Using web scraping techniques to derive co-authorship data: insights from a case study

Utilizzo di tecniche di web scraping per derivare reti di co-authorship: evidenze da un caso studio

Domenico De Stefano, Vittorio Fuccella, Maria Prosperina Vitale, Susanna Zaccarin

Abstract The aim of the present contribution is to discuss the first results of the application of web scraping techniques to derive co-authorship data among scholars. A semi-automatic tool is adopted to retrieve metadata from a platform introduced for managing and supporting research products in Italian universities. The co-authorship relationships among Italian academic statisticians will be used as basis to analyze updated collaborations patterns in this scientific community.

Abstract *Il presente contributo riporta i primi risultati delle procedure di web scraping utilizzate per ottenere dati sulla co-autorship tra studiosi. Uno strumento semi-automatico è stato utilizzato per estrarre i metadati delle pubblicazioni da una piattaforma online introdotta per la gestione dei prodotti della ricerca delle università italiane. Le relazioni di co-authorship tra gli statistici italiani saranno utilizzate come base per individuare i patterns di collaborazione recenti.*

Key words: Bibliographic data, Web scraping, Network data mining

Domenico De Stefano
Department of Political Science, University of Trieste e-mail: ddestefano@units.it

Vittorio Fuccella
Department of Informatics, University of Salerno e-mail: vfuccella@unisa.it

Maria Prosperina Vitale
Department of Economics and Statistics, University of Salerno e-mail: mvitale@unisa.it

Susanna Zaccarin
Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste e-mail: susanna.zaccarin@deams.units.it

1 Introduction

Several bibliographic sources are available online to retrieve co-authorship relations to analyze scientific collaboration among groups of scholars. Usually, scientific collaboration studies are based on international databases (e.g. ISI-WoS or Scopus) containing high-impact publications. Whereas the interest is to describe collaboration patterns among scholars involved in a field and/or affiliated to an institution, these sources can provide a partial coverage of their scientific production. Hence, it emerges the need to integrate these results by using local bibliographic archives. In addition, the complexities related to the scraping data process from heterogeneous online sources at international and national level, often including different data format, are well-recognized (7; 8; 6). Data mining tools have been introduced for this topic highlighting the strengths and weaknesses of each of them (7).

In this scenario, scientific production of the Italian academic scholars can be retrieved from general and thematic international bibliographic archives as well as national ones (2). To describe collaboration in the national academic community, publications can be collected from individual web pages ("sito docente"), managed by the Italian Ministry of University and Research (MIUR) and the Cineca consortium. Unfortunately, the access to this database is not freely available, due to privacy policies.

The recent introduction of the Institutional Research Information System (IRIS) developed by Cineca consortium seems to furnish a unique platform in Italy for managing and supporting research in academic and research institutions. Within this system, it is available an open archive module for the repository of the research products allowing the storage, the consultation and the enhancement of these outputs. Thanks to this tool, the affiliated universities can access to a system able to communicate with the national (i.e. "sito docente" of MIUR) and international databases for the management and dissemination of scholars' scientific publications. Considering the list of institutions affiliated to the IRIS platform,[1] 65 Italian universities out of 97 [2] (in which 67 are public universities, 19 private universities and 11 private online universities) adopt the IRIS platform for publications data storage.

The aim of the present contribution is to discuss the main results of the first step of web scraping procedures, before to obtain clean data to derive co-authorship relationships among scholars. A semi-automatic tool is adopted to retrieve publication metadata from the IRIS platform with the purpose of automatically extract the data from the system in order to obtain a good coverage of the author scientific production and reducing the manual adjustments to manage errors. The derived co-authorship relationships among Italian academic statisticians, considering the publications until 2017, will be used as basis to derive updated collaborations patterns in this scientific community. The contribution offers also the possibility to compare

[1] For detail see the IRIS webpage https://www.cineca.it/en/content/IRIS-institutional-research-information-system.

[2] For detail see http://www.miur.gov.it/istituzioni-universitarie-accreditate.

the results with the previous network analysis carried on the same target population with data updated to 2010 (2; 3).

The paper is organized as follows. Section 2 reports details on the population under analysis and on the web scraping procedure adopted to extract data from the IRIS system. Section 3 traces the directions of the further network analysis.

