funkce, zánik funkce, funkce, adresa, jméno a příjmení, případně název).

Informace o členech orgánů neobsahují data narození u osob, případně IČO u firem, není tak možné spolehlivě párovat členy orgánů napříč všemi subjekty z rejstříku.

Archiv z 9. 2. 2018 obsahuje údaje o 978 653 subjektech.

1.3.2 Kvalita dat

Téměř 13 % subjektů má u adresy sídla uvedeno neplatné adresní místo RÚIAN a neúplné údaje adresy.

1.4 Cribis

Cribis od společnosti CRIF – Czech Credit Bureau je datový zdroj s daty o subjektech získaných ze zdrojů jako: insolvenční rejstřík, sbírka listin, ARES, ... [13]

1.4.1 Popis dat

Cribis je jediný neveřejný zdroj dat použitý v této práci. Díky společnosti CRIF je možné použít data popisující vazby mezi subjekty a osobami.

Součástí jsou i data ze slovenských rejstříků a zahraniční subjekty, které mají vztahy se subjekty z ČR a SR.

Export z databáze byl vytvořen v únoru 2018 a obsahuje data 6 267 108 subjektů, 6 683 730 osob a 18 051 928 vazeb 114 různých druhů.

1.5 Veřejné zakázky

Věstník veřejných zakázek [14] je jednotným místem pro uveřejňování základních informací o veřejných zakázkách.

Podle zákona o zadávání veřejných zakázek [15] jsou zadavateli veřejných zakázek tyto subjekty:

- veřejní zadavatelé:
  - Česká republika,

  - Česká národní banka,
  - státní příspěvkové organizace,
  - územní samosprávné celky nebo jejich příspěvkové organizace,
  - právnické osoby založené nebo zřízené za účelem uspokojování potřeb veřejného zájmu, které nemají průmyslovou nebo obchodní povahu, a jiný veřejný zadavatel ji převážně financuje, může v ní uplatňovat rozhodující vliv nebo jmenuje nebo volí více než polovinu členů v jejím statutárním nebo kontrolním orgánu,

- zadavatelé:
  - osoby, které k úhradě nadlimitní nebo podlimitní veřejné zakázky použijí více než 200 000 000 Kč, nebo více než 50 % peněžních prostředků, poskytnutých z rozpočtu veřejného zadavatele, rozpočtu Evropské unie nebo veřejného rozpočtu cizího státu s výjimkou, kdy je veřejná zakázka plněna mimo území Evropské unie,

- zadavatelé sektorových zakázek.

Podle předmětu veřejné zakázky dělí zákon veřejné zakázky na druhy:

- veřejná zakázka na dodávky,
- veřejná zakázka na služby,
- veřejná zakázka na stavební práce.

Podle předpokládané hodnoty veřejné zakázky se určí její režim:

- nadlimitní veřejná zakázka (zakázka, jejíž předpokládaná hodnota je vyšší než limit stanovený nařízením vlády, v současnosti se jedná o 3 686 000 Kč u zakázky na dodávky nebo služby a 142 668 000 Kč u zakázky na stavební práce),
- podlimitní veřejná zakázka,
- veřejná zakázka malého rozsahu (zakázka, jejíž předpokládaná hodnota je rovna nebo nižší než 2 000 000 Kč u zakázky na dodávky nebo služby a 6 000 000 Kč u zakázky na stavební práce).

Zadavatelé jsou povinni dodržovat při zadávání veřejných zakázek tyto čtyři zásady:

- zásadu transparentnosti (předem stanovená kritéria výběru),
- zásadu přiměřenosti (parametry zadávacího řízení přiměřené charakteru veřejné zakázky),

- zásadu rovného zacházení (stejné podmínky pro všechny potencionální dodavatele),
- zásadu zákazu diskriminace (všichni zájemci mají stejnou příležitost zakázku získat). [16]

Veřejné zakázky malého rozsahu není zadavatel povinen zadat v zadávacím řízení, je ale povinen dodržet předchozí zásady. To může vést ke snaze rozdělit zakázku na více částí a vyhnout se tak povinnosti zadání v zadávacím řízení.

Ve výjimečných případech je možné zakázku zadat v jednacím řízení bez uveřejnění (JŘBU). Tento způsob je podle [17] nejméně transparentní a ve většině případů je osloven pouze jeden dodavatel. JŘBU je možné využít například v těchto případech:

- v předchozím otevřeném řízení nebyly podány žádné vhodné nabídky,
- z technických či uměleckých důvodů může být splněna pouze určitým dodavatelem,
- je nezbytné ji zadat v krajně naléhavém případě, který zadavatel nezpůsobil a ani jej nemohl předpokládat,
- jde o dodatečné dodávky od téhož dodavatele a změna dodavatele by znamenala nepřiměřené technické potíže,
- zakázka je navázána na soutěž o návrh.

Zadavatelé jsou povinni zveřejňovat výzvy k podání nabídky na profilu zadavatele, jehož odkaz zaregistrovali v informačním systému o veřejných zakázkách. Profil zadavatele může zadavatel vytvořit na svých webových stránkách nebo využít služeb jako jsou Portál pro vhodné uveřejnění [18], Profil zadavatele [19], Tender arena [20] apod.

1.5.1 Popis dat

Data o veřejných zakázkách se získávají pomocí nástroje public-contracts [21]. Nástroj v první fázi uloží URL adresy profilů zadavatelů, ve druhé fázi parsuje údaje zveřejněné na těchto profilech a ukládá je do PostgreSQL [22] databáze. Databáze obsahuje následující tabulky:

- `entity` (název subjektu, IČO, DIČ, typ subjektu, veřejný zadavatel),
- `contract` (název zakázky, zadavatel, kód zakázky, identifikátor zakázky na profilu zadavatele, druh zadávacího řízení, stav zadávacího řízení, rok stažení profilu zadavatele),
- `submitter` (vazba mezi subjektem zadavatele a zakázkou),

- `candidate` (vazba mezi subjektem uchazeče a zakázkou, nabídková cena),
- `supplier` (vazba mezi subjektem dodavatele a zakázkou, cena),
- `subsupplier` (vazba mezi subjektem subdodavatele a dodavatele).

Poslední aktualizace databáze proběhla v prosinci 2017.

1.5.2 Kvalita dat

Profily zadavatelů nemají jednotný formát a některé z nich nástroj nezvládne zpracovat.

Některá IČO v databázi obsahují mezery mezi skupinami číslic a bílé znaky na začátku nebo konci.

1.6 Registr územní identifikace, adres a nemovitostí

RÚIAN [23] je jedním ze čtyř základních registrů veřejné správy ČR. Evidují se v něm územní prvky: stát, kraje, obce, katastrální území, stavební objekty, adresní místa, ...

1.6.1 Popis dat

RÚIAN umožňuje stažení adresních míst ve formátu CSV. Soubory jsou rozdělené podle obcí a obsahují tyto údaje:

- kód adresního místa v informačním systému územní identifikace,
- kód obce,
- název obce,
- kód městské části,
- název městské části,
- kód městského obvodu Prahy,
- název městského obvodu Prahy,
- kód části obce,
- název části obce,
- kód ulice,
- název ulice,

- typ stavebního objektu (číslo popisné nebo číslo evidenční),
- číslo popisné,
- číslo orientační,
- znak čísla orientačního,
- PSČ,
- souřadnice Y v systému S-JTSK (uvedeno v metrech),
- souřadnice X v systému S-JTSK (uvedeno v metrech),
- platnost od (pokud jsou data převedena z informačního systému územní identifikace, je uvedeno 1. 7. 2011).

Kapitola 2

Předzpracování dat

Tato kapitola se v první části zabývá výběrem technologií pro uložení dat a návrhem datového modelu. V druhé části je popsáno čištění dat z dříve uvedených zdrojů a jejich nahrání do databáze vytvořené podle navrženého datového modelu.

2.1 Uložení předzpracovaných dat

Pro prohledávání vztahů mezi entitami v datech je výhodné mít je uložena ve formě grafu. Uzly grafu mohou představovat subjekty, osoby, smlouvy, zakázky a adresy. Hrany grafu mohou vyjadřovat vztah mezi entitami (společník, dodavatel veřejné zakázky, plátce smlouvy, ...).

2.1.1 Relační databáze

Graf je možné uložit ve standardní relační databázi, uzly a jejich vlastnosti jako tabulky a hrany s vlastnostmi jako vazební tabulky. Procházení grafu je možné pomocí operace JOIN. Pokud je použit MERGE JOIN a sloupce používané pro JOIN jsou seřazeny, je jeho složitost $O(N + M)$, kde N je počet uzlů v tabulce a M počet hran ve vazební tabulce. Přičemž v orientovaném multigrafu může být M řádově větší než N.

2.1.2 Grafové databáze

Grafová databáze je databázový stroj navržený pro uložení uzlů a hran. Díky tomu umožňuje reprezentovat složité vztahy mezi daty v přirozené formě. Každý uzel obsahuje seznam hran, procházení je tak možné v čase $O(1)$. Grafové databáze také často umožňují efektivní použití algoritmů z teorie grafů k nalezení vztahů, které nejsou na první pohled patrné. [24]

Využití grafové databáze je typicky výhodné, pokud data obsahují velké množství M:N vazeb, je třeba rychlé procházení mezi uzly nebo dotazování

založené na vztazích mezi daty. Nabízejí také větší flexibilitu (nemusejí mít pevně definované schéma). Naopak se grafové databáze nehodí na agregace počítané přes všechna data.

Existují různé grafové databáze (například Neo4j [25], OrientDB [26], ArangoDB [27], JanusGraph [28], ...), mezi sebou se liší podporou schéma, zálohování, replikace, škálování, ..., ale i dotazovacími jazyky (některé databáze sdílejí stejný dotazovací jazyk, ale neexistuje jednotný standard).

Neo4j

Neo4j byla zvolena zejména kvůli její rozšířenosti. První verze byla vydána v roce 2007 a dnes patří mezi nejpopulárnější grafové databáze. Dalším důvodem byly existující komponenty v Talend Open Studio for Big Data [29] a rozšíření umožňující připojení z vizualizačního nástroje Gephi [30].

Graf v Neo4j se skládá z uzlů a hran. Každý uzel může mít typ a libovolný počet atributů (klíč-hodnota). Všechny hrany jsou orientované, každá musí mít typ, počáteční uzel a koncový uzel. Stejně jako uzly mohou mít hrany libovolný počet atributů. Orientaci hran je při procházení grafu možné ignorovat, rychlost procházení je stejná pro oba směry. Databáze by měla být schopna pracovat na běžném hardware až s miliardami uzlů.

K datům je možné přistupovat pomocí Traversal framework nebo dotazovacího jazyka Cypher. Cypher vychází z jazyka SQL, původně byl vyvinut pro Neo4j, ale rozšířil se i do některých dalších grafových databází.

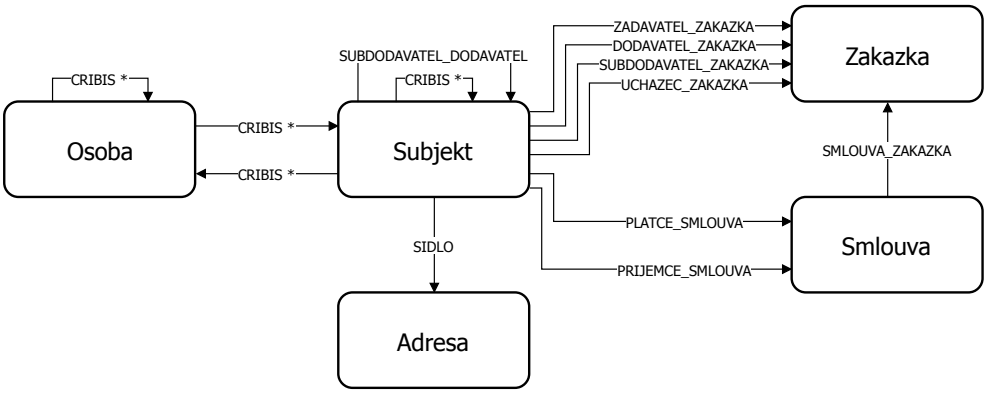
V kódu 1 je uvedena ukázka dotazu pro získání filmů s titulem začínajícím na „T“, každý film má vytvořené atributy titul a seznam jmen herců, kteří v něm hrají, výstup je seřazen podle titulu a omezen na prvních 10 záznamů.

Kód 1 Ukázka dotazu v jazyce Cypher [25].

```
MATCH (actor:Person)-[:ACTED_IN]->(movie:Movie)
WHERE movie.title STARTS WITH "T"
RETURN movie.title AS title, collect(actor.name) AS cast
ORDER BY title ASC LIMIT 10;
```

2.2 Datový model

S ohledem na zvolené technologie byl navržen datový model pro uložení dat ze zdrojových systémů. Datový model je popsán na obrázku 2.1, graf se skládá z 5 typů uzlů a 45 typů hran, jednotlivé uzly a hrany jsou dále podrobně popsány. Schéma není v Neo4j pevně definováno, a tak je možné přidávat k uzlům další vypočtené vlastnosti.



Obrázek 2.1: Datový model.

2.2.1 Uzly

Subjekt

Uzel reprezentuje právnickou nebo fyzickou osobu, která má přidělené IČO.

- `ico` – IČO (klíč subjektu)
- `nazev` – název
- `zapis` – datum zápisu do veřejného rejstříku
- `vymaz` – datum výmazu z veřejného rejstříku
- `zakladni_kapital` – základní kapitál
- `insolvence` – probíhá insolvenční řízení?

Osoba

Uzel reprezentuje osobu zapsanou ve veřejných rejstřících (společník, jednatel, člen správní rady, ...), podnikatele nebo politika (člen zastupitelstva, starosta, předseda politické strany, ...).

- `osoba_id` – umělý identifikátor z databáze Cribis (klíč osoby)
- `jmeno` – jméno
- `prijmeni` – příjmení
- `cele_jmeno` – spojené jméno a příjmení
- `datum_narozeni` – datum narození

Smlouva

Uzel reprezentuje smlouvu z registru smluv.

- `smlouva_id` – umělý identifikátor z Hlídače Státu (klíč smlouvy)
- `predmet` – předmět smlouvy
- `datum_uzavreni` – datum uzavření smlouvy
- `cas_zverejneni` – datum a čas uveřejnění smlouvy v registru smluv
- `cena_bez_dph` – cena bez DPH
- `cena_s_dph` – cena s DPH
- `cena_vypoctena` – vypočtená cena s DPH
- `chyba_identifikace_smluvni_strany` – chyba „identifikace smluvní strany“ nebo „chybí identifikace smluvní strany“
- `chyba_chybne_strany_smlouvy`
- `chyba_stejne_strany_smlouvy`
- `chyba_neexistujici_ico`
- `chyba_vadne_ico`
- `chyba_smlouva_uzavrena_s_nespolehlivym_platcem_dph`
- `chyba_firma_vznikla_az_po_podpisu_smlouvy`
- `chyba_firma_vznikla_kratce_pred_podpisem_smlouvy`
- `chyba_budouci_datum_uzavreni`
- `chyba_neplatny_datum_uzavreni_smlouvy`
- `chyba_nulova_hodnota_smlouvy`
- `chyba_neplatna_cena`
- `chyba_zaporna_cena_bez_dph`
- `chyba_zaporna_cena_s_dph`
- `chyba_bez_dph_s_dph`
- `chyba_chybi_predmet_smlouvy`
- `chyba_necitelnost_smlouvy`

Zakázka

- `zakazka_id` – umělý identifikátor v databázi (klíč zakázky)
- `nazev` – název veřejné zakázky
- `druh` – druh zadávacího řízení

Adresa

- `ruian` – RÚIAN kód (klíč adresy)
- `ulice` – ulice
- `cp` – číslo popisné
- `co` – číslo orientační
- `obec` – obec
- `psc` – PSČ

2.2.2 Hrany

Plátce-smlouva

Hrana od subjektu, který je plátcem smlouvy, ke smlouvě.

Příjemce-smlouva

Hrana od subjektu, který je příjemcem smlouvy, ke smlouvě.

Zadavatel-zakázka

Hrana od subjektu, který je zadavatelem veřejné zakázky, k veřejné zakázce.

Dodavatel-zakázka

Hrana od subjektu, který je dodavatelem veřejné zakázky, k veřejné zakázce. Atribut `cena` obsahuje informaci o smluvní hodnotě veřejné zakázky.

Subdodavatel-zakázka

Hrana od subjektu, který je subdodavatelem veřejné zakázky, k veřejné zakázce.

Uchazeč-zakázka

Hrana od subjektu, který je uchazečem o veřejnou zakázku, k veřejné zakázce. Atribut `cena` obsahuje informaci o nabídkové ceně.

Subdodavatel-dodavatel

Hrana od subjektu, který je subdodavatelem veřejné zakázky, k subjektu, který je dodavatelem veřejné zakázky.

Cribis

Souhrnné označení hran z databáze Cribis. Původních 114 typů hran bylo zjednodušeno na 36 typů. Hrany s podobným významem byly sloučeny (například „člen dozorčí rady“, „člen revizní komise“ a „člen kontrolní komise“ byly sloučeny na „člen kontrolního orgánu“) a hrany získané z jiných zdrojů („dodavatel veřejné zakázky“) byly odstraněny.

Sídlo

Hrana od subjektu k adrese jeho sídla.

2.3 Integrace dat

Integrace dat je proces, při kterém se kombinují data z různých zdrojových systémů. V průběhu integrace je nutné se vypořádat s nekvalitou, různými formáty dat a nekonzistencí mezi zdroji dat.

2.3.1 Nástroje

Pro usnadnění integrace dat existují nástroje poskytující připravené komponenty pro načtení, transformaci a uložení dat.

Přehled nástrojů sestavuje pravidelně organizace Gartner, nástroje hodnotí podle jejich vyspělosti a inovativnosti. Podle Gartner Magic Quadrant for Data Integration Tools [31] byly v roce 2017 nejlépe hodnocené nástroje firem Informatica, IBM, SAP, SAS, Talend a Oracle.

