Analýza struktur v komplexních sítích zaměřená na posouzení kvality vybraných metod

Ondřej Chovanec, Vedoucí práce: doc. Mgr. Miloš Kudělka, Ph.D.

Fakulta elektrotechniky a informatiky, Vysoká škola Báňská – Technická univerzita Ostrava

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

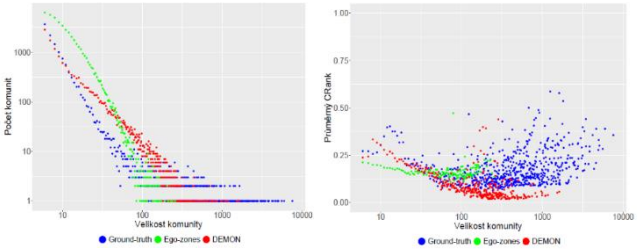
Motivace a cíl práce

Sítě hrají důležitou roli v mnoha oblastech informatiky a nás zajímalo jejich uplatnění při analýze dat. Cílem této práce bylo provést experiment s výběrem algoritmů týkajících se detekce překrývajících se komunit v sítích a experiment s metodami konstrukce sítí z vektorových dat. Pomocí získaných výsledků těchto experimentů jsme následně porovnali použité algoritmy a metody mezi sebou a určili jejich kvalitu, výhody, nevýhody a vhodné případy užití.

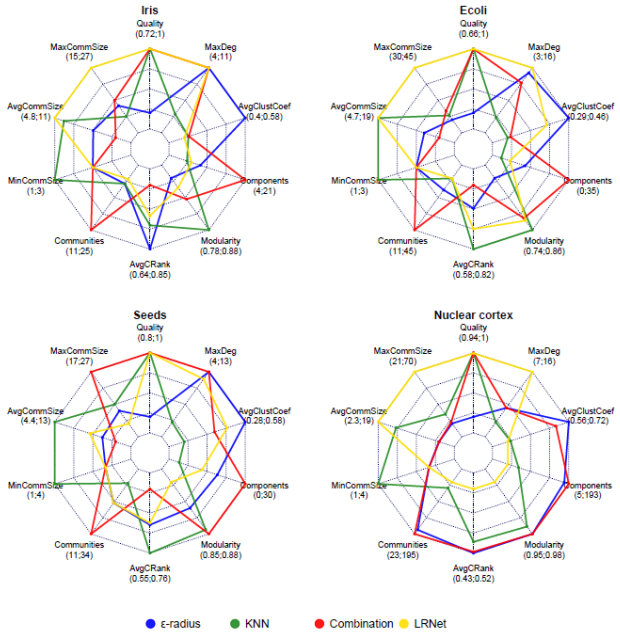
Experiment s algoritmy detekce překrývajících se komunit

Komunity se hledají v sítích, abychom našli skupiny uzlů v síti s podobnými vlastnostmi a mohli si vytvořit představu o její struktuře. Nás zajímaly algoritmy detekce **překrývajících se** komunit, jelikož ty jsou mnohem lepší pro zachycení reálné struktury sítě. Jenomže pro hledání komunit není vždy jednoduché vybrat vhodnou metodu a následně ohodnotit kvalitu nalezených komunit.

Využili jsme definovaných **Ground-truth** komunit a algoritmů **Ego-zones** a **DEMON**. Pro posouzení kvality jimi nalezených komunit jsme použili měr **Modularitu**, **Conductance** a **CRank**. Každá z použitých měr využívá jiných vlastností komunit pro lepší možnosti hodnocení. Algoritmy jsme mezi sebou porovnávali podle hodnot těchto měr vzhledem k velikostem komunit a také podle dalších obecných vlastností, jako jsou např. počet nalezených komunit, distribuce velikosti komunit atd.



Obrázek: Distribuce velikostí komunit (vlevo) a průměrný CRank (vpravo) pro síť DBLP

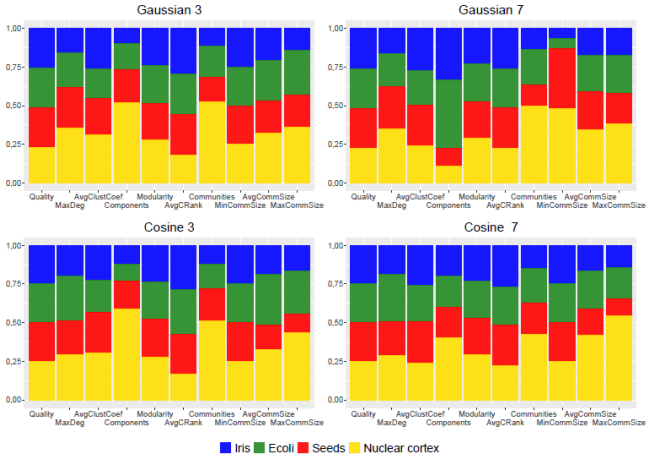


Obrázek: Porovnání vlastností sítí získaných převodem z vektorových dat pro kosinovou podobnost a průměrný stupeň 7

Experiment s metodami konstrukce sítí z vektorových dat

Druhá část práce se věnuje metodám konstrukce sítě z vektorových dat. Požadavkem standardně je, aby vzniklá struktura sítě co nejvíce odpovídala vektorovým datům. Výhodou sítí je například možnost vizualizace dat, což pro vektorová data s větší dimenzí není možné. Na síť potom můžeme použít klasické postupy analýzy sítě a z výsledků si udělat úsudek o struktuře dat.

Experimentovali jsme s metodami ***ε*-radius**, **KNN**, **LRNet** a **kombinací** *ε*-radius s KNN. Pomocí těchto metod jsme vytvořili několik sítí o různých průměrných stupních z několika datasetů. Následně jsme porovnávali strukturní vlastnosti získaných sítí. Počítali jsme také kvalitu sítí pomocí námi definované účelové funkce berouci v potaz podobnost propojených uzlů.



Obrázek: Vlastnosti sítí získaných pro LRNet