2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Comparison of co-authorship networks across scientific fields using motifs

Sarvenaz Choobdar, Pedro Ribeiro, Sylwia Bugla and Fernando Silva
CRACS and INESC-TEC
Faculdade de Ciencias, Universidade do Porto, Portugal
Email:{sarvenaz,pribeiro,syl-via,fds}@dcc.fc.up.pt

***Abstract*—Comparing scientific production across different fields of knowledge is commonly controversial and subject to disagreement. Such comparisons are often based on quantitative indicators, such as papers per researcher, and data normalization is very difficult to accomplish. Different approaches can provide new insight and in this paper we focus on the comparison of different scientific fields based on their research collaboration networks. We use co-authorship networks where nodes are researchers and the edges show the existing co-authorship relations between them. Our comparison methodology is based on network motifs, which are over represented patterns, or subgraphs. We derive motif fingerprints for 22 scientific fields based on 29 different small motifs found in the corresponding co-authorship networks. These fingerprints provide a metric for assessing similarity among scientific fields, and our analysis shows that the discrimination power of the 29 motif types is not identical. We use a co-authorship dataset built from over 15,361 publications inducing a co-authorship network with over 32,842 researchers. Our results also show that we can group different fields according to their fingerprints, supporting the notion that some fields present higher similarity and can be more easily compared.**

***Keywords*-network comparison; co-authorship network; collaboration pattern; network motifs; motif profile;**

I. Introduction

Understanding the similarities and differences in the process of scientific production of different fields of knowledge is one of the traditional debates in science [1]. Co-authorship networks emerge as a powerful concept to gain a new insight in this area. They are one of the most active and well studied form of social networks and their statistical properties were already being studied in the mid-1970s [2]. We believe that the analysis and comparison of co-authorship networks, specific to each scientific field, can help in understanding the differences in research production across different fields.

Network comparison can be defined as the process of contrasting two or more networks, and is a common approach in complex networks analysis. The main goal is to find topological similarities that might explain, for example, equivalent functionality. Another possible goal is to "align" networks, by finding groups of nodes or edges that are likely to have a similar "position" or function in different networks. Most of the existing network comparison and alignment methods were designed for biological data, and make use of the biological context in order to analyze the data [3], [4]. A survey of the existing network comparison methodologies from a biological perspective can be found in [5].

Regarding more specific topological comparison methodologies, the most frequent and traditional approach is to use global graph metrics such as degree distribution. This is the approach considered, for example, by Newman [6]. He analyzes three different scientific fields using properties such as the number of co-authors per paper, shortest path between two authors, or clustering coefficient. Recently, local properties have been gaining increasing attention, and this is precisely the approach we take in this paper, by looking for characteristic collaboration patterns in the form of different subgraphs.

There are several possible methodologies for incorporating subgraphs in a network comparison framework. Pržulj introduced "graphlet degree distributions" (GDD) [7], by defining graphlets as small, connected, induced subgraphs. GDD counts the number of graphlets in which the node participates, trying to generalize the degree distribution definition. Kuchaiev et al. used graphlets to propose a new global network alignment algorithm relying solely on topology [8]. They define graphlet degrees *signatures* for each node, by counting its participation in k-sized graphlets, for a varying k. Milenković et al. further improved this alignment approach by using a greedy "seed-and-extend" approach, leading to an optimal alignment algorithm [9] by using the Hungarian algorithm [10].

Network motifs are another possible angle, and can be defined as small subgraphs that appear in a network at significantly higher frequencies than what would be expected in randomized networks [11]. They can really encapsulate and characterize the structure of networks, and motif profiles have been shown to be a very powerful measurement [12]. The correlation of small network motifs (of sizes 3 and 4) with citation frequencies in co-authorship networks has already been the subject of a study [13].

In this paper, we compare co-authorship networks in different scientific fields based on their motif profiles. Our goal is to discover the set of network motifs that best characterize and discriminate a certain scientific domain. Motif shapes are correlated to the collaboration patterns found in a co-authorship network, and we divide and group the networks based on their similar collaboration patterns. Our results show that this approach can identify relevant patterns and that we are able to distinguish different models of collaboration across the scientific fields. We are also able to discover the motifs

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.34

TABLE I: Statistical properties of the considered co-authorship networks.

