Ví policie, kdo vás zavraždí?

Data mining a mapy společenské sítě

S rostoucím zapojením počítačů do všech lidských činností začínají být získané záznamy o našem životě nepříjemně přesné, tím spíše, že jejich sběr často probíhá bez našeho vědomí. Co o společnosti prozrazují data a komu se vyplatí je zpracovávat?

HONZA ŠÍPEK

V průzkumu zaměřeném na sexuální chování můžete lhát o počtu orgasmů na jeden sex. Těžko ale zakryjete, kolikrát a kde jste platili kartou nebo komu adresujete e-maily. Stroje sbírají data s trpělivou přesností. Sofistikovaná analýza ohromného množství dat („big data“), které se také říká „data mining“, dolování dat, zajímá v první řadě banky a pojišťovny, které přemýšlejí, zda vám mají navýšit úvěrový limit nebo zvednout pojistku. A marketingové specialisty! Když se v roce 2004 k USA blížil hurikán Frances, ozkoušeli informační mágové obchodního řetězce Wal-Mart svůj nový trumf: za pomoci data miningu zjistit, co by lidé mohli před katastrofou kupovat. Analyzovali data z prodejů před nedávným hurikánem Charley a překvapeně zjistili, že kromě baterek a pršiplášťů seznamu dominuje pivo a „pop-tarts“, taštičky plněné ovocnou náplní. Nakonec se rozhodli svým databázím věřit – obsahovaly totiž 460 terabytů dat, odhadem dvojnásobek obsahu tehdejšího internetu – a vyslali hurikánu vstříc náklaďáky s pivem, taštičkami a výbavou pro přežití. Vyplatilo se.

Začátky informační společnosti

Hitlerovské Německo bylo bez nadsázky první informační společností. Americká IBM nejenže dodávala Hitlerovi stroje na zpracování děrných štítků, které umožnily bleskové předválečné sčítání lidu a identifikaci Židů, ale skrze svou německou dceřinou společnost mu poskytovala servis i po celou válku. Její stroje zefektivnily evidenci židovského majetku, logistiku transportů i „konečné řešení“. Československá StB používala počítače v omezené míře, většinu dat zpracovávala v papírové podobě, dosahovala v tom však obdivuhodné dokonalosti. Estébáci potřebovali obzvlášť informace o sociálních vazbách a k vytvoření dojmu, že o člověku vědí vše, využívali i zdánlivě nenápadné detaily z běžného života. Archivy východoněmecké Stasi vydaly ručně kreslenou mapu jedné takové sociální sítě. K jejímu vytvoření byly nejspíš zapotřebí stovky hodin sledování v terénu, zprávy udavačů, riskantní nasazování štěnic, odposlechy a rešeršní práce v archivech. Dnes podobné grafy vytvářejí počítače automaticky a za pár okamžiků.

Matematickou teorii grafů založil Leonhard Euler, když v roce 1736 elegantně vyřešil tehdy oblíbený hlavolam: zda je možné přejít všech sedm mostů v pruském Královci a vrátit se do původního místa, aniž by bylo nutné nějaký most přejít dvakrát. Euler všechny části pevniny oddělené vodou nakreslil jako body (uzly) a mosty jako jejich spojnice (vazby). Úvahou dospěl k tomu, že uzel se třemi vazbami musí být pouze výchozí nebo konečný, přičemž všechny uzly v jeho grafu mají tři vazby, a tedy že problém se sedmi mosty nemá řešení. Podobně můžeme zobrazit vztah molekul v buňce, síť sexuálních partnerů nebo třeba celou lidskou společnost. Každý člověk je pak uzlem a to, že někoho zná, vytvoří vazbu, spojnici mezi dvěma uzly.

Říká se, že od každého člověka (uzlu) na planetě nás dělí průměrně šest přeskoků (vazeb). Kolega zná člověka, který pomáhal imigrantům z Iráku, ten jednal s iráckým ambasadorem, s nímž údajně jednal i Mohammad Atta, a ten s někým dalším, kdo byl v kontaktu s bin Ládinem. Většina lidí udržuje zhruba podobné množství kontaktů, pak se ale vyskytují lidé, kteří jich mají tisíce. Obecně se pro ně používá pojem centrum nebo hub. Novinář Malcolm Gladwell ve své knize *Bod zlomu* (The Tipping Point, 2000; česky 2006) popisuje tyto lidi jako doslova posedlé shromažďováním kontaktů a seznamováním dalších lidí, kterým jde spíš o komunikaci samotnou než o její obsah. Informace (stejně tak kapavka nebo módní slovo) se pak sítí šíří nejprve poměrně normálně z uzlu na uzel, dokud nenarazí na hub – odtud se rozprskne do všech stran a nabere na rychlosti. Taková síť je poměrně odolná vůči výpadkům jednotlivých uzlů, ale při odstranění vysoce propojených uzlů se může snadno rozpadnout na vzájemně oddělené ostrůvky. Jinými slovy: víte, koho zavřít, až bude hrozit, že se něco semele (o těchto sítích zvaných bezškálové vyšla v češtině v roce 2005 pod názvem *V pavučině sítí* průkopnická kniha matematika Alberta-László Barabásiho).