Talend Open Studio for Big Data

Nástroj Talend Open Studio for Big Data byl zvolen zejména kvůli open-source verzi, dobrému hodnocení a podpoře všech použitých databází a formátů souborů.

2.3.2 Popis datových integrací

Cílem integrace dat je propojení dat z jednotlivých zdrojových systémů a jejich uložení do grafové databáze.

Obrázek 2.2: Spojení dat z ARES.

Obrázek 2.3: Spojení dat z RÚIAN.

Zpracování malých souborů

Data ze systémů ARES a RÚIAN jsou uložena ve velkém množství malých souborů. Data ze systému ARES se skládají z 978 653 souborů o průměrné velikosti 7 kB a data ze systému RÚIAN ze 6 258 souborů o průměrné velikosti 55 kB. Zejména při opakovaném načítání dat, například při vývoji, je výhodné soubory spojit a dále pracovat pouze s jedním souborem.

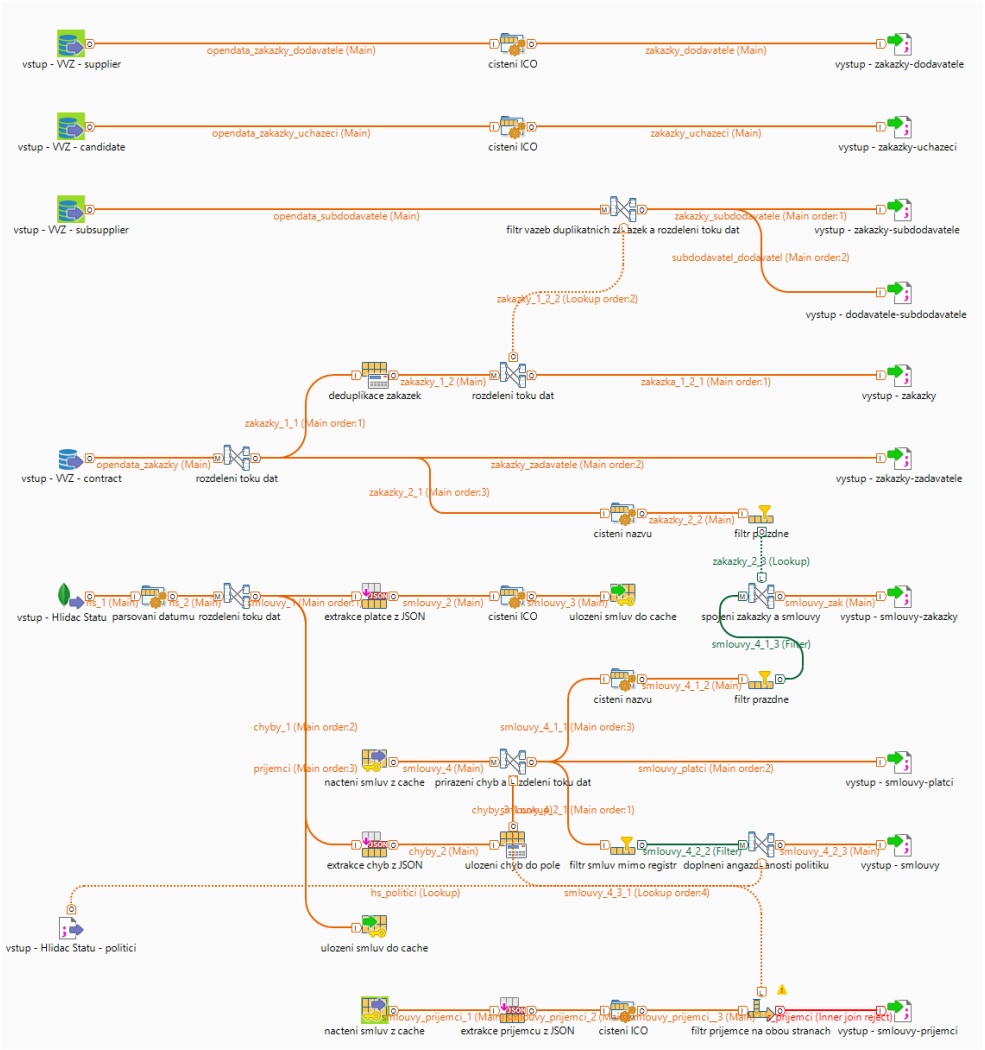
Na obrázku 2.2 je popsané schéma spojení dat ze systému ARES. Nejprve je vytvořen seznam XML souborů v adresáři se zdrojovými soubory, poté jsou z každého z nich načteny požadované údaje a jsou zapsány do výstupního souboru ve formátu CSV. Obdobně probíhá spojení dat i ze systému RÚIAN, na obrázku 2.3, jedinou výjimkou je formát zdrojových dat, která jsou ve formátu CSV místo XML.

Práce s velkými soubory

Data z Hlídače Státu jsou uložena v souboru ve formátu JSON, který má velikost přes 44 GB. Komponenta `tFileInputJSON` vyžaduje načtení celého souboru do operační paměti. Tento problém byl vyřešen načtením dat do souborové databáze MongoDB, ke které lze přistupovat pomocí komponenty `tMongoDBInput`.

Propojení s Neo4j

Nahrání dat do Neo4j databáze je z možné různými způsoby. Pomocí komponenty v Talend Open Studio bylo nahrávání pomalé a není podporována aktuální verze Neo4j. Druhou možností je import dat z CSV souborů, ten navíc umožnil rozdělit integraci na dvě nezávislé úlohy, které je možné spouštět samostatně a ušetřit tak čas při změnách.



Obrázek 2.4: Integrace dat z registru smluv a věstníku veřejných zakázek.

Integrace dat z registru smluv a věstníku veřejných zakázek

Integrace dat z registru smluv a věstníku veřejných zakázek je znázorněna na obrázku 2.4. Význam jednotlivých komponent je dále popsán.

Vstup – Hlídač Státu Připojení k MongoDB databázi s daty z Hlídače Státu. Pomocí dotazu `{platnyZaznam: true}` jsou vyfiltrovány pouze platné záznamy (poslední verze nezrušených smluv).

Parsování datumu Čas uveřejnění smlouvy se v datech vyskytuje ve dvou různých formátech. Prvním je `yyyy-MM-dd'T'HH:mm:ssXXX` a druhým

`yyyy-MM-dd'T'HH:mm:ss`. Funkce pro parsování datumů umožňuje použít pouze jeden formát, pro parsování je tedy nutné vytvořit vlastní kód v jazyce Java.

Rozdělení toku dat V těchto komponentách se data rozdělují do více toků, do každého postupují pouze potřebné atributy.

Extrakce plátce z JSON Informace o plátci smlouvy jsou v atributu uloženy ve formátu JSON, v této komponentě se z JSON vyextrahuje IČO plátce.

Čištění IČO Některá IČO obsahují mezery nebo jim chybí počáteční nuly, v těchto komponentách je provedeno odstranění bílých znaků a doplnění nulami na 8 číslic.

Uložení smluv do cache V toku dat se nemůže vyskytovat cyklus, proto jsou smlouvy uloženy do cache a později znovu načteny.

Extrakce chyb z JSON Nalezené chyby jsou v atributu uloženy jako pole ve formátu JSON, z jednotlivých chyb je vyextrahován jejich název. Výstupem jsou dvojice (ID smlouvy a název chyby).

Uložení chyb do pole Dvojice z předchozího kroku jsou agregovány podle shody ID smlouvy a názvy chyb jsou uloženy do atributu typu seznam.

Načtení smluv z cache Načtení smluv dříve uložených do cache.

Přiřazení chyb a rozdělení toku dat Ke smlouvám je připojen seznam jejich chyb.

Filtr smluv mimo registr Smlouvy, které nepocházejí z registru smluv, jsou odfiltrovány podle chyby „smlouva nespadá pod registr smluv“.

Vstup – Hlídač Státu – politici Načtení souboru s ID smluv, u jejichž příjemců byly nalezeny vazby na politiky.

Doplnění angažovanosti politiků Ke smlouvám je doplněn příznak udávající, zda u příjemce smlouvy existuje vazba na politiky.

Výstup – smlouvy Uložení smluv do CSV souboru.

Výstup – smlouvy-plátci Uložení vazby mezi smlouvou a plátcem smlouvy (ID smlouvy a IČO plátce) do CSV souboru.

Extrakce příjemců z JSON Příjemci smlouvy jsou uloženy v atributu stejným způsobem jako chyby, vyextrahována jsou jejich IČO.

Filtr příjemce na obou stranách Odstranění vazeb, kde je příjemce zároveň plátcem smlouvy.

Výstup – smlouvy-příjemci Uložení vazby mezi smlouvou a příjemcem (ID smlouvy a IČO příjemce) do CSV souboru.

Vstup – VVZ – contract Načtení dat o veřejných zakázkách z tabulky `contract` v PostgreSQL databázi. Data jsou pomocí SQL dotazu doplněna o IČO zadavatele z tabulky `entity`.

Výstup – zakázky-zadavatelé Uložení vazby mezi veřejnou zakázkou a zadavatelem (ID zakázky a IČO zadavatele) do CSV souboru.

Čištění názvu V těchto komponentách je název veřejné zakázky a předmět smlouvy převeden na malá písmena a očištěn o bílé znaky na začátku a na konci.

Filtr prázdné Odfiltrovány jsou veřejné zakázky a smlouvy s prázdným názvem/předmětem a prázdným IČO zadavatele/příjemce.

Spojení zakázky a smlouvy V případě shody názvu veřejné zakázky s předmětem smlouvy a IČO zadavatele s IČO plátce, je vytvořena vazba mezi smlouvou a zakázkou.

Výstup – smlouvy-zakázky Uložení vazby mezi smlouvou a veřejnou zakázkou (ID smlouvy a ID zakázky) do CSV souboru.

Deduplikace zakázek U veřejných zakázek, které se v datech vyskytují opakovaně, je ponechána poslední verze.

Výstup – zakázky Uložení veřejných zakázek do CSV souboru.

Vstup – VVZ – subsupplier Načtení subdodavatelů veřejných zakázek z tabulky `subsupplier` v PostgreSQL databázi. Data jsou pomocí SQL dotazu doplněna o IČO subdodavatele, IČO dodavatele a ID veřejné zakázky.

Filtr vazeb duplikátních zakázek a rozdělení toku dat Odfiltrování vazeb odstraněných duplikátních veřejných zakázek.

Výstup – zakázky-subdodavatelé Uložení vazby mezi veřejnou zakázkou a subdodavatelem (ID zakázky a IČO subdodavatele) do CSV souboru.

Výstup – dodavatelé-subdodavatelé Uložení vazby mezi dodavatelem veřejné zakázky a subdodavatelem (IČO dodavatele a IČO subdodavatele) do CSV souboru.

Vstup – VVZ – supplier Načtení dodavatelů z tabulky `supplier` v PostgreSQL databázi. Data jsou pomocí SQL dotazu doplněna o IČO dodavatele.

Výstup – zakázky-dodavatelé Uložení vazby mezi veřejnou zakázkou a dodavatelem (ID zakázky, IČO dodavatele a smluvní hodnota) do CSV souboru.

Vstup – VVZ – candidate Načtení uchazečů z tabulky `candidate` v PostgreSQL databázi. Data jsou pomocí SQL dotazu doplněna o IČO uchazeče.

Výstup – zakázky-uchazeči Uložení vazby mezi veřejnou zakázkou a uchazečem (ID zakázky, IČO uchazeče a nabídková cena) do CSV souboru.

Integrace dat z Cribis, ARES a veřejného rejstříku.

Integrace dat z databáze Cribis, ARES a veřejného rejstříku je znázorněna na obrázku 2.5. Jednotlivé komponenty jsou dále popsány.

Vstup – Cribis – company Načtení dat subjektů z CSV souboru.

Filtr zahraničních subjektů Vyfiltrovány jsou pouze subjekty, které mají uvedené české IČO. Slovenské IČO má stejný formát a čísla by nebyla unikátní.

Vstup – Veřejný rejstřík Načtení JSON souboru s daty získanými z veřejného rejstříku.

Vstup – RÚIAN – spojené Načtení CSV souboru se spojenými adresními místy z RÚIAN.

Vstup – ARES – spojené Načtení CSV souboru se spojenými daty z ARES.

Spojení ARES s RÚIAN K subjektům, které mají v ARES uvedený platný RÚIAN kód sídla, je doplněna adresa. Původní adresa získaná z ARES je odstraněna, aby bylo zajištěno, že stejná adresní místa budou mít tu samou adresu zapsanou stejným způsobem.

Spojení údajů z rejstříku Údaje z databáze Cribis jsou doplněny o údaje z veřejného rejstříku a ARES. Jako identifikátor pro spojení se používá IČO.

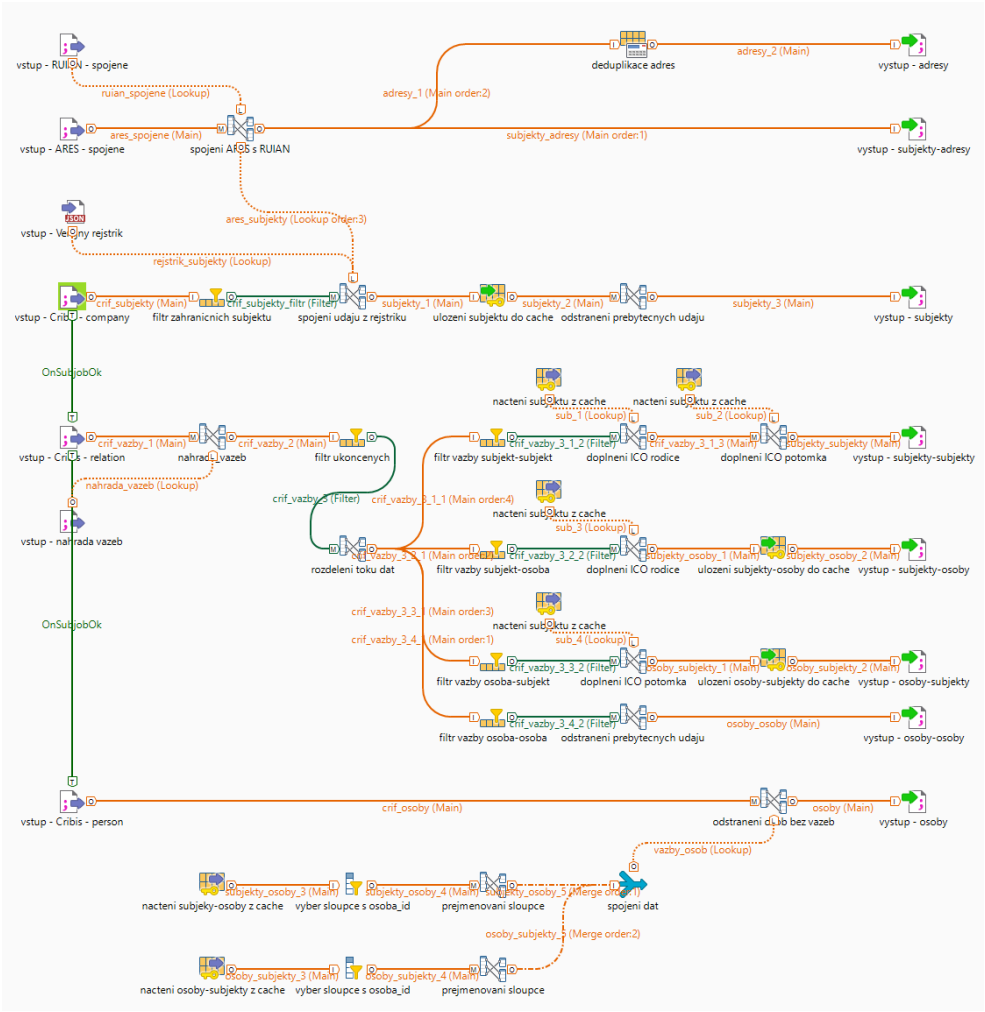
Uložení subjektů do cache Uložení subjektů do cache.

Odstranění přebytečných údajů Odstranění umělého identifikátoru subjektu.

Výstup – subjekty Uložení subjektů do CSV souboru.

Deduplikace adres Odstranění opakujících se sídel podle RÚIAN kódu.

Výstup – adresy Uložení sídel subjektů do CSV souboru.



Obrázek 2.5: Integrace dat z ARES, veřejného rejstříku a Cribis.

Výstup – subjekty-adresy Uložení vazeb mezi subjektem a sídlem (IČO subjektu, RÚIAN kód) do CSV souboru.

Vstup – Cribis – relation Načtení vazeb z CSV souboru.

Vstup – náhrada vazeb Načtení seznamu vazeb z databáze Cribis s jejich novými názvy.

Náhrada vazeb Přejmenování vazeb podle načteného seznamu.

Filtr ukončených Odstranění vazeb zaniklých před rokem 2013.

Rozdělení toku dat V těchto komponentách se data rozdělují do více toků, do každého postupují pouze potřebné atributy.

Filtr vazby subjekt-subjekt Vyfiltrování vazeb mezi dvěma subjekty.

Filtr vazby subjekt-osoba Vyfiltrování vazeb od subjektu k osobě.

Filtr vazby osoba-subjekt Vyfiltrování vazeb od osoby k subjektu.

Filtr vazby osoba-osoba Vyfiltrování vazeb mezi dvěma osobami.

Načtení subjektu z cache Načtení dříve uložených subjektů z cache.

Doplnění IČO rodiče Umělý identifikátor rodiče vazby je nahrazen IČO.

Doplnění IČO potomka Umělý identifikátor potomka vazby je nahrazen IČO.

Uložení subjekty-osoby do cache Uložení vazeb od subjektu k osobě do cache.