Network name	Num Edges	Num Nodes	Avg Degree	Clustering Coefficient
Agricultural Sciences	2099	1086	3.866	0.706
Biology & Biochemistry	6242	3029	4.121	0.734
Chemistry	8409	3284	5.121	0.796
Clinical Medicine	31751	5884	10.792	0.867
Computer Science	2806	1731	3.242	0.717
Economics & Business	945	292	6.473	0.784
Engineering	7069	3294	4.292	0.756
Environment/Ecology	3269	1740	3.757	0.762
Geosciences	2379	1082	4.397	0.785
Immunology	9984	2393	8.344	0.776
Materials Science	3727	1707	4.367	0.727
Mathematics	1171	911	2.571	0.626
Microbiology	3561	1819	3.915	0.767
Molecular Biology & Genetics	13205	3879	6.808	0.819
Multidisciplinary	3297	1229	5.365	0.771
Neuroscience & Behavior	4295	1896	4.531	0.781
Pharmacology & Toxicology	3550	1815	3.912	0.665
Physics	6107	2226	5.487	0.789
Plant & Animal Science	12010	4166	5.766	0.814
Psychiatry/Psychology	5717	1203	9.505	0.800
Social Sciences, general	6549	2078	6.303	0.732
Space Science	42609	1680	50.725	0.916

with more discriminating power, whose combination in the right proportion constitutes a fingerprint of the network, that could can be used to infer the scientific field based on its co-authorship network.

II. Comparison of co-authorship networks across scientific fields

A. Data

In this paper, we use a set of 15,361 publications authored by researchers from the University of Porto, ranging from 2003 to 2011. These are publications drawn from ISI Thompson Web of Knowledge and they induce a co-authorship network with 32,842 nodes (researchers). The association between publication authors and researchers at the university is done automatically by a specialized name identification system built by our group [14].

There are many possible categorizations for academic scientific fields. In the present study, we use the ISI categories [1] that divides science domains into 22 subjects or fields. An undirected co-authorship network is built for each field in which nodes are the authors of publications in that field, and edges are created whenever two different researchers appear as co-authors of the same publication.

Table I shows the statistical properties of the 22 co-authorship networks created for each field, including number of edges, nodes, average degree and clustering coefficient.

B. Network Motif Mining

We use g-tries [15] to find the motif profiles of each network. G-tries are a tree-shaped data structure designed to store collections of subgraphs and to find their frequency on a larger graph. By encapsulating and identifying common sub-substructures, and by using symmetry breaking conditions, g-tries give us a very efficient subgraph counting algorithm.

For the purposes of this work, we consider all possible undirected subgraphs of sizes 3 to 5. We apply the same process, as described next, for each scientific field. We start by computing the frequency of the subgraph set in the original co-authorship network. We then produce a large set of similar random networks and compute the frequency of the subgraphs in these networks. These randomized networks keep the exact same degree sequence of the original network, with each node preserving its degree, but with different connections. For that purpose, we use a Markov-Chain methodology in which we start with the original network and repeatedly swap the endpoints of edges, preserving their degree.

We use 100 random networks for each scientific field and compute the significance of each subgraph by comparing its frequency in the original network and in the randomized networks. The significance of a subgraph G_k is measured in terms of a z-score, as depicted next, where $\overline{f_{random}}$ and $\sigma(f_{random})$ are respectively the average and standard deviation of the frequency in the randomized networks.

$$z\text{-}score_k = \frac{f_{original}(G_k) - \overline{f_{random}(G_k)}}{\sigma(f_{random}(G_k))}$$

The motif profile of a network consists of the values of the z-score for each subgraph type.

C. Significance analysis of Motifs

Before comparing the networks, we first assess if all the motif types from size 3 to 5 have the same importance in terms of differentiating co-authorship networks. In other words, the question is to know which subgraphs (or motifs) are more particular and can better differentiate scientific fields. To answer this, we examine the discrimination power of each

[1] http://sciencewatch.com/about/met/fielddef/

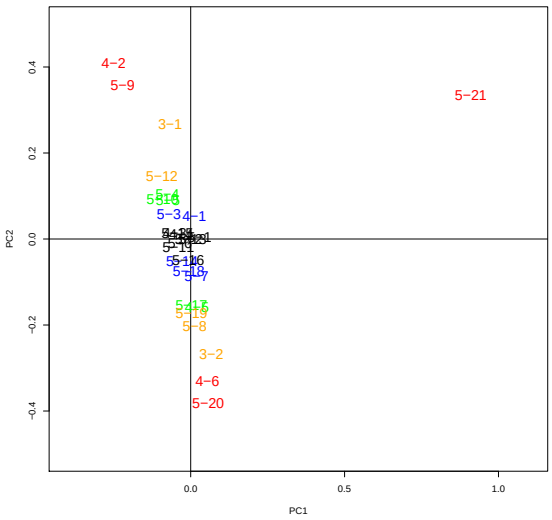


Fig. 1: Coefficients of network motifs in the two first principal components.

