1 Konstrukce sítí z vektorových dat

Metody konstrukce sítí pracují se sítěmi, kde vrchol reprezentuje určitá data, které jsou propojeny hranou. Dělí se na:

- Nezávislé na úloze - nepotřebují označená data
- Závislé na úloze - používají jak označená tak neoznačená data pro vytváření sítí

Sítě vytváříme z vektorových dat hlavně kvůli vizualizaci. Je třeba definovat funkci podobnosti, která nám umožní určit jak jsou si data podovná a umíme určit matici podobnosti. Tato podobnost řídí to, jestli je mezi daty hrana. Funkce: Gaussův kernel, kosínova podobnost, korelace, euklidovská vzdálenost. Gaussův kernel: $K(x, y) = exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$.

Pokud v grafu není hrana tak tam není, v sítí si nejsme jistí.

Vlastnosti sítí, realného světa: Small world (proměrná nejkratší cesta je $<= \log N$), bezškálovost (distribuce stupňů podle mocnineho dělení), komunity. Metody transformace vektorových dat na sítě (složitost obecně alespoň $> O(n^2)$):

- $\epsilon - radius$ - hrana je mezi vrcholy jestliže jejich vzdálenost je $< \epsilon$
- $k - NN$ - každý vrchol má k nejbližších sousedů
- kombinace obou - jestliže je v ϵ okolí více než k vrcholů vem je, jinak $k - NN$
- b-matching - oproti $k - NN$ zaručuje, že každý vrchol má stejný počet hran, vyvážené sítě, nemá vlastnosti realného světa
- NNN - přidává shlukování
- LRNet - vytváří vážené grafy a lokální reprezentativitu (zaleží na počtu sousedů a počtu vrcholů pro které ja daný vrchol nejbližším sousedem)

2 Vzorkování sítí (sampling)

Proces, při kterém se analyzuje část dat, určitý vzorek sítě s cílem získat relevantní informace o celé sítí, důvody:

- Nemáme přístup k celé sítí
- Síť je moc velká
- Síť se nestále mění
- Menší síť se dá lépe vizualizovat

Musíme zjístit jaká velikost vzorku je vhodná a reprezentativní. Vzorky se dají brát jako zmenšené sítě, sítě v minulosti (ty co se mění), lokální podragfy. Jak získat vzorek:

- Plný přístup ke grafu - náhodný výběr vrcholů nebo hran
- Omezený přístup ke grafu - umožněno jen procházení vrcholů postupně, nevleze se do paměti, metody založené na procházení

- Stream dat - vhodné pro dynamické sítě

Sampling obecně generuje 3 vzory:

- Řídké grafy (náhodný výběr)
- Relativně kompaktní grafy (seed-based sampling)
- Velké grafy s malým stupněm, např. $d = 1$

2.1 Metody založené na náhodném výběru

Random Node Sampling - Vyber vrchol a jeho sousedy s pravděpodobností p (sousedy s p?), chceme li n' hran: $p = \frac{n'}{n+2m}$

Random Edge Sampling - Vyber hranu s uniformní pravděpodobností p, metoda nemění relativní četnost hran

2.2 Metody založené na náhodném prohledávání grafu

Snowball Sampling - Pro počáteční vrchol (seed) a vzdálenost l, vyber všechny vrcholy a jejich sousedy ve vzdálenosti l od seedu. Dobře popisuje strukturu okolí seedu. Ovlivněno stupněm seedu.

Random Walk - náhodná procházka po vrcholech. V každém krokue je pravděpodobnost návrata do počátku, opakování do požadované velikosti. Výsledný vzorek je souvislý.

Random Jump - jako RW ale s pravděpodobností skočí zcela jinam

Forest Fire - kombinace snowball a RW

2.3 Zhodnocení samplingu

Porovnání vlastností originální sítě a vzorků. Kolmogorovův-Smirnovův test, testuje zda 2 náhodné veličiny pocházejí ze stejného rozdělení pravděpodobnosti, případně zda náhodná proměnná má předpokládané teoretické rozdělení. Nulová hypotéza říká, že dva výběry odpovídají stejnému rozdělení.

3 Shlukování

Používá se matice podobnosti, hodnoty jsou kladné, matice symetrická. Matice stupně, má na diagonále stupeň vrcholů jinak všude 0. Normalizovaná matice sousednosti, hodnoty řádku jsou vyděleny stupněm vrcholu řádku. Vlastní číslo matice A je číslo λ, lteré splňuje $|A - \lambda I| = 0$ vlastní vektory $(A - \lambda I)v = 0$, I je identity matice.

3.1 Dělení grafu

"Rozříznutí"grafu, které optimalizuje určitou funkci. Vrcholy ve stejné komponentě, by si měly býti podobné, vrcholy z různých komponent by měli být odlišné. Váha řezu je suma váh hran, které vedou mezi komponentami. *Volume* komponenty je suma váh hran, které začínají nebo končí v komponentě.

Ratio Cut se snaží minimalizovat sumu podobnosti komponenty k ostatním vrcholům mimo komponentu. Normalized Cut je podobný Ratio ale dělí váhu komponent jejich objemem.

Spektrální shlukovací algoritmus, používá vlastní čísla a vektory.

Kernighan-Lin

- Rozdělíme vrcholy do 2 skupin určené velikosti, např. náhodně

- Pro každou dvojicí vrcholů z různých skupin vypočítáme jak moc by se změnilo *cut size*, kdybychom je přehodili (cut size = počet hran mezi skupinami)
- dvojici, která nejvíce zmenší *cut size* přehodíme
- Opakujeme, ale vyměněné vrcholy už nesmíeme prohazovat
- Když už jsme přehodili všechny dvojice, vybereme tu iteraci, kde je nejméně hran mezi skupinami

Velikost obou skupin zůstává stejná, minimalizujeme počet hran mezi nimi. Algortimus je dobré spustit vícekrát.

4 Detekce komunit

Komunita je hustě propojený podgraf v síti. Komunita je skupina, kde každá zná každého. **Silná komunita** je taková, kde každý vrchol komunity ma více hran s ostatními členy komunity než s vnějškem. **Slabá komunita**, celkový vnitřní stupeň je větší než celkový vnejší stupeň. Komunity se mohou překrývat.

Sociální sítě mají přirozenou komunitní struktůru. Chceme zjistit zda síť má komunitní struktůru, velikosti komunit, počet komunit a kam patří daný vrchol. Detekce komunit je snadná jen v řídkých grafech. Metody hledání komunit jsou např. spektrální metod, používající kliky (Clique Percolation Method), metody založené na heuristikách (modularita). Obecně shora dolů, zdola nahoru, nepřekrývající a překrývající se komunity.

Metody zdola nahoru se zaměřína jednotlivce a zkoumají jak je zakotven v překrávajících se skupinách. Shora dolu hledají slabá místa, např. hrany, které by mohli odstranit.

Soudržnost komunity ovlivňuje vzájemné propojení, kompaktnost (malá vzdálenost v komunitě), hustota a oddělení od zbytku sítě.

Shlukování na základě podobnosti:

- každý vrchol je komunitou
- spoj 2 nejvíce podobné komunity
- přepočítej podobnost (single,complete,average)
- opakuj dokud není 1 komunita

4.1 Kliky

Klika grafu je takový podgraf nějakého (neorient.) grafu, který je úplným grafem. Kliky se mohou překrývat a výskyt kliky v grafu reprezentuje velkou soudržnost nějaké skupiny. Maximální klika je klika, kterou nelze rozšířit o další sousední vrchol. Největší klika je klika největší možné velikosti v daném grafu Klikovost grafu je velikost největší kliky.

N-Klika - je možno je definovat vrcholy jako členy kliky, pokud jsou připojeny ke každému jinému členu skupiny, ve vzdálenosti větší než jedna, obvykle se používá vzdálenost N=2.

K-plex – je maximální podmnožina množiny n vrcholů taková, že každý její vrchol je incidentní s alespoň $n - k$ vrcholy. 1-plex je klasická klika.

K-core - je maximální podmnožina vrcholů takových, že každý vrchol je incidentní s alespoň k vrcholy této podmnožiny, tj. každý vrchol má stupeň alespoň k. Nalezení: odstraňujeme všechny

vrcholy se stupněm $< k$, toto se opakuje dokud takové vrcholy existují. Nakonec zůstane množina K-core.

Metody detekce komunit

- **Clique Percolation Method** - 2 k-kliky jsou přilehlé, jestliže sdílejí $k-1$ vrcholů. Komunita je maximální spojení k-klik, které jsou propojené přilehlými k-klikami
  1. Odstraní se všechny úplné podgrafy (kliky), které nejsousoučástí větších klik
  2. vytvoří se matice překryvu klik
  3. naleznou se k-klik komunity
- **Girvan-Newman betweenness clustering**
  1. Spočítá se edge betweenness všech hran
  2. Odstraní se hrana s největší edge betweenness
  3. Opakuje se dokud nedosáhneme určené modularity, nebo nějaká hrana má betweenness větší než určitý threshold
- **Louvain** - Heuristická metoda chamtivé optimalizace
  1. Každý vrchol je komunita
  2. Pro každý vrchol vyzkoušej jej přesunout asi do jiné komunity a přepočti modularitu
  3. Přesuň vrchol tak ať je maximalizovana modularita
  4. Opakuj dokud se zlepšuje modularita
  5. Sjednoť vrcholy do komunit, váha hran se sjednotí
  6. Opakuj dokud nedosáhneme maximální modularitu
- Zaine - komunita je určena podle fitness funkce, která určuje ostrost R hranice komunity. Hranice je ostrá pokud má méně spojení hranicí a zbytkem grafu, než je počet spojení z hranice do komunity.
  1. Algoritmus začíná od jednoho vrcholu, takže na začátku hranice B a komunita D obsahuje pouze první vrchol a plášť S obsahuje všechny jeho sousedy.
  2. V každém kroku algoritmus spočítá ostrosti R pro každý vrchol v S.
  3. Poté se vrchol s nejvyšší hodnotou R (v případě, že nová R je vyšší než současná R) přidá ke komunitě a tři množiny vrcholů (D, B, S) jsou aktualizovány
  4. Algoritmus opakuje tento proces, dokud se R zvyšuje

Modularita Míra kvality nalezených komunit, vetší modularita je lepší, M=0, jedna komunita=celá síť. Pozitivní modularita znamená, že komunita má více hran než je očekáváno v rámci sítě, je silně propojená.

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

5 Modely sítí

5.1 Komunitní modely sítí

Triádový uzávěr indikuje preferenční připojování.

Holme-Kim model rozšíření BA modelu. S pravděpodobností p je provedena formace triády namísto PA. Formace triády znamená, že je vytvořena hrana k náhodnému sousedovi naposledny spojeného vrcholu.

Bianconi model rozšíření BA modelu. První hrana nového vrcholu je napojena na náhodný vrchol sítě u,(náhoda záleží na stáří vrcholu), druhá hrana je s pravděpodobností p napojena na náhodného souseda u, jinak k náhodnému vrcholu sítě. Další hrany stejně jako druhá hrana.

Komunitní modely lze rozšířit o fitness, které hraje roli při preferenčním připojování.

Link selection model - v každém kroku přidáme 1 vrchol a vybereme náhodnou hranu sítě, ze které vybereme 1 konec a na ten napojíme nový vrchol. Větší stupeň vrcholu, znamená větší šanci, že se na něj někdo napojí.

Copying model nový vrchol v se s pravděpodobností p připojí na náhodný vrchol sítě u a s pravděpodobností $1-p$ se připojí na náhodného souseda u, kopíruje hranu od u do tohoto souseda.

Barabasi-Albertův model se dá rozšířit o stárnutí vrcholů, vnitří hrany, mazání hran.

5.2 Temporální sítě

Hrany jsou aktivni v nebo od určitého časového okamžiku. Můžeme pracovat s časovými razítky, intervaly, frekvencí. Např. tranzitivita v temporálních sítích nemusí platit. Ve statických sítích platí, že pokud vrchol A je sousedem B a B je sousedem C, pak A a C jsou dostupné přes vrchol B. V temporálních sítích, v případě, že hrana (A, B) je aktivní pouze v pozdější době, než hrana (B, C), potom A a C jsou nedostupné. Síťě v různých časových okamžicích se dají považovat za vícevrstvé sítě. Reprezentace nejčastě jako trojice, dvojice hrany a časové razítko.

- Time-respecting (TR) path - cesta s neklesajícím časem
- Oblast vlivu vrcholu i - vrcholy dosažitelné po TR cestách z vrcholu i
- Reachability ratio - průměrný počet vrcholů, které jsou součástí oblastí vlivů všech vrcholů v síti
- Source set i - vrcholy, ze kterých je dostupný vektor i po TR cestách
- Trvání nejkratší cesty - délka TR cesty mezi i měřená časem
- Latence - nejkratší TR cesta (nejkratší čas) z i do j
- Vzdálenost - nejkratší TR cesta (nejmenší počet hran) z i do j

5.3 Vícevrtsvé sítě

Skládájí se z více vrstev, kde vrstva je síť. Aktor je jednotlivec, reprezentovaný vrcholem ve vrstvě. Aktor může být ve více vrstvách. Vrchol je teda reprezentace aktora. Vrstvy mohou být propojeny mezivrstvovýma hranama. Dalo by se reprezentovat i heterogeníma vazbama. Vícevrstvá síť se dá převést na jednovrstvou pomocí projekce, sploštění, zaleží jestli budeme sčítat váhu hran.

Sploštění sjednotí všechny vrcholy aktéra do 1, hrana mezi aktéry je tehdy pokud byla v nějaké vrstvě. Dá se počítat s váhami vrstev a hran.

Degree centralita je počet hran daného aktora v množině vrstev a mezi nimi. Sousedé aktora v množině vrstev je množina aktorů, se kterými je v rámci těchto vrstev náš aktor spojen. Neighborhood centralita je jejich počet. Redundance připojení je 1-poměr neighborhood a degree. p.s.$xN(a, L) = |N(a, L) \setminus N(a, \mathbf{L} \setminus L)|$.

Vzdálenost mezi aktéry je velmi složitá záležitost. Cesta se zapisuje do matice L hodnota L_{ij} značí délku cesty z vrstvy i do vrstvy j. Můžeme pozorovat více cest, kde se budou lišit délky v různých vrstvách, takže záleží jak jsou jednotlivé vrstvy důležité.

Closenesss a Betweenness centralita je definována přes náhodnou procházku.

Relevance aktéra v množině vrstev L je definována jako poměr neighborhood centrality aktéra v L, ku neighborhood centralitě aktéra ve všech vrstvách. Jak je aktér relevantní ve zvolených vrstvách.

6 Procesy v sítích

Netwrok resillience – porozumění šíření poruch, chceme pochopit roli sítí, kterou hrají při zajišťování robustnosti složitého systému. Ukázat, že struktura sítí hraje zásadní roli ve schopnosti systému přežít náhodná selhání nebo záměrné útoky. Zkoumání role sítí při vzniku kaskádových poruch jako zničujících jevů, které se často vyskytují v reálných systémech.

Inverze Perkolace - Náhodné selhání (každý vrchol/hrana odstraněna s pravděpodobností), útok na síť (cílené odebírání nejdůležitějších vrcholů).

Náhodné poruchy - bezškálové sítě se nerozpadají po náhodném odstranění nějaké konečné podmnožiny vrcholů. Náhodně musíme odstranit téměř všechny uzly, aby došlo k rozpadu největší komponenty. Tj. k rozpadu dochází velmi pomalu. V bezškálové síti máme mnohem více vrcholů s malým stupněm než hubů (center). Proto náhodné odstranění vrcholu vede k tomu, že se odstraňují převážně vrcholy s malým stupněm. Vrcholy s malým stupněm přispívají k robustnosti sítě.

Je nepravděpodobné, že odstraněním jediného centra se síť rozpadne, protože ji mohou zbývající centra ještě držet pohromadě. Nicméně po odstranění několika málo center se síť rozpadne do malých shluků.

Modely epidemie, SI, SIS, SIR.

7 Korelace v sítích

Sociální sítě – lidé mají tendenci se sdružovat se sobě podobnými lidmi, citační sítě – citují se články ze stejné oblasti. assotrative mixing (homophily) – výběrové slučování (nebo také homofilní (homotypické) vazby). Opakem je disassorative mixing (heterofilní (heterotypické) vazby) – Internet

Diskrétní (kategoriální) atributy (pohlaví, národost, rasa), skalárání (věk, plat).

Síť je assortative (výběrová, nenáhodná), pokud významná část hran spojuje vrcholy stejného typu, tj. např. se stejnou hodnotou atributu, s obdobným stupněm, apod.

Modularita a určuje míru propojenosti podobných vrcholů (vrcholů stejného typu). Už jsme viděli u komunit. Pozotivní hodnoty, jestliže je více hran mezi vrcholy stejného typu.