UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Analýza sieti sesterských miest

Projekt z predmetu Veda o sieťach

Vypracovali:

Rudolf Nosek: Získanie a spracovanie dát

Daryna Kalinchuk Yuriivna: Programovanie výpočtu štatistík

Adam Barčák: Programovanie výpočtu štatistík

Lukáš Macko: Napísanie záverečného reportu a tvorba prezentácie

Úvod

Sesterské mestá predstavujú akúsi formu právnej alebo spoločenskej dohody medzi dvoma odlišnými mestami. Tieto mestá majú väčšinou prepojenie v historickom vývoji, podobné demografické charakteristiky alebo zameranie cestovného ruchu. Výhodou sesterských miest je ich vzájomné utužovanie kultúrnych a obchodných väzieb, ale aj podpora cestovného ruchu. V tomto projekte sa pozrieme na podrobnejšiu analýzu siete týchto sesterských miest v rámci celého sveta. Analýzu vykonáme pomocou výpočtov rôznych štatistických ukazovateľov a grafickým znázornením siete. Pre lepšiu interpretáciu výsledkov, porovnáme niektoré vypočítané hodnoty aj s inými známymi sieťami. Výpočty a vykresľovanie grafov realizujeme prostredníctvom programovacieho jazyka Python a softvéru R. Pri volení a výpočtoch štatistík sme vychádzali z prednášok predmetu Veda o sieťach.

Predstavenie a grafická vizualizácia dát

Dáta potrebné pre analýzu siete sesterských miest sme získali pomocou knižníc `request` a `BeautifulSoup` programovacieho jazyka Python z webových stránok Wikipédie jednotlivých miest. Štruktúra dát je nasledovná: v riadkoch sú uvedené jednotlivé názvy miest a k nim sú v stĺpcoch premenné, ktoré prezentujeme v Tabuľke 1. Hlavičku dát je možné nájsť v priloženom programovom kóde.

