

# Assignment 03: Link Analysis

Michelle Gybels  
Anaïs Ools

November 2016

## 1 Aanpak

Net zoals bij de vorige assignments wordt ter voorbereiding het bestand `dblp50000.xml` ingelezen. Vervolgens wordt met behulp van de klasse `XMLParserPODS` een nieuwe file `PodsBig.txt` gegenereerd, welke enkel de auteurs bevat van PODS-publicaties. Hierin staan de auteurs van eenzelfde publicatie op eenzelfde regel gegroepeerd. Deze file wordt uiteindelijk als input aan het programma `Assignment03.py` meegegeven.

Tijdens het inlezen van de auteurs per PODS-publicatie wordt met behulp van de `networkx`-library een gewogen graaf opgesteld. Hierbij is iedere knoop een auteur en zijn de gewichten van de bogen het aantal keer dat de twee verbonden auteurs samen een publicatie geschreven hebben. Daarnaast wordt een hashmap bijgehouden met de auteurs als key waarbij voor iedere auteur een tupel van de vorm (`[Publication count], [PageRank], [Authority Score]`) opgeslagen wordt.

De publication count wordt tijdens het inlezen geüpdatet. Vervolgens wordt met behulp van de functies `pagerank()` en `hits()` uit de `networkX`-library voor iedere auteur de pagerank en authority score berekend. De bekomen waarden worden uiteindelijk naar de file `authorImportance_BIG.csv` weggeschreven.

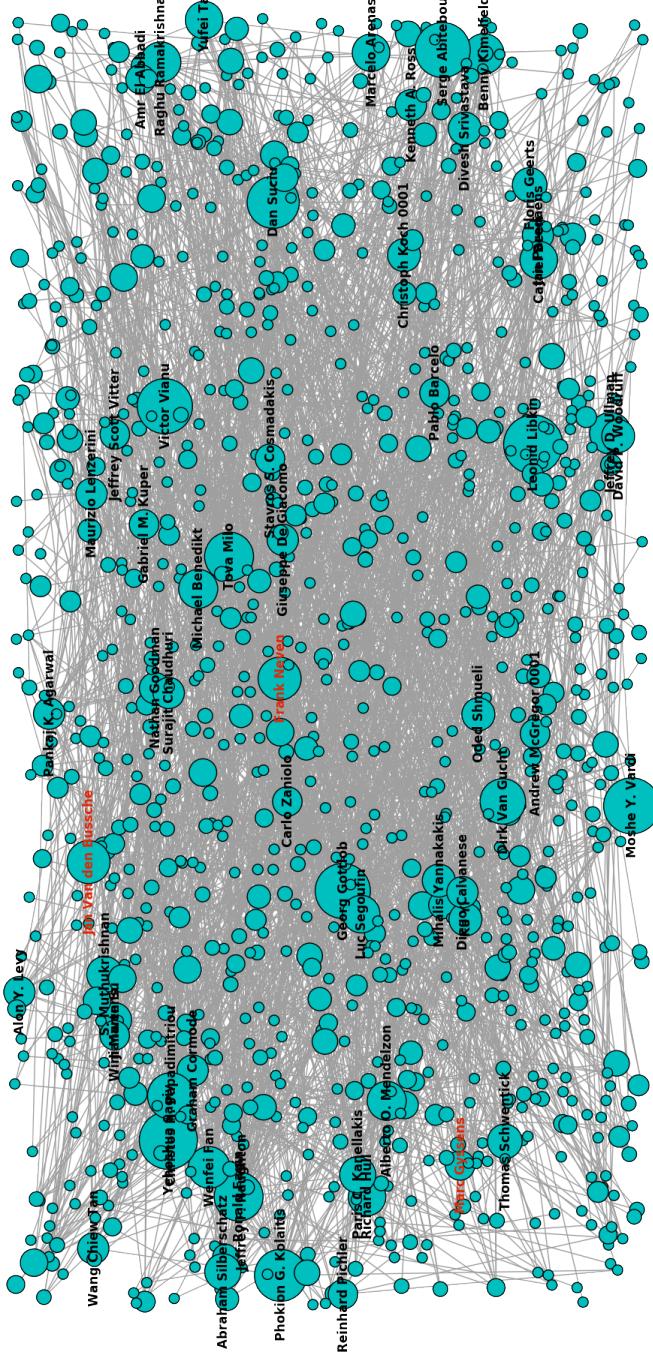
Tijdens de implementatie van het bovenstaande zijn geen moeilijkheden naar voor gekomen. De enige “moeilijkheid” was het opstellen van de graaf, wat vergemakkelijkt werd door gebruik te maken van de `networkX`-library. Ook de andere waarden konden met behulp van deze library eenvoudig berekend worden.

