DumboAsAService
PredictLikes: Blog like-profil klaszterezés

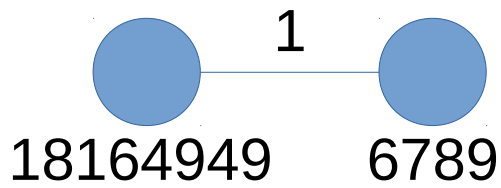
Nótai István (notaiistv@gmail.com)
Szakállas Dávid(david.szakallas@gmail.com)
Mátyás-Barta Csongor(mbcsongor@gmail.com)

2015. december 8.

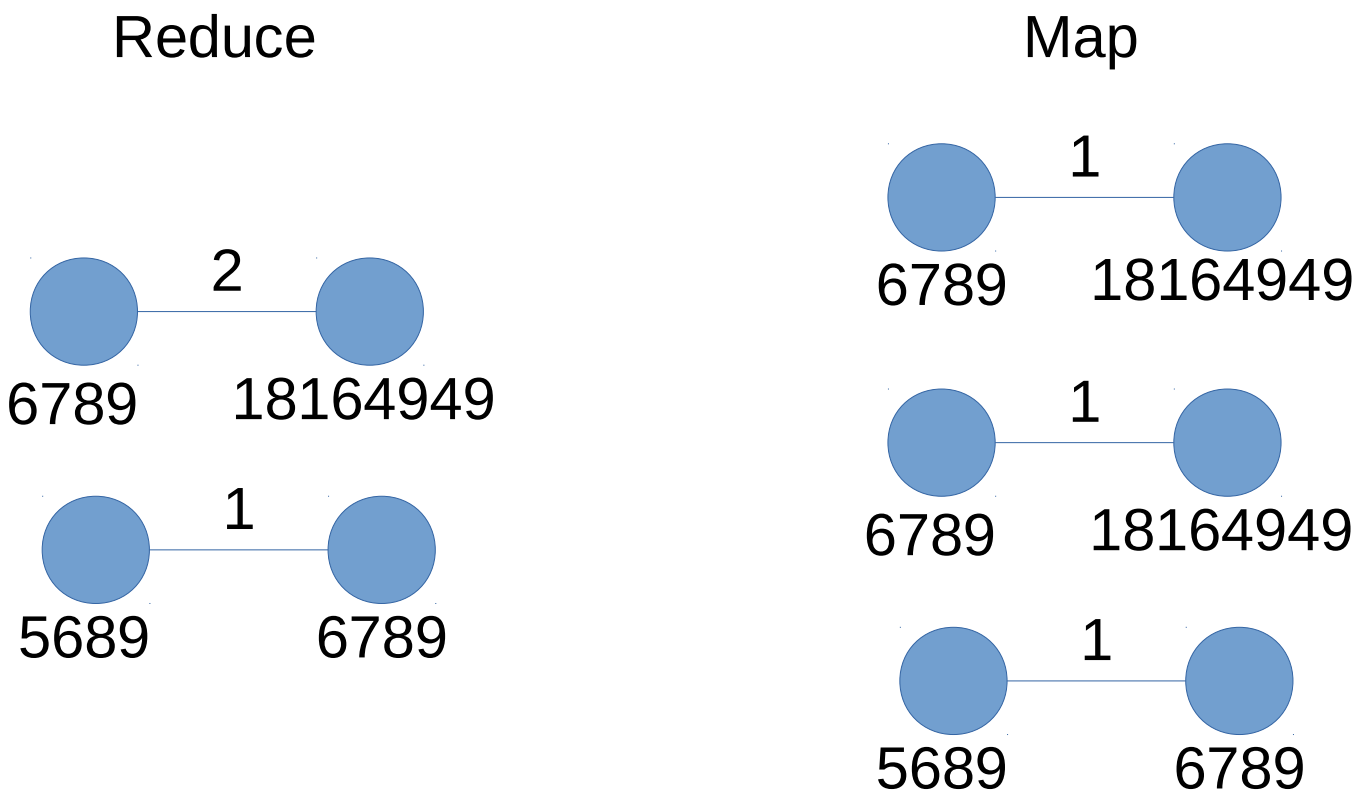
Feladat értelmezés

PredictLikes - blogok "like-profil" alapján klaszterezése és a csoportosítás vizualizációja:
Like profil a mi értelmezésünkben = Ha két blognak van közös likeolója, akkor ők hasonló blogok.

