Identifying Research Talent Using Web-Centric Databases

The Network Institute

Anca Dumitrache[1], Paul Groth[2], Peter van den Besselaar[3]
Network Institute, VU University Amsterdam

VU UNIVERSITY AMSTERDAM

[1]anca.dumitrache@student.vu.nl, [2]p.t.groth@vu.nl, [3]p.a.a.vanden.besselaar@vu.nl

Goal: Analyze both online and offline networks of scholarly publications, to determine which can be used for discovering and predicting valuable research work.

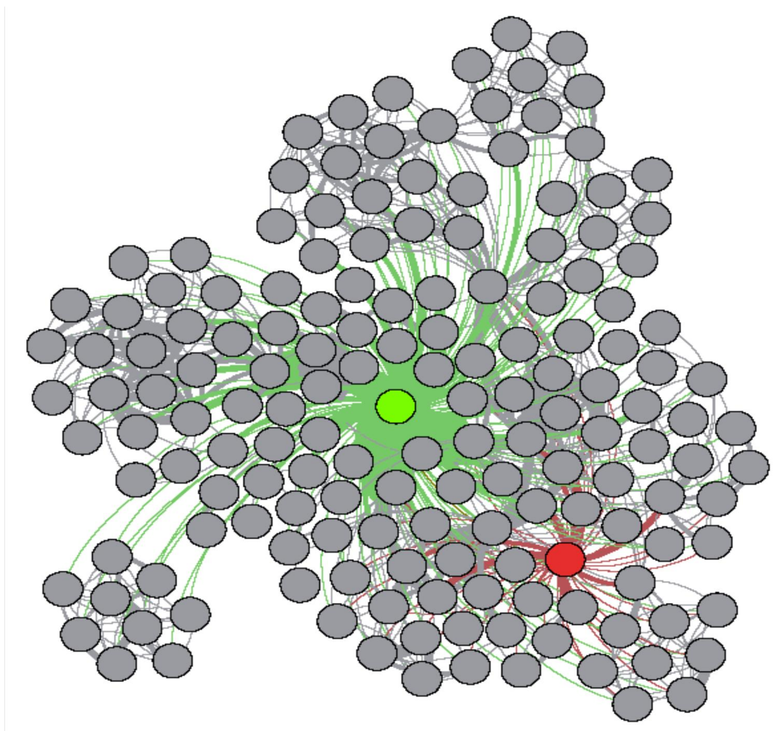
Solution: Study the relation between an early career researcher, and their former PhD supervisor, to determine their **independence** in research.

Approach: Construct a dataset of researcher-supervisor pairings, and get their publications from Google Scholar (online), and Web of Science (offline); define independence indicators to describe the data.

Indicator 1: The **co-author network** is the graph of authors with whom the young researcher has published papers.

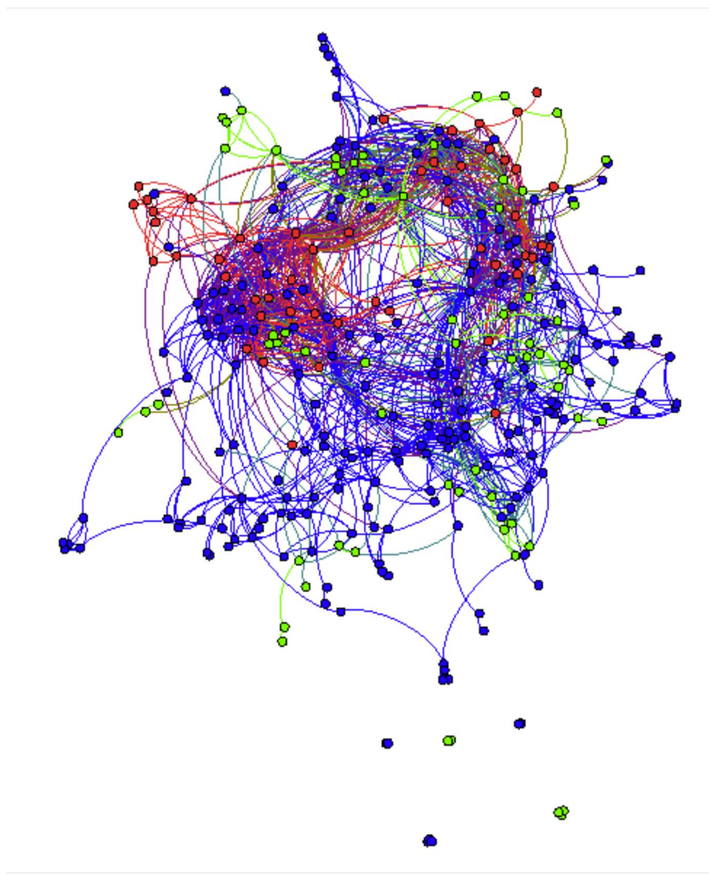
Properties:

- **eigenvector centrality** of the supervisor;
- **clustering coefficient** of the supervisor.



Example network, with the young researcher (green node), and supervisor (red node), highlighted.

Indicator 2: The **topic network of the researcher** is the combined graph of papers for the young researcher and the supervisor. Edges are based on title similarity.



Example network, with the papers of the young researcher (green nodes), supervisor (blue nodes), and joint publications (red nodes), highlighted.

Results: Online sources have larger datasets, with a wider scope than offline ones; independence is more prevalent in online data.

Network property	Google Scholar	Web of Science
Papers retrieved	128	24
Eigenvector centrality of supervisor	0.12	0.94
Clustering coefficient of supervisor	0.18	0.61

Example results for the same researcher-supervisor pairing.