## 2 Resultaat van het programma

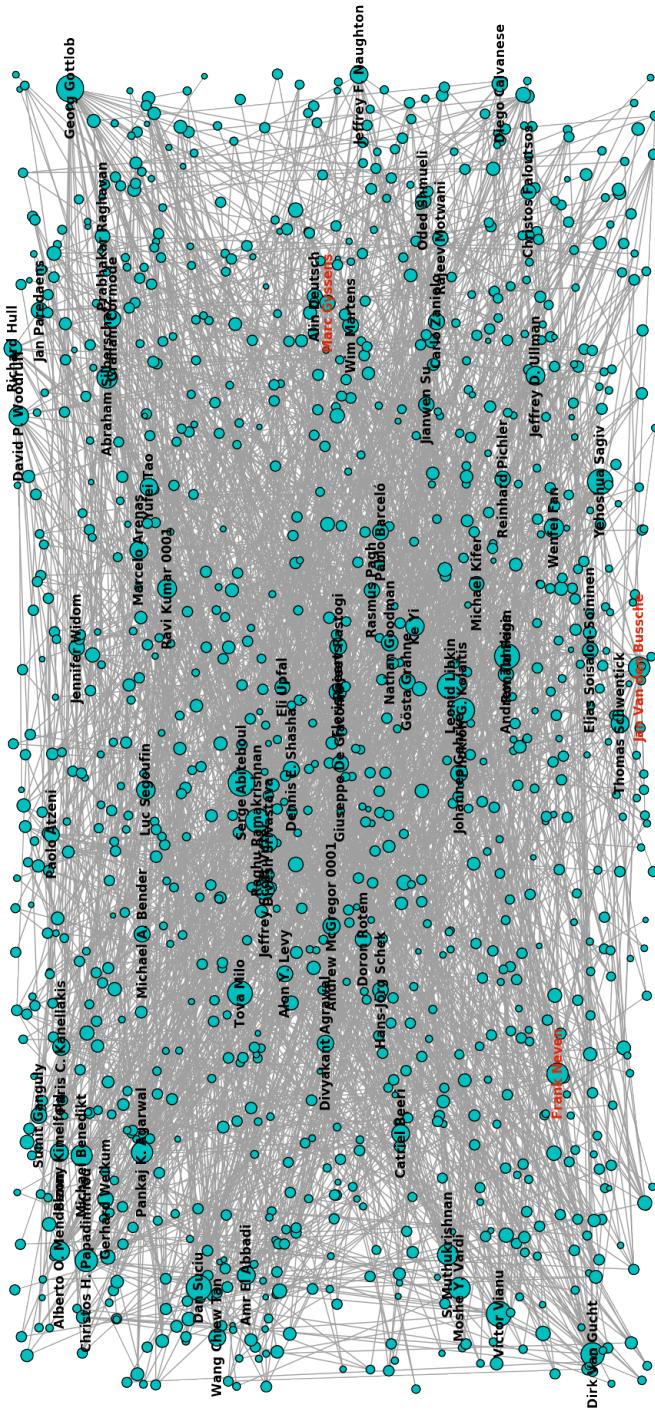
Het programma berekent uiteindelijk de publication count, pagerank en authority score van elke auteur op basis van hun aantal collaboraties met andere auteurs en visualiseert deze gegevens met behulp van een graaf. Figuur 1 doet dit voor de publication count, Figuur 2 voor de pagerank en Figuur 3 voor de authority score. In iedere graaf wordt de grootte van een node bepaald op basis van de waarde van de node. Ook kan een percentage als threshold meegegeven worden. Indien een node een grotere waarde heeft dan een bepaald procent van de waardes, aangegeven door de threshold, krijgt deze node een label in de visualisatie. Daarnaast wordt het label van een kleur voorzien indien de betreffende auteur afkomstig is van Universiteit Hasselt.

De resultaten werden eveneens naar een bestand weggeschreven. We verwijzen de lezer naar het bijgevoegde bestand `authorImportance.csv`. Hieronder worden enkele visualisaties van deze gegevens besproken.

