

Applying Gini Coefficient to quantify Scientific Collaboration in Researchers Network

Giseli Rabello Lopes, Roberto da Silva,
José PALAZZO M. de Oliveira

Instituto de Informática
Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre, Brazil

WIMS'11

inf
INSTITUTO
DE INFORMÁTICA
UFRGS

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

Summary

- Introduction
- Basic Approaches to SNA
- Gini Coefficient
- Gini Coefficient applied on SNA
- Case Study
- Conclusions

Introduction

- The Social Network Analysis (SNA) is based on the assumption that the **importance of relationships among the interaction units** is a **central point** to the evaluation and analysis of social interaction
- Some fundamental concepts used on SNA include actors and relational ties
 - **Actors** are social entities that have social linkages modeled by the social network.
 - Actors are linked to other actors by **relational ties**

Introduction

- The increasing interest in researching on Social Networks was encouraged by the popularization of online social networks (e.g., Facebook, Orkut)
- Another example of the social networks is a co-authorship social network
 - Actors → authors
 - Relational ties → presence of at least one co-authored paper between two authors
 - Data sources for the construction → Digital libraries as DBLP

Basic Approaches to SNA

- The basic approach to analyze a network (SN) is to represent it as a graph $G:=(N,E)$ where:
 - nodes $n \in N$ (actors)
 - edges $e \in E$ (relational ties)
- Examples of SNA approaches:
 - Degree Centrality, Closeness Centrality, Density, Diameter, Clustering Coefficient

Gini Coefficient

- Measure of statistical dispersion
- Commonly used to evaluate the inequality of income and wealth distributions
- Definitions:
 - **Lorenz curve** is a graphical that represents the cumulative distribution of a probability density function
 - **Gini coefficient** is calculated as the area between the perfect equality line (*identity*) and the observed Lorenz curve

Gini Coefficient applied on SNA

- This application is possible because the relational ties (edges) between actors (nodes) can be viewed as a distribution of amounts corresponding to the possible pairs of actors on the SN
- For each pair of actors, the value associated is:
 - Zero: if there is no relationship between them
 - Otherwise: weight of the relationship between them

Case Study

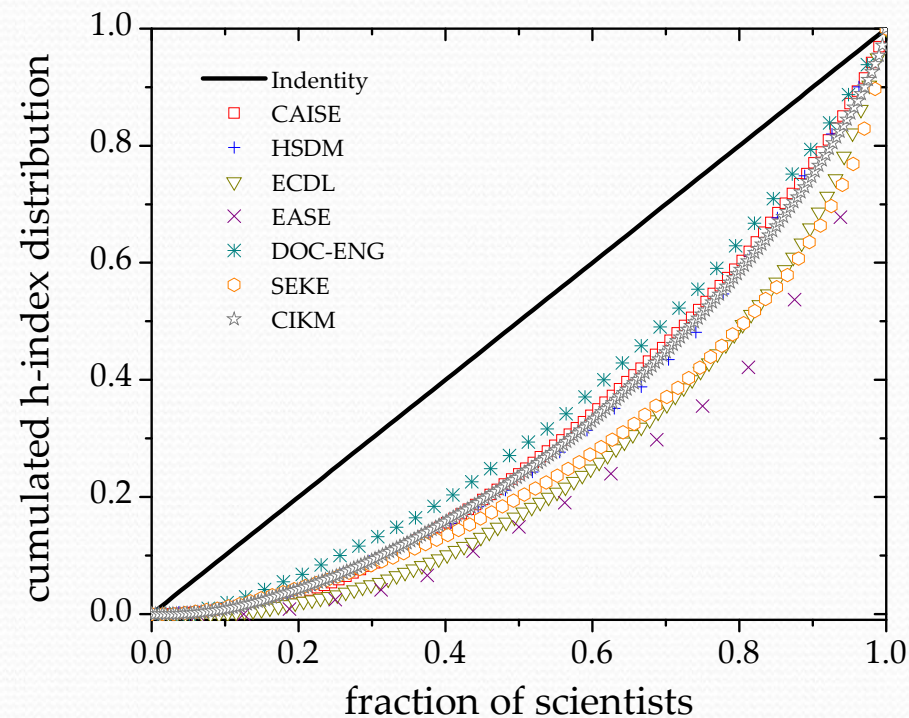
- Apply the Gini coefficient in the context of SNA, more precisely in a Co-Authorship Social Network
- Weights:

$$w_{ij} = \frac{n_{ij}}{n_i} \quad n_i = \sum_{j=1}^n n_{ij}$$

- n_{ij} denotes the number of common papers between the pair of the neighbors $\langle i, j \rangle$
- n_i is the number of papers of author i

