A Comparative Study of Group Profiling Techniques in Co-Authorship Networks

João Emanoel Ambrósio Gomes[1,2], Ricardo B. C. Prudêncio[2] and André C. A. Nascimento[3]
[1]Centro de Informática, Universidade Federal de Pernambuco (UFPE)
[2]Instituto Federal do Sertão de Pernambuco (IFSertão-PE), Campus Serra Talhada
[3]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco (UFRPE)
Email: {jeag, rbcp, acan}@cin.ufpe.br

***Abstract*—Group profiling methods aim to construct a descriptive profile for communities in complex networks. The application of such methods in the analysis of co-authorship networks enables us to move forward in understanding the scientific communities, leading to new approaches to strengthen and expand scientific collaboration networks. This task is similar to the document cluster labeling task, which encourages the adaptation of cluster labeling methods for group profiling problems. In this work, we present a comparative study of group profiling and cluster labeling algorithms in a co-authorship network. A qualitative survey was conducted to evaluate the generated profiles, as well as the pros and cons of different profiling strategies, were analyzed with concrete examples. The results demonstrated a similar performance of both group profiling and cluster labeling methods.**

I. Introduction

It has long been realized that the analysis of co-authorship graphs can help us to understand the structure and evolution of corresponding academic societies. Communities play a crucial role in co-authorship networks, since they reflect the basis of collaboration networks among authors. Though widely studied in its static and evolutionary aspects, little attention has been devoted to the exploration of the nature of such communities [1]. Even though social network analysis have been frequently using data from scientific collaboration to analyze these relations, most of these studies are focused on the prediction of new relationships between collaborators, i.e., link predicton [2].

There are several reasons that can lead the community formation in co-authorship networks [3]. Thus, an important question such "how scientific communities in a co-authorship network are built?", enable us to move forward in understanding the structure of this type of network. The investigation of their characteristics and peculiarities may eventually lead to new approaches to strengthen and expand scientific collaboration networks.

According to [4], a network can be divided into three regions: *singletons* (nodes that do not interact with other nodes), isolated communities, and a giant connected component. Many of real-world networks are composed by isolated communities, and the natural interconnection of such groups along time is rare. However, the characterization of such groups can enable an external agent to encourage the relations among similar groups, i.e., with the same interests. The process of extraction of descriptive attributes from a group of people is referred to as group profiling [5].

Given a network partitioned in communities, where each node is represented by a set of features, a group profiling task focuses in the automatic selection of the most descriptive user features for each group (Figure 1).

A descriptive analysis of communities can be done in basically three distinct ways [5]: Aggregation-based Group Profiling (AGP), Differentiation-based Group Profiling (DGP) and Egocentric Differentiation-based Group Profiling (EDGP). In the first, the objective is to find feature values that are most likely to occur within the group, ignoring the rest of the network; both the second and the third strategies aim to select features which differentiate one group from the others in the network. In the differentiation approach, all other users of the network are considered, while in the egocentric approach, only its neighbors (i.e., the *fringe*[1]) are taken into account.

A previous comparative study [7], considered only one method of each approach: Bi-standard separation (BNS) [8] (DGP and EDGP) and TF [9] (AGP). The authors conclude that the aggregation of individual features is applicable only in a relatively noise-free environment. But, if profiles are built over noisy attributes, such as user blog posts or self-reported interests, differentiation-based approaches consistently outperform the aggregation-based approach. Although it showed good results, the egocentric approach was less accurate than global differentiation methods. In [6], a new DGP method was proposed, the Wilcoxon Rank Sum Test (WRS), in a numeric attribute context. The studies on group profiling are still very limited. For example, previous works[7], [6], considered only one DGP method, analysing only BNS and WRS, respectively.

Regarding objectives, group profiling is very similar to a more established class of algorithms, known as cluster labeling. The main differences between them are the source of information (instance features for the first and both network and node features for the latter) and the process of identification of groups (traditional unsupervised learning, e.g., clustering, and network community detection). In fact, it is relevant to analyze the performance of consolidated methods of cluster labeling in the context of group profiling. Despite

[1]The fringe of a community P is defined as the set of all vertices not in P, that have at least one connection to members of P.