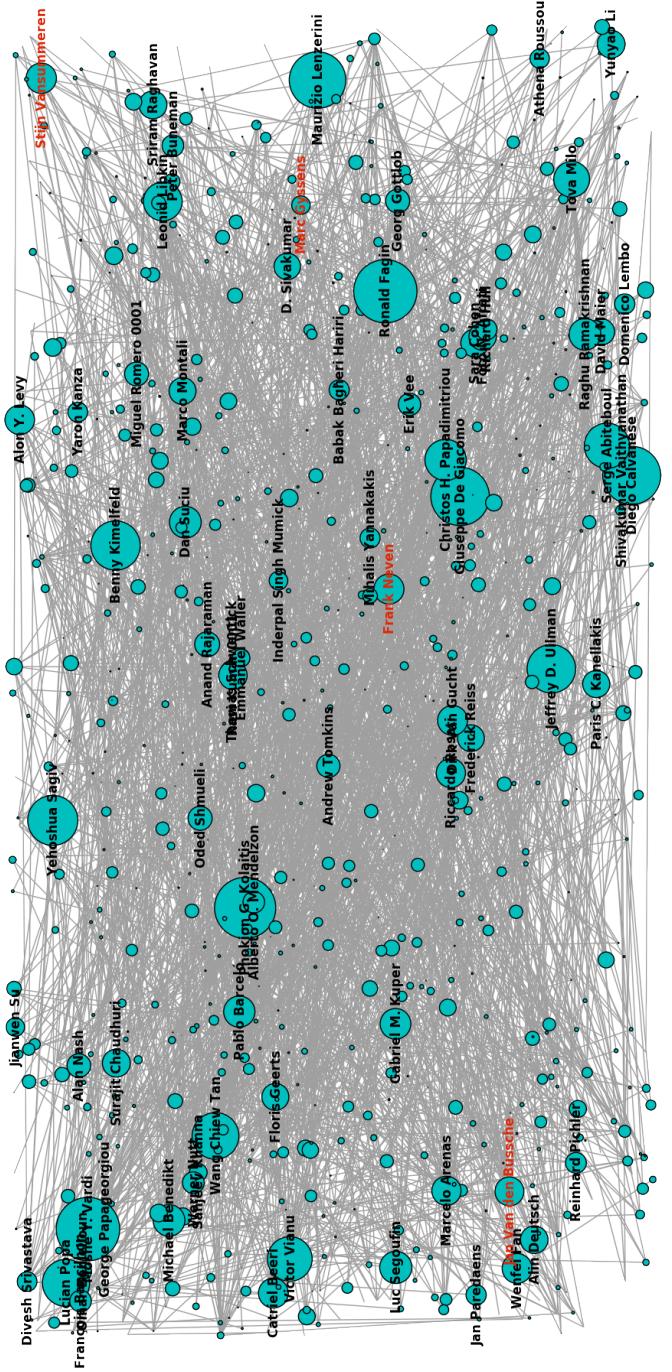
Uit Figuur 4 kunnen we afleiden dat de pagerank recht evenredig is met het aantal publicaties. Wanneer we de authority vergelijken met de publication count, te zien in Figuur 5, is het verband minder opvallend, maar toch valt op te merken dat een auteur met een lager aantal publicaties meer kans heeft op een lagere authority score. Figuur 6 toont aan dat een lagere pagerank leidt tot een lagere authority en omgekeerd, echter hier ook met enkele uitzonderingen. Tot slot tonen Figuur 7 en Figuur 8 de top 20 van respectievelijk pagerank waardes en authority waardes. Deze namen zijn ook terug te vinden in de graaf van deze waardes, nl. Figuur 2 en Figuur 3.