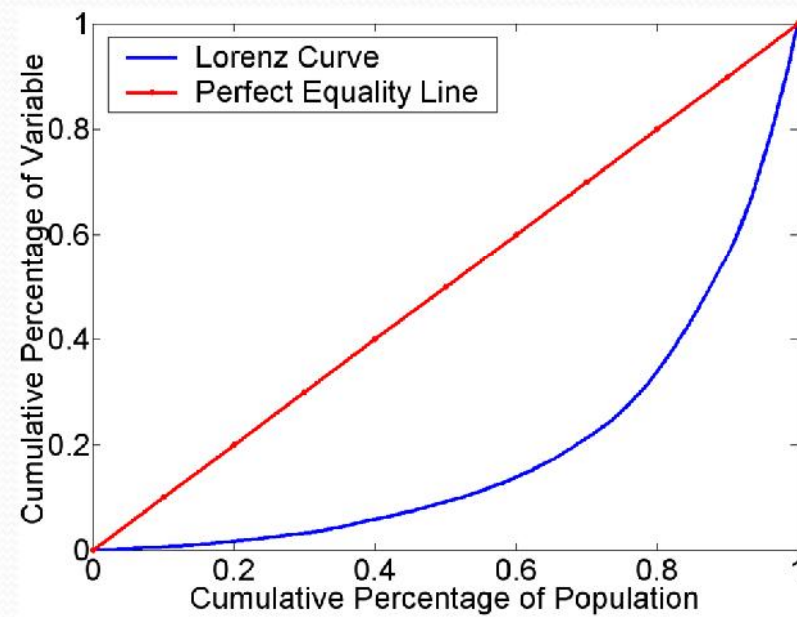
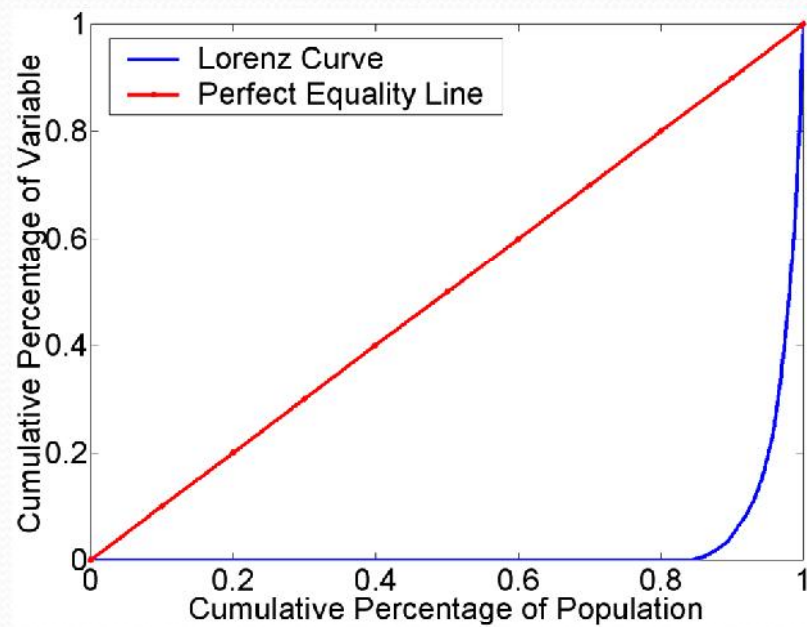
Gini Coefficient

- Example: Lorenz curves for the h-index distributions of researches in conferences of Software Engineering