2 Web scraping techniques to extract publications of Italian statisticians

The present contribution deals with the retrieval of proper data to construct co-authorship network in Statistics starting from the IRIS online platform. In particular, we focus on the academic statisticians in Italy, that is, those scientists classified as belonging to one of the five subfields established by the governmental official classification: Statistics (Stat), Statistics for Experimental and Technological Research (Stat for E&T), Economic Statistics (Economic Stat), Demography (Demo), and Social Statistics (Social Stat). The target population is composed of the 721 statisticians who have permanent positions in Italian universities, as recorded in the MIUR database at July 2017. Table 1 reports the composition of the statisticians by Statistics subfields, gender, academic ranking and university geographic location. A comparison with the number of scholars by Statistics subfields with a permanent position in March 2010 is also given (Table 1 in De Stefano et al. 2, p. 373).

Table 1 Italian academic statisticians by Statistics subfields (%) in 2017 and (total) and 2010 comparison. Source: MIUR, March 2010 and July 2017

	All	Stat	Stat for E&T	Economic Stat	Demo	Social Stat
Gender						
Female	47.2	49.4	30.0	37.2	58.6	47.7
Male	52.8	50.6	70.0	62.8	41.4	52.3
Academic ranking						
Researcher	33.8	33.7	45.0	33.1	38.6	27.7
Associate professor	38.3	39.2	35.0	35.2	34.3	44.6
Full professor	27.9	27.1	20.0	31.7	27.1	27.7
University geographic location						
North	42.7	46.3	15.0	38.6	40.0	40.0
Center	25.7	24.0	20.0	31.7	27.1	23.1
South	31.6	29.7	65.0	29.7	32.9	36.9
Total (July 2017)	721	421	20	145	70	65
Total (March 2010)	792	443	30	160	85	74
Relative difference 2017-2010 (%)	-9.0	-5.0	-33.3	-9.4	-17.6	-12.2

Starting from scraping data techniques, a semi-automated tool based on two main steps is used to retrieve the publication metadata of the population of Italian aca-

demic statisticians in the IRIS platform. Indeed, each author has a page from which it is possible to access to the data of his/her publications. The tool is implemented in Java. Besides Java standard libraries to download Web pages, the Tagsoup library[3] is used for parsing well-formed or even unstructured and malformed HTML. This tool is programmed with the aim of automatically extract the data from the system obtaining a good coverage of the author publications and reducing the manual adjustments to manage errors or uncertainty conditions.

The input information is a table containing references (name, surname and academic institution) of the 721 statisticians.

In the *first step*, the URL of IRIS page is retrieved for each author. It is worth noting that each institution hosts a different deployment of the system, thus each statistician is linked to the index page of the IRIS deployment of his/her institution. Then a query is launched on a specific search by author interface available on the system. The interface responds by outputting a Web page containing a list of authors indicated by name and surname, each associated with a link to the author's page. The last name of the author is used as a query string and both author's name and surname are considered to match an item in the list. In the case of a single match, the link is directly captured. In case of no match or multiple matches, the procedure returned an error. As a result of the first step, the complete database of publications records for each author is available. The author's page contains the list of publications of which the person is co-author. If an author has more than 100 publications, these are necessarily split into multiple pages. Each publication in the list is associated to a link to a new page containing the details of the publication (title, authors, venue, year of publication and various identifiers –URL, DOI, ISI codes WoS and Scopus and so on).

In the *second step*, the proposed web scraping procedure retrieved and followed the links of each author publication in order to download these metadata. In a few cases, it will be necessary to manually retrieve the link to the author's publication pages to check the aforementioned errors and to integrate the retrieved metadata for the authors not found by the tool. The complete database of publication records for each author derived at step 2 reports information for the entire population of statisticians. It could contain many duplications in presence of publications co-authored by statisticians affiliated to different institutions. Indeed, in each IRIS system the same publication can be reported with a different format. To manage this issue, it is therefore necessary a further phase of record linkage, in which the records corresponding to the same publication are automatically reconciled. Where available, the correspondence of any of the aforementioned authors' identifiers allowed a sure identification of the match. In case of lack of unique identifiers, the records can be reconciled using information related to the title, list of authors and year of publication, following the procedure described in Fuccella et al. (3).