Tabuľka 1: Štruktúra dát spolu s vysvetlivkami premenných

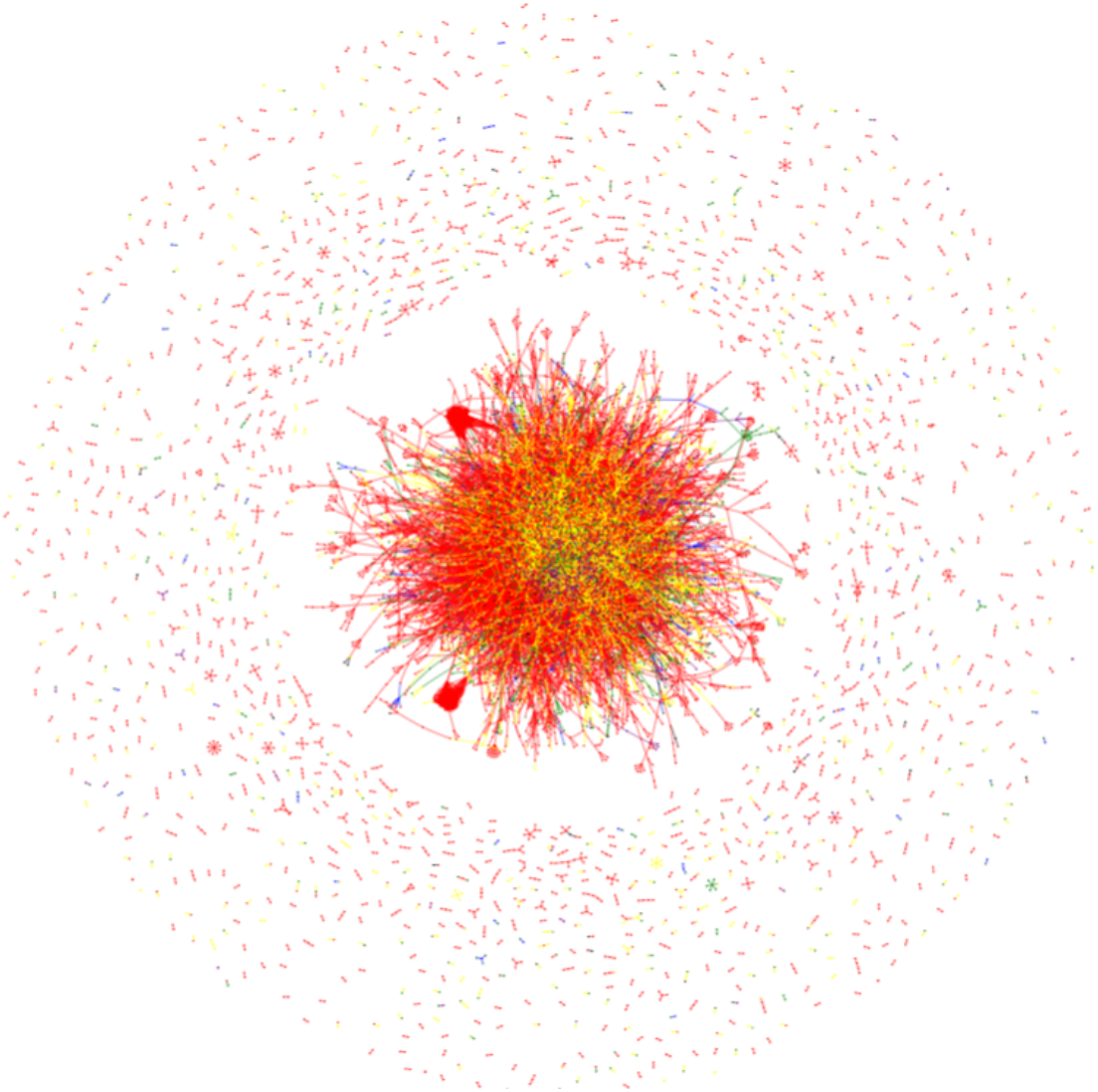
Názov premennej	stručná vysvetlivka
`sister_city`	názov sesterského mesta
`country_of_sc`	názov krajiny sesterského mesta
`country`	názov krajiny skúmaného mesta
`continent`	kontinent skúmaného mesta
`continent_of_sc`	kontinent sesterského mesta

Počet skúmaných miest v jednotlivých kontinentoch uvádzame v Tabuľke 2.

Tabuľka 2: Počet skúmaných miest pre jednotlivé kontinenty

Kontinent	Európa	Ázia	Afrika	Severná Amerika	Oceánia	Južná Amerika	Spolu
Počet miest	8 752	1 700	433	1 805	213	362	13 265

Predtým než vypočítame jednotlivé štatistiky, vykreslíme graf reprezentujúci túto sieť pomocou knižníc `pyvis` a `networkx`. V tomto grafe reprezentujú vrcholy jednotlivé mestá a hrany medzi jednotlivými mestami symbolizujú, že mestá sú sesterské. Tento graf je prirodzene neorientovaný, keďže ak je napríklad Bratislava sesterské mesto s Viedňou, tak aj Viedeň je sesterské mesto s Bratislavou.