To je skutečná podstata sporu o „data retention“, povinnosti telekomunikačních operátorů uchovávat data o proběhlé komunikaci. Její obránci argumentují tím, že se neuchovává samotný obsah vašeho hovoru nebo esemesky, ale jenom záznam o tom, že komunikace proběhla. To ale úplně stačí. Specializovaný software typu Analyst's Notebook dokáže z výpisu hovorů (nebo třeba bankovních operací) vytvořit podrobnou mapu sociální sítě. Ve shlucích se společně ocitnou lidé, kteří k sobě patří, aniž by spolu komunikovali přímo – mají-li jistý počet společných známých. Tato zbraň má oboustranné ostří: nespokojení občané si mohou například z obchodního rejstříku vytahat data o firmách podezřelých z korupce a zobrazit jejich pavučinu a napojení na politiky.

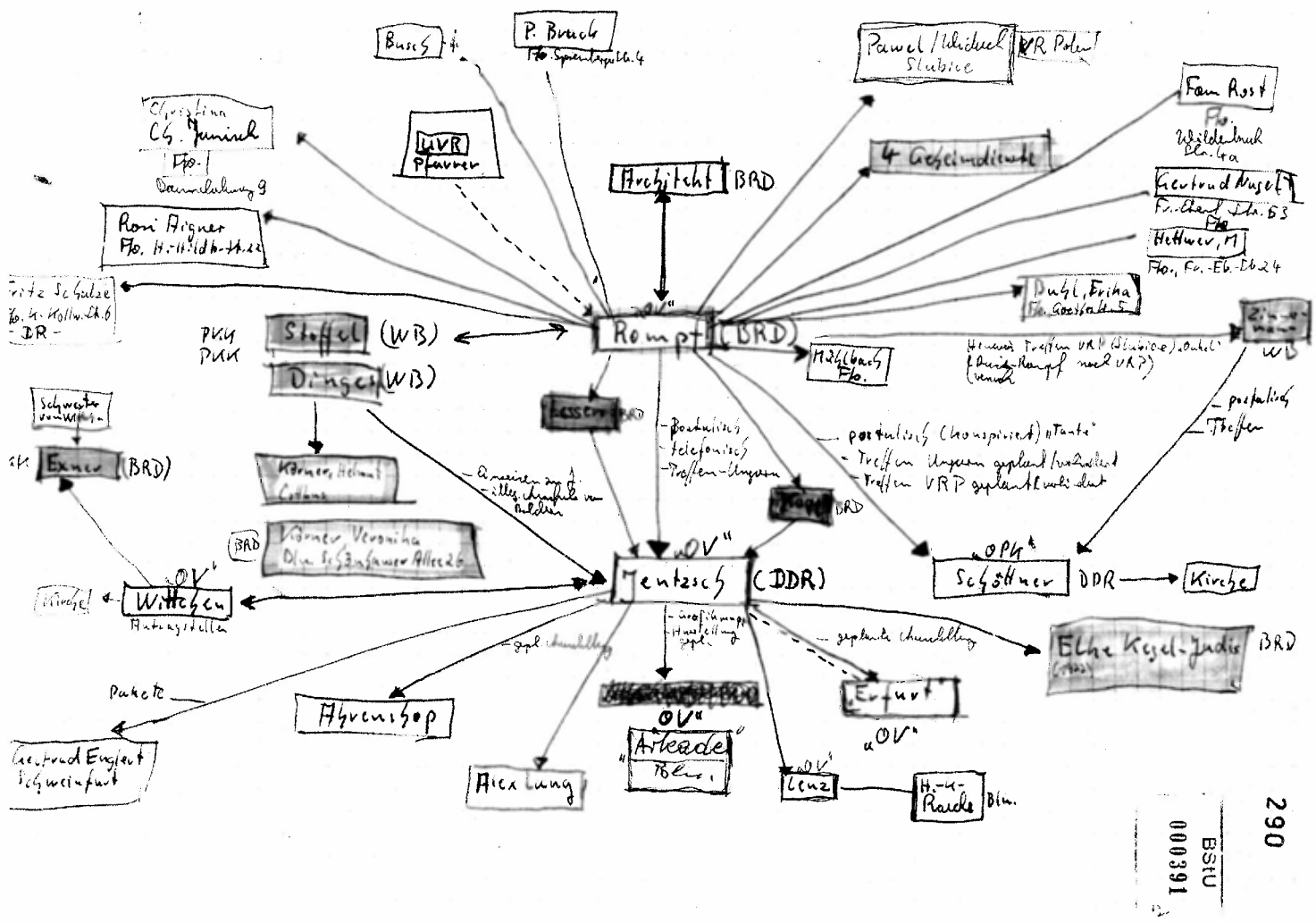
Předvídání zločinu

Ve známé povídce Philipa K. Dicka *Minority Report* z roku 1956 pracuje hlavní hrdina v policejním oddělení „precrime“, které má na starosti předpovídat zločiny a eliminovat potenciální pachatele. Dnes už nejde o fikci. Náhodná kontrola dokladů na ulici, statistiky trestných činů v lokalitě, ale i záznamy z dopravních kamer, jež umějí číst poznávací značky automobilů – to vše řádek po řádku plní policejní databáze. Prvním nápadem bylo zanášení těchto informací do jakýchsi map zločinu. Přidáme-li k nim údaje o čase, můžeme data z minulosti extrapolovat do budoucnosti – podobně jako se počítačově modeluje předpověď počasí – a získat místo a čas, kde s vyšší pravděpodobností dojde ke zločinu.

Podobné mapy sestavili už před lety například v Santa Cruz, Memphisu nebo Richmondu. V Chicagu, kde se sociální problémy přetavily do alarmujícího množství vražd, získala místní policie několikamilionový grant na projekt mapování zločinu, jenž se Dickově temné vizi blíží ještě víc. Inspirovala se výzkumy yaleského sociologa Andrewa Papachrista, který v akademické sféře analyzoval sociální sítě lidí zapletených do násilných zločinů a zjistil, že pravděpodobnost vaší vraždy je tím vyšší, čím blíže jste v sociální síti k někomu, kdo byl zavražděn (a stejně to funguje i s vrahy). Policie najala bývalého armádního analytika Milese Wernicka a začala zpracovávat data ze svých velice rozsáhlých databází. „Nákaza násilím“ je podle Papachrista podobná šíření nemocí – důležité je, s kým se stýkáme. Podstatné proto byly sociální vazby, jejichž zdrojem je například společné zatčení při drobnějším deliktu, ale i to, že vás s někým policisté potkali na ulici a tuto událost zaznamenali. Policie pak ze získané sociální sítě vyrobila „hotlist“ asi 420 lidí, kteří by mohli v budoucnu spáchat vraždu, nebo se naopak stát její obětí. Dotyčné začala obcházet. Zatím se prevencí myslí jenom domluva, případně nabídka služeb sociálních pracovníků, nikoliv zatýkání. A policejní velitel Jonathan Lewin doufá, že tento příklad ukáže celému světu, jak se to má dělat.

V České republice jsou zatím podobné výzkumy, zdá se, v počátcích a je to spíš sama veřejnost, kdo podobné analýzy vytváří. Velký kus práce na analýzách odvedl pečlivý průkopník datové žurnalistiky Jan Cibulka, který se svým týmem nejen sestavil mapu českého zločinu, ale srovnával jeho míru například s výskytem heren nebo laciných bytů. Problémem jsou nepřesná zdrojová data – statistiky se například evidují ke konkrétním policejním okrskům, nikoliv ke geografickým lokalitám, a je třeba je složitě přepočítávat na hustotu obyvatel, aby data vůbec dávala smysl. Online mapu kriminality aktualizuje sdružení ProPolice/Otevřená společnost, ostravskou „heatmapu“ nejnebezpečnějších míst sestavil výzkumný tým VŠB. Nejvíce na ní „svítí“ Stodolní ulice a její okolí.

Na jakých strategiích pracuje česká policie, která je považována za spíše byrokratickou a nepružnou, se můžeme jen domnívat. Jisté však je, že software typu Analyst's Notebook, který vykreslí mapu sociální sítě třeba z bankovních operací nebo výpisů od operátora, používají české tajné služby již dlouhá léta. Na druhé straně oceánu byly publikovány četné studie sociálních sítí ve vztahu k analýze terorismu, poháněné obrovskými finančními prostředky investovanými po 11. září. A právě nějaký druh „vazby“ k osobě podezřelé z terorismu je podle Williama Binneyho, odpadlíka NSA, důvodem k zařazení na seznam osob, jejichž komunikace se sleduje a analyzuje. Stačí, že jste si s někým vyměnili mail. Nebo s jeho známým.



Mapa sociální sítě, kterou z archivu Stasi získala novinářka Julia Angwin, zachycuje 64 vazeb sledované osoby na lidi, místa a schůzky. Foto BStU

Autor je dokumentarista.