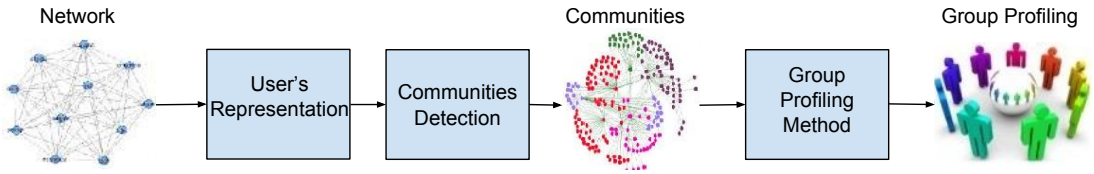

Fig. 1. An overview of the group profiling strategy [6].

some initial studies on the analysis of communities in co-authorship networks [3], there is no previous study of group profiling in such networks.

To address the limitations of previous studies, this work presents an evaluation of differentiation-based methods for the characterization of co-authorship network communities. In such setting, network nodes represent authors, endowed with side information, which is encoded as textual attributes extracted from the titles and abstracts of the published papers. About the information used to differentiate the groups, two distinct classes of methods were considered in the present study: network and non-network (i.e., cluster labeling) based methods. The first class of methods incorporate network structure in the process of group profiling. In this work, two distinct network-based methods were considered: the BNS [8], and a variation of the WRS defined in [6]. Non-network methods comprise methods that do not consider any network information in the group profiling task, e.g., cluster labeling strategies. From this class, two distinct methods were considered: a method based only on term frequencies, Term Frequency - Inverse Document Frequency (TF-IDF) [9], and the *Chi-Squared Selection* (χ^2) [10].

Experiments were performed using data collected from the ArXiv repository[2]. This dataset is maintained by Cornell University, and contains bibliographical records about thousands of pre-print scientific papers. In this study, a subset of Artificial Intelligence related articles were considered, from which a co-authorship network was extracted. A community detection algorithm [11] was then applied, resulting in a total of 10 communities, which were considered in the Group Profiling strategies. In order to evaluate the final profiles assigned by the four methods considered in this study, a total of 340 responses were collected in a survey, where each response corresponds to the profile that best represented a given community, among four distinct profiles.

The remaining of this paper is organized as follows: problem statement is firstly defined in Section II. In Section III, the methods considered in this study are described in more detail, followed by Section IV, in which the experimental methodology is presented. In Section V, the results and the discussion are presented. Finally, in Section VII presents some conclusions and point some future works.

II. Problem Statement

In this section, a formal description of the characterization of densely connected subgraphs (communities) problem is given. To this end, attribute data are modeled together with graph data. Formally, the input is a graph $G = (V, E)$ with vertices $V = \{\mathbf{v}_1, \mathbf{v}_1, ..., \mathbf{v}_n\}$ and edges $E \subseteq V \times V$. Additionally, each vertex is represented by a d-dimensional vector, $A \in \mathbb{D}^d$, where $\mathbb{D}$ comprises the attribute domain (e.g., $\{0, 1\}$ or $\mathbb{R}$). We assume an undirected graph without self-loops, i.e. $(v, u) \in E \iff (u, v) \in E$ and $(u, u) \notin E$.

A group/community is represented by a subgraph $P = (V_P, E_P)$, where $V_P \subseteq V$, $E_P \subseteq V_P \times V_P$ e $E_P \subseteq E$. In such setting, the characterization of a given group is defined as the vector of attributes $\mathbf{c}_P \in \mathbb{D}^k, k \leq d$. This way there is a total of $\binom{d}{k}$ distinct characterizations for each group. The objective is to select the best k descriptive attributes for each partition $\mathbf{c}_P$. For such, one can define a scoring function f, to assign the importance (i.e., descriptive score) for each attribute in a given partition, and then select the top-k scored attributes.