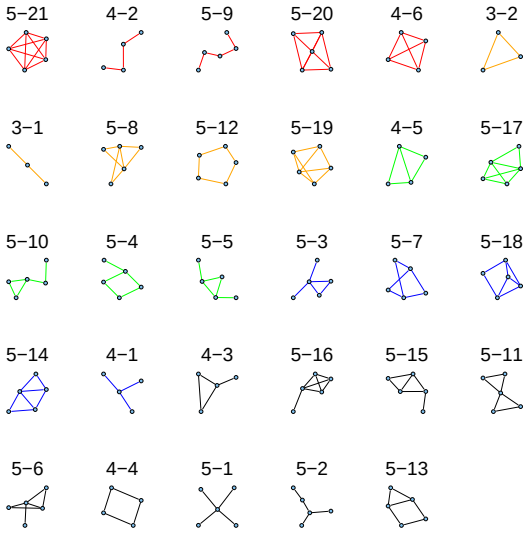


Fig. 2: Network motifs ranked by their importance in collaboration models.

motif. We use principle component analysis (PCA) [16] to rank the motifs according to their significance in co-authorship networks. PCA is a simple, non-parametric method for revealing hidden information in the data. The primary benefit of this method is dimension reduction. However, it is also useful for measuring the importance of each dimension in explaining the variability of a data set.

Our PCA analysis over motif profiles of co-authorship networks shows that two principle components can capture 100 percent of variation in the motif profiles of the networks. The coefficients along each component provide the relative importance of each motif. The higher the coefficient, the more important the motif is. Fig. 1 shows these coefficients for the two first components. We ordered the motif types regarding their importance as depicted in Fig. 2. The measure for the importance of a motif is the squared sum of its coefficients, for the first and the second principle components. The first five most distinguishing motifs are (5-21), (4-2), (5-9), (5-20) and (4-6), highlighted in red color in the figures. Motifs having comparable importance are depicted in the same color.

D. Network Comparison

We compare co-authorship networks from different scientific fields to determine similarity among fields in terms of collaboration patterns. Co-authorship analysis enables the identification of groups of scientists and their model of collaboration across research disciplines. In order to be able to compare the networks, we determine the fingerprint of each co-authorship network, one for each scientific field, as a motif profile [12], which basically is a vector of the normalized z-scores of the motifs of sizes 3 to 5.

The networks are then compared against their motif profiles and for the purpose of this comparison, the distance of two networks i and j is calculated as:

$$D(i,j) = \sqrt{\sum_k (S_{i,k} - S_{j,k})^2} \ .$$

In our calculation, we use normalized values of motif significance to remove the effect of high values, by dividing them by their averages:

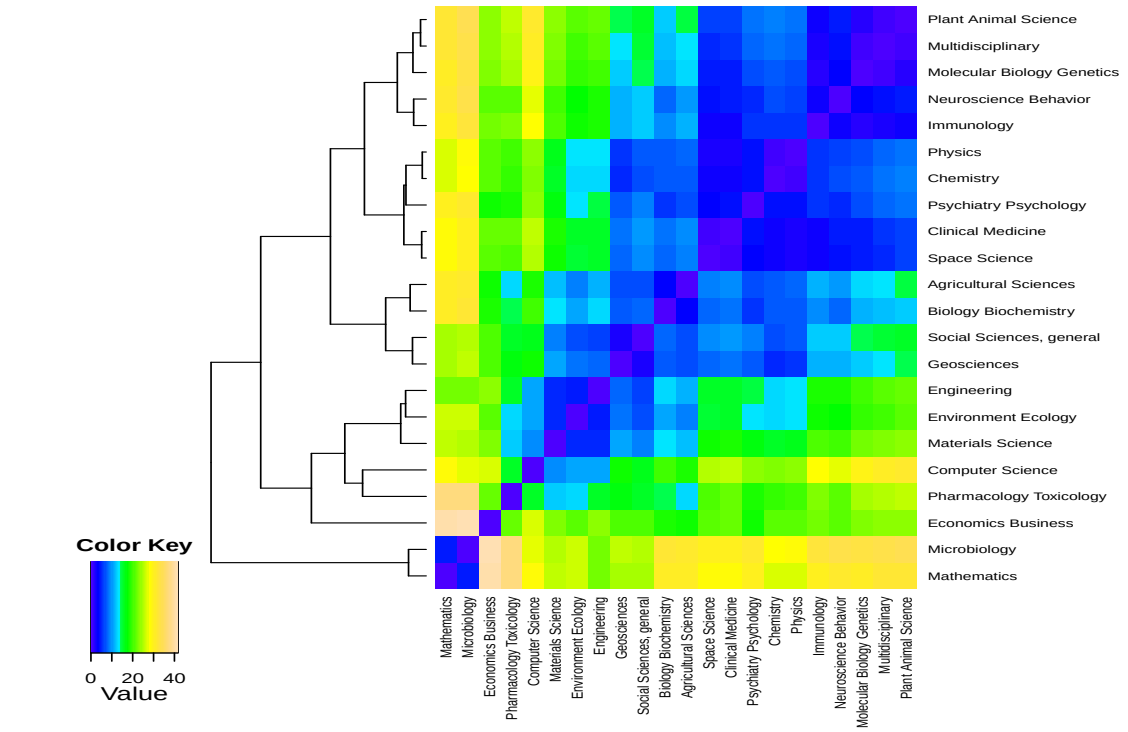
$$S_{i,k} = z\text{-}score_{i,k} / \overline{z\text{-}score_i}$$

