RAW DATA STRUCTURE APPENDIX

Scopus Data Table

The main contribution of this working paper is on the research data. The available research data of the Ecuador give us the paper title, the authors, the year of publication, and the university affiliation. Using those variables, we can build a dataset with the location of the publication, and the available text. The available text is necessary to classify the paper as a industry based on the ISIC4 code. The following table is an example of the raw publication dataset, we add the ID variable as identifier of each paper

ID	Year	Province	City	Authors	Text
ECOTEC1	2018	GUAYAS	GUAYAQUIL	Espinoza...	Ethi...
ECOTEC25	2018	GUAYAS	GUAYAQUIL	Loor P...	Anal...
EPN23	2018	PICHINCHA	QUITO	Silva M....	Biopol...
EPN41	2018	PICHINCHA	QUITO	Douillet...	Revisiti...
EPN61	2018	PICHINCHA	QUITO	Sirunyan A.M...	Obser...

International Standard Industrial Classification

This is the common code used in economics to classify an economic activity

The Similarity Table

The similarity say how similar is each paper to each ISIC4 category. The similarity is a value between 0 and 1, and this relationship is got by SpaCy Python Library. The table 3 shows the first 10 values of the similarity

ID	A0150	C1074	C1420	C2021
ECOTEC0	0.317848	0.548013	0.446434	0.549180
ECOTEC25	0.320490	0.610176	0.641187	0.616872
EPN23	0.387861	0.613222	0.596166	0.638505
EPN41	0.507463	0.829084	0.650273	0.792887
EPN61	0.425886	0.675037	0.567544	0.686002

Publications Network Table

The publication Network Table contains the information of the publications interactions. It contains the publications variables, an the identifier of the publication and the publication that replicates the other.

It is the first part of the table, and represent the initial node of the network because each variable contains the letter *i* before the varaible name. The other part of the table is, uses *o* before the varaible name:

It is possible to have multiples authors from different institutions, however, those observations will be different and the iid to **iid**. The variable **pid** makes a publication unique. Moreover, it is possible to see the same pid for more than one iid in the table. At the beginning of the table will be add an interaction id. This makes the observation unique over the hole dataset.

The **code** variables refers to the industrial classification of the paper. Those are the most similar industry to the paper information, and the investigator can see the matching process on the *PUBLICATIONS DATA APPENDIX*.

The other variables are to show the location of the institution, or to show the author of the publication. The **iauthor:** The author of the publication. It could exists more than one iiauthor per ipid

Industrial Network

The industrial network contains data from the ecuadorian service taxation institutions: *Servicio de Rentas Internas* (SRI). The observations are aggregated by ecuadorian province, and the table shows the year and the product ISIC4 code. As in the publications table, the table divide the variables into *i,o* as input-output