III. Methods

A brief overview of the differentiation methods considered in this study is presented in this section: a profiling strategy based on TF-IDF [9] features, χ^2 [10], WRS [6] and BNS [7]. Also, a modification of the original method proposed in [6] to the context of text-related attributes is described.

A. TF-IDF

TF-IDF is a numeric statistic usually used in information retrieval and text mining applications. The objective is to identify how important a word is to a document in a collection or corpus. In the group profiling context, labels are generated by the combination of intra-cluster (TF) and inter-cluster (IDF) features. In this way, the importance of a given feature/word A to label the group (P) is given by:

$$f_{TF-IDF}(A, P) = tf_{A,P} \times idf_A, \tag{1}$$

where $tf_{A,P}$ corresponds to the frequency of occurrences of feature A in group P, and idf_A is a measure of the general importance of the feature obtained by weighting how distintive the feature A is, in relation to the rest of the network (G'). Thus, attributes that present high (local) frequency in P, and simultaneously have low frequencies in G' (overall low frequency) receive higher scores. It is important to note that, for textual attributes, we have as input a term frequency matrix (see [9] for a more detailed description).

B. Chi-Square Selection (χ^2)

The Pearson's chi-squared is a statistical test used to evaluate how likely it is that the occurrence of an event matches the initial expectations [8]. In particular, it can be used to

[2]https://arxiv.org/

determine whether two events, X and Y, are statistically independent. In the case of differential cluster labeling, let X be a variable associated with the membership in a group (P) and Y a variable associated with the presence of a feature (A). Let also both P and A be binary variables (0 or 1), the Pearson's chi-square equation can be rewritten as follows:

$$f\chi^2 = \sum_{P\in 0,1} \sum_{A\in 0,1} \frac{(O_{P,A} - E_{P,A})^2}{E_{P,A}}, \quad (2)$$

where, $O_{1,0}$ is the observed number of nodes in a particular group that do not contain a certain feature, and $E_{1,0}$ the expected number of nodes in a particular group but don't contain a certain feature. The initial assumption is that the two events are independent, so the expected co-occurrence probabilities can be calculated by multiplying individual probabilities

$$E_{1,0} = N * P(P = 1) * P(A = 0), \quad (3)$$

where N is the total number of nodes in the network. Here we find the dependence score of the feature (A) in the group (P) and in the rest of the network (G'), noting that for the rest of the network we have G' instead of P in the Equations 2 and 3; by selecting the attributes (C_P) with greater reliance score on group compared to the rest of the network. In this method for textual attributes we considered as input a terms binary matrix.

C. Bi-standard separation (BNS)

The BNS [7] proposes an optimization strategy for the group profiling problem. The method adopt binary classification concepts in order to differentiate a given partition from the rest of the network. It considers the group (P) as the positive instances (denoted "+") and the other nodes that do not belong to the group (G') as the negative instances (denoted as "−"). In this scenario, for a given A, true positive (*tp*) corresponds to the number of positive instances containing feature A; true negative (*tn*) to the number of negative instances not containing feature A; false positive (*fp*) the number of negative instances containing feature A, and, false negative (*fn*) to the number of positive instances not containing feature A.

The true positive rate (*tpr*) is viewed as the conditional probability of a given feature (i.e., A) occur in a group, while the false positive rate (*fpr*) is the conditional probability of a feature to occur outside the group:

$$tpr = P(A|+) = \frac{tp}{tp + fn}, \quad (4)$$

$$fpr = P(A|-) = \frac{fp}{fp + tn} \quad (5)$$

Then, the most relevant attributes to describe a given group are the ones with the highest k tpr values. In other words, feature A should better explain the positive class rather than the negative class. The calculation of the score a feature applying BNS is defined as:

$$f_{BNS} = \quad |F^{-1}(tpr_A) - F^{-1}(fpr_A)|, \quad (6)$$

where F^{-1} is the inverse cumulative probability function of a standard normal distribution [8] and $tpr_A \geq fpr_A$. Essentially, only those features that frequently appear in one group but rarely outside the group are selected.