Uložení osoby-subjekty do cache Uložení vazeb od osoby k subjekty do cache.

Výstup subjekty-subjekty Uložení vazeb mezi dvěma subjekty (IČO subjektu, IČO subjektu) do CSV souboru.

Výstup subjekty-osoby Uložení vazeb od subjektu k osobě (IČO subjektu, ID osoby) do CSV souboru.

Výstup osoby-subjekty Uložení vazeb od osoby k subjektu (ID osoby, IČO subjektu) do CSV souboru.

Výstup osoby-osoby Uložení vazeb mezi dvěma osobami (ID osoby, ID osoby) do CSV souboru.

Vstup – Cribis – person Načtení osob z CSV souboru.

Načtení subjekty-osoby z cache Načtení dříve uložených vazeb z cache.

Načtení osoby-subjekty z cache Načtení dříve uložených vazeb z cache.

Výběr sloupce s osoba_id Výběr pouze sloupce s ID osoby.

Přejmenování sloupce Přejmenování obou sloupců s ID osoby (`rodic_id` a `potomek_id`) na stejný název.

Spojení dat Spojení ID osob vyskytujících se u obou druhů vazeb.

Odstranění osob bez vazeb Odstranění osob, které nejsou uvedené u žádné vazby se subjektem.

Výstup – osoby Uložení osob do CSV souboru.

2.4 Import do Neo4j

Import CSV souborů je možné provést dvěma způsoby.

Prvním způsobem je import pomocí Cypher příkazu `LOAD CSV`, ten je ale vhodný pro soubory do 10 MB a některé soubory mají až 450 MB. [25]

Druhým způsobem je import pomocí nástroje neo4j-admin, který je součástí instalace Neo4j. Ukázka spuštění importu je uvedena v kódu 2. Při importu je také možné spojit více CSV souborů, to je možné využít pro uložení hlaviček do samostatných souborů.

Typ uzlů nebo hran v souboru je možné specifikovat při importu nebo načíst z CSV souboru, ze sloupce označeného `LABEL` (pro uzly) nebo `TYPE` (pro hrany).

Kód 2 Ukázka spuštění importu CSV souborů do Neo4j.

```
neo4j-admin import --database graph.db --id-type string ^
--ignore-missing-nodes=true ^
--nodes:Adresa h_adresy.csv,adresy.csv ^
--nodes:Osoba h_osoby.csv,osoby.csv ^
--nodes:Subjekt h_subjekty.csv,subjekty.csv ^
--relationships h_osoby_subj.csv,osoby_subj.csv ^
--relationships h_osoby_osoby.csv,osoby_osoby.csv ^
--relationships h_subj_osoby.csv,subj_osoby.csv ^
--relationships h_subj_subj.csv,subj_subj.csv ^
--relationships:SIDLO h_subj_adresy.csv,subj_adresy.csv
```

Ukázka hlavičky uzlu je uvedena v kódu 3. Pomocí `ID` je specifikován sloupec s identifikátorem uzlu (typ uzlu v závorce nemusí být uveden, identifikátor je potom globální). `LABEL` označuje sloupec udávající typ uzlu, může být také nastaven parametrem při importu (uzel může mít více typů).

Ukázka hlavičky hrany je uvedena v kódu 4. `START_ID` a `END_ID` specifikují sloupce s identifikátory počátečních a koncových uzlů. Pomocí `TYPE` je

Kód 3 Ukázka hlavičky uzlu subjekt pro import do Neo4j.

```
ico:ID(Subjekt),nazev,zak_kapital:FLOAT,insolvence:BOOLEAN
```

zvolen sloupec udávající typ hrany, podobně jako u uzlu může být nastaven parametrem při importu.

Kód 4 Ukázka hlavičky hrany subjekt-osoba pro import do Neo4j.

```
popis:TYPE,rod_ico:START_ID(Subjekt),pot_id:END_ID(Osoba)
```

Sloupce, které nemají být importovány lze označit `IGNORE`. Podobně lze také nastavit datové typy atributů, podporované jsou tyto typy: `NUMBER`, `INTEGER`, `FLOAT`, `STRING` a `BOOLEAN`. [25]

Import souborů o celkové velikosti 1,7 GB trvá tímto způsobem přibližně jednu minutu. Po importu byly ještě odstraněny subjekty, osoby, adresy a zakázky, které neměly žádnou hranu.

Kapitola 3

Analýza dat

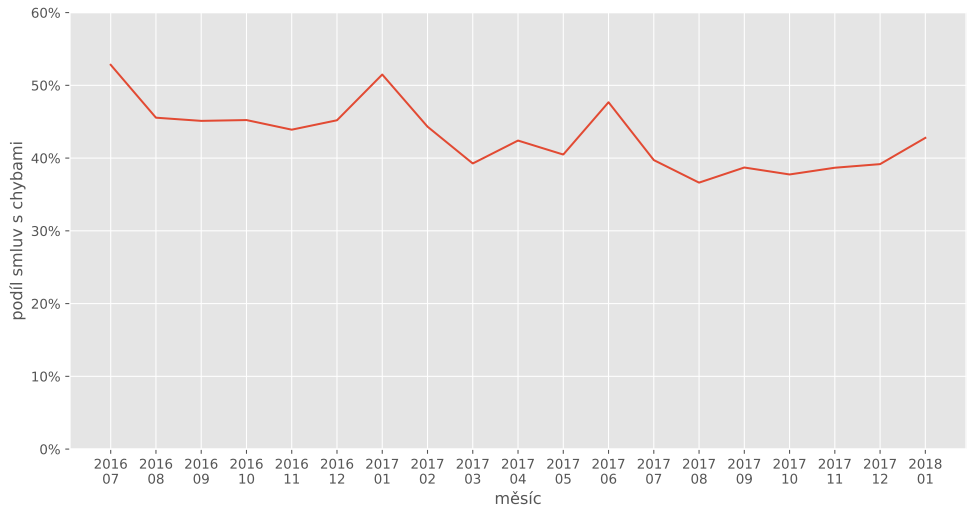
V této kapitole je provedena analýza uložených dat, tak aby podle ní bylo později možné navrhnout příznaky popisující smlouvy.

3.1 Chyby detekované Hlídačem Státu

Alespoň jednu chybu detekovanou Hlídačem Státu obsahuje 413 284 smluv, to tvoří téměř 42 % z celkového počtu 984 197 smluv. Zastoupení jednotlivých chyb je uvedeno v tabulce 3.1.

Tabulka 3.1: Zastoupení chyb detekovaných Hlídačem Státu.

Chyba	Počet smluv	Podíl smluv
Identifikace smluvní strany	125 853	12,79 %
Chybné strany smlouvy	7 111	0,72 %
Stejné strany smlouvy	16 939	1,72 %
Neexistující IČO	1 515	0,15 %
Vadné IČO	569	0,06 %
Sml. uzav. s nespolehlivým plátcem DPH	9	0,00 %
Firma vznikla až po podpisu smlouvy	3 785	0,38 %
Firma vznikla krátce před podpisem sml.	1 982	0,20 %
Budoucí datum uzavření	3 266	0,33 %
Neplatný datum uzavření smlouvy	2 919	0,30 %
Nulová hodnota smlouvy	206 266	20,96 %
Neplatná cena	43 201	4,39 %
Záporná cena bez DPH	711	0,07 %
Záporná cena s DPH	638	0,06 %
Bez DPH = s DPH	57 535	5,85 %
Chybí předmět smlouvy	315	0,03 %
Nečitelnost smlouvy	74 250	7,54 %



Obrázek 3.1: Vývoj chybovosti smluv.

Podíl smluv s chybami je znázorněný na obrázku 3.1. Největší chybovost byla v prvním měsíci (červenec 2016), po zavedení sankcí (červenec 2017) došlo pouze k malému poklesu. Průměrná chybovost do června 2017 byla 45 %, od července 2017 39 %.

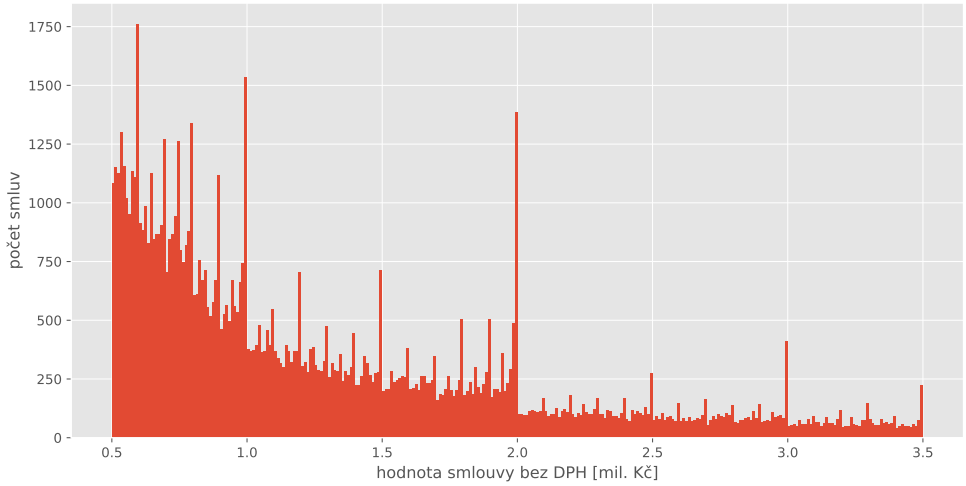
Vazba příjemce smlouvy na politiky (členství politiků v orgánech subjektu nebo sponzoring politických stran) byla nalezena u 78 799 (8 %) smluv.

3.2 Hodnoty smluv

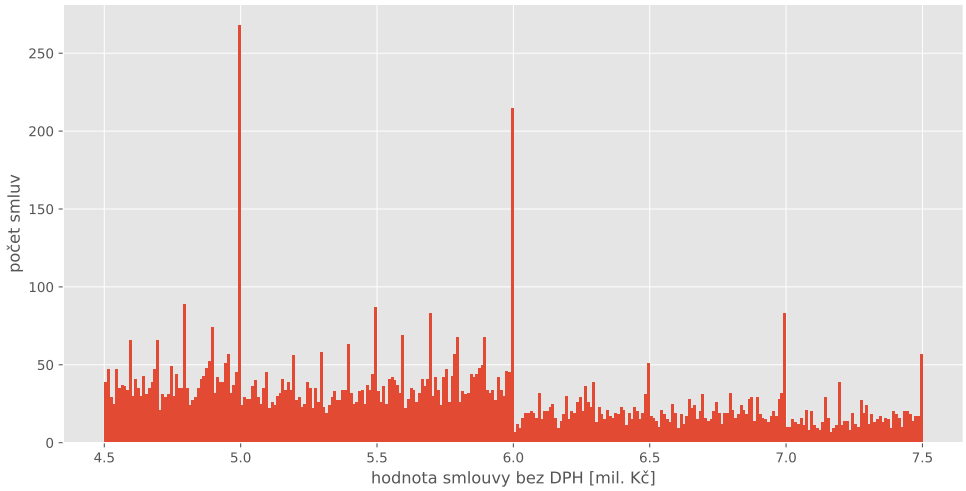
Průměrná hodnota smlouvy bez DPH je 12 992 920 Kč a medián 127 971 Kč. 72 642 smluv je v registru smluv uveřejněno, přestože jejich hodnota bez DPH nepřesahuje 50 000 Kč.

Obrázky 3.2 a 3.3 znázorňují počty smluv s hodnotami pohybujícími se okolo hranice pro veřejné zakázky malého rozsahu. Tato hranice je různá pro zakázky na stavební práce a ostatní zakázky. Smlouvy není možné rozdělit na stavební a ostatní, uvedeny jsou tedy počty všech smluv. U obou hranic je patrný pokles počtu smluv s hodnotou přesahující hranici a zároveň prudký nárůst počtu smluv s hodnotou těsně pod hranicí – v intervalech $(1\,990\,000, 2\,000\,000]$ Kč a $(5\,990\,000, 6\,000\,000]$ Kč.

Některé subjekty mají v těchto intervalech velké množství smluv, které jsou často uzavřené se stejným příjemcem, někdy i ve stejný den, a nesou tak znaky dělení veřejných zakázek. Podle [32] bylo do 15. 2. 2018 zveřejněno 448 smluv s hodnotou mezi 1 999 000 Kč a 1 999 999 Kč a celkovou hodnotou přes 895 milionů Kč bez DPH. V intervalu od 1 990 000 Kč do 2 000 000 Kč

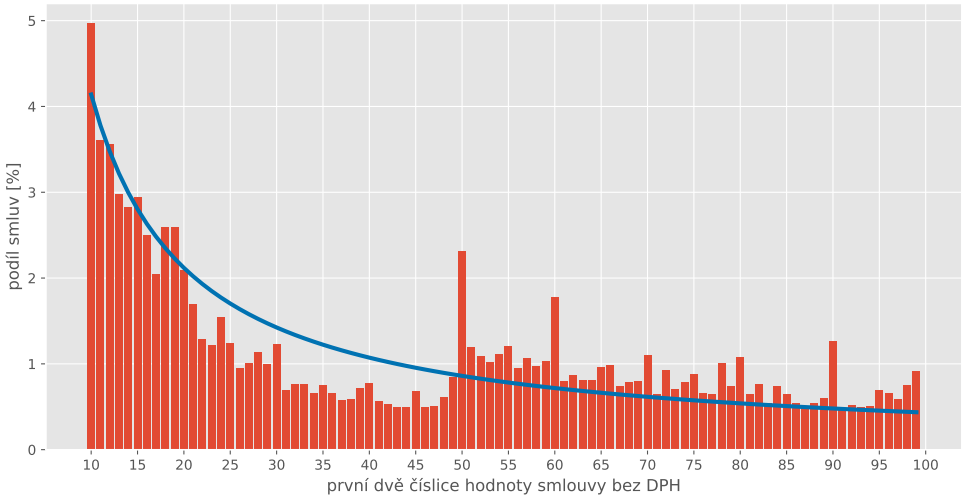


Obrázek 3.2: Histogram hodnot smluv okolo hranice 2 000 000 Kč.



Obrázek 3.3: Histogram hodnot smluv okolo hranice 6 000 000 Kč.

bylo nejvíce smluv nalezeno u Vojenských lázeňských rekreačních zařízení – 237 smluv uzavřených pouze s 67 různými příjemci. Na dalším místě je až Hlavní město Praha, které má takovýchto smluv 48, ale s 39 různými příjemci. Pod hranicí 6 000 000 Kč mají nejvíce smluv Brněnské komunikace a. s. – 28 smluv s 13 různými příjemci.



Obrázek 3.4: Rozložení hodnot smluv podle Benfordova zákona (všechny smlouvy).

3.2.1 Benfordův zákon

Benfordův zákon [33] říká, že v mnoha souborech číselných dat se čísla začínající číslicemi 1, 2, 3, ... vyskytují častěji než čísla začínající číslicemi jako 7, 8 a 9.

Pravděpodobnost výskytu čísel podle první číslice je popsána vzorcem:

$$P(d) = \log\left(1 + \frac{1}{d}\right),$$

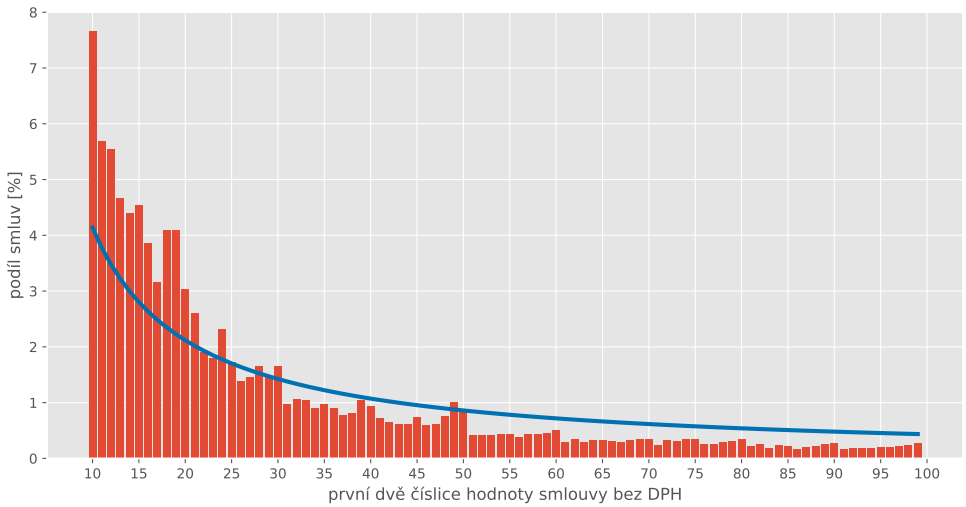
kde $d \in \{1, 2, \ldots, 9\}$. Zákon lze ale zobecnit i pro více počátečních číslic.

Zákon se často využívá pro odhalení neobvyklých částek při detekci podvodů.

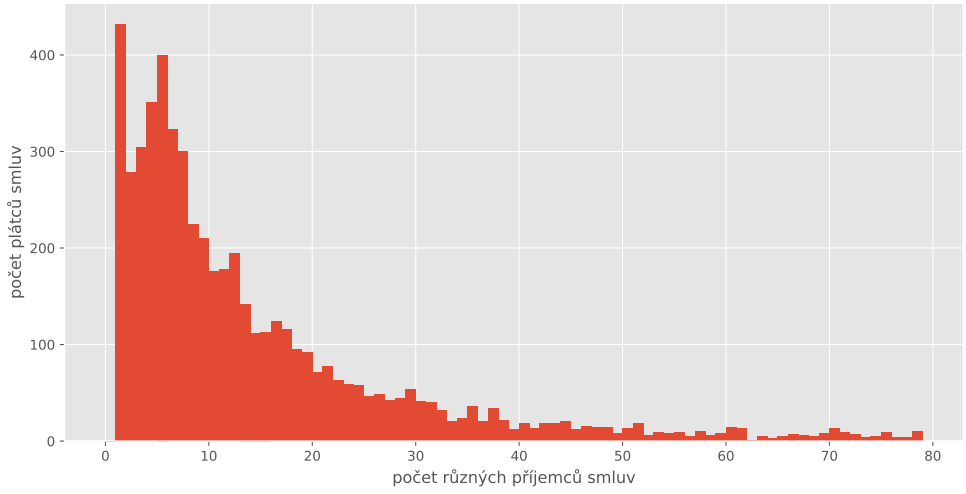
Rozložení hodnot smluv podle prvních dvou číslic je znázorněno na obrázku 3.4, toto rozložení je zkresleno povinností zveřejňovat smlouvy až od hodnoty 50 000 Kč. Na obrázku 3.5 je znázorněno rozložení pro smlouvy s hodnotou 100 000 Kč a vyšší. Modrá čára znázorňuje očekávanou hodnotu podle Benfordova zákona.