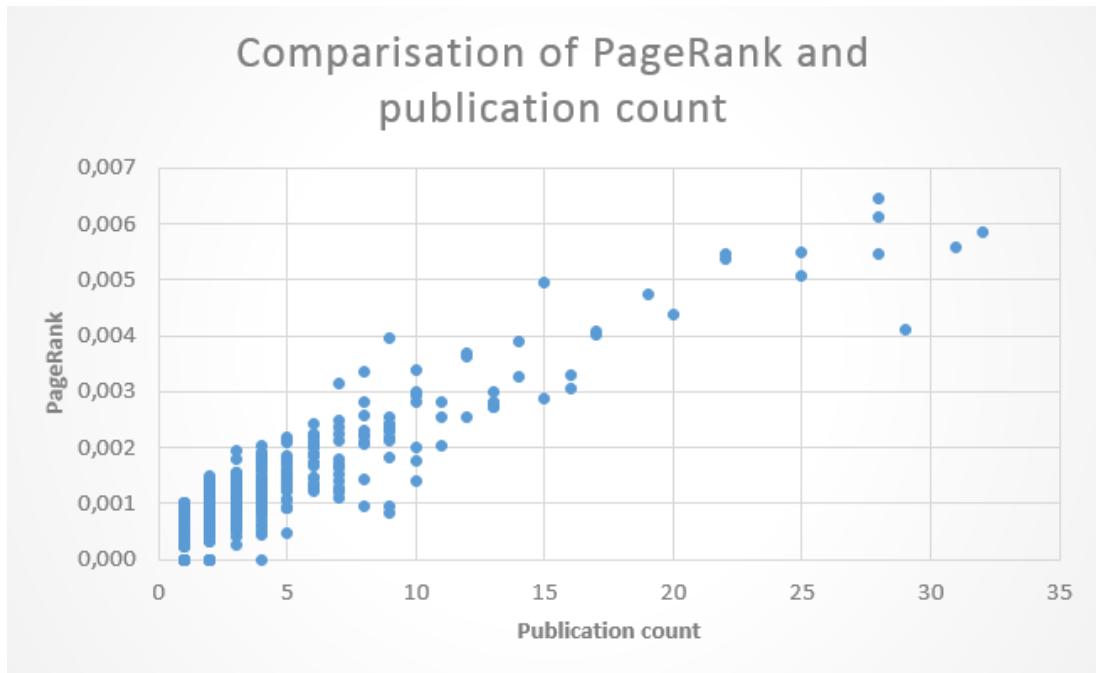
Figuur 1: Graaf van het aantal publicaties per auteur



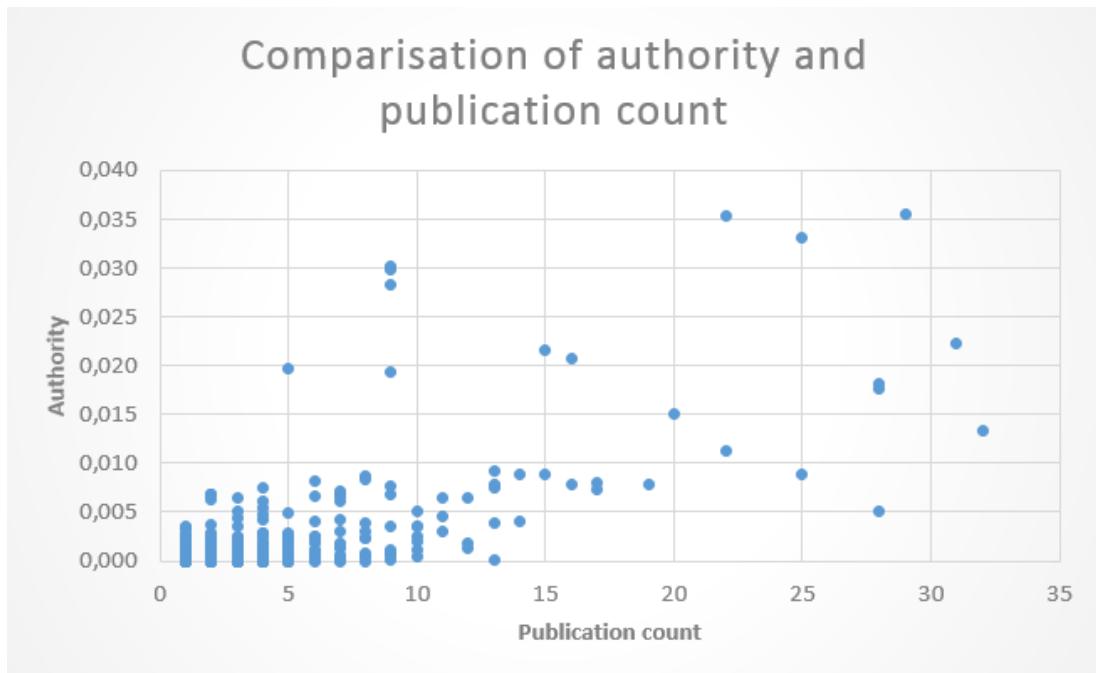
Figuur 2: Graaf van de pagerank van auteurs