D. Wilcoxon Rank Sum Test (WRS)

In [6], the Wilcoxon test was adapted to extract group profiles in social networks. The approximate *p*-value of the test is computed by a z-statistic [12]. In the context of group profiling, the z-statistic can be calculated as follows: given the sizes of one partition P and of the rest of the network G' (i.e., $G' = G - P$), where $|P| < |G'|$:

$$Z \quad = \quad \frac{R - \mu_R}{\sigma_R}, \quad (7)$$

where R is the sum of ranks of the feature (A) values in the partition P, and μ_R and σ_R are respectively the mean and standard deviation.

Unlike [7], the approach proposed in [6] has as input distributions of attributes (numerical data). In the present work, an adaptation is proposed, in order to make use of textual features/terms. As the purpose is to characterize the group, a constraint is added to the initial selection of attributes, i.e., the average of the attribute in the group (m_P) must be larger than the comparison sample (e.g., the rest of the network, $m_{G'}$), i.e., $m_P > m_{G'}$. Thus, the z-statistic compares the feature values distributions of two distinct vertex sets, in a way that the k features with smaller *p*-values(score) are selected (C_P).

IV. Experiment Setup

This section describes the co-authorship network data set, the community detection procedure and the evaluation strategy to analyze the profiles assigned by the considered group profile methods.

A. Data Set

As aforementioned, to conduct a group profiling study, a suite of related data on individual attributes is necessary. Hence, we selected the arXiv co-authored network data in our case study. Maintained by Cornell University, this data set contains millions of bibliographical records and pre-print scientific papers, mostly in mathematics, computer science, biology, finance and statistics.

Seeking to facilitate the meeting of qualified evaluators for experiments, only papers in the field of Artificial Intelligence, published between 2012 and 2014, were considered in this study. In the constructed network, each node represents an author, and two nodes are connected if they have co-authored at least one paper. The resulting network contained 1850 authors with 2560 relationships. Major network statistics are presented in Table I.

B. Authors' Representation

Each network node (i.e., author) is associated with side information, describing the articles he has authored. To do so, several text processing procedures were performed. Firstly, all published papers of each author were collected and combined into a single document. Then, a series of pre-processing steps were applied for each document: tokenization; removal of stopwords[3]; stemming (reducing inflected (or sometimes derived) terms to their word stem, base or root form); removal of non-nouns and non-adjectives [13]; and, finally, composition of n-gram composed words ($n = \{1, 2, 3\}$), increasing the attribute vector (A) for $3d - 3$ dimensions.

C. Community Detection

Since no explicit community has been defined in arXiv co-authored network yet, the application of external algorithms to identify communities groups was mandatory. Before that, all singletons (i.e., nodes that do not interact with anyone) were removed. Then, the Gephi [14] implementation of the Multi-level Aggregation Method [11], was applied to the resulting network, resulting in a total of 439 identified groups. This high number of communities is mainly because of the high number of small groups of nodes with very few connections, not attaching to any other larger community.

Groups that had fewer than ten users were removed since they were considered too small and irrelevant for the study. Also, the communities were filtered according to their density values, resulting in 10 remaining groups. Table I shows the final network statistics.

TABLE I
NETWORK RELATED MEASURES OF THE INITIAL ARXIV CO-OCCURRENCE NETWORK AND AFTER FILTER APPLICATION

Measure	Original	Filtered
#Authors	1850	372
#Links	2560	654
Link Density	0.001	0.009
Average Link	2.768	3.516
Diameter	19	19
Number of Groups	439	10

The final communities can be visualized in Figure 2. The statistics of each group are presented in Table II, including the size, density, average degree and cohesion of each group. Cohesion was calculated based on the cosine metric [15], i.e., the average similarity of each paper of the group with all other articles. It is also possible to identify in the table groups that are more dense than others, for example, groups 145 and 134.