{"uid": "34168956",
"likes":
[{"blog": "18164949"},
{"blog": "6789"}]}



MapReduce - Éllista épités



Reduce: csak a 10-nél nagyobb súlyú éleket írjuk ki

Community Search

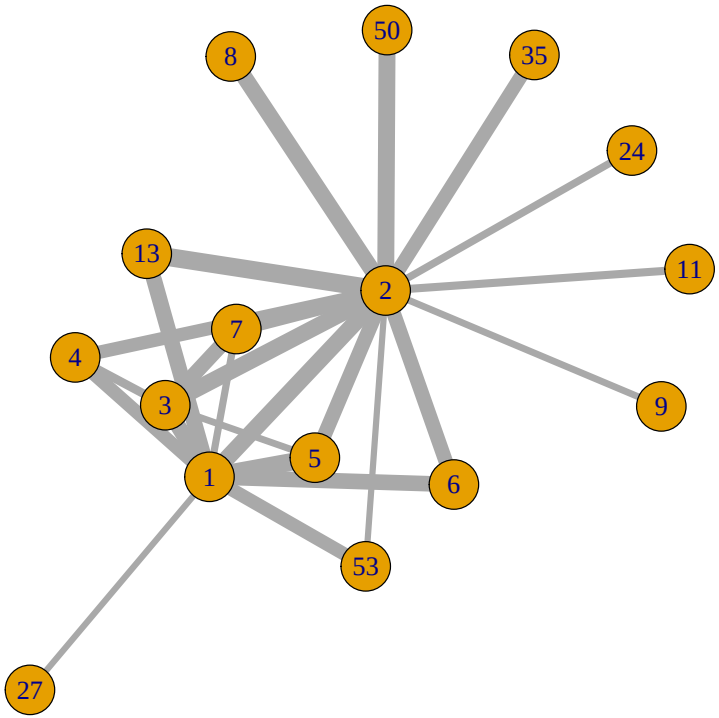
Fastgreedy community search [1]

- bottom-up hierarchikus megközelítés
- kezdetben minden csomópont egy közösség
- greedy: melyik két közösség összevonása eredményezné a legnagyobb *modularitás* növekedést lokálisan
- rezolúciós hiányosság: a gráf egészéhez képest relatív kis közösségek mindig összevonódnak

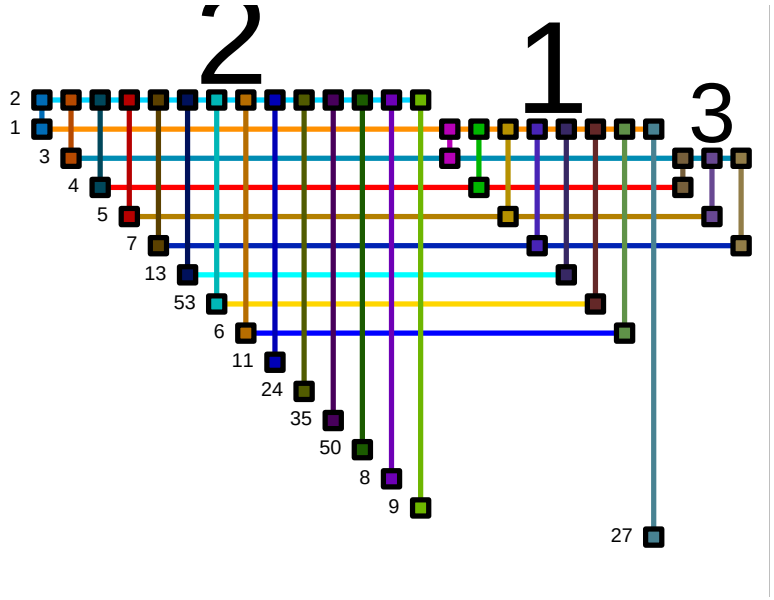
Végeredmény és vizualizáció

Közel 7000 csomópont és 700k él $\rightarrow$ 54 közösség

igraph



BioFabric



Használt eszközök

Éllista épités

- local Hadoop MapReduce(Java)
- 15GB RAM

Gráf feldolgozás

- R
- igraph R csomag [2]
- BioFabric R port [3]

Köszönjük a figyelmet!

Hivatkozások

1. http://arxiv.org/abs/cond-mat/0408187
2. http://igraph.org/r/
3. https://github.com/wjrl/RBioFabric/blob/master/R/bioFabric.R