Drahomira Herrmannova (drahomira.herrmannova@open.ac.uk) & Petr Knoth (petr.knoth@open.ac.uk)

Knowledge Media Institute

The Open University

Semantometrics: Fulltext-based measures for analysing research collaboration

Q: Would you rate the **quality** of a movie based **only** on the number of views?

Aim

Up to date many studies of scientific citation, collaboration and co-authorship networks have focused on the concept of **cross-community ties**.

We explore how ***Semantometrics*** can help in understanding the nature of the cross-community ties and in characterising the types of research collaboration in scholarly publication networks.

Semantometrics

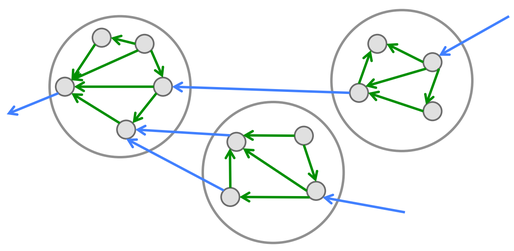
In contrast to the existing research evaluation metrics such as Bibliometrics, Altmetrics or Webometrics, which are based on measuring the number of interactions in the scholarly network, Semantometrics build on the premise that **fulltext** is needed to understand the value of publications.

Cross-community ties

Up to date, many studies have focused on the concept of cross-community ties. However, these studies have predominantly been concentrating on analysing citation and collaboration networks without considering the content of the analysed publications.

Cross-community ties

Links between communities



Cross-community ties

- The importance of cross-community ties
  - In citation networks, cross-community citation patterns are characteristic for high impact papers [Shi. et al., 2010]
  - Same holds true in case of cross-community scientific collaboration [Newman, 2004; Lambiotte and Panzarasa, 2009]

How to identify cross-community ties?

- From citation/coauthorship network
  - E.g. betweeness centrality
- From fulltext
  - Semantic similarity
- Different types of collaboration when writing a paper
  - Emerging vs. established
  - Interdisciplinary vs. intradisciplinary
  - Etc.

Emerging and established collaboration

- Endogamy
  - In social sciences: the practice or tendency of marrying within a social group
  - In research: collaboration within a group of authors
  - Higher endogamy = more frequent collaboration

$$endo(A) = \frac{|d(A)|}{|\bigcup_{a \in A} d(\{a\})|} \qquad endo(p) = \frac{\sum_{x \in L(p)} endo(x)}{|L(p)|}$$

p Publication
A Set of authors
d(A) Papers coauthored by authors in *A*
L(p) Set of all subsets with at least two authors of *p*

Inter- and intra-disciplinary collaboration

- Semantic distance of publication authors
  - Higher author distance indicates more distant communities
  - Calculated using *cosine similarity* on *tf-idf* term-document vectors created from document fulltexts
  - Author publication record considered as a single text

$$a\_dist(p) = \frac{1}{|A(p)| \cdot (|A(p)| - 1)} \sum_{a_1 \in A(p), a_2 \in A(p), a_1 \neq a_2} dist(a_1, a_2)$$

p Publication
A(p) Set of authors of *p*

Types of research collaboration

	Low endogamy	High endogamy
High author distance	Emerging interdisciplinary collaboration	Established interdisciplinary collaboration
Low author distance	Emerging expert collaboration	Expert group

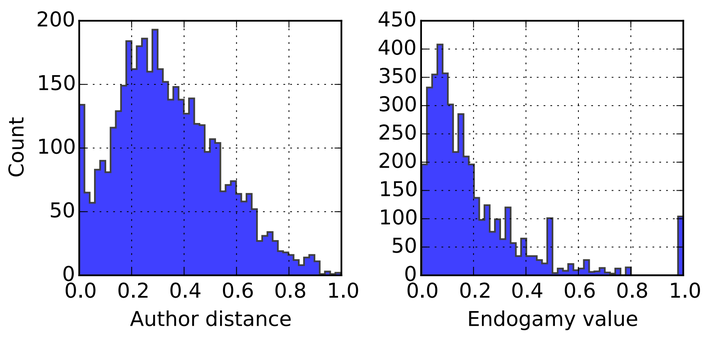
Experiment

- What is the distribution of the four different types of collaboration in scholarly literature?
- CORE (core.ac.uk) used as a dataset
  - Cross-discipline
  - Enables sampling by authors and institutions
  - Selected sample
    - Fulltext documents from Open Research Online repository (ORO)
    - All other fulltext publications of the authors from ORO found in CORE

Dataset statistics

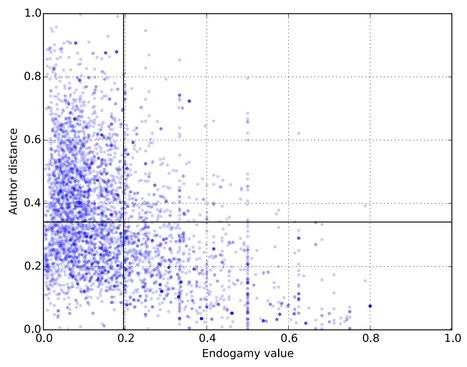
Fulltext articles from ORO	4,207
Number of authors	8,473
Average number of publications per author	7.61
Max number of publications per author	310
Average number of authors per publication	4.31
Max number of authors per publication	25
Average number of received citations	0.30
Average number of collaborators	80.23
Total number of publications	30,484

Endogamy and author distance distribution



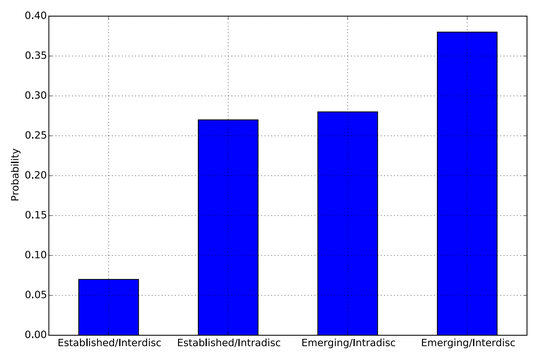
Relation between author distance and endogamy

Author dist. and endogamy of the analysed publications



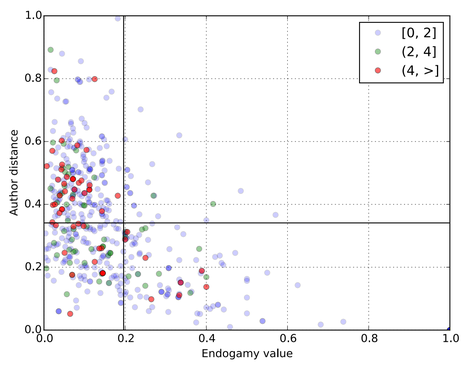
Types of research collaboration

Probability of appearance per type



Types of research collaboration and "impact"

Types of collaboration and number of received citations



References

Xiaolin Shi, Jure Leskovec, and Daniel A Mcfarland (2010) *Citing for High Impact.*

M. E. J. Newman (2004) *Coauthorship networks and patterns of scientific collaboration.*

R. Lambiotte and P. Panzarasa (2009) *Communities, know- ledge creation, and information diffusion*.