D. Evaluation

To achieve a clear notion of quality of the generated labels, all competing approaches were evaluated under the same conditions, i.e., the same network, communities and node representation. As there is no guideline or gold standard for the evaluation of the detected profiles, we resort to a human blind selection of the best labels for each group. Each evaluator was presented to a form containing:

1) The titles of the ten most cohesive papers[4] in the group (with a link to the abstract in arXiv web page). This selection was necessary since it is impossible for evaluators to consider all papers simultaneously[5];
2) A table with the generated profiles, i.e., the ten most representative terms, detected by each method (one per column).
3) Evaluation question: "Based on these articles, which method produced the best profile for the group?"
4) Finally, a space for the selection of the best method.

On each evaluation page, the four methods were denoted as "Method I", "Method II", "Method III" and "Method IV", as well as the presentation order of group profiles was randomized for each page, to avoid the bias associated with the method names.

TABLE II
STATISTICS ON GROUPS

Group	Size	Average Degree	Density	Cohesion
6	41	2.829	7.1%	0.248
80	28	2.929	10.8%	0.384
104	68	3.735	5.6%	0.307
116	28	3.929	14.6%	0.363
134	60	3.167	5.4%	0.299
145	14	3.429	26.4%	0.377
151	18	3.333	19.6%	0.346
153	53	3.321	6.4%	0.337
156	29	3.724	13.3%	0.352
256	33	3.273	10.2%	0.316

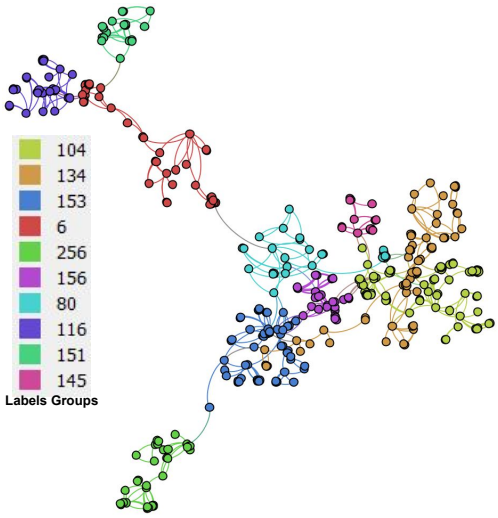

Fig. 2. Resulting network and detected communities.

[3]Stopwords is a list of all non-informative terms in a document, usually composed of prepositions, articles, adverbs, numbers, pronouns and punctuation.

[4]One paper is considered cohesive if it presents great similarity to the content found in the group.

[5]As we notice in one pilot study, subjects tend to assign random ratings if the task takes too long.

V. Experiment Results and Discussion

A total of 34 people with diverse backgrounds (undergraduate, graduate students, university faculty) participated in the survey, which resulted in 340 evaluations. The percentage of answers where each method was marked as "best labels" is presented in Table III.

TABLE III
PERCENTAGE OF ANSWERS WHERE EACH METHOD WAS MARKED AS "BEST LABELS" (AVERAGE OVER ALL GROUPS).

WRS	Chi-Square	BNS	TF-IDF
23.82%	27.65%	17.66%	**30.9%**

Although TF-IDF method is one of the simplest methods, it achieved the highest overall performance in the survey. In contrast, the BNS method obtained the lowest overall performance. The methods WRS and Chi-Square were quite similar, in average. The WRS was pointed as the best profile in groups 134 and 151, while the Chi-Square was the best in groups 104, 153 and 156 (Details Figure 3). The group 145 did not have one major best profile among raters, presenting equal rates for methods WRS, BNS, and TF-IDF (i.e., good profiles generated). This is precisely the group of higher density, which demonstrates the importance of this metric in the performance of all methods since this represents the degree of connectivity of the authors group.