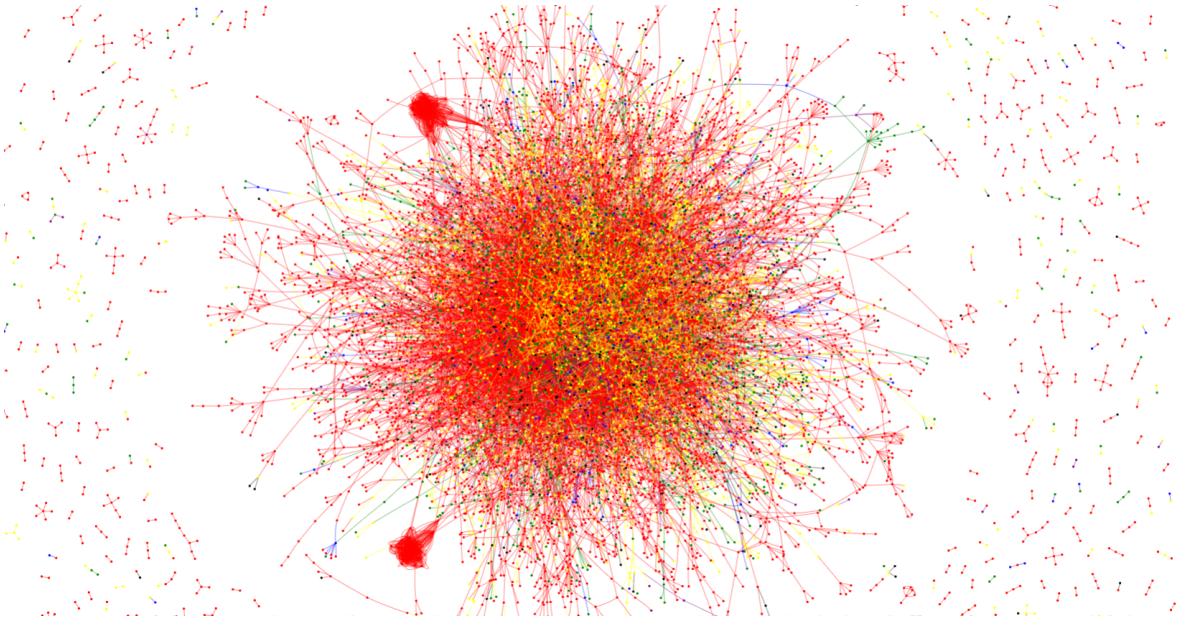


Obr. 1: Grafická vizualizácia siete sesterských miest

V tomto grafe sme farebne rozlíšili mestá, podľa toho z akého kontinentu pochádzajú.

Mestá nachádzajúce sa v Afrike sú znázornené čiernou farbou, mestá z Ázie zelenou farbou, mestá z Európy červenou farbou, severo-americké mestá sú vyznačené žltou farbou, juho-americké mestá fialovou farbou a mestá nachádzajúce sa v Oceánií Modrou farbou.