The distance metrics of networks are used to group them. A suitable approach is to use hierarchical clustering, since it has the advantage of providing a general picture of all possible groupings for the objects at different distance level over other clustering methods. Therefore, the desired grouping can be derived by using the similarity, or distance, between networks. We use agglomerative hierarchical clustering methods [17] to divide networks into homogeneous groups holding similar patterns of collaboration. This method begins with as many clusters as objects. Two clusters are separated by the smallest distance and are successively merged until only one cluster remains.

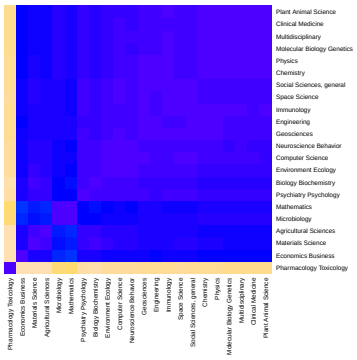
E. Results

Fig. 3 shows the overall results of the hierarchical clustering process for co-authorship networks. Networks are grouped together if they have a similar motif profile, meaning that the same subgraphs are more significant in these networks.

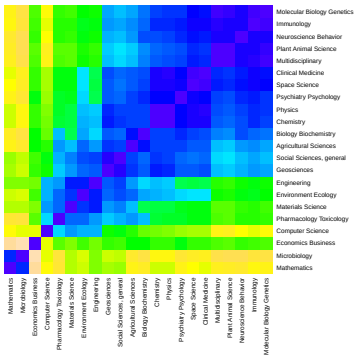
In section II-C, we found that the most important motif types have five nodes, mostly revealing the variation in the motif profiles of co-authorship networks. To further confirm this result, cluster analysis is done for different combination of motif sizes, by discarding one size at a time. In this way, we also compare networks against motif sets $\{3,4\}$, $\{3,5\}$, and $\{4,5\}$, and the results can be seen in Fig. 3. By leaving out motifs of size 5 (Fig. 3b), almost no grouping is derived, meaning that the structure of networks is not different



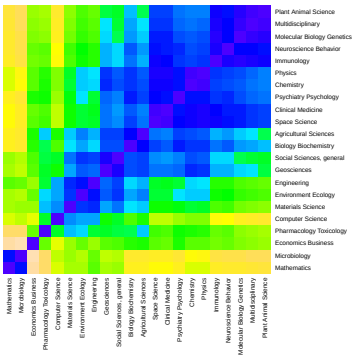
(a) Motifs of sizes 3 to 5.



(b) Motifs of sizes 3 and 4.



(c) Motifs of sizes 3 and 5.



(d) Motifs of sizes 4 and 5.

Fig. 3: Similarity matrix of scientific fields.

regarding motif types from size 3 and 4. By incorporating the motifs of size 5 in the comparison as depicted in Fig. 3c and Fig. 3d, the groupings appear. This shows that going for larger motif sizes does indeed provide more information and that even going for a small increment in the number of nodes may result in new and insightful information. In this case the difference between sizes 4 and 5 is highly significant.

To have a comprehensive comparison of co-authorship networks, we included all 29 motif types in the hierarchical analysis. Four groups of scientific fields are more distinct in this clustering, depicted in Fig. 3a. Drilling down to the motif profiles for the four identified groups, as depicted in Fig. 4, we can observe that the patterns of collaboration vary across the fields. Clearly, the motif profile distinguishes the networks from each other, and research communities in each group follow different models of collaboration.