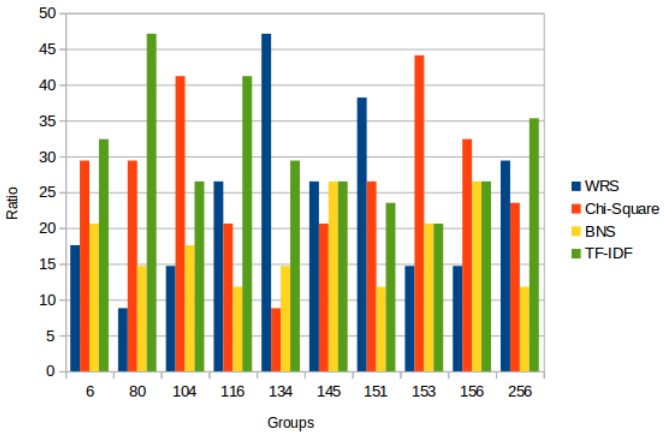

Fig. 3. Bar graph with the results for each group.

To assess whether there were statistical differences between the methods, the non-parametric Friedman test [16] was applied ($\alpha = 0.05$). Although this is a relatively conservative test, it allows the comparison of multiple methods, checking whether there is statistically significant difference between them. Since no significant differences were detected between the methods ($p = 0.753$), the post-hoc test (Nemenyi test) was not applied. This can be in part explained by the low number of samples (i.e., groups) used as input for the test. However, this demonstrates the feasibility of the methods for group characterization (or group profiling) in complex networks making use of textual attributes.

In order to check if the profiles indicated by each method are similar or not, the generated profiles of each method were also compared. For such, the Jaccard similarity coefficient [15] was applied, which returns a similarity measure between finite sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. This analysis showed that all methods returned similar profiles in average, with TF-IDF and Chi-Square being slightly more different (Table IV). It is important to note that, for some groups, these two methods returned even more similar profiles ($S_{Jaccard} = 0.4280$).

TABLE IV
JACCARD COEFFICIENT OVER GENERATED PROFILES.

	WRS	CHI-SQUARE	BNS
CHI-SQUARE	0.0052	-	-
BNS	0.0000	0.0491	-
TF-IDF	0.0458	0.2300	0.0105

To perform an analysis of the output of different methods, here we show two concrete examples: groups 80 and 134. Table V presents profiles extracted to describe these groups based on titles and abstracts of authors' papers. The features are sorted by alphabetic order, seeking to facilitate a comparison between the profiles generated by each method.

Group 80 has 28 authors and is the most cohesive group among all (Table II). It can also be noted that the characteristics selected by WRS and BNS (Table V) do not have very much relevance for the group description, for example, 'evidence', 'fault', 'positive', 'convention', 'continuous', 'control', 'time', 'utility' and so forth. However, the Chi-Square and the TF-IDF methods have many terms in common, as already indicated by the Jaccard similarity coefficient. Among the terms selected by Chi-Square method, one can find 'Bayesian', 'Bayesian networks', 'probabilistic', reflecting studies in Bayesian learning. This turns out to slightly less specific in the profile generated by the TF-IDF method, with the addition of other terms, such as 'models' and 'probability'. Thus, one can conclude that the group 80 is a machine learning community based on probabilistic models, focused on the study of Bayesian networks and Stochastic and dynamic models of learning.

On the other hand, group 134 is the second largest group, with 60 authors and the lowest density (5.4%). The WRS method has proved robust in the characterization of this group, despite the low density, detecting good characteristics such as: 'DEC-POMDPs'[6] (decentralized partially observable Markov decision process), 'MER'[7] (Most Relevant Explanation), 'observable Markov decision', 'optimal policy' (multi-agent system algorithm), 'multiagent' and 'heuristic'. All other methods generated surface profiles, abstracting the central content of the group, with small differences observed in the TF-IDF profile, which adds some more specific terms as 'linear programming'

[6]Is a very general model for coordination among multiple agents

[7]method to automatically identify the most relevant target variables in forming its explanation.

TABLE V
PROFILES FOR GROUPS 80 AND 134 IN ARXIV.