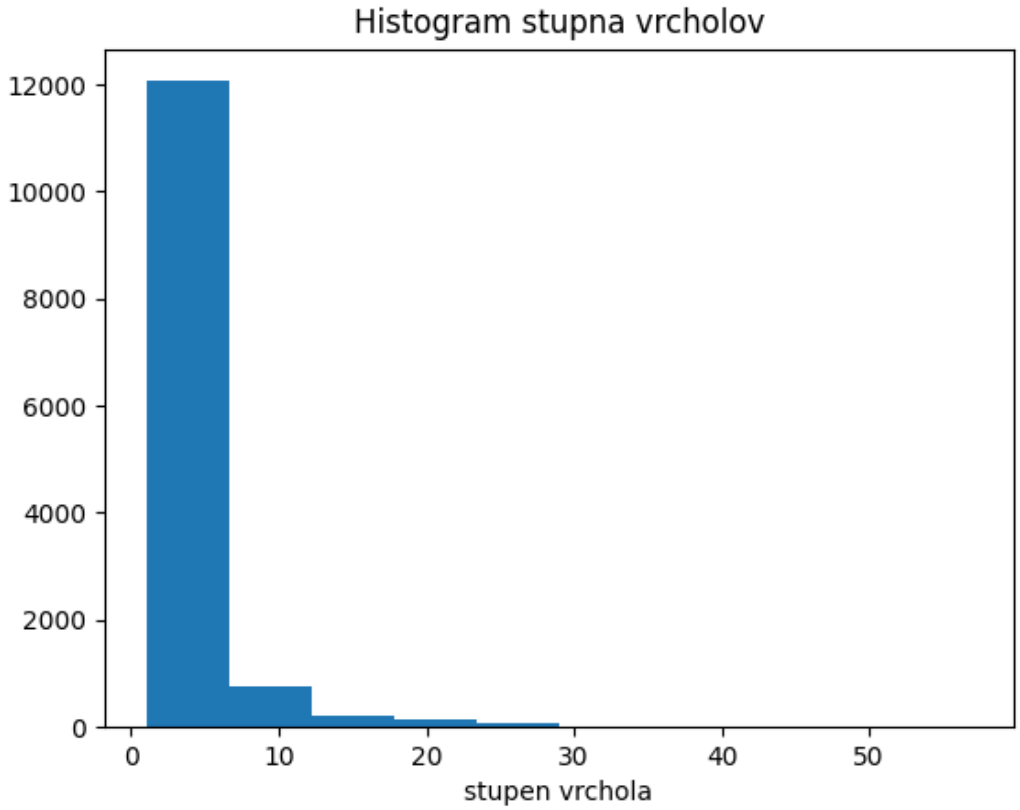
Z Obr. 1 môžeme pozorovať, že graf má jeden veľký komponent (stred grafu s množstvom hrán). Nachádzajú sa tu hlavne mestá z Európy a Severnej Ameriky, ktoré sú navzájom sesterské, a z malej časti mestá z ostatných kontinentoch. Môžeme preto predpokladať, že tento komponent tvorí samostatnú veľkú sieť sesterských miest. Je možné si taktiež všimnúť dva červené uzly na dolnom a hornom okraji veľkého komponentu. Keďže sú vyznačené červenou farbou ide o európske sesterské mestá. Konkrétne sa jedná o organizáciu, ktorá združuje mestá Európskej únie s názvom Douzelage a o Charta Európskych vidieckych spoločenstiev. Na okrajoch grafu sú mestá, ktoré majú málo sesterských miest a tvoria veľmi malé komponenty grafu. V istých prípadoch ide len o dve mestá, ktoré sú navzájom sesterské, pričom ani jedno z týchto miest nemá iné sesterské mesto. Pre lepší náhľad do veľkého komponentu grafu, uvádzame priblíženú verziu grafu.



Obr. 2: Priblíženie veľkého komponentu z grafickej vizualizácie siete

Štatistiky súvisiace so stupňami vrcholov

Výpočet štatistík, ktoré súvisia stupňami vrcholov sú dôležité pri charakterizácií siete. Stupeň vrchola v grafe reprezentuje počet hrán vychádzajúcich z vrchola. V našom prípade to môžme interpretovať ako počet sesterských miest pre dané mesto (vrchol). Najvyšší stupeň vrchola dosiahlo z pozorovaných miest mesto Rio de Janeiro, konkrétne so stupňom vrchola 57. Znamená to teda, že mesto Rio de Janeiro má 57 sesterských miest a tento počet je najväčší spomedzi všetkých pozorovaných miest. Pre porovnanie s ostatnými mestami uvádzame histogram stupňa vrcholov.



Obr. 3: Histogram stupňa vrcholov pre sieť sesterských miest

Ako ďalšie sme spočítali priemerný stupeň vrchola tejto siete. Táto štatistika z časti hovorí o tom ako dobre je sieť prepojená. Hodnotu priemerného stupňa vrchola je možné spočítať ako dvojnásobok počtu hrán vydelený celkovým počtom vrcholov. V našom prípade táto hodnota vyšla 2,78, čo značí, že v priemere ma každé mesto približne 3 sesterské mestá po zaokrúhlení nahor.

Posledná štatistika súvisiaca so stupňom vrchola, ktorú sme zrátali je hustota.

