VYJÁDŘENÍ ŠKOLITELE K DISERTAČNÍ PRÁCI

Téma práce:	**Data Mining in Social Network Analysis**
Author práce:	**Mgr. Peter Zvirinský**, KTIML MFF UK, Praha
Školitelka:	**doc. RNDr. Iveta Mrázová, CSc.**, KTIML MFF UK, Praha

---

Aktuálnost tématu práce

Zadluženost (ať už fyzických anebo právnických osob) v České republice nabývá obrovských rozměrů. Jen v hlavním městě se týká exekuce desítek tisíc lidí a obdobný trend je patrný i v dalších krajích. Extrémně vysoký počet osob v exekuci se pak projevuje v nižší účasti ve volbách a zvyšuje úspěch extrémně pravicových a populistických stran. Exekuce rodičů zároveň negativně ovlivňují školní úspěchy jejich dětí, a to více než nezaměstnanost nebo nižší vzdělání rodičů. Uspokojení věřitelů přitom může být pod deseti procenty z hodnoty dluhů.

Poměrně značné množství dlužníků řeší své finanční závazky v rámci insolvenčního řízení. Jeho průběh a podmínky upravuje insolvenční zákon. Údaje o jednotlivých insolvenčních řízeních a jejich aktérech obsahuje Insolvenční rejstřík ČR, který je veřejně přístupný online. Uložená data jsou jak ve strukturované, tak také v nestrukturované podobě (obrazové, časové ale zejména textové údaje). Mnohé z informací se ale plně projeví až ve formě vztahů mezi jednotlivými aktéry řízení, ty pak vytvářejí tzv. sociální síť.

Údaje z Insolvenčního rejstříku ČR je možné využít při analýze zadluženosti české společnosti. Výzvou ovšem zůstává nutnost efektivně a spolehlivě extrahovat velké objemy dat a předzpracovat je do podoby vhodné pro následující zpracování technikami datové analýzy, strojového učení a analýzy sociálních sítí. Problémem může být i extrémně řídké zastoupení u dat určitého typu a adekvátní interpretace získaných výsledků.

Cíl práce

Záměrem práce tedy bylo zmapovat procesy související s platební neschopností a insolvenčním řízením pomocí metod analýzy sociálních sítí. Aby bylo možné lépe zohlednit průběh insolvenčního řízení v čase, snažil se disertant využít při modelování těchto procesů principů dynamických sociálních sítí. Velké množství užitečných informací ovšem obsahují i nestrukturované údaje připojené k probíhajícím řízením např. ve formě naskenovaných textových dokumentů. Kvůli obohacení studované insolvenční sociální sítě bylo dalším cílem práce extrahovat z těchto dokumentů vhodné údaje a využít je při analýze této sítě. Posledním cílem práce bylo navrhnout způsob, jak odhadnout vývoj vlivu jednotlivých uzlů v rámci dané sociální sítě.

Metody zpracování

Po zevrubném úvodu do řešené problematiky a formulaci cílů prováděného výzkumu se rešeršní část textu zabývá modelem tzv. sociálních sítí a studiem

metod vhodných pro jejich analýzu. Kapitola 3 podrobně popisuje problematiku insolvenčního zákona a Insolvenčního rejstříku ČR. Specifikaci pokročilejších přístupů pro netriviální extrakci informací z insolvenčního rejstříku se věnuje Kapitola 4 a Kapitola 5 představuje hlavní ideje frameworku GraphSlices, který autor implementoval pro snazší modelování insolvenční sociální sítě.

Experimentální část disertace nabízí řešení vybraných úloh z oblasti analýzy insolvenčního rejstříku – zejména charakteristiky typických forem zadlužení, jeho průběhu a dosahované výše. Nové poznatky přinášejí i přístupy navržené pro odhad vlivu a jeho možného vývoje u jednotlivých subjektů insolvenčního řízení. Závěrečná část práce shrnuje dosažené výsledky a nabízí náměty pro další výzkum v řešené oblasti. Vlastní text je napsaný v angličtině, má pěknou grafickou úpravu a je doplněný celou řadou ilustrativních obrázků, tabulek a grafů.