Group 80				Group 134			
WRS	Chi-Square	BNS	TF-IDF	WRS	Chi-Square	BNS	TF-IDF
convention	bayesian	continuo	algorithms	DEC-POMDPs	algorithms	algorithms	basis
effort	bayesian networks	control	approximate	heuristic	approach	approach	circuit
engineering	belief	empirical	bayesian	MRE	decision	decision	concept
evidence	inference	expert	bayesian network	multiagent	functions	heuristics	hierarchy
fault	method	method	inference	observable markov decision	optimality	method	kernel
inference bayesian	networks	powerful	method	online	papers	model	linear programming
knowledge engineering	papers	simple	models	optimal policy	representations	networking	practice
positive	probabilistic	system	networks	policy	sampling	optimal	reward
sensitivity analysis	probability	time	probabilistic	search	search	planning	suitable
SPI	utility	utility	probability	speedups	value	policy	value function

and 'kernel'. Based mainly on the profile generated by the WRS method, we can interpret the group 134 as a community focusing on studies of Knowledge-Based Agents and multi-agents systems, considering the technical applications such as: MER, Markov models and the use of heuristics functions. This small analysis encourages further studies on the influence of density on the performance of group profiling methods.

VI. CONCLUSIONS AND FUTURE WORK

In this work, a comparative study with four differentiation-based group profiling methods, WRS, TF-IDF, BNS and X^2, was performed. To do so, profiles of 10 distinct co-authorship groups were generated automatically. The results demonstrated a very similar performance of the four evaluated methods, although some group characteristics, such as cohesion and density community, had an influence on the performance of some methods.

To the best of our knowledge, this was the first time that cluster labeling techniques were compared against group profiling algorithms. It also introduces group profiling techniques to the analysis of co-authorship scientific collaboration data. These insights help to explain why the authors connect and interact with them in authorship network, may eventually lead to new approaches to strengthen and expand scientific collaboration networks.

As future work, adaptations of the evaluated group profiling methods under an egocentric differentiation perspective can be done, minimize the imbalance among the classes, as well as improvements to the textual preprocessing.

ACKNOWLEDGMENT

The authors would like to thank CNPq (Brazilian Agency) for its financial support.

REFERENCES

[1] D. Martin-Borregon, L. Aiello, P. Grabowicz, A. Jaimes, and R. Baeza-Yates, "Characterization of online groups along space, time, and social dimensions," *EPJ Data Science*, vol. 3, no. 1, 2014.

[2] L. L and T. Zhou, "Link prediction in complex networks: A survey," *Physica A*, vol. 390, no. 6, p. 11501170, 2011.

[3] M. Savić, M. Ivanović, M. Radovanović, Z. Ognjanović, A. Pejović, and T. Jakšić Krüger, *ICT Innovations 2014: World of Data*. Cham: Springer International Publishing, 2015, ch. Exploratory Analysis of Communities in Co-authorship Networks: A Case Study, pp. 55–64.

[4] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 611–617.

[5] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno, "Topic taxonomy adaptation for group profiling," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 4, pp. 1:1–1:28, Feb. 2008.

[6] J. Gomes, R. B. C. Prudêncio, L. Meira, A. A. Filho, A. C. A. Nascimento, and H. Oliveira, "Group profiling for understanding educational social networking," in *The 25th International Conference on Software Engineering and Knowledge Engineering, Boston, MA, USA, June 27-29, 2013.*, 2013, pp. 101–106.

[7] L. Tang, X. Wang, and H. Liu, "Group profiling for understanding social structures," *ACM Trans. Intell. Syst. Technol.*, vol. 3, pp. 15:1–15:25, 2011.

[8] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.

[9] P. Treeratpituk and J. Callan, "Automatically labeling hierarchical clusters," in *Proceedings of the 2006 International Conference on Digital Government Research*, ser. dg.o '06. Digital Government Society of North America, 2006, pp. 167–176.

[10] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[11] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, p. 8, 2008.

[12] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, pp. 80–83, 1945.

[13] A. Barrera and R. Verma, *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Combining Syntax and Semantics for Automatic Extractive Single-Document Summarization, pp. 366–377.

[14] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," 2009.

[15] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," 2007.

[16] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, "Learning probabilistic relational models," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, ser. IJCAI '99, 1999, pp. 1300–1309.