Big Data Analysis
Group 6

Cedric Mingneau (1232611)
William Thenaers (1746077)
2019-12-20

Graph Mining

1. *Results*

- De *betweenness centrality* kan telkens teruggevonden worden in de outputfiles in de *visualisation* map. De bestanden `interval_confpods_*` bevatten een lijstje met de "Top 10 central authors" en hun score. Dit voor periodes startende vanaf 1990 over 5 jaar, met 2 jaar overlap.
- Ook de communities staan opgelijst in deze bestanden. Als we nagaan waar de namen zich bevinden in de afbeelding van de graaf, zien we inderdaad verschillende auteurs samen gegroepeerd die samen publiceerden.
- De auteurgraaf is telkens bijgevoegd als png.

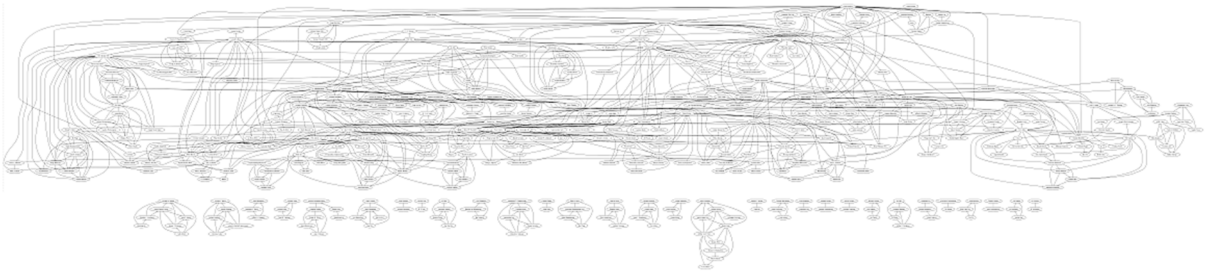
Voorbeeld voor periode 2014-2019:

```
Top 10 authors: (by pagerank)
   2% David P. Woodruff
   1% Benny Kimelfeld
   1% Dan Suciu
   1% Leonid Libkin
   1% Wim Martens
   1% Pablo Barcel
   1% Georg Gottlob
   1% Frank Neven
   1% Piotr Indyk
   1% Ke Yi 0001
```

```
Top 10 central authors: (by centrality)
  12812.2 Ke Yi 0001
  11573.3 Piotr Indyk
  11507.5 Sariel Har-Peled
   8009.1 David P. Woodruff
   4700.2 Dan Suciu
   3868.9 Pierre Bourhis
   3720.6 Benny Kimelfeld
   3358.9 Leonid Libkin
   3240.5 Samuel McCauley
   3172.0 Victor Vianu
```

Merk op dat er een groot aantal auteurs in de top 10 zoals hier afgebeeld dezelfde rang hebben en dus kunnen er mogelijk andere auteurs in deze lijst komen te staan na nieuwe uitvoering van het programma.

Voor de 47 communities (volgens het algoritme) in periode 2006-2016, zie `visualisation/interval_confpods_2006-2016.txt`. In Figuur 1 een verkleinde weergave van de volledige graaf.



Figuur 1 Preview van de graaf in interval_confpods_2006-2016.png

2. *Approach*

- Zoals in de opgave aangegeven, gebruikten we de Stanford SNAP library voor Python. Na een nieuwe ParsingStrategy toe te voegen (`AuthorStrategyYearGraphs`), volgde de implementatie van een wrapper rond de SNAP `PUNGraph` (Undirected Graph).
- De nieuwe strategy houdt twee dictionaries bij met de gevonden jaren als key en een graaf van auteurs als value. In het ene dictionary zitten enkel jaren en auteurs met `PODS` publicaties en in de andere de overige auteurs. Bij publicaties met meerder auteurs, krijgen de auteurs edges naar elke andere auteur in dezelfde publicatie.

Cedric Mingneau (1232611)
William Thenaers (1746077)

- Na het parsen van de XML, voegen we ook nog edges toe voor `PODS` auteurs die samen publiceerden buiten `PODS` publicaties.
- De resultaten (top auteurs met pagerank, top centrale auteurs met in-betweenness en communities met Girvan-Newman methode) worden telkens naar een bestand weggeschreven. De volledige graaf voor een jaar of periode werd ook als png opgeslaan door te exporteren naar een DOT-representatie en GraphViz dit om te laten zetten naar een afbeelding.
- Op dit moment worden de top 10 auteurs telkens bepaald in `PODS`. Maar de top-X en conferentie filter zijn beide parameters die gewijzigd kunnen worden bij de creatie van de `GraphMiner` klasse of het oproepen van een `plot` functie.

3. Difficulties

- Eerst werden enkel auteur in `PODS` verbonden met elkaar. Maar later lazen we dat de gerefereerde paper ook sprak over publicaties die deze auteurs samen maakten buiten de `PODS` conferentie. Deze werden dan later ook toegevoegd.