3.3 Diverzifikace příjemců smluv

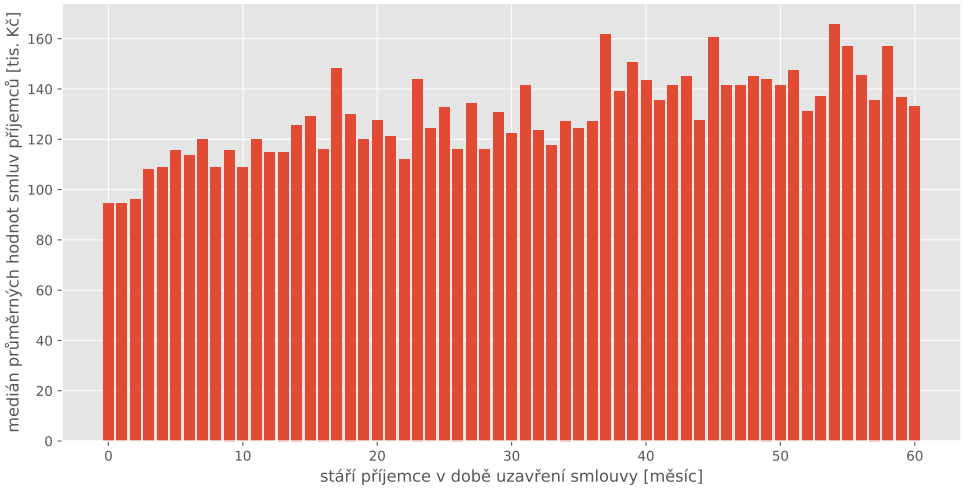
68 % subjektů má pouze jediného příjemce smluv. Na obrázku 3.6 je znázorněn počet subjektů v závislosti na počtu různých příjemců jejich smluv. Subjekty s méně než pěti smlouvami, byly odfiltrovány.



Obrázek 3.5: Rozložení hodnot smluv podle Benfordova zákona (smlouvy s hodnotou 100 000 Kč a vyšší).



Obrázek 3.6: Histogram počtu různých příjemců smluv.



Obrázek 3.7: Medián průměrných hodnot smluv podle stáří příjemce.

3.4 Stáří příjemců smluv

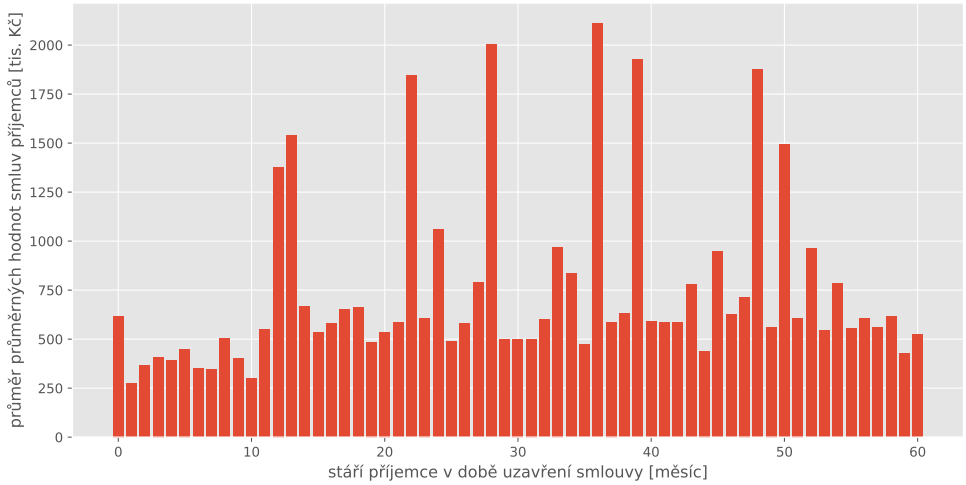
U každé smlouvy je vypočteno stáří příjemce v době uzavření smlouvy, toto číslo je zaokrouhleno na celé měsíce a pro každého příjemce je určena průměrná hodnota smluv se stejným stářím. Medián těchto průměrných hodnot je znázorněn na obrázku 3.7, je patrné, že se stářím příjemců se hodnoty smluv zvyšují. Pokud se místo mediánu použije průměr, jak je znázorněno na obrázku 3.8, je průměrná hodnota smluv příjemců založených před méně než měsícem přibližně dvojnásobná oproti průměrné hodnotě smluv uzavřených během prvního roku od založení, to poukazuje na výskyt extrémně vysokých hodnot smluv.

Zahrnuty jsou pouze společnosti s ručením omezeným, akciové společnosti a veřejné obchodní společnosti, u některých dalších subjektů neodpovídá, z důvodu legislativních změn, datum zápisu do rejstříku datu založení.

3.5 Základní kapitál příjemců smluv

Zákon o obchodních korporacích [34] definuje základní kapitál jako souhrn všech vkladů. Povinnost vytvářet základní kapitál mají kapitálové společnosti (společnosti s ručením omezeným a akciové společnosti). Od 1. 1. 2014 je minimální výše základního kapitálu společnosti s ručením omezeným stanovena na 1 Kč, předchozí minimální výše základního kapitálu byla 200 000 Kč.

Podle [35] jsou u společností se základním kapitálem 1 Kč veškerá aktiva financována z cizích zdrojů a při jakékoliv ztrátě tak hrozí riziko insolvence, typicky ve formě předlužení (závazky společnosti převyšují její majetek).



Obrázek 3.8: Průměr průměrných hodnot smluv podle stáří příjemce.

Na obrázku 3.9 je znázorněn počet příjemců smluv podle výše jejich základního kapitálu. Nejčastěji se vyskytuje základní kapitál ve výši 200 000 Kč. Druhou nejčastější výší základního kapitálu je 100 000 Kč (mininimální výše do 31. 12. 2000). Společnosti založené před 1. 1. 2001 si mohly ponechat nižší základní kapitál než 200 000 Kč, ale při jakékoliv změně jej musely navýšit alespoň na toto minimum. Každý sloupec má šířku 2 000 Kč, v intervalu $[0, 2\,000)$ Kč je nejčastějším základním kapitálem 1 000 Kč, následují 1 Kč, 100 Kč a 2 Kč (minimální vklad každého společníka musí být alespoň 1 Kč).

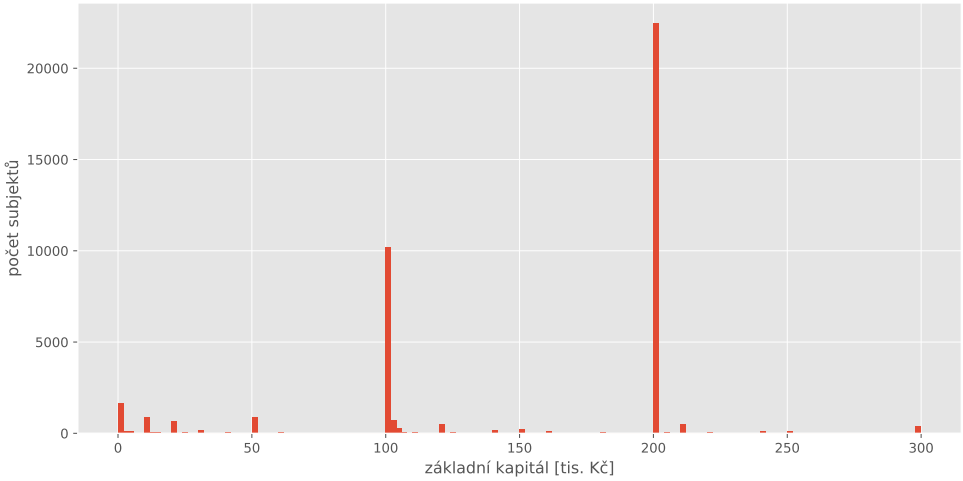
3.6 Insolvence příjemců smluv

U příjemců 1 292 smluv probíhalo v době stahování dat insolvenční řízení. Subjekty tedy nemusely být v insolvenci přímo v době uzavření smlouvy, příznak ale může poukazovat na špatné finanční zdraví v této době.

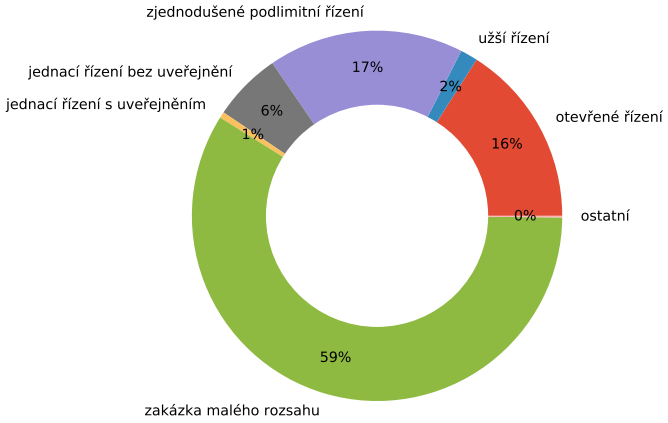
3.7 Veřejné zakázky

Zastoupení jednotlivých druhů veřejných zakázek je znázorněno na obrázku 3.10. Největší podíl (59 %) mají zakázky malého rozsahu. Veřejné zakázky s jednacím řízením bez uveřejnění tvoří 6 %.

Z obrázku 3.11 je patrné, že veřejné zakázky s užším řízením, otevřeným řízením a zjednodušeným podlimitním řízením mají průměrně více než 3,5 uchazečů. Naopak u zakázek s jednacím řízení bez uveřejnění se většinou vybírá pouze z jedné nabídky. Průměrný počet uchazečů se počítá pouze ze zakázek,

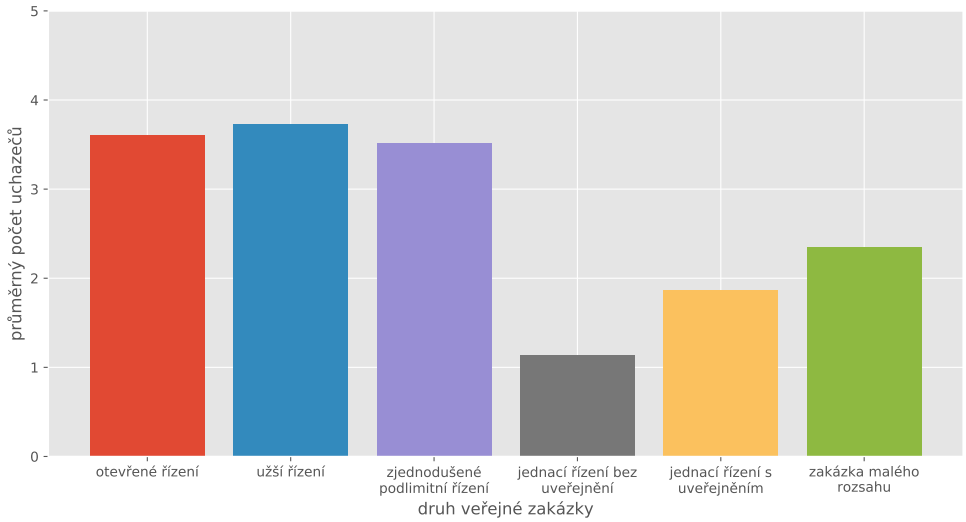


Obrázek 3.9: Histogram základního kapitálu příjemců smluv.



Obrázek 3.10: Podíl druhů veřejných zakázek.

u kterých je znám alespoň jeden uchazeč (18 %). Dodavatel je známý u 33 % zakázek, u zbylých zakázek se jej buď nepodařilo stáhnout z profilu zadavatele, nebo byla zakázka neúspěšná.



Obrázek 3.11: Průměrný počet uchazečů o zakázku.

3.8 Vlastnosti grafu

3.8.1 Centralita

Míra centrality se používá pro identifikaci klíčových uzlů grafu. Původně vznikla pro analýzu sociálních sítí, ale později se rozšířila do dalších oblastí. Centrality je možné rozdělit na dvě skupiny: založné na stupních uzlů (degree centralita, eigenvector centralita) a na nejkratších cestách v grafu (closeness centralita, betweenness centralita).

Následující popisy jednotlivých centralit vycházejí z [36].

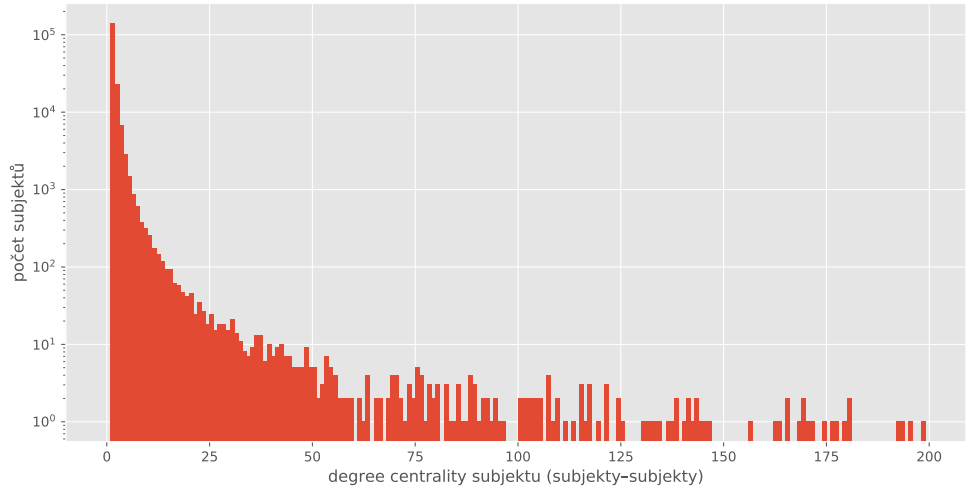
Degree centralita

Nejjednodušší mírou je degree centralita, která je rovna stupni uzlu, v případě orientovaného grafu se určuje vstupní a výstupní degree centralita jako vstupní a výstupní stupeň uzlu. Čím vyšší je stupeň uzlu, tím více má uzel v síti zdrojů informací a informace se k němu dostane rychleji.

Definice 1 (Degree centralita)**.** V neorientovaném grafu je degree centralita uzlu i $(i = 1, 2, 3, \ldots, N)$ definována jako:

$$c_i^D = k_i = \sum_{j=1}^{N} a_{ij},$$

kde k_i je stupeň uzlu i a a_{ij} počet hran mezi uzly i a j.



Obrázek 3.12: Degree centralita subjektů propojených s ostatními subjekty.

V orientovaném grafu jsou vstupní a výstupní degree centrality uzlu definovány jako:

$$c_i^{D^{in}} = k_i^{in} = \sum_{j=1}^{N} a_{ji}, \qquad c_i^{D^{out}} = k_i^{out} = \sum_{j=1}^{N} a_{ij},$$

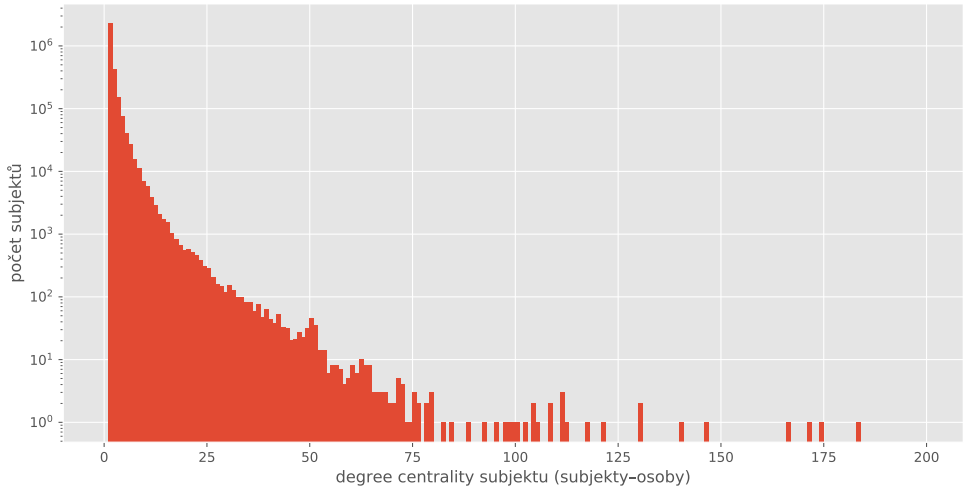
kde k_i^{in} (k_i^{out}) je vstupní (výstupní) stupeň uzlu i a a_{ji} (a_{ij}) je počet hran z uzlu j do uzlu i (z uzlu i do uzlu j).

Degree centralita byla vypočítána pro různé podgrafy.

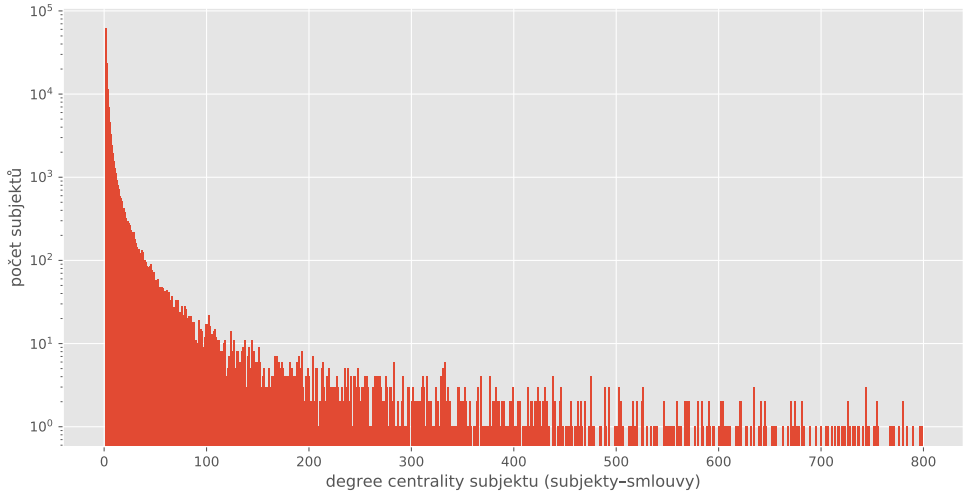
Prvním podgrafem je graf obsahují subjekty a vazby mezi nimi. Největší centralitu (8 526) má „Sdružení hasičů Čech, Moravy a Slezska“. Mezi deset uzlů s největší centralitou patří spolky s velkým množstvím pobočných spolků, firmy poskytující ready-made společnosti a odborové svazy. Rozdělení centralit je znázorněno na obrázku 3.12.

Další podgraf je složen ze subjektů a osob. Centralita v tomto případě představuje počet osob se vztahem k subjektu. Největší centralitu (516) má „Statutární město Ostava“. V prvních deseti je například společnost poskytující carsharing financovaná pomocí investičního crowdfundingu přes platformu Fundlift [37] nebo veřejná obchodní společnost vlastnící sdílený les. Rozdělení centralit je znázorněno na obrázku 3.13.

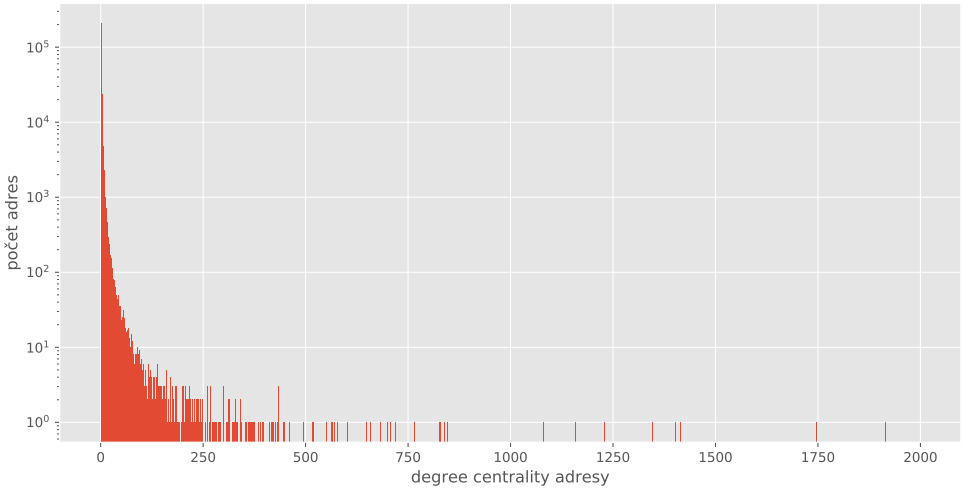
V grafu, vytvořeném přidáním hran představujících smlouvy mezi subjekty, udává centralita počet smluv subjektu. Největší počet smluv má „PHOENIX lékařský velkoobchod s.r.o.“ (55 167), v prvních deseti jsou dále státní podniky, úřady, nemocnice a lékárny. Rozložení centralit je znázorněno na obrázku 3.14.



Obrázek 3.13: Degree centralita subjektů propojených s osobami.



Obrázek 3.14: Degree centralita subjektů propojených smlouvami.



Obrázek 3.15: Degree centralita adres propojených se subjekty.

Poslední podgraf obsahuje adresy a subjekty. Centralita udává počet subjektů sídlících na stejné adrese. Největší počet subjektů (4 336) sídlí na adrese „Rybná 716/24, Praha 1“, kde firma „Simply Office s.r.o.“ provozuje virtuální sídlo a nabízí prodej ready-made společností. Všechny adresy mezi prvními deseti jsou virtuální sídla. Rozdělení centralit je znázorněno na obrázku 3.15.

Příliš vysoký počet firem sídlících na stejné adrese poukazuje na virtuální sídla. Virtuální sídla využívají firmy, které sídlo nepotřebují a nechtějí jako sídlo uvádět adresu majitele, chtějí se zviditelnit exkluzivní adresou, ale i využít menší pravděpodobnost kontrol finančních úřadů ve velkých městech (podle [38] provádí finanční úřad v Šumperku kontrolu firmy průměrně každých 18 let, zatímco v Praze 2 je to 284 let). Firmy s virtuálními sídly bývají také využívány pro karuselové podvody nebo sponzoring politických stran. Podle Bisnode [39] měla v červenci 2017 přibližně čtvrtina nespolehlivých plátců DPH virtuální sídlo a téměř polovina měla sídlo v Praze. Na některých virtálních sídlech byl přitom v minulosti podíl nespolehlivých plátců DPH ke všem plátcům DPH přes 70 % [40].

Eigenvector centralita

Na rozdíl od degree centrality zohledňuje eigenvector centralita také stupeň sousedních uzlů. Nejdůležitější uzly jsou ty, které jsou propojené s velkým množstvím důležitých uzlů.

Eigenvector centralitu je možné použít pouze v silně souvislém grafu. Uzly, které nemají žádné vstupní hrany mají nulovou centralitu a nemají tedy vliv na centralitu svých sousedů.

Definice 2 (Eigenvector centralita)**.** V neorientovaném spojitém grafu je eigenvector centralita c_i^E uzlu i definována jako:

$$c_i^E = u_{1,i},$$

kde $u_{1,i}$ je i-tá složka u_1, vlastního vektoru přidruženého k největšímu vlastnímu číslu λ_1 matice sousednosti A, splňujícího $Au_1 = \lambda_1 u_1$.

V orientovaném silně souvislém grafu jsou vstupní a výstupní eigenvector centrality $C_i^{E^{in}}$ a $C_i^{E^{out}}$ uzlu i rovné i-té složce vlastních vektorů přidružených k největším vlastním číslům λ_1 matic A^T a A.

α-centralita

α-centralita řeší problém eigenvector centrality přičtením vnitřní centrality ke každému uzlu. pomocí parametru α je možné nastavit poměr důležitosti struktury grafu oproti přičtenému vektoru.

Definice 3 (α-centralita)**.** V grafu (neorientovaném nebo orientovaném) je α-centralita c_i^α uzlu i definována jako i-tá složka vektoru:

$$c^\alpha = (I - \alpha A^T)^{-1} I,$$

kde I je jednotková matice a $0 < \alpha \leq 1$.

Closeness centralita

Closeness centralita je založena na myšlence, že jedinec, který je blízko k ostatním jedincům je důležitý, protože s nimi může rychle interagovat. Klíčové uzly mají malou vzdálenost ke všem ostatním uzlům.

Definice 4 (Closeness centralita)**.** Ve spojitém grafu je closeness centralita uzlu i definována jako převrácená hodnota součtu vzdáleností z uzlu i do ostatních uzlů:

$$c_i^C = \frac{1}{\sum_{j=1}^{N} d_{ij}}.$$

Betweenness centralita

Betweenness centralita identifikuje jedince, kteří propojují skupiny jedinců a mohou tak mít vliv na ostatní. Centralita je založena na předpokladu, že komunikace probíhá vždy po nejkratších cestách. Nejdůležitější uzly se vyznačují tím, že leží na velkém množství nejkratších cest mezi všemi uzly.

Definice 5 (Betweenness centralita)**.** Ve spojitém grafu, je betweenness centralita uzlu i definována jako:

$$c_i^B = \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\substack{k=1 \\ k \neq i,j}}^{N} \frac{n_{jk}(i)}{n_{jk}},$$

kde n_{jk} je počet počet nejkratších cest z uzlu j do uzlu k a $n_{jk}(i)$ počet nejkratších cest z uzlu j do uzlu k, které procházejí uzlem i.

PageRank

PageRank je algoritmus vyvinutý společností Google pro výpočet důležitosti stránek používané při vyhledávání. Algoritmus je založený na předpokladu, že na důležité stránky vede více odkazů, přitom záleží i na důležitosti odkazujících stránek. Matematicky je PageRank pravděpodobnost navštívení uzlu při náhodném procházení grafu s možností přechodu do libovolného uzlu.

Princip PageRank je obecný a lze jej aplikovat na grafy z různých oblastí. V současnosti se využívá například v neourovědě, při doporučování sledování uživatelů Twitteru, předpovědi vytíženosti ulic, sémantické analýze textu, analýze sítí proteinů nebo vlivu vědeckých publikací. [41]

Definice 6 (PageRank)**.** Pro stránku A, na kterou odkazují stránky $T_1, \ldots, T_n$ je PageRank definován jako:

$$\mathrm{PR}(A) = (1-d) + d\left(\left(\frac{\mathrm{PR}(T_1)}{\mathrm{C}(T_1)}\right) + \ldots + \left(\frac{\mathrm{PR}(T_n)}{\mathrm{C}(T_n)}\right)\right),$$

kde parametr $d \in [0,1]$ je *damping factor* (obvykle se používá $d = 0{,}85$) a $\mathrm{C}(A)$ je počet odchozích odkazů ze stránky A.

PageRank tvoří pravděpodobnostní rozdělení, součet PageRank všech stránek je roven jedné.

PageRank lze spočítat iterativně (na počátku mají všechny stránky nastavenou stejnou pravděpodobnost). [42]

Problém při výpočtu PageRanku způsobují tzv. dead ends (uzly bez výstupních hran) způsobující, že všechny uzly grafu mají na konci výpočtu PageRank 0, a spider traps (skupina uzlů, které nemají žádné výstupní hrany mimo skupinu) způsobující přesunutí veškerého PageRanku do této skupiny. Tyto problémy řeší damping factor představující pravděpodobnost d pokračování v procházení grafu, $1 - d$ je pravděpodobnost přechodu na náhodný uzel. [43]

PageRank byl využit pro identifikaci nejdůležitějších subjektů, hrany mezi nimi byly vytvořeny ve směru od plátce smlouvy k příjemci. Výhodné by bylo použít součty hodnot smluv jako váhy hran, ale tato možnost zatím není v knihovně Neo4j Graph Algorithms [41] implementována.

Mezi deset uzlů s nejvyšší hodnotou PageRank patří dodavatelé tepelné energie (z Frýdku-Místku a Žatce), nemocnice (Fakultní nemocnice sv. Anny v Brně a Fakultní nemocnice Plzeň), univerzity (Vysoké učení technické v Brně, Univerzita Pardubice, Univerzita Karlova a České vysoké učení technické v Praze) a města (Statutární město Brno).

3.8.2 Analýza komunit

Knihovna Neo4j Graph Algorithms [41] implementuje několik algoritmů pro detekci komunit.

Detekce komunit probíhala na grafu subjektů a jejich hran, do kterého byly přidány hrany, pokud jsou uzly propojeny pomocí osob nebo sídla.

Clustering Coefficient

Clustering Coefficient měří pravděpodobnost, že sousedé uzlu jsou také sousedé. Průměrný clustering coefficient udává hustotu grafu. Počítá se z počtu trojúhelníků, kterých je uzel součástí.

Průměrný clustering coefficient grafu je 0,11. Pro srovnání, uživatelé sociální sítě Facebook [44] se 100 přáteli měli v roce 2011 clustering coefficient 0,14 (přibližně pětkrát více než v roce 2008). [45]

Label Propagation

Algoritmus Label Propagation je založen na šíření identifikátoru uzlu sítí. Každému uzlu je přidělen unikátní identifikátor, při každé iteraci se identifikátor uzlu rozšíří do svých sousedů a změní se na identifikátor, který má nejvíce sousedů. Uzly, které mají po zkonvergování stejné identifikátory, patří do stejné komunity.

V grafu bylo nalezeno 1 996 836 komunit, jejichž velikost se pohybuje od 1 do 7 842. Průměrná velikost komunity je 1,59 a medián 1. Počty komunit podle velikosti jsou znázorněny na obrázku 3.16.

Největší komunity tvoří subjekty okolo firem poskytujících virtuální sídla a prodávající ready-made společnosti nebo velké spolky (hasiči, zahrádkáři).

Louvain

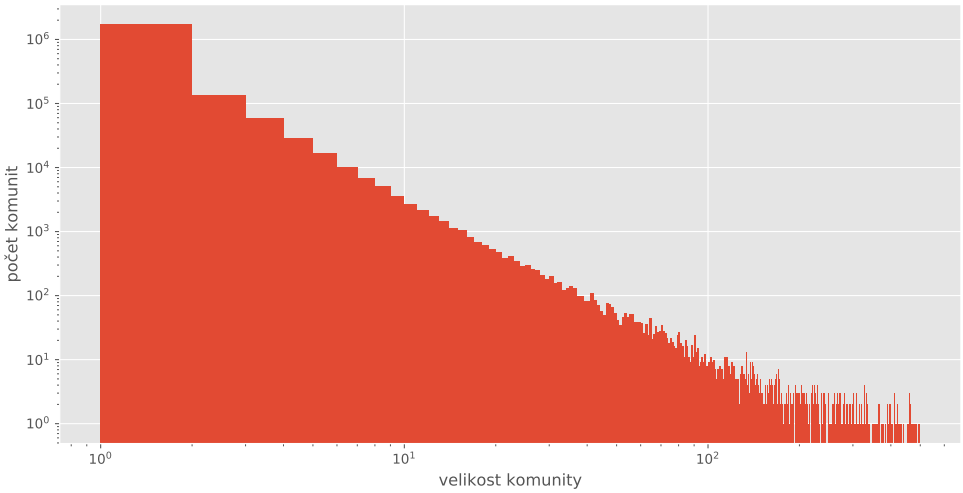
Algoritmus je založen na heuristice pro maximalizaci modularity. V prvním kroku jsou vytvořeny malé komunity lokálně maximalizující modularitu. Ve druhém kroku je vytvořen graf skládající se z uzlů vytvořených sloučením komunit. Oba kroky se opakují, dokud se zvyšuje modularita.

V grafu bylo nalezeno 1 863 218 komunit o velikosti od 1 do 10 749. Průměrná velikost komunity je 1,70 a medián 1. Počty komunit podle velikosti jsou znázorněny na obrázku 3.17.

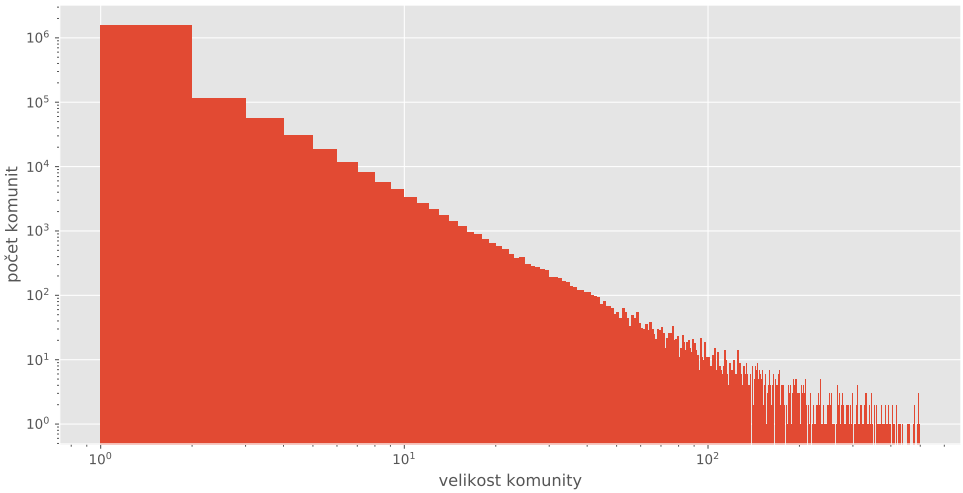
Největší komunity jsou podobné komunitám nalezeným pomocí Label Propagation. První čtyři největší komunity tvoří subjekty okolo stejných firem a na pátém místě je pouze jiná firma prodávající ready-made společnosti.

Souvislé komponenty

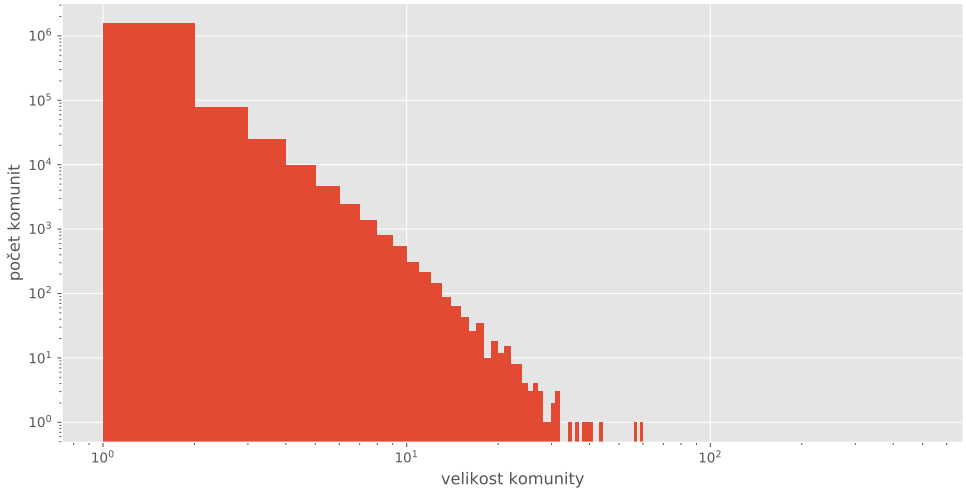
Graf obsahuje 1 695 025 souvislých komponent. Největší komponenta obsahuje 1 253 308 subjektů, ostatní komponenty se skládají z 1 až 59 subjektů. Průměrná



Obrázek 3.16: Velikost komunit podle Label Propagation.