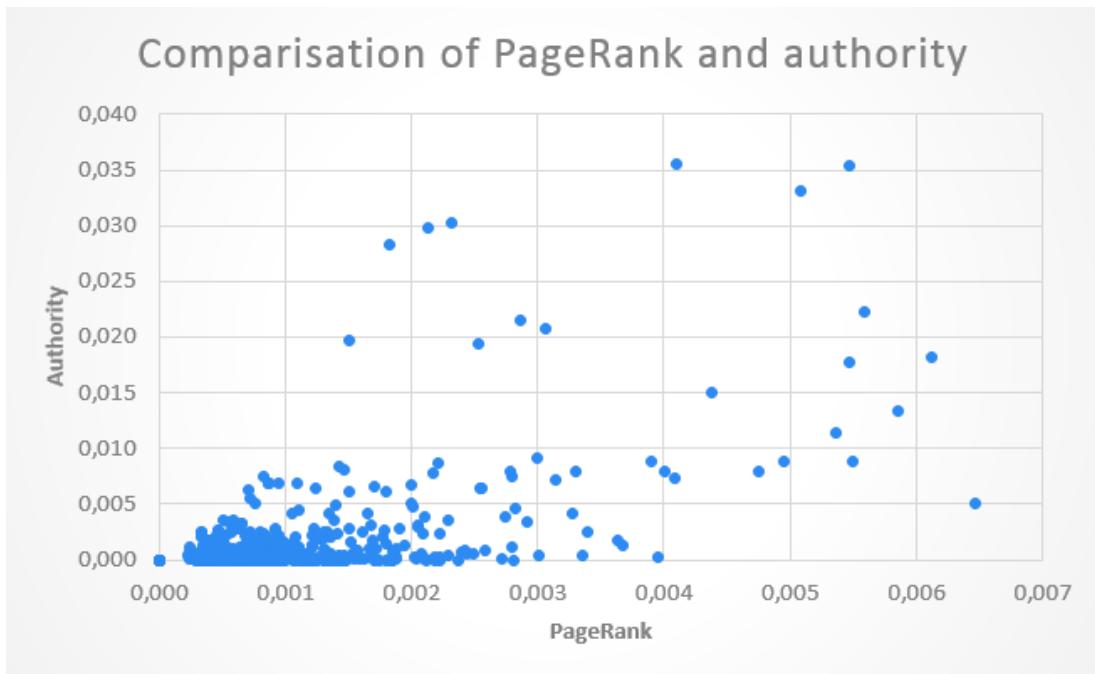
Figuur 3: Graaf van de authority score van auteurs



Figuur 4: Visualisatie van pagerank en publication count



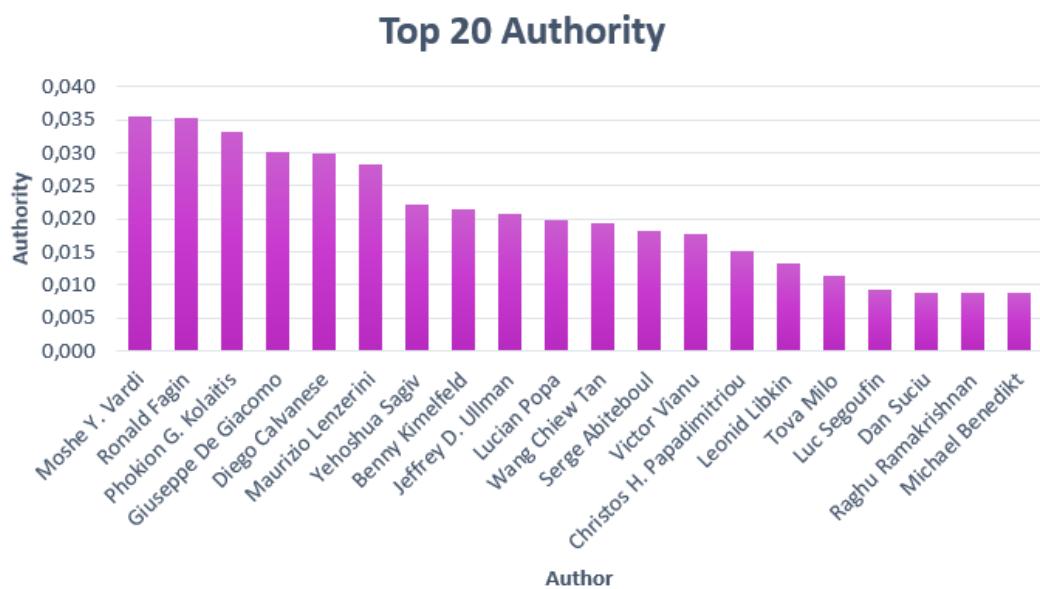
Figuur 5: Visualisatie van authority en publication count



Figuur 6: Visualisatie van pagerank en authority



Figuur 7: Top 20 auteurs met hoogste pagerank



Figuur 8: Top 20 auteurs met hoogste authority