Stanford Network Analysis Project – zdroj dat a nástrojů pro analýzu sociálních sítí

Aleš Vomáčka

Analýza sociálních sítí představuje relativně mladý přístup ke zkoumání sociální reality. Tato technika, která se inspiruje Morenovou sociometrií, je založena na analýze relačních dat, tedy vazeb mezi aktéry. Tvoří tak alternativu ke klasickým technikám, jež se zaměřují na zkoumání atributů respondentů (např. jejich úrovně vzdělání). Analýza sociálních sítí sice pronikla do českého prostředí již na sklonku 90. let [Kusá 1997; Buštíková 1999], dodnes je ale spíše okrajovou záležitostí. V posledním desetiletí zájem o tuto techniku rostl a objevily se studie aplikující analýzu sociálních sítí na témata integrace migrantů [Vašát, Bernarnd 2015], regionální politiky [Skála 2013; Vajdová et al 2010] nebo politického aktivismu [Mazák, Diviák 2018]. V omezené míře je rozvíjena i metodologie [Toušek 2009; Diviák 2017]. Analýza sociálních sítí ale v českém prostředí stále zůstává relativně neznámou alternativou k tradičnějším způsobům analýzy dat. Tomu odpovídá fakt, že v České republice v současné době neexistuje datový archiv poskytující data využitelná pro síťovou analýzu. V zahraničí ovšem takové projekty najdeme.

Jednou z celosvětově nejvýznamnějších organizací pro rozvoj analýzy sociálních sítí je *Stanford Network Analysis Project* (SNAP[1]). Tento projekt, který spatřil světlo světa roku 2004, sídlí na americké Stanfordské univerzitě pod vedením Jura Leskovece a jeho týmu a plní dvě důležité funkce pro rozvoj zkoumání sociálních sítí.

První z nich je spravování datového archivu *Stanford Large Network Dataset Collection*, druhou pak vývoj nástrojů pro analýzu tohoto typu dat. Archiv obsahuje volně ke stažení více než 50 datasetů rozdělených do 17 tematických kategorií. Většina dat původně pochází z internetových komunit, jedná se tedy o sociální sítě uživatelů *Facebooku*, *Twitteru*, *Wikipedie* nebo *YouTube*. K nalezení jsou zde ale i (anonymizovaná) data o e-mailové komunikaci jedné (kvůli anonymizaci nejmenované) evropské výzkumné instituce nebo sítí citací na portálu *Arxiv.org*.

Speciální pozornost si od sociologů zaslouží právě data z prostředí internetových komunit. Vzhledem k rostoucímu významu internetové komunikace a relativně malým nákladům na pořízení online dat představují internetové komunity atraktivní předmět sociálního výzkumu. Pro příklady konkrétního využití těchto dat je možné nahlédnout do seznamu prací publikovaných členy SNAP. Například práce Althoffa, Jindala a Leskovece [2017] mapuje vztah mezi aktivitou ve fyzickém a online světě pomocí dat ze self-trackingové sportovní aplikace. Dále je možné zmínit práce Kumara, Hamiltona, Leskovece, Jurafskyho [2018] o konfliktech na internetovém fóru *Reddit* nebo spotřebním chování na stránce *Pinterest* od autorů Frankowskiho a Leskovece [2016].

[1] snap.stanford.edu

Jak již název archivu napovídá, většinu sítí lze označit za velké. Počet uzlů, tj. aktérů, v dostupných datech se pohybuje řádově v tisících až milionech s odpovídajícím počtem vazeb. Datasety obsahují informace o povaze sítí, tedy zda obsahují orientované vazby, zda jsou bipartitní (tj. obsahují více typů uzlů, například jedince a stránky, které tito jedinci navštěvují) nebo zda obsahují atributy. Dále je dostupný základní popis struktury sítě, mimo jiné počet vazeb a uzlů, průměrný shlukový koeficient, počet trojúhelníků v síti a poloměr. Samozřejmostí jsou pro všechna data informace o místě a době sběru. Všechny tyto údaje umožňují rychlé nalezení vhodných dat podle preferencí výzkumníka. Archiv sice nemá propracovanější filtrovací systém, vzhledem k malému počtu datasetů ale není problém vyhledávat ručně.

Kromě vlastních dat obsahuje SNAP i odkazy na řadu dalších datových archivů uchovávajících relační data. Zvláštní zmínku si zaslouží zejména datový archiv spravovaný týmem vyvíjejícím software *Pajek* a kolekce dat shromážděná Markem Newmannem. Tyto dva zdroje obsahují velké množství dat o malých sociálních skupinách, tedy něco, co *Stanford Large Network Dataset Collection* postrádá. SNAP tak představuje důležitý rozcestník v systému archivů dat využitelných pro analýzu sociálních sítí.