Hustotu v sieti možno interpretovať ako proporcia existujúcich hrán vzhľadom nato koľko ich reálne môže existovať v sieti, zo čoho vyplýva, že hodnota hustoty je z intervalu $\langle 0, 1\rangle$. Prirodzene, čím je táto hodnota bližšia 0, tým sú vrcholy v grafe menej pospájané. Hustotu sme počítali ako priemerný stupeň vrchola siete vydelený počtom vrcholov v sieti -1. Táto hodnota vyšla veľmi malá, konkrétne 0,000209. Dá sa teda tvrdiť, že sieť sesterských miest nie je veľmi prepojená, čo zapríčiňuje veľké množstvo komponentov súvislosti, ktorým sa venujeme v ďalšej časti projektu. Ná výpočet štatistík v tejto časti projektu sme používali knižnicu `networkx` v Pythone.

Štatistiky súvisiace s komponentami siete

Na základe grafického znázornenia siete, ktoré reprezentuje Obr. 1 sme skonštatovali, že graf obsahuje niekoľko komponentov. V tejto časti sa pozrieme bližšie na tieto komponenty, čo nám umožní lepšie charakterizovať túto sieť. Ako prvé sme vyrátali celkový počet komponentov súvislosti v tejto sieti. Táto hodnota vyšla celkom vysoká, konkrétne 1 400. Je to spôsobené najmä tým, že táto sieť obsahuje veľa miest, ktoré majú len jedno resp. dve sesterské mestá a zároveň žiadne iné.

Podľa prednášok z predmetu Veda o sieťach v neorientovanom grafe existuje obrovský komponent, ktorý zaberie veľké percento vrcholov. V sieti sesterských miest je jeden spomínaný obrovský komponent súvislosti (stred Obr. 1), ktorý má veľkosť 9 725, čiže tento komponent tvorí až 9 725 miest z celkového počtu pozorovaných miest 13 265. Na základe týchto hodnôt sme sa pozreli aký podiel tvorí tento komponent vzhľadom na celú sieť, teda vydelením týchto dvoch čísel sme dostali hodnotu 0,7332. Interpretovať to môžeme tak, že obrovský komponent zahŕňa 73,32 % miest z celej sieti sesterských miest. Pre lepšie interpretovanie tejto hodnoty vykonáme jej porovnanie s vybranými inými známymi reálnymi sieťami. Údaje sme čerpali z prednášok predmetu Veda o sieťach. Napríklad v sieti spoluautorstva v oblasti matematiky je podiel obrovského komponentu 0,822 a v sieti proteínových interakcií 0,689, kým v sieti internetu je táto hodnota 1. Z vymenovaných sietí sa z hľadiska podielu obrovského komponentu najviac podobá sieť sesterských miest na sieť proteínových interakcií. Aby sme lepšie videli z akých kontinentov pochádzajú mestá v tomto obrovskom komponente vyrátali sme ich počty. Tieto počty pre jednotlivé kontinenty prezentujeme v Tabuľke 3 spolu

s pomermi vzhľadom na celú sieť.

Tabuľka 3: Počet miest v obrovskom komponente spolu s pomerom vzhľadom na celú sieť

Kontinent	Európa	Ázia	Afrika	Severná Amerika	Oceánia	Južná Amerika
Počet miest	6 228	1 408	388	1 237	156	308
Pomer	0,7116	0,8282	0,8960	0,6853	0,7324	0,8508

Na základe Tabuľky 3 sa dá tvrdiť, že v obrovskom komponente, čo sa týka počtu, nachádza najviac miest z Európy. Pokiaľ sa pozrieme na pomer jednotlivých miest z kontinentov vzhľadom na celú sieť, je možné konštatovať, že 89,60 % percent miest z Afriky sa nachádza v obrovskom komponente, čo je najvyššia hodnota spomedzi všetkých. Mimo tohto komponentu je teda minimum miest z Afriky.

V rámci tohto obrovského komponentu sme vyrátali aj priemernú najkratšiu cestu medzi každými dvoma mestami. Táto hodnota vyšla 7,320 a značí, že v priemere sa z nejakého mesta v obrovskom komponente do iného mesta v obrovskom komponente dostaneme cez približne 7 hrán po zaokrúhlení. Výpočet všetkých štatistík, ktoré súvisia s komponentami siete sme spočítali opäť pomocou knižnice `networkx`.