Obrázek 3.17: Velikost komunit podle Louvain.



Obrázek 3.18: Velikost komunit podle souvislých komponent.

velikost komponenty je 1,87 a medián 1. Počty komponent podle velikosti jsou znázorněny na obrázku 3.18.

V největší komponentě mají největší degree centralitu uzly jako: PHOENIX lékárenský velkoobchod, s.r.o., Česká pošta, s.p. a Úřad práce České republiky.

Power law

V mnoha grafech bylo pozorováno, že velikost komunit má power-law rozdělení. Pokud se graf takového rozdělení vykreslí s logaritmickým měřítkem na ose X i Y, tvoří data přímku. [46]

Power-law splňují i komunity detekované předchozími algoritmy.

Kapitola 4

Detekce podezřelých smluv

Tato kapitola se v první části věnuje výběru příznaků pro popis smluv. V druhé části jsou smlouvy s příznaky vizualizovány pomocí Self organizing map. Poslední část popisuje algoritmy používané pro detekci anomálií.

4.1 Řešení pro veřejné zakázky

Odhalování korupce u veřejných zakázek pomocí data miningu se věnuje projekt DIGIWHIST [47]. Analyzovaná data z 32 evropských zemí a institucí Evropské unie zveřejňují na serveru Opentender [48].

Detekce je založena na třech předpokladech:

- korupce je odchylka od normy,
- expertní znalost je nutná, ale ne dostatečná,
- korupce je kategorická (nelineární).

Veřejné zakázky mají pět fází, každá z nich je zranitelná různými technikami korupce:

- analýza požadavků (zbytečné požadavky, požadavky zvýhodňující konkrétního dodavatele),
- návrh procesu (přizpůsobení kritérií způsobilosti, zneužití formálních a administrativních požadavků, dlouhodobé a složité smlouvy, nastavení lhůty pro podání nabídek),
- příprava zadání a zveřejnění (selektivní poskytování informací, neuveřejnění výzvy pro podání nabídek, úprava výzvy, složitý přístup k dokumentům),
- vyhodnocení a výběr vítěze (zrušení řízení, opakované porušení pravidel, nerovné hodnocení),

- provedení (změna smluv, zneužívání dodatků ke smlouvám).

Indikátory jsou nejprve navržené s využitím expertních znalostí a následně ověřeny statistickými metodami. Jako indikátory využívají například:

- počet uchazečů,
- zveřejnění řízení,
- délku doby zveřejnění,
- stáří společnosti,
- sídlo společnosti (daňové ráje),
- změny podílu na trhu po volbách,
- propojení uchazečů,
- pravděpodobnost výhry uchazeče,
- pravidelné střídání vítěze,
- rozdíl mezi nejnižší a druhou nejnižší nabídkou,
- rozsah nabídkových cen.

Například podíl zakázek s jedinou nabídkou se napříč Evropou liší. Nejvyšší podíl těchto zakázek je v Polsku, Chorvatsku a Maďarsku, naopak nejnižší v Lichtenštejnsku, Finsku a Švýcarsku. Tento výsledek koreluje s indikátorem Control of Corruption vydávaného světovou bankou.

4.2 Příznaky

Na základě předchozí analýzy a příznaků používaných v projektu DIGIWHIST byly vybrány následující příznaky:

- `cena_vypoctena` – cena vypočtená,
- `chyba_firma_vznikla_kratce_pred_podpisem_smlouvy` – firma vznikla krátce před podpisem smlouvy,
- `chyba_necitelnost_smlouvy` – nečitelnost smlouvy,
- `chyba_nulova_hodnota_smlouvy` – nulová hodnota smlouvy,
- `chyba_smlouva_uzavrena_s_nespolehlivym_platcem_dph` – smlouva uzavřena s nespolehlivým plátcem DPH,

- `clustering_coefficient_prijemce` – clustering coefficient příjemce smlouvy,
- `pagerank_prijemce` – PageRank příjemce smlouvy,
- `shoda_community_stran` – shoda komunit smluvních stran (podle Louvain),
- `stupen_prijemce` – stupeň příjemce smlouvy,
- `stupen_sidla_prijemce` – počet subjektů se stejným sídlem jako má příjemce smlouvy,
- `politici` – angažovanost politiků nebo sponzoring politických stran,
- `pomer_ceny_prumer_platce` – podíl hodnoty smlouvy a průměrné hodnoty smluv plátce,
- `pomer_ceny_prumer_prijemce` – podíl hodnoty smlouvy a průměrné hodnoty smluv příjemce,
- `pomer_stejny_prijemce` – poměr počtu smluv plátce se stejným příjemcem,
- `suma_pod_hranici_stejny_prijemce` – součet hodnot smluv s hodnotou nižší než 2 mil. Kč se stejným příjemcem,
- `pocet_uchazecu_zakazky` – počet uchazečů o veřejnou zakázku propojenou se smlouvou,
- `cena_smlouva_zakazka` – podíl hodnoty smlouvy a nabídkové ceny dodavatele zakázky,
- `napojeni_na_spolecnika_prijemce` – osoba spojená s plátcem je společníkem/akcionářem/komandistou/komplementářem příjemce,
- `pomer_spoluprace_prijemce` – maximální poměr veřejných zakázek, o které se příjemce ucházel se stejným uchazečem,
- `pomer_uspesnosti_prijemce` – poměr získaných veřejných zakázek příjemce,
- `pomer_jrbu_prijemce` – poměr veřejných zakázek příjemce získaných v jednacím řízení bez uveřejnění,
- `zakazka_jrbu` – smlouva propojena s veřejnou zakázkou v jednacím řízení bez uveřejnění,
- `insolvence_prijemce` – příjemce smlouvy je v insolvenci,

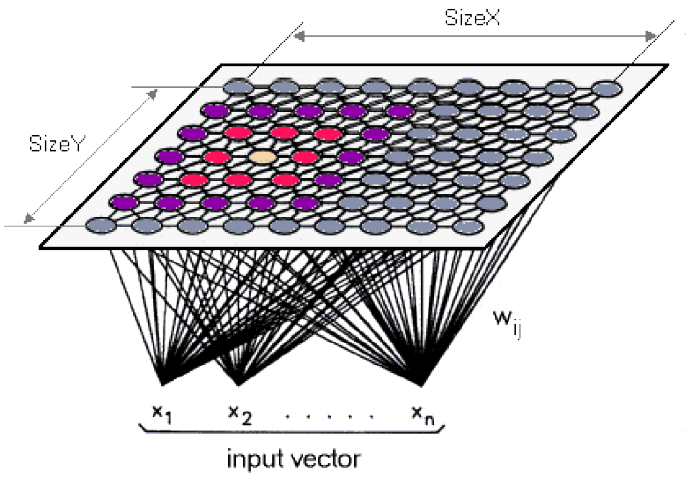
- `zakladni_kapital_prijemce` – základní kapitál příjemce smlouvy,
- `pocet_uchazecu_zakazky_jeden` – veřejná zakázka propojená se smlouvou má jen jednoho uchazeče,
- `cena_smlouva_zakazka_pres` – hodnota smlouvy přesahuje nabídkovou cenu dodavatele o více než 25 % (rezerva pro nezapočítané DPH),
- `stari_prijemce_novy` – příjemce smlouvy vznikl méně než rok před uzavřením smlouvy,
- `cena_bez_dph_pod_2m` – hodnota smlouvy je těsně pod hranicí 2 mil. Kč,
- `cena_bez_dph_pod_6m` – hodnota smlouvy je těsně pod hranicí 6 mil. Kč,
- `nizky_zakladni_kapital` – základní kapitál příjemce smlouvy je nižší než 10 000 Kč,
- `vikend` – smlouva byla uzavřena o víkendu,
- `benford` – z-skóre rozdílu počtu smluv podle prvních dvou číslic od předpokládaného počtu podle Benfordova zákona.

Mezi příznaky nebyly zařazeny některé chyby detekované Hlídačem Státu (upozorňují spíše na chybné vložení smluv do registru smluv) a korelované příznaky (komunity detekované různými algoritmy).

4.2.1 Korelace příznaků

Vyšší absolutní hodnota Pearsonova korelačního koeficientu byla nalezena mezi těmito příznaky:

- `cena_smlouva_zakazka` a `cena_smlouva_zakazka_pres` (0,77),
- `pocet_uchazecu_zakazky` a `pocet_uchazecu_zakazky_jeden` (-0,53),
- `pagerank_prijemce` a `stupen_prijemce` (0,47),
- `pagerank_prijemce` a `pomer_stejny_prijemce` (0,45),
- `chyba_nulova_hodnota_smlouvy` a `benford` (-0,41),
- `stupen_prijemce` a `pomer_stejny_prijemce` (0,4).



Obrázek 4.1: Struktura SOM [49].

4.3 Self organizing map

Self organizing map (SOM) je neuronová síť využívající kompetitivní učení k aproximaci mnohodimenzionálních dat pomocí malého počtu reprezentantů, které je možné zobrazit v prostoru o dvou nebo třech dimenzích. Platí, že sousedé v původním prostoru zůstanou blízko u sebe i v novém prostoru. Struktura SOM je znázorněna na obrázku 4.1.

Algoritmus se skládá z pěti kroků:

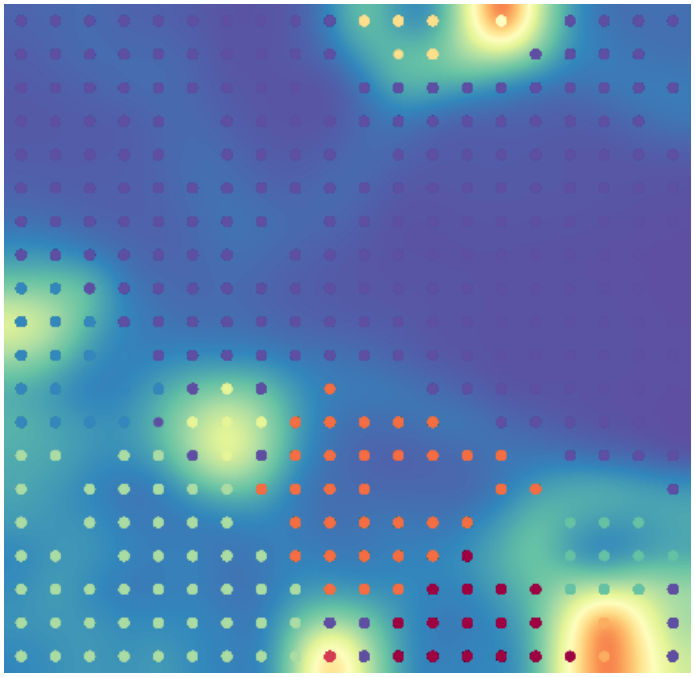
1. náhodná inicializace vah,
2. náhodný výběr vektoru z dat,
3. nalezení nejbližšího reprezentanta – Best Matching Unit (BMU),
4. aktualizace vah tak, že se BMU posune k vybranému vektoru a s ním i okolní neurony (velikost posunutí určuje gaussovská funkce vzdálenosti od BMU),
5. opakování kroku 2 dokud není splněno kritérium pro zastavení.

Pro vizualizaci neuronů se používá U-matice (matice vzdáleností váhového vektoru reprezentanta k váhovým vektorům sousedů). Vzdálenost je vyjádřena barvou, velká vzdálenost typicky tmavou a malá vzdálenost světlou. Tmavé oblasti tak tvoří hranice mezi clustery. [50]

Na nalezené reprezentanty je možné aplikovat clustering algoritmy. Pro clustering bylo použito aglomerativní shlukování s počtem clusterů stanoveným podle vizualizace U-matice na 11. Výsledná U-matice s reprezentanty roztříděnými do 11 clusterů je na obrázku 4.2 (velká vzdálenost je vyjádřena červeně

a malá modře). Podle vizualizací jednotlivých komponent lze určit specifické vlastnosti clusterů:

- modrý (většina smluv, nemají společný příznak, různé části mají nulovou hodnotu, vysoký clustering coefficient příjemce, v orgánech příjemce jsou politici, existuje napojení na vlastníka příjemce nebo má příjemce vysokou úspěšnost u veřejných zakázek),
- světle modrý (příjemci smluv získaly velký podíl zakázek v jednacím řízení bez uveřejnění a mají vysokou úspěšnost v získávání zakázek, část smluv má nulovou hodnotu a nejsou strojově čitelné, u některých smluv je hodnota vyšší než nabídková cena dodavatele zakázky),
- světle zelený (plátci smluv mají vysoký podíl smluv s jedním příjemcem, příjemci mají vyšší PageRank a stupeň, hodnota smluv je častěji těsně pod hranicí 2 mil. Kč a mají vyšší z-score odchylky od předpokladu podle Benfordova zákona),
- zelený (smlouvy uzavřené o víkendu, častěji mají nulovou hodnotu a o veřejnou zakázku se ucházel jen jeden uchazeč),
- zelenožlutý (příjemci smluv mají vysoký clustering coefficient, stejné sídlo sdílí s příjemcem velké množství subjektů, častěji jsou příjemci subjekty založené před méně než jedním rokem před uzavřením smlouvy a mají nízký základní kapitál),
- žlutý (hodnota smluv je těsně pod hranicí 2 mil. Kč, častěji nejsou strojově čitelné a mají vyšší z-score odchylky od předpokladu podle Benfordova zákona),
- světle oranžový (smluvní strany patří do stejné komunity, častěji má plátce přímé napojení na vlastníka příjemce a smlouvy nejsou strojově čitelné),
- oranžový (hodnota smlouvy převyšuje nabídkovou cenu dodavatele veřejné zakázky, příjemci mají vysokou úspěšnost u veřejných zakázek, příjemci smluv jsou častěji v insolvenci a hodnota smluv je těsně pod hranicí 6 mil. Kč),
- červenooranžový (smlouvy nejsou strojově čitelné),
- červený (smlouvy s vysokou hodnotou, která převyšuje průměrnou hodnotu smluv plátce i příjemce, veřejné zakázky mají vyšší počet uchazečů),
- tmavě červený (smlouvy mají vyšší clustering coefficient, příjemce se často uchází o veřejné zakázky se stejnými subjekty a má vyšší úspěšnost, jeden z neuronů v clusteru reprezentuje smlouvy uzavřené s nespolehlivými plátci DPH).



Obrázek 4.2: Vizualizace U-matice.

4.4 Detekce anomálií

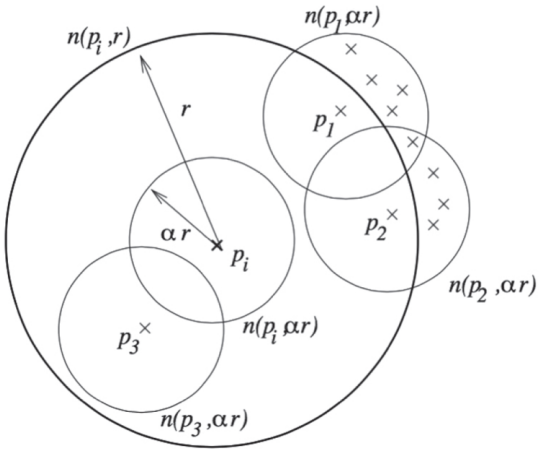
Detekce anomálií má podle [51] využití v mnoha oblastech, patří mezi ně například: kybernetická bezpečnost (ochrana soukromí, detekce malware a podvodných e-mailů), finančnictví (detekce podvodů s platebními kartami, zjišťování schopnosti splácet úvěry, predikce bankrotu), zdravotnictví (diagnóza EKG nebo radiografických snímků, monitoring pacientů, vyhledávání mutací virů a bakterií), obrana (detekce neobvyklého chování osob), zabezpečení domácnosti (zpracování dat ze senzorů), průmysl (kontrola kvality, chování zákazníků a zaměstnanců).

V [51] jsou uvedeny následující algoritmy.

4.4.1 Metody založené na vzdálenosti a hustotě

Vzdálenost od ostatních dat

Nejjednodušší metoda detekce anomálií je založena na předpokladu, že data jsou jednodimenzionální a mají normální rozdělení se střední hodnotou μ a standardní odchylkou ω. Jako anomálie se považují data s velkou vzdáleností od středu rozdělení, obvykle se používá hranice $\mu + z\sigma$, kde $z = 3$. Další možností je Grubbsův test, který určuje z v závislosti na množství dat N



Obrázek 4.3: Princip algoritmu LOCI [52].

a hladině významnosti α podle vzorce:

$$z = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-1}}{N-2+t^2_{\frac{\alpha}{2N},N-1}}},$$

kde t je kritická hodnota Studentova rozdělení.

Pro data s více dimenzemi je možné provést detekci v každé dimenzi samostatně nebo data transformovat do jedné dimenze.

Předpokládá se, že data jsou symetricky rozdělená a tvoří jeden cluster. Pokud existuje více clusterů, je nutné data přiřadit ke clusterům a provádět výpočet v rámci nich.

Local Correlation Integral (LOCI)

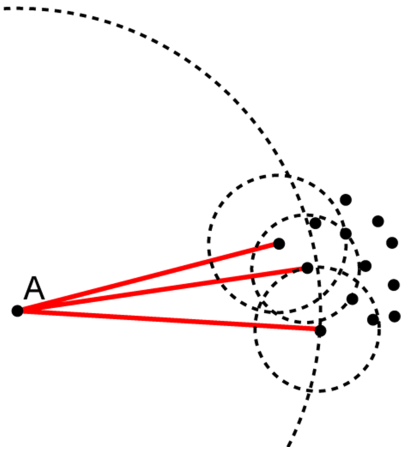
Algoritmus určí pro každý bod Multi-granularity Deviation Factor (MDEF), za anomálie se považují body s vysokou hodnotou.