Je nicméně třeba zmínit dvě potenciální nevýhody SNAP archivu. Vzhledem k lokaci sídla SNAP není překvapivé, že většina dat je amerického původu. Archiv proto nemusí být pro evropské výzkumníky tak užitečný jako pro jejich americké kolegy. Najdou se však výjimky, kromě již zmíněné e-mailové komunikace evropské výzkumné agentury například síť komunikací na slovenské stránce *Pokec*. Data jsou také relativně stará, většina pochází z období mezi roky 2000 a 2010, jejich věcný přínos pro současný výzkum tak může být problematický. Archiv proto pravděpodobně využijí zejména metodologicky zaměření výzkumníci pro vytváření měřicích nástrojů.

Jak již bylo zmíněno, archivace dat není jedinou činností SNAP. Jeho členové také vyvíjejí nástroje pro analýzu relačních dat v jazycích C++ a Python zvané *Snappy*. Tyto nástroje jsou vyvíjeny speciálně pro manipulaci, analýzu a vizualizaci velkých sítí. Nástroje obsahují podrobnou dokumentaci a k dispozici je přímo na internetových stránkách i několik návodů pro základní využívání nástrojů. K dispozici jsou i sylaby kurzů vyučovaných členy projektu, které mohou sloužit jako inspirace pro další výzkum. SNAP navíc pod svou hlavičkou organizuje velké množství dílčích projektů. Je možné zmínit projekty *Model-based Approach to Detecting Densely Overlapping Communities in Networks* a *The Structure of Political Media Coverage as Revealed by Quoting Patterns*. Pro všechny tyto projekty jsou kromě výsledků analýz dostupná i data a kód použitý pro jejich zpracování. Stránky projektu obsahují i odkazy na další software využitelný pro analýzu sociálních sítí, zejména *Pajek* a *NodeXL*. Výhodou těchto softwarů je větší uživatelská přívětivost, než jakou poskytuje *Snappy* v rozhraní *Python*.

Jak bylo řečeno na začátku, analýza sociálních sítí patří k rozvíjejícím se specializacím v oboru sociálních věd. Institucionální podpora, včetně dostupnosti dat pro její aplikaci, je proto omezená. Nicméně díky rozvoji výpočetní techniky, umožňující zpracování velkého množství informací, a pokroků v teorii grafů se analýza sociálních sítí stává stále populárnější a spolu

s rostoucí popularitou se rozvíjí i infrastruktura kolem relačních dat, představovaná i *Stanford Network Analysis Project*.

ZDROJE:

Althoff, Tim, Pranav Jindal, Jure Leskovec. 2017. „Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior". Pp. 537–546 in de Rijke, Maarten, Milad Shokouhi (eds.) *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining.* New York: ACM.

Buštíková, Lenka, 1999. „Analýza sociálních sítí". *Sociologický časopis* 35 (2):193–206.

Diviák, Tomáš. 2017. „Ekvivalence a blokové modelování v analýze sociálních sítí". *Naše společnost* 15 (1): 27–40.

Frankowski, Dan, Jure Leskovec. 2016. „Understanding Behaviors that Lead to Purchasing: A Case Study of Pinterest". Pp. 531–540 in Krishnapuram, Balaji, Mohak Shah (eds.) *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM.

Kusá, Zuzana. 1997. Analýza sociálnych sietí a jej miesto v sociologickom skúmaní. Sociológia 29 (5): 479–504.

Mazák, Jaromír, Tomáš Diviák. 2018. „Transactional activism without transactions: network perspective on anti-corruption activism in the Czech Republic". *Social Movements Studies* 17 (2): 203–218.

Skála, Vít. 2013. „Využití analýzy sociálních sítí pro odkrývání mocenských vztahů ve volených orgánech. Analýza vazeb mezi zastupiteli Kraje Vysočina". *Acta Politologica* 5 (1): 46–68.

Srijan Kumar, William Hamilton, Jure Leskovec, Dan Jurafsky. 2018. „Community Interaction and Conflict on the Web". Pp. 933–943 in Champin, Pierre-Antoine, Fabien Gandon, Lionel Médini (eds.) *Proceedings of the 2018 World Wide Web Conference.* Geneva: International World Wide Web Conferences Steering Committee.

Toušek, Laco. 2009. „Problematika vytváření relačních dat: příklad analýzy sociálních sítí bezdomovců". *Antropowebzin* 2 (3): 35–42.

Vajdová, Zdenka, Josef Bernard, Jana Stachová, J, Daniel Čermák. (2010). „Síť institucionálních aktérů rozvoje malého města". *Czech Sociological Review* 46 (2): 281–299.

Vašát, Petr, Josef Bernard. 2015. „Forming Communities or Social Integration? A Personal Network Analysis of Ukrainian Immigrants in Pilsen". *Czech Sociological Review* 51 (2): 199–225.