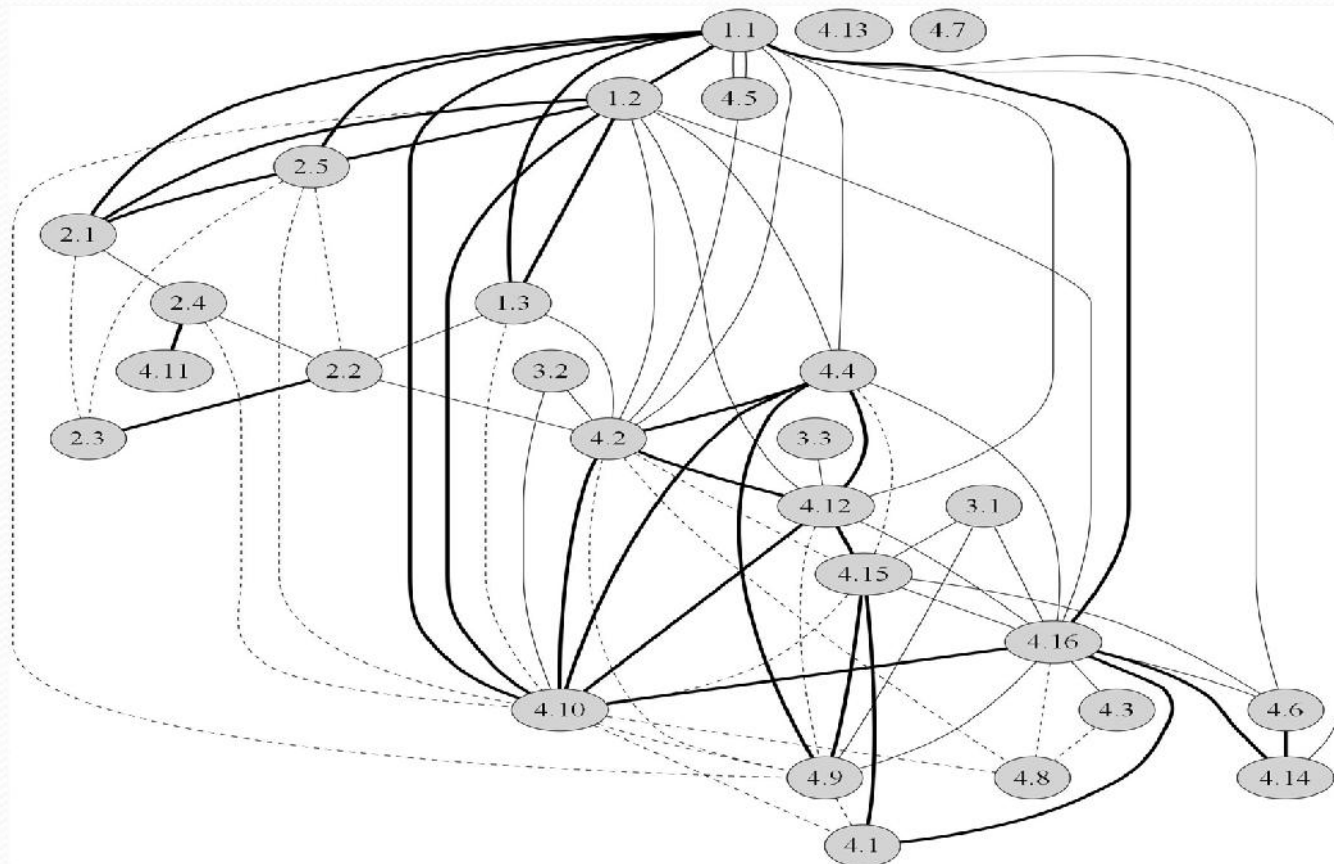
Case Study - Dataset

- Publication data obtained at DBLP on August 03, 2010
- Network composed by 27 researchers (professors of graduate programs in CS) involved in a research project of the INWeb (CNPq)
 - This project begins at year 2008
- Two different time intervals:
 - Until year 2007 (SN2007)
 - Until year 2010 (SN2010)
- To establish a comparative among the collaborations before and after the project implementation



Case Study - Social Networks

- Social Network comparative



Legend:

- Intensified
- - - - - Initiated
- Not modified

#Id	Institutions
1	UFAM
2	UFRGS
3	CEFET-MG
4	UFMG

Case Study - Evaluations

- Global Analysis (1)
- Cooperative Analysis (2.1 and 2.2)

Analysis	Distribution	Gini Coefficient
(1)	SN2007_allPairs	0.9471
	SN2010_allPairs	0.9327
(2.1)	SN2007_Pairs2007	0.5824
	SN2010_Pairs2007	0.5735
(2.2)	SN2007_Pairs2010	0.7009
	SN2010_Pairs2010	0.6160

Coherent results obtained because the SN2010 is more connected than SN2007. Moreover, high values were obtained to both networks indicating unequal distributions. Collaboration network of INWeb is still very disconnected (in relation to one totally connected network) in both considered time intervals.

Case Study - Evaluations

- Global Analysis (1)
- Cooperative Analysis (2.1 and 2.2)

Analysis	Distribution	Gini Coefficient
(1)	SN2007_allPairs	0.9471
	SN2010_allPairs	0.9327
(2.1)	SN2007_Pairs2007	0.5824
	SN2010_Pairs2007	0.5735
(2.2)	SN2007_Pairs2010	0.7009
	SN2010_Pairs2010	0.6160

This difference reflects the improvement in the homogeneity of weight distributions occurred between the researchers that already collaborate before the beginning of the INWeb project. The collaborations between the researchers that are intensified contributed to a more equalitarian distribution in SN2010 than in SN2007.