MDEF se určí jako:

$$\text{MDEF}(p,r,\alpha) = 1 - \frac{n(p,\alpha r)}{\hat{n}(p,r,\alpha)},$$

kde p je bod z dat, r a $\alpha \in (0,1)$ jsou parametry, $n(p,\alpha r)$ je počet bodů ve vzdálenosti menší než αr od p a $\hat{n}(p,r,\alpha)$ je průměrný počet bodů ve vzdálenostech menších než αr od jednotlivých bodů nacházejících se ve vzdálenosti menší než r od p.

Princip algoritmu je znázorněn na obrázku 4.3. p_i je aktuálně vyhodnocovaný bod a p_1, p_2, p_3 jsou body vzdálené od p_i méně než r.



Obrázek 4.4: Princip algoritmu LOF [53].

Nejbližší sousedé

Algoritmus vychází z k-NN a jako anomálie označuje body s velkou vzdáleností ke svým k nejbližším sousedům.

Local Outlier Factor (LOF)

Podle vzdálenosti ke k nejbližším sousedům algoritmus určí lokální hustotu bodu. Lokální hustota se porovná s lokálními hustotami sousedů a jako anomálie se označí body, které mají oproti sousedům nízkou lokální hustotu.

Princip algoritmu je znázorněn na obrázku 4.4. Poloměry kružnic vyjadřují lokální hustoty bodů.

Connectivity-Based Outlier Factor (COF)

LOF selhává v případech, kdy je hustota odlehlého bodu podobná hustotě jeho sousedů. COF se od LOF liší výpočtem lokální hustoty, místo vzdálenosti ke k nejbližším sousedům se počítá délka nejkratší cesty spojující bod s těmito sousedy.

INFLuential Measure of Outlierness by Symmetric Relationship (INFLO)

Algoritmus používá množinu Reverse Nearest Neighborhood (RNN) definovanou jako:

$$\mathcal{RN}_k(p) = \{q : q \in \mathcal{D} \wedge p \in \mathcal{N}_k(q)\},$$

kde $\mathcal{D}$ je množinou bodů dat a $\mathcal{N}_k(q)$ je množina k nejbližších sousedů bodu q.

Množina $\mathcal{RN}_k(p)$ obsahuje body, které mají bod p v množině k nejbližších sousedů. Na rozdíl množiny $\mathcal{N}_k(p)$ nemusí mít $\mathcal{RN}_k(p)$ k prvků a může být i prázdná.

Jako anomálie se označují body, které nepatří mezi nejbližší sousedy svých sousedů.

4.4.2 Metody založené na pořadí

Rank-Based Detection Algorithm (RBDA)

Místo vzdálenosti k ostatním bodům používá algoritmus pořadí bodů od nejbližšího k nejvzdálenějšímu. Obvykle poskytuje lepší výsledky než algoritmy založené na vzdálenostech a hustotě, zejména pokud skupiny bodů nemají stejnou hustotu. Pro každý bod p se určí množina k nejbližších sousedů ($\mathcal{N}_k(p)$) a bod p je ohodnocen průměrem pořadí určených podle vzdáleností od jednotlivých bodů z množiny $\mathcal{N}_k(p)$.

Body s vysokým průměrným pořadím se považují za anomálie.

4.4.3 Metody založené na clusteringu

NC-clustering

Výhodou algoritmů založených na clusteringu je jejich rychlost. NC-clustering je modifikací algoritmu DBSCAN, oba mají parametr pro určení minimálního počtu bodů v clusteru, ale dosažitelnost bodu se místo velikostí okolí určuje počtem nejbližších bodů. Dalším rozdílem je, že hraniční body nespadají do clusteru. Body p a q jsou propojené, pokud p je dosažitelné z q a naopak. Spojitá komponenta C tvoří cluster, pokud je každý bod v C dosažitelný z alespoň dvou jiných bodů z C a počet bodů v C není menší než minimální počet definovaný parametrem.

Za anomálie se považují body nezařazené do žádného clusteru.

Kapitola 5

Vyhodnocení

V této kapitole je provedeno ruční vyhodnocení výsledků použitých algoritmů a jejich srovnání.

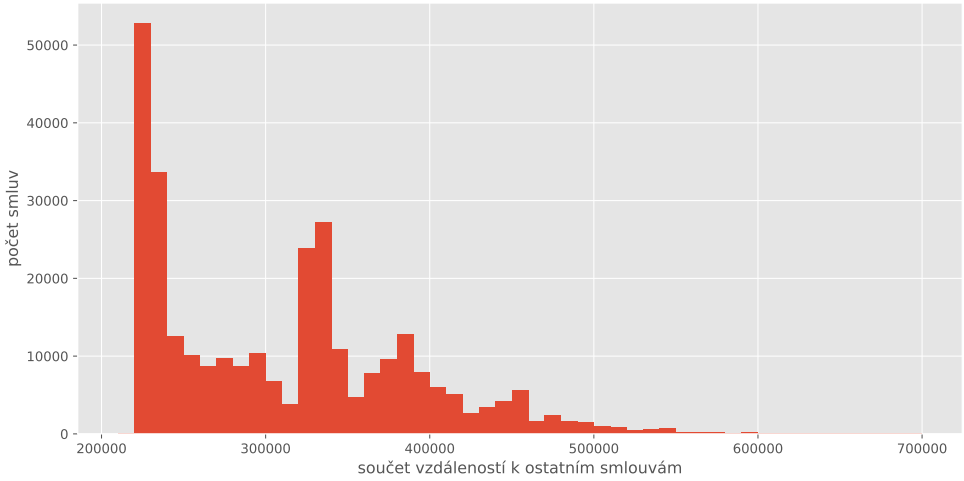
5.1 LOF

Jako první byl použit algoritmus LOF implementovaný v knihovně scikit-learn [54]. Výpočet není náročný a bylo možné jej provést pro všechny smlouvy. Parametr k byl nastaven na 1 000 a podíl odlehlých hodnot na 0,001 %. Pro výpočet vzdálenosti se používá Minkowského vzdálenost s parametrem $p = 2$ (odpovídá Euklidovské vzdálenosti).

Nalezené smlouvy jsou popsány v tabulce 5.1.

Tabulka 5.1: Smlouvy nalezené pomocí LOF.

ID	Plátce	Příjemce	Poznámka
167589	Piknik box s.r.o.	Statutární město Brno, městská část Brno-střed	
112497	Zenagel	Fakultní nemocnice Královské Vinohrady	smluvní strana není identifikována pomocí IČO
110177	Swedish Orphan Biovitrum AB	Fakultní nemocnice Královské Vinohrady	smluvní strana není identifikována pomocí IČO (zahraniční subjekt)
2938234	ČSS, z.s. - sportovně střelecký klub DUKLA Plzeň	Česká republika - Ministerstvo obrany	
1558998	Vodafone Czech Republic a.s.	Statutární město Brno	
1821	Město Chotěboř	Kraj Vysočina	příjemce je sponzorem politické strany
4225336	Daiichi Sankyo, Inc.	Fakultní nemocnice Hradec Králové	smluvní strana není identifikována pomocí IČO (zahraniční subjekt)
4351596	Bartůněk Miloš	Státní pozemkový úřad	smluvní strana není identifikována pomocí IČO (fyzická osoba)
1505798	New Karolina Office Development	Statutární město Ostrava	smluvní strana není identifikována pomocí IČO
2530714	Město Bystřice nad Pernštejnem	Kraj Vysočina	příjemce je sponzorem politické strany



Obrázek 5.1: Počet smluv v závislosti na vzdálenosti k ostatním smlouvám.

U těchto smluv nebyly nalezeny žádné podezřelé znaky, spíše jde o drobné odlišnosti, chybně vyplněné údaje a falešné hrozby v případě subjektů, které nemají přidělené IČO.

5.2 Vzdálenost od ostatních dat

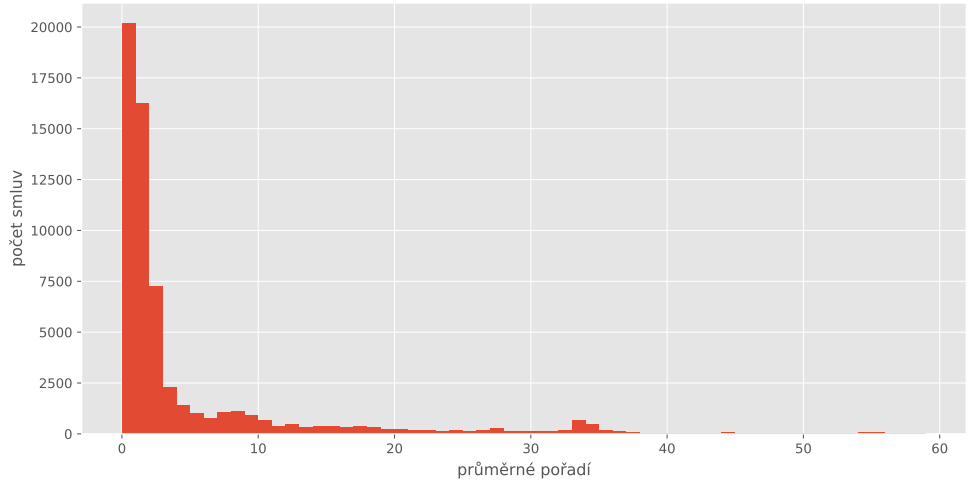
Dalším použitým algoritmem byla vzdálenost od ostatních dat. Pro každou smlouvu byl vypočten součet Euklidovských vzdáleností ke všem ostatním smlouvám. Tento výpočet není paměťově náročný, ale časová složitost algoritmu je $O(n^2)$ a výsledek se podařilo získat pouze pro 300 000 nejnovějších smluv. Algoritmus je možné snadno paralelizovat a neměl by být problém ani s postupným přidáváním nových smluv. Výhodou oproti LOF je možnost porovnávání míry odlišnosti smluv.

Rozdělení počtu smluv podle vzdálenosti k ostatním smlouvám je znázorněno na obrázku 5.1. Smluv, jejichž vzdálenost se od střední hodnoty odlišuje o více než 3σ, bylo nalezeno 1 591. V tabulce 5.2 je popsáno 17 smluv odlišujících se o více než $4{,}5\sigma$.

Některé nalezené smlouvy mají znaky, které lze považovat za podezřelé, často se ale jedná jen o nájemní smlouvy, výjimkou je například smlouva 4134012. Smlouvy s Dopravním podnikem hl. m. Prahy, a.s. na straně příjemce jsou pravděpodobně neobvyklé zejména vysokým počtem smluv a PageRank příjemce, který je způsoben jeho častým výskytem na obou stranách smluv, dále u něj je vysoký podíl veřejných zakázek v jednacím řízení bez uveřejnění způsobený chybějícími daty o dalších zakázkách.

Tabulka 5.2: Smlouvy nalezené podle vysoké vzdálenosti od ostatních smluv.

ID	Plátce	Příjemce	Poznámka
4215816	Dům kultury Akord Ostrava - Zábřeh, s.r.o.	Divadlo Verze s.r.o.	utajená hodnota smlouvy, příjemce vznikl 12 dní před uzavřením smlouvy
2562002	Správa železniční dopravní cesty, státní organizace	Veolia Energie ČR, a.s.	utajená hodnota smlouvy, text není strojově čitelný, člen dozorčího orgánu sponzorem politické strany, vysoká úspěšnost příjemce u veřejných zakázek
4311392	Ústav chemických procesů AV ČR, v.v.i.	PharmaCan s.r.o.	utajená hodnota smlouvy, nepřímá vazba na sponzora politické strany
2400198	ABOUT ME s.r.o.	Dopravní podnik hl. m. Prahy, a.s.	utajená hodnota, člen statutárního orgánu sponzorem politické strany
2079974	M-TRAFIK s.r.o.	Dopravní podnik hl. m. Prahy, a.s.	
3201206	ČSAD MHD Kladno a.s.	Dopravní podnik hl. m. Prahy, a.s.	utajená hodnota smlouvy, člen dozorčího orgánu sponzorem politické strany
4370968	Městský dům kultury Sokolov, příspěvková organizace	DIVADLO VERZE s.r.o.	text není strojově čitelný, příjemce vznikl 5 dní před uzavřením smlouvy
3603696	Město Hodonín	Lázně Hodonín, Lázně Hodonín, s.r.o.	utajená hodnota smlouvy, člen dozorčího orgánu sponzorem politické strany, plátce je zřizovatelem a společníkem příjemce, jeden z příjemců založen 26 dní před uzavřením smlouvy
704489	Údržba silnic s.r.o.	Krajská správa a údržba silnic Středočeského kraje, příspěvková organizace	člen statutárního orgánu sponzorem politické strany, všechny smlouvy plátce mají stejného příjemce
4344668	Severočeská vodárenská společnost a.s.	SčVK, a.s.	utajená hodnota smlouvy, v různých orgánech je 8 osob sponzorujících politické strany nebo jsou politiky, plátce je akcionářem příjemce a vyskytují se u nich stejné osoby, vysoká úspěšnost příjemce u veřejných zakázek
4134012	CENDIS, s.p.	Sapphire advise s.r.o.	text není strojově čitelný, hodnota těsně pod hranicí 2 mil. Kč, příjemce vznikl 4 dny před uzavřením smlouvy, hodnota více než 3× převyšuje průměrnou hodnotu smluv plátce
4313384	Okresní autobusová doprava Kolín	Dopravní podnik hl. m. Prahy, a.s.	utajená hodnota, text není strojově čitelný
3102286	Dům kultury Střelnice Rumburk, příspěvková organizace	ŠMITEC s.r.o.	utajená hodnota, příjemce vznikl 47 dní před uzavřením smlouvy, virtuální sídlo
2191926	TIPSPORT a.s.	Dopravní podnik hl. m. Prahy, a.s.	člen statutárního orgánu je sponzorem prezidentského kandidáta
4061904	Základní škola a Mateřská škola, Praha 6, Bílá 1	T-candy.cz, s.r.o.	utajená hodnota, text není strojově čitelný, příjemce vznikl 40 dní před uzavřením smlouvy, na stejné adrese sídlí 229 subjektů
2292714	To & Mi Vdf. spol. s r.o.	Ředitelství silnic a dálnic ČR	utajená hodnota, sponzoring politické strany, poslanec v dozorčí radě
2400338	ABOUT ME s.r.o.	Dopravní podnik hl. m. Prahy, a.s.	utajená hodnota, člen statutárního orgánu sponzorem politické strany



Obrázek 5.2: Počet smluv v závislosti na odlišnosti určené algoritmem RBDA.

5.3 RBDA

Jako poslední byl implementován algoritmus RBDA. Jeho nevýhodou je nutnost uložení matice o velikosti $O(n^2)$, která znemožnila použití pro více než 60 000 smluv. Podobně jako u vzdálenosti od ostatních dat je možné porovnávat míru odlišnosti smluv mezi sebou nebo výsledek přepočítat na uživatelsky přívětivý údaj (skóre od 0 do 100 apod.).

Rozdělení počtu smluv podle průměrného pořadí je znázorněno na obrázku 5.2. V tabulce 5.3 je popsáno 15 nejvíce odlišných smluv.

Všechny nalezené smlouvy spojuje vznik příjemce v krátkém období před uzavřením smlouvy. První, nejvíce odlišná smlouva, je shodná s předchozím algoritmem. Velkou část tvoří nájemní smlouvy s prohozenými smluvními stranami nebo smlouvy o zřízení pracovního místa dotovaného Úřadem práce, u kterých je nízké stáří příjemce přirozené. Algoritmus pravděpodobně zvyšuje váhu spojitých kritérií tím, že pořadí ovlivňují i nepatrné rozdíly.

5.4 Srovnání algoritmů

Pro detekci byly použity tři různé algoritmy. Podle ručního vyhodnocení dosahuje nejlepších výsledků vzdálenost od ostatních dat. Pro zvýšení přesnosti je možné na výsledky jednotlivých algoritmů aplikovat ensembling, konkrétně bagging.

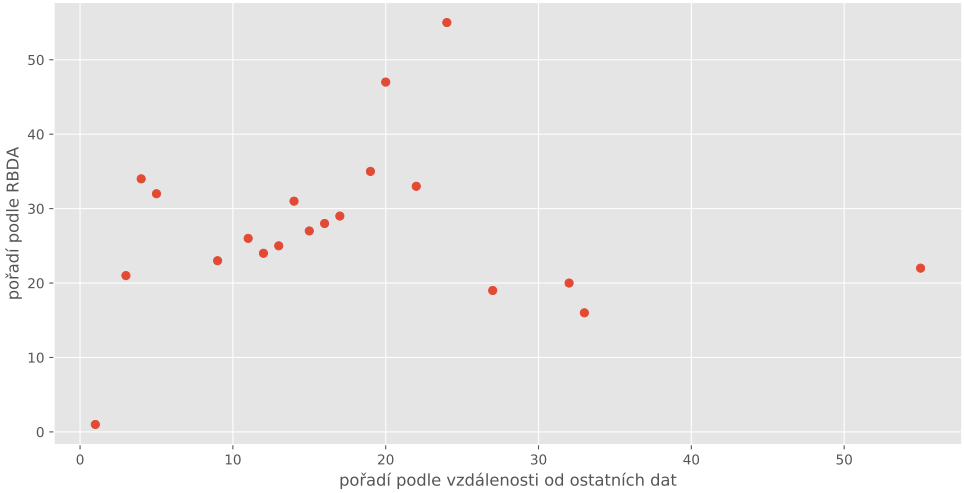
Všechny algoritmy byly spuštěny pro shodnou množinu smluv (60 000 smluv) a nastaveny tak, aby bylo jako anomálie považováno 0,1 % nejvíce odlišných smluv.

Tabulka 5.3: Smlouvy nalezené pomocí RBDA.

ID	Plátce	Příjemce	Poznámka
4215816	Dům kultury Akord Ostrava - Zábřeh, s.r.o.	Divadlo Verze s.r.o.	utajená hodnota smlouvy, příjemce vznikl 12 dní před uzavřením smlouvy
315845	Úřad práce České republiky	FLOW FINANCE s.r.o.	příjemce vznikl 53 dní před uzavřením smlouvy
4294924	Městská část Praha 6	Keepsmile s.r.o.	příjemce vznikl 43 dní před uzavřením smlouvy, prohozené strany smlouvy
4044820	Správa železniční dopravní cesty, státní organizace	DEBRNÍK s.r.o.	příjemce vznikl 25 dní před uzavřením smlouvy, prohozené strany smlouvy
4187312	Muzeum Českého krasu, příspěvková organizace	Stavby Schwarz s.r.o.	příjemce vznikl 36 dní před uzavřením smlouvy
1037009	Město Žatec	Společenství vlastníků Mládežnická 2737	příjemce vznikl 51 dní před uzavřením smlouvy, prohozené strany smlouvy
4125416	Univerzita Jana Evangelisty Purkyně v Ústí nad Labem	Pedologický institut s.r.o.	příjemce vznikl 1 den před uzavřením smlouvy,
402117	Město Havlíčkův Brod	ENERGY ENGINEERING s.r.o.	příjemce vznikl 49 dní před uzavřením smlouvy
3992048	Město Jeseník	KETCH UP s.r.o.	příjemce vznikl 47 dní před uzavřením smlouvy, virtuální sídlo
331049	MOSTECKÁ BYTOVÁ, a.s.	Správa budov Most s.r.o.	příjemce vznikl 18 dní před uzavřením smlouvy, základní kapitál příjemce je 1 Kč
4078316	Základní škola a mateřská škola Ústavní, Praha 8, Hlivická 1	Penzion Mileta s.r.o.	příjemce vznikl 36 dní před uzavřením smlouvy
3991244	Město Vizovice	Traffic Design s.r.o.	příjemce vznikl 46 dní před uzavřením smlouvy
3516050	Úřad práce České republiky	Jan Házl & syn s.r.o.	příjemce vznikl 42 dní před uzavřením smlouvy
4059040	GANEL s.r.o.	Správa nemovitostí města Šlapanice s.r.o.	příjemce vznikl 12 dní před uzavřením smlouvy, subjekty sídlí mají shodné sídlo
4075692	Město Orlová	Sociální služby města Orlová, příspěvková organizace	plátce je zřizovatelem příjemce, příjemce vznikl 13 dní před uzavřením smlouvy

Množiny smluv identifikovaných jednotlivými algoritmy mají prázdný průnik (množina smluv nalezená algoritmem LOF má prázdný průnik s oběma ostatními). Zbývající algoritmy označily shodně 1/3 smluv.

Srovnání pořadí smluv nalezených pomocí vzdálenosti od ostatních dat a RBDA je znázorněno na obrázku 5.3. Mezi pořadími je jen velmi slabá korelace (0,05).



Obrázek 5.3: Srovnání pořadí smluv nalezených pomocí vzdálenosti od ostaních dat a RBDA.

Závěr

Cílem této práce bylo navrhnout metriky pro usnadnění nalezení podezřelých smluv v registru smluv. Pomocí zvoleného algoritmu bylo takto označeno přibližně 0,5 % smluv vhodných k další kontrole. Tyto smlouvy nesou větší množství rizikových příznaků, ale bez pochopení obsahu smlouvy a zasazení do kontextu není možné učinit další rozhodnutí, tato část se patrně ani v budoucnu zcela neobejde bez práce například investigativních novinářů.

Nejprve bylo nutné najít datové zdroje obsahující využitelné informace a zjistit význam poskytovaných dat. S ohledem na neexistenci API a různá omezení přístupu byly použity tyto zdrojové systémy: Hlídač Státu, Veřejný rejstřík a Sbírka listin, ARES, databáze Cribis, Věstník veřejných zakázek a profily zadavatelů, Registr územní identifikace, adres a nemovitostí.

Pro data ze zdrojových systémů byl navržen datový model, byla vyčištěna, integrována a uložena do databáze. Z důvodů snadné práce se vztahy mezi jednotlivými entitami a implementovaných algoritmů pro analýzu grafů byla pro uložení dat zvolena grafová databáze Neo4j.

Na základě provedené analýzy dat a inspirace projektem DIGIWHIST bylo navrženo 32 příznaků popisujících smlouvy. Pro získání přehledu byla data vizualizována pomocí Self organizing map.

Podezřelé smlouvy se identifikují algoritmy pro detekci anomálií. Vyzkoušeny byly tři různé algoritmy: Local outlier factor, vzdálenost od ostatních dat a Rank-Based detection algorithm. Smlouvy označené jako anomálie s největší jistotou byly ručně ověřeny a bylo zjištěno, že nejlepších výsledků bylo dosaženo určením anomálií podle vysoké vzdálenosti k ostatním smlouvám.

Dalšímu rozšíření a praktickému využití výsledků v současnosti nejvíce brání omezení některých veřejných zdrojů dat, neevidování části subjektů ve veřejných rejstřících a chybějící data narození osob v datech přístupných přes API. Využití pouze veřejných zdrojů dat je sice možné i v současnosti, ale vyžaduje získání dat z dalších zdrojů (živnostenský rejstřík, seznam politických stran, rejstřík škol a školských zařízení, ...) a snižuje přesnost (párování osob bez dat narození). Algoritmus není paměťově náročný a je snadno paralelizovatelný, a tak by při

provedení počátečního výpočtu na výkonném hardware a průběžné aktualizaci neměl být problém s nasazením. V průběhu vzniku této práce byl také Hlídač Státu rozšířen o další funkce, například o určování typu přiloženého dokumentu a oblasti, které se dokument týká. Díky tomu by mohlo být možné přesněji detekovat anomálie.

Literatura

[1] Zákon č. 340/2015 Sb., o zvláštních podmínkách účinnosti některých smluv, uveřejňování těchto smluv a o registru smluv (zákon o registru smluv). In *Sbírka zákonů*, 24. 11. 2015. ISSN 1211-1244.

[2] Hlídač Státu z.ú.: Hlídač státu [online]. [cit. 2018-02-24]. Dostupné z: `https://hlidacstatu.cz`

[3] Elasticsearch B.V.: Elasticsearch [software]. [přístup 2018-02-26]. Dostupné z: `https://elastic.co/products/elasticsearch`

[4] Správa základních registrů: Registr osob [online]. [cit. 2018-03-10]. Dostupné z: `http://szrcr.cz/registr-osob`

[5] Ministerstvo vnitra České republiky: Registr smluv [online]. [cit. 2018-03-10]. Dostupné z: `https://smlouvy.gov.cz`

[6] Ministerstvo vnitra České republiky: Registr smluv - Zveřejnění záznamu [online]. [cit. 2018-03-15]. Dostupné z: `https://portal.gov.cz/formulare/registr-smluv-zverejneni-zaznamu/online/ACFSTTS3QFYBNOGBPKJHVN4ELMFHBTNW`

[7] Mazancová, H.: Nadbytečný, řekl právník Budvaru o registru smluv. Senátorům napsal zdarma návrh na jeho zrušení [online]. *iROZHLAS*, [cit. 2018-02-28]. Dostupné z: `https://irozhlas.cz/zpravy-domov/navrh-na-zruseni-registru-smluv-senatorum-ho-zdarma-napsal-pravnik-budvaru_1711070600_hm`

[8] Zákon č. 304/2013 Sb., o veřejných rejstřících právnických a fyzických osob. In *Sbírka zákonů*, 12. 9. 2013. ISSN 1211-1244.

[9] Ministerstvo spravedlnosti České republiky: Technická specifikace pro předávání digitalizovaných listin do Sbírky listin veřejného rejstříku [online]. [cit. 2018-02-12]. Dostupné z: `https://or.justice.cz/ias/ui/specifikaceSL`

[10] Zákon č. 563/1991 Sb., o účetnictví. In *Sbírka zákonů*, 12. 12. 1991. ISSN 1211-1244.

[11] Ministerstvo spravedlnosti České republiky: Podmínky provozu a využívání údajů veřejného rejstříku na síti Internet [online]. [cit. 2018-02-12]. Dostupné z: `https://or.justice.cz/ias/ui/podminky`

[12] Ministerstvo financí České republiky: Administrativní registr ekonomických subjektů [online]. [cit. 2018-03-02]. Dostupné z: `http://wwwinfo.mfcr.cz/ares`

[13] CRIF - Czech Credit Bureau, a. s.: Cribis [online]. [cit. 2018-03-02]. Dostupné z: `https://mcribis.cz`

[14] Ministerstvo pro místní rozvoj České republiky: Věstník veřejných zakázek [online]. [cit. 2018-03-03]. Dostupné z: `https://vestnikverejnychzakazek.cz`

[15] Zákon č. 134/2016 Sb., o zadávání veřejných zakázek. In *Sbírka zákonů*, 19. 4. 2016. ISSN 1211-1244.

[16] Ministerstvo pro místní rozvoj České republiky: Portál o veřejných zakázkách a koncesích [online]. [cit. 2018-03-03]. Dostupné z: `http://portal-vz.cz`

[17] Bartoň, D.: *Veřejné zakázky – jednací řízení*. Rigorózní práce, Masarykova univerzita, Právnická fakulta, Katedra obchodního práva, 2014.

[18] QCM, s.r.o.: Portál pro vhodné uveřejnění [online]. [cit. 2018-03-15]. Dostupné z: `https://vhodne-uverejneni.cz`

[19] profilzadavatele.cz, s.r.o.: Profil zadavatele [online]. [cit. 2018-03-15]. Dostupné z: `https://profilzadavatele.cz`

[20] Tender systems s.r.o.: Tender arena [online]. [cit. 2018-03-15]. Dostupné z: `https://tenderarena.cz`

[21] Profinit EU s.r.o.: public-contracts [software]. [přístup 2018-03-03]. Dostupné z: `https://github.com/profinit/public-contracts`

[22] PostgreSQL Global Development Group: PostgreSQL [software]. [přístup 2018-03-03]. Dostupné z: `https://postgresql.org`

[23] Český úřad zeměměřický a katastrální: RÚIAN [online]. [cit. 2018-03-03]. Dostupné z: `http://cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/RUIAN.aspx`

[24] Cox, G.: Introduction to Graph Databases [online]. [cit. 2018-03-09]. Dostupné z: https://compose.com/articles/introduction-to-graph-databases

[25] Neo4j, Inc.: Neo4j [online]. [cit. 2018-03-10]. Dostupné z: https://neo4j.com

[26] Callidus Software Inc.: OrientDB [software]. [přístup 2018-03-15]. Dostupné z: https://orientdb.com

[27] ArangoDB GmbH: ArangoDB [software]. [přístup 2018-03-15]. Dostupné z: https://arangodb.com

[28] The Linux Foundation: JanusGraph [software]. [přístup 2018-03-15]. Dostupné z: http://janusgraph.org

[29] Talend Inc.: Talend Open Studio [software]. [přístup 2018-03-10]. Dostupné z: https://talend.com/products/talend-open-studio

[30] Gephi Consortium: Gephi [software]. [přístup 2018-03-10]. Dostupné z: https://gephi.org

[31] Informatica LLC.: 2017 Gartner Magic Quadrant for Data Integration Tools [online]. [cit. 2018-03-13]. Dostupné z: https://informatica.com/data-integration-magic-quadrant.html

[32] Eibl, M.; Chromý, M.; Leyer, P.: Rekordmani „z ruky“ aneb miliarda korun mimo zákon [online]. [cit. 2018-03-30]. Dostupné z: https://transparency.cz/rekordmani-z-ruky-aneb-miliarda-korun-mimo-zakon

[33] Durtschi, C.; Hillison, W.; Pacini, C.: The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting*. Dostupné z: https://pdfs.semanticscholar.org/1020/696451732ce203b219c19fdc31ef1ab8d8c8.pdf

[34] Zákon č. 90/2012 Sb., o obchodních korporacích a družstvech. In *Sbírka zákonů*, 25. 1. 2012. ISSN 1211-1244.

[35] Komancová, R.: Základní kapitál od 1.1.2014 a související dopady do vlastního kapitálu [online]. [cit. 2018-03-30]. Dostupné z: http://fucik.cz/publikace/zakladni-kapital-od-1-1-2014-a-souvisejici-dopady-do-vlastniho-kapitalu

[36] Latora, V.; Nicosia, V.; Russo, G.: *Complex Networks*. Cambridge University Press, první vydání, 7 2017, ISBN 978-1-107-10318-4.

[37] Fundlift, s.r.o.: Fundlift [online]. [cit. 2018-03-30]. Dostupné z: https://fundlift.cz

[38] Chaloupská, M.: Firmy často využívají virtuální pražské adresy. Na jednom místě jich sídlí i 1500 [online]. *iROZHLAS*, [cit. 2018-04-03]. Dostupné z: https://irozhlas.cz/zpravy-domov/firmy-casto-vyuzivaji-virtualni-prazske-adresy-na-jednom-miste-jich-sidli-i-1500-_201412010745_mhromadka

[39] Bisnode Česká republika, a.s.: Virtuální sídla jsou plná nespolehlivých plátců DPH [online]. [cit. 2018-04-03]. Dostupné z: https://bisnode.cz/o-bisnode/o-nas/novinky/virtualni-sidla-jsou-plna-nespolehlivych-platcu-dph

[40] Morávek, D.: Pozor na virtuální adresy, na některých je většina plátců DPH nespolehlivých [online]. *Podnikatel.cz*, [cit. 2018-04-03]. Dostupné z: https://podnikatel.cz/clanky/pozor-na-virtualnich-adresy-je-na-nich-znacne-vyssi-podil-nespolehlivych-platcu

[41] Jaderberg, J.; Needham, M.: Neo4j Graph Algorithms User Guide [online]. [cit. 2018-03-26]. Dostupné z: https://neo4j-contrib.github.io/neo4j-graph-algorithms

[42] Rogers, I.: The Google Pagerank Algorithm and How It Works [online]. [cit. 2018-03-26]. Dostupné z: http://cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm

[43] Ullman, J. D.: Web Search [online]. [cit. 2018-04-13]. Dostupné z: http://infolab.stanford.edu/~ullman/mining/websearch.pdf

[44] Facebook Inc.: Facebook [online]. [cit. 2018-03-27]. Dostupné z: https://facebook.com

[45] Ugander, J.; Karrer, B.; Backstrom, L.; aj.: The Anatomy of the Facebook Social Graph. *CoRR*, ročník abs/1111.4503, 2011. Dostupné z: http://arxiv.org/abs/1111.4503

[46] Stegehuis, C.; van der Hofstad, R.; van Leeuwaarden, J. S. H.: Power-law relations in random networks with communities. *Phys. Rev. E*, ročník 94, Jul 2016, doi:10.1103/PhysRevE.94.012302. Dostupné z: https://link.aps.org/doi/10.1103/PhysRevE.94.012302

[47] Tanis, D.; Fazekas, M.: Mining public procurement data for corruption [online]. Oct 2017, [cit. 2018-04-02]. Dostupné z: http://digiwhist.eu/wp-content/uploads/2017/10/Athens-presentation_2nd-part.pdf

[48] Open Knowledge Foundation Deutschland e.V.: Děláme veřejné zakázky transparentnější [online]. [cit. 2018-04-02]. Dostupné z: https://opentender.eu

[49] Epina GmbH: Kohonen Network - Background Information [online]. [cit. 2018-04-30]. Dostupné z: http://lohninger.com/helpcsuite/kohonen_network_-_background_information.htm

[50] Kordík, P.: SOM. Fakulta informačních technologií, České vysoké učení technické v Praze, 2016.

[51] Mehrotra, K. G.; Mohan, C. K.; Huang, H.: *Anomaly Detection Principles and Algorithms (Terrorism, Security, and Computation)*. Springer, první vydání, 2017, ISBN 978-3-319-67526-8.

[52] Nun, I.; Protopapas, P.; Sim, B.; aj.: Ensemble Learning Method for Outlier Detection and its Application to Astronomical Light Curves. *The Astronomical Journal*, 9 2016. Dostupné z: http://iopscience.iop.org/article/10.3847/0004-6256/152/3/71

[53] Wikimedia Commons: Basic idea of LOF [online]. 2010, [cit. 2018-04-22]. Dostupné z: https://commons.wikimedia.org/wiki/File:LOF-idea.svg

[54] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; aj.: Scikit-learn: Machine Learning in Python [software]. [přístup 2018-04-20]. Dostupné z: http://scikit-learn.org

Příloha A

Seznam použitých zkratek

API Application Programming Interface

ARES Administrativní registr ekonomických subjektů

COF Connectivity-Based Outlier Factor

CSV Comma-separated values

ČR Česká republika

DBSCAN Density-based spatial clustering of applications with noise

DPH Daň z přidané hodnoty

HTML HyperText Markup Language

IČO Identifikační číslo osoby

INFLO INFLuential Measure of Outlierness by Symmetric Relationship

JŘBU Jednací řízení bez uveřejnění

JSON JavaScript Object Notation

LOCI Local Correlation Integral

LOF Local Outlier Factor

MDEF Multi-granularity Deviation Factor

OCR Optical Character Recognition

PDF Portable Document Format

PSČ Poštovní směrovací číslo

RBDA Rank-Based Detection Algorithm

RNN Reverse Nearest Neighborhood

RÚIAN Registr územní identifikace, adres a nemovitostí

S-JTSK Systém jednotné trigonometrické sítě katastrální

SOM Self organizing map

SQL Structured Query Language

SR Slovenská republika

VVZ Věstník veřejných zakázek

XML Extensible Markup Language

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text`
  - `DP_Staněk_Jan_2018.pdf` ..... text práce ve formátu PDF