The aforementioned procedure allowed us to obtain a good coverage of the target population of authors in the first data extraction phase from IRIS platform. The metadata of the publications of around 80% of all statisticians is available from the

[3] For details see https://hackage.haskell.org/package/tagsoup.

extraction made in April 2018. After a manual check to integrate the retrieved metadata, we improved the coverage rate for all subfields (Table 2). This result seems almost in line with the authors' coverage obtained by using three different archives (Table 2 in De Stefano et al. 2, p. 374).

The tool returned zero publications for some authors, especially belonging to the Statistics and Economic Stat subfields. These errors are mainly due to the absence of the IRIS platform in some institutions (mainly private and online universities) and limitations to its access (i.e., a password is required to access to the system). Errors happened also when the author search returned more than one match.

Table 2 Statisticians found before and after manual check and descriptive statistics of the retrieved publications by Statistics subfields

	Authors			Publications				
	Found	Manual check	% found	Min.	Max.	Median	Average	Dev.ST.
All	555	82	88.3	1	333	50.0	56.4	38.0
Stat	319	54	88.6	1	292	49.0	54.9	34.6
Stat for E&T	18	2	100.0	8	126	46.0	55.5	34.1
Economic Stat	110	11	83.4	2	196	42.0	45.2	29.1
Demo	55	8	90.0	12	314	55.0	68.2	48.6
Social Stat	53	7	92.3	8	333	68.5	75.6	50.1

Focusing on the retrieved bibliographic data, several issues affect the data quality. First, the IRIS platform content is different and independent for each university and there is no automatic procedure that allows to match the same publication co-authored by authors enrolled in different institutions. Therefore for each co-authored publication, a number of duplication of this product equal to the number of the co-authors hired by different universities can be found. Second, the imputation of some crucial fields for the network construction –e.g. the name of the external authors– is up to the individual researcher and for this reason it can happen that same author names can be typed in several ways.

Both product and author name duplications should be addressed before the co-authorship network construction by adapting to this context the procedure proposed in Fuccella et al. (3).

3 Final remarks and future lines of research

The availability of the IRIS archive is certainly a promising tool but a lot of issues must be managed during the data collection process from this source. The unavailability of IRIS platform in some universities, the private access to the IRIS system in some cases, the different publication data format, and the presence of more than one record found for the same author are examples of errors obtained after the extraction of information. In addition, record linkage and author disambiguation processes

need to be taken into account to obtain co-authorship data among scholars affiliated in different universities.

After the data cleaning to reconcile publication records, the detection of duplicates and the recognition of internal and external authors of the same publications, the co-authorship networks will be derived. In line with the findings discussed in the previous contributions, the advancements of the current study will be devoted to two main directions of analysis. First, the interest is in discovering clusters of scholars through community detection algorithms, comparing results from two well-known community detection algorithms (4; 1), and a new proposed method based on an adaptation of modal clustering procedure (5). Second, the stability of research groups and of collaboration behaviors will be analyzed in order to capture the effect of research assessment exercises, introduced in Italy to evaluate researchers and their scientific production.

References

[1] Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. **2008**, P10008 (2008)

[2] De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. Social Networks. **35**, 370-381 (2013)

[3] Fuccella, V., De Stefano, D., Vitale, M. P., Zaccarin, S.: Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians. Scientometrics. **107**, 167-184 (2016)

[4] Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences. **99**, 7821-7826 (2002)

[5] Menardi, G., De Stefano, D.: Modal clustering of social network. In: Cabras, S. and Di Battista, T. and Racugno, W. (eds) Proceedings of the 47th SIS Scientific Meeting of the Italian Statistical Society, CUEC Editrice, Cagliari (2014)

[6] Mitchell, R.: Web scraping with Python: collecting data from the modern web. Packt Publishing, Birmingham (2015)

[7] Murthy, D., Gross, A., Takata, A., Bond, S.: Evaluation and development of data mining tools for social network analysis. In: Ozyer, T., Erdem, Z., Rokne, J., Khoury, S. (eds.) Mining Social Networks and Security Informatics, pp. 183-202. Springer, Dordrecht (2013)

[8] Vargiu, E., Urru, M.: Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artificial Intelligence Research. **2**, 44-54 (2013)