Modularita

Aby sme zistili prepojenosť tejto siete v rámci jednotlivých kontinentov, vypočítali sme aj modularitu, pričom na výpočet sme opäť použili knižnicu `networkx`. Sieť sesterských miest sme teda rozdelili na kontinenty a sledovali sme ako dobre je sieť prepojená v rámci týchto kontinentoch. Hodnota modularity vyšla 0,1706, pričom túto hodnotu sme zrátali pomocou balíku `community_louvain` v programovacom jazyku Python. Túto hodnotu, môžeme interpretovať tak, že sieť je len mierne prepojená v rámci jednotlivých kontinentoch, veľa miest má teda sesterské mesto z iného kontinentu. Ak by táto hodnota vyšla 1, znamenalo by to, že sieť je prepojená iba v rámci jednotlivých kontinentov.

PageRank

PageRank vrchola resp. mesta vyjadruje jeho dôležitosť (popularitu) v rámci celej siete. Zo zaujímavosti sme vypočítali hodnoty PageRankov pre jednotlivé mestá z našej siete pomocou algoritmu v programovacom jazyku Python z knižnice `networkx`. Najvyššiu hodnotu PageRanku dosiahlo mesto Laredo zo štátu Texas v Severnej Amerike. Podľa webovej stránky: `https://en.wikipedia.org/wiki/Laredo,_Texas#Sister_cities`, Laredo sponzoruje Medzinárodný festival sesterských miest Laredo, kde sa konajú rôzne výstavy ohľadom turizmu, obchodu a kultúry, pričom zúčastniť sa jej môžu všetky sesterské mestá Laredo. V roku 2004 dokonca toto podujatie získalo cenu za najlepší celkový program od Sister cities international.

Distribúcia stupňa vrcholov

V poslednej časti tohto projektu sme sa pozreli na distribúciu stupňa vrcholov, teda či sa stupne vrcholov správajú podľa nejakého pravdepodobnostného rozdelenia. Vykonali sme to v softvéri R pomocou balíku `fitdistr` a knižnice `VGAM`. Vyskúšali sme tri diskretizované spojité pravdepodobnostné rozdelenia, keďže jednotlivé stupne vrcholov nadobúdajú hodnoty z množiny prirodzených čísel. Konkrétne sa jednalo o exponenciálne rozdelenie, Paretovo rozdelenie a zeta rozdelenie. Žiaľ všetky tieto testy dopadli tak, že ani jedno z týchto pravdepodobnostných rozdelení nepopisuje dobre stupne vrcholov v našej sieti. Z uvedených pravdepodobnostných rozdelení však rozdelenie stupňov vrcholov popisuje najlepšie Paretovo rozdelenie.

Záver

V tomto projekte sme podrobne analyzovali sieť sesterských miest. Analýzu sme vykonávali prostredníctvom grafického znázornenia a výpočtu rôznych sieťových štatistík, pomocou čoho sa nám podarilo popísať rôzne trendy, ktoré charakterizujú túto sieť. Zistili sme, že sieť má pomerne nízky priemerný stupeň vrchola a hustotu, z čoho sme usúdili, že sieť sesterských miest nie je veľmi prepojená. Je to spôsobené tým, že má veľa komponentov, avšak, nachádza sa tu jeden obrovský komponent, ktorý zahŕňa

až 73,32 % miest z celej siete. Na základe výpočtov sme aj usúdili, že sieť nie je príliš prepojená v rámci jednotlivých kontinentov, čo potvrdila vypočítaná hodnota modularity. V ďalšej časti projektu sme riešili dôležitosť jednotlivých miest v sieti pomocou PageRanku, kde najvyššiu hodnotu malo mesto Laredo, ku ktorému sme uviedli aj pár zaujímavostí. Záver projektu bol venovaný distribúcií stupňa vrcholov, kde sme usúdili, že žiadne pravdepodobnostné rozdelenie z uvedených nepopisuje rozdelenie stupne vrcholov dôkladne.