Překlepy se v práci objevují jen zřídka (např. na ř. 4 Algoritmu 1 na str. 57 – „scapeIsir" namísto „scrapeIsir," v definici 2.7.4 – „timestmap" namísto „timestamp" anebo na ř. 8 kapitoly 2.4.3 – „leave" namísto „leaf"). Některé z pojmů by bylo vhodné zavést přesněji (např. značení použité ve vztahu 2.2 a význam zvolených parametrů pro charakter modelované sítě, dále pojem tzv. betweenness centrality pro hrany anebo grafový isomorfismus).

Disertaci doplňuje rozsáhlá bibliografie čítající 140 položek, implementovaný software a odkaz na repozitář s daty extrahovanými z insolvenčního rejstříku (https://github.com/zviri/czech-insolvency-dataset). Přesto by bylo vhodné seznam použité literatury doplnit, např. o použité monografie B. Liu (Web Data Mining) a A.-L. Barabásiho (Network Science).

Dosažené výsledky

Posuzovaná disertace je výrazně aplikačně orientovaná, její podstatou je návrh, detailní popis a použití vhodné metodologie pro zpracování záznamů z Insolvenčního rejstříku ČR. Zvolený postup přirozeným způsobem kombinuje klasické metody dobývání znalostí pro extrakci znalostí ze strukturovaných dat a z textu s novějšími technikami analýzy sociálních sítí a algoritmy strojového učení. Použité přístupy práce dále rozvíjí směrem k vyšší spolehlivosti a snazší interpretabilitě získaných výsledků se zřejmým přesahem do dalších oblastí computer science, ale i společenských věd. Ke zpracování údajů z insolvenčního rejstříku použil disertant povětšinou vlastní software, zejména

- systém IRES pro extrakci informací ze záznamů a dokumentů uložených v insolvenčním rejstříku; extrahovaná data (CID ~ Czech Insolvency Dataset) zahrnují údaje o více než 370 000 insolvencích z let 2008 až 2022,
- framework GraphSlices pro práci s (dynamickými) sociálními sítěmi a
- modul PredictionModel pro odhad budoucího vývoje ve studované sociální síti.

Do diskuse mám následující otázky:

1. Pro detekci komunit je v Kapitole 6.1. použitý dříve nezmiňovaný algoritmus MOOM (Method of Optimal Modularity). Popište jeho princip.

2. Posuďte možnosti integrace grafových DB technologií (např. Neo4j a knihovny Graph Data Science) do implementovaného frameworku GraphSlices pro analýzu (dynamických) sociálních sítí.
3. Vysvětlete základní principy funkce tzv. grafových neuronových sítí a posuďte možné přínosy jejich využití při analýze dat z Insolvenčního rejstříku ČR.

Závěr

Předkládaná práce splňuje požadavky kladené na disertační práci. Uchazeč se v práci zabývá společensky i odborně vysoce aktuální problematikou analýzy sociálních sítí, které vznikají v rámci insolvenčních řízení. Dosažené výsledky vznikly v rámci dvou grantových projektů a jsou podpořeny šesti články zveřejněnými většinou na mezinárodních recenzovaných konferencích. Některé z publikací již byly citované. Součástí práce je také software implementovaný pro praktické ověření vlastností navrhovaných přístupů, zde na datech z Insolvenčního rejstříku ČR. Stanovené cíle se tedy disertantovi podařilo splnit.

Těžiště předložené práce vidím zejména v návrhu několika nových technik podstatných pro efektivní analýzu charakteru zkoumaných insolvenčních řízení a vzájemných vztahů mezi jejich aktéry. Navržené přístupy vycházejí z moderních principů používaných při dobývání znalostí v rámci sociálních sítí. Dají se ale použít i pro řešení jiných úloh z oblasti strojového učení anebo umělých neuronových sítí. Odborný přínos práce přitom spočívá i ve snadné interpretaci extrahovaných znalostí. Předloženou práci proto doporučuji k obhajobě a po jejím úspěšném obhájení doporučuji udělit Mgr. Petrovi Zvirinskému titul Ph.D.

V Praze, 14. 3. 2024

doc. RNDr. Iveta Mrázová, CSc., KTIML MFF UK