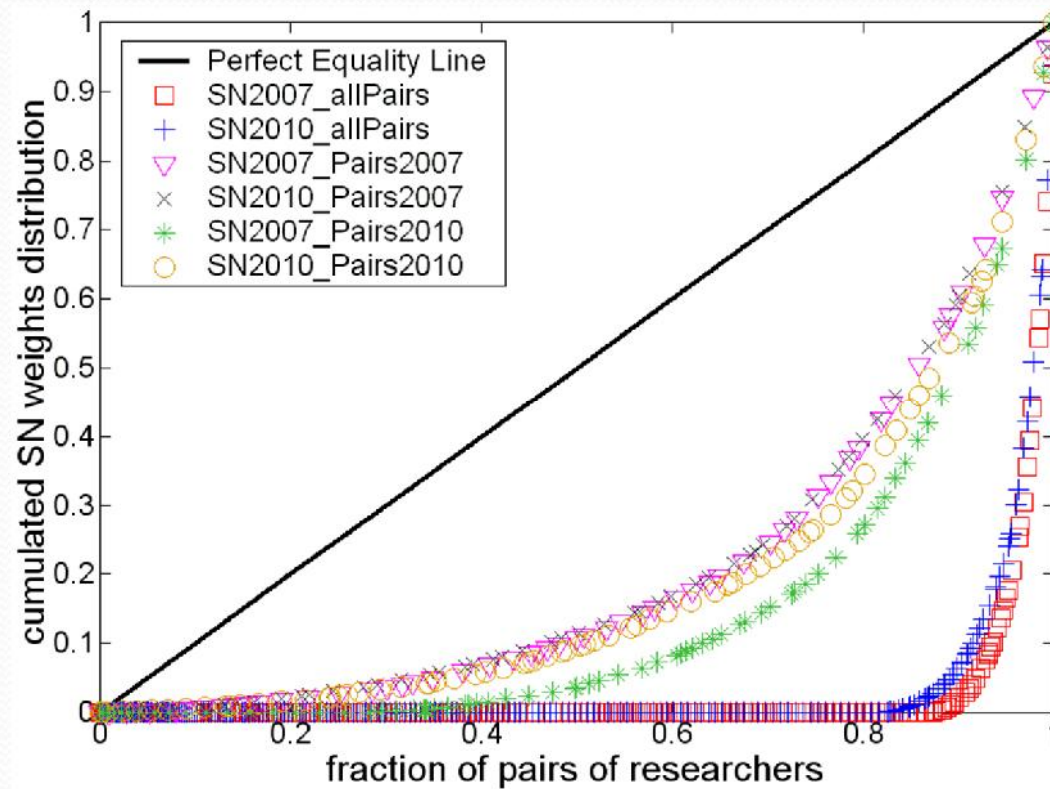
Case Study - Evaluations

- Global Analysis (1)
- Cooperative Analysis (2.1 and 2.2)

Analysis	Distribution	Gini Coefficient
(1)	SN2007_allPairs	0.9471
	SN2010_allPairs	0.9327
(2.1)	SN2007_Pairs2007	0.5824
	SN2010_Pairs2007	0.5735
(2.2)	SN2007_Pairs2010	0.7009
	SN2010_Pairs2010	0.6160

The results indicate unequal distributions in both SNs. It is correct to both networks because: (i) in the SN2007 there are pairs of authors that are not connect and the distribution of weights is nonequalitarian, and (ii) in the SN2010, some already existent collaborations are intensified (increased weights) while new collaborations emerged (first publications co-authored, probably low weights), then the distribution of weights is nonequalitarian in values.

Case Study - Lorenz Curve



Conclusions

- We demonstrated the possibility of uses of the Gini Coefficient to analyze weighted Social Networks
- The initial results pointed out indices of the validity and utility of this approach
- In further works, we intend to expand the study and to develop more evaluations

Applying Gini Coefficient to quantify Scientific Collaboration in Researchers Network

Giseli Rabello Lopes

- grlopes@inf.ufrgs.br

Roberto da Silva

- rdasilva@inf.ufrgs.br

J. Palazzo M. de Oliveira

- palazzo@inf.ufrgs.br

Instituto de Informática
Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre, Brazil

WIMS'11

inf
INSTITUTO
DE INFORMÁTICA
UFRGS

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico