Liben-Nowell a Kleinberg model šíření

David Liben-Nowell a Jon Kleinberg ve své studii *Tracing information flow on a global scale using internet chain-letter* data zkoumali šíření řetězových emailů. Konkrétně se jednalo o několik internetových petic, z nichž nejvýznamnější byla na protest proti válce v Iráku z let 2002 a 2003. Data byla sesbírána z veřejného internetového archivu emailů.

Za normálních okolností je téměř nemožné legálně vystopovat cestu řetězového emailu. Jelikož na rozdíl od sociálních sítí emailové zprávy jsou soukromé a příjemce emailu se musí sám dobrovolně rozhodnout, zda zprávu zveřejní. Petice je však unikátní v tom, že stačí relativně malé procento příjemců, které zprávu zveřejní pro rekonstrukci šíření. Toto je dáno tím, že lidé, kteří zprávu nezveřejní v celé své podobě, ale chtějí petici podpořit, připojí své jméno jako podpis k emailu a přepošlou jej dál. Jejich jméno je pak zveřejněno někým jiným.

Model popisuje reakci příjemců na zaslaný email. Příjemce s určitou pravděpodobností na email reaguje anebo jej ignoruje. Pokud se rozhodne reagovat, tak nejčastěji na něj pouze odpoví, ale nepřepošle ho dál. Pokud jej přepošle dál, pak existuje šance, že ho zveřejní. Unikátní vlastností tohoto modelu je také fakt, že reakce na zprávu probíhá v reálném čase, ne v okamžiku přijetí. V reálném světě komunikace také neprobíhá okamžitě, ale trvá určitý čas, než příjemce zareaguje.

Tento model se odlišuje od ostatních modelů tím, že je odvozen od šíření konkrétních emailů. Velmi důležitým faktem, který zdůrazňují i sami autoři studie je, že výsledná struktura propagace petice velmi připomíná strom, což je překvapivé vzhledem k vysoké rozvětvenosti a všeobecné struktuře emailové sítě.

Struktura šíření ostatních modelů je často odvozená od šíření informací po sociálních sítích anebo dokonce od epidemiologických modelů, které jsou mnohem lépe dokumentovatelné díky jejich veřejné povaze. Žádná struktura však ani vzdáleně nepřipomíná strom. Rozdílné jsou však i struktury sítí. Sociální sítě jsou typicky *scale-free*, ale u emailové sítě je toto tvrzení komplikovanější.

Vyvstává tedy otázka, proč je tato studie tolik odlišná od ostatních. Nabízí se několik vysvětlení. Ostatní modely zkrátka neberou v potaz strukturu emailové sítě, která je poměrně unikátní oproti jiným sítím v reálném světě. Odlišnost je způsobená realistickým časem odpovědi. Model může být specifický pro šíření petic, ne však všech emailů. Anebo je model nepřesný či dokonce chybný.

Poslední tvrzení může znít velmi nepravděpodobně u takto hojně citované studie. Nicméně v takto málo neprozkoumané oblasti je nutné myslet i na tuto možnost. Toto tvrzení podporuje i fakt, že výsledky studií o šíření emailů jsou často velmi rozdílné a někdy si i částečně odporují.

Implementace modelu

David Liben-Nowell a Jon Kleinberg vychází ve své implementaci pouze ze jmen, která byla obsažena v seznamu příjemců petice. Autoři našli 1123 kopií petice, z nichž vyextrahovali podpisy a odvodili potenciální propojení příjemců. Každé jméno reprezentuje uzel v grafu. Algoritmus k nim doplňuje hrany a s určitou pravděpodobností je označuje jako viditelné. Z výsledné struktury je pak zpětně rekonstruován viditelný graf. Tento viditelný graf je často stromem.

Tento přístup má tu výhodu, že se zakládá na 100 % reálných datech. Nevýhodou je, že tato data nejsou úplná, z principu nelze zachytit všechny instance přeposlání zprávy a také nepopisují přesně strukturu emailové sítě.

Z výše uvedených důvodů jsem při implementaci tohoto modelu zvolil lehce odlišný přístup. Ponechal jsem všechny základní parametry. Algoritmus simuluje běh reálného času a rychlosti odpovědi se náhodně generují pomocí distribuční funkce. Operuji se třemi pravděpodobnostmi jejichž hodnoty jsou nastaveny v podobném rozmezí jako v originální studii. Jedná se o šance na reakci na zprávu (65 %), na odpověď odesílateli (90–95 %) a zveřejnění (20–25 %). Jako výchozí data jsem však vzal grafy reálných emailových sítí. Výchozí uzel pošle všem svým sousedům zprávu a ti na ni odpovídají pomocí výše uvedených principů, toto se opakuje, dokud zpráva nedosáhne svého maximálního potenciálu a všichni příjemci na ni nějakým způsobem zareagovali anebo ji ignorovali. Po skončení běhu se také zpětně rekonstruuje viditelný graf.

Díky tomuto přístupu je možné sledovat vliv struktury sítě na šíření. Nevýhodou je, že každá síť může být definovaná jinak a nebere v potaz vztahy mezi příjemcem a odesílatelem. Grafy typicky nemají žádné váhy hran, pokud dva lidé pošlou jednu zprávu, v grafu se vytvoří hrana. To samé nastane, pokud si dva kamarádi posílají zprávy pravidelně. Dá se předpokládat, že reálná pravděpodobnost reakce se liší i podle různých demografických indikátorů a osobních vztahů příjemce a odesílatele.

Galton – Watson model

Pro model šíření od Davida Liben-Nowella a Jona Kleinberga (dále jen LNK model) jsou typické hluboké stromy s nízkou rozvětveností, které tento algoritmus generuje. Tuto strukturu lze replikovat Galton – Watsonovým modelem. Jedná se o překvapivě velice jednoduchý postup pro generování stromů, který byl publikován již v roce 1875 a jeho původní účel nemá nic společného s emaily ani šířením zpráv.

Algoritmus generuje stromy pomocí pravděpodobnostního rozdělení počtu dětí. Začne graf s jedním uzlem, který označí jako kořen stromu a náhodně vygeneruje počet dětí, podle předem dané distribuce. Mezi rodiče a potomka vloží hranu a proces opakuje pro každý další vložený uzel. Proces končí po dosažení předem daného počtu generací anebo pokud všechny uzly aktuální generace mají počet potomků roven 0.

Pro původní LNK model, kde *k* se rovná počtu dětí je distribuce následovná.

Obsah obrázku text, snímek obrazovky, řada/pruh, Písmo

Popis byl vytvořen automaticky

Tento model asi nelze považovat za plnohodnotnou simulaci šíření řetězových emailů, nabízí však zajímavou perspektivu. Lze s ním operovat v porovnávání výsledků ostatních modelů díky nezávislosti na struktuře sítě a dalších reálných datech.

Opakované běhy s výše uvedenou distribucí, také dobře vizualizují skutečnost, že rozšíření e-mailu velkému množství příjemců není nemožné, ale je velmi málo pravděpodobné. Většina běhů skončí s malým počtem uzlů, před dosažením limitu počtu generací.

Struktura sítí

Teorie sítí je velmi komplexní disciplínou a pro potřeby této práce je nutné definovat několik pojmů a uvést do kontextu vlastnosti a rozdíly různých sítí, převážně sítě emailové a sociální.

Centrální uzel

Jako centrální uzel, anglicky *hub node*, se označuje takový uzel, který má vysoké množství sousedů*.* Počet sousedů, který definuje toto označení nemusí být pevně definovaný. Pro potřeby této práce a implementaci algoritmů jsem jej nastavil na 50. Centrální uzly jsou zásadní pro generování emailů s velikým dosahem. Pokud se zároveň jedná výchozí bod šíření, dramaticky se pak navyšuje počet příjemců pro všechny modely.

Bezškálová síť

Bezškálová síť, anglicky *scale-free network*, se vyznačuje vysokým počtem uzlů, které mají malé množství navazujících hran, typicky 1. Zároveň se většina hran koncentruje okolo několika málo centrálních uzlů. Rozložení stupňů uzlů je tedy nerovnoměrné a přibližně se odvíjí od Paretova rozdělení.

Struktura emailové versus sociální sítě