More specifically, we can see that in the first group, including Chemistry and Physics, the most significant pattern of collaboration is in the form of a clique (5-21), meaning that there exists a large number of "clusters" in which every author has a publication with the others, and also that the size of co-authorship communities is larger than in the other

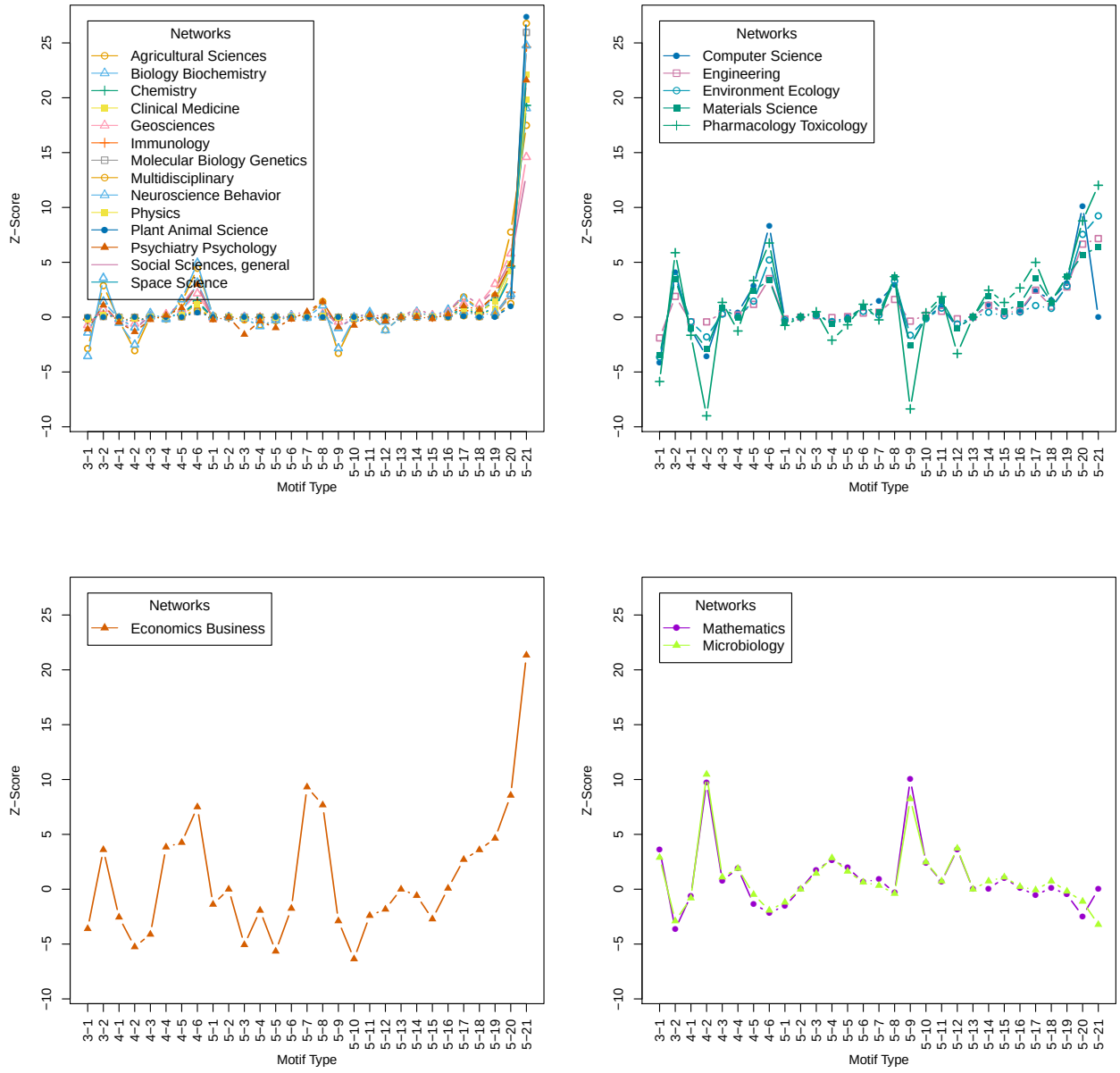


Fig. 4: Motif profile of co-authorship networks from different scientific fields.

fields. The second significant model in this group is also a clique, but of size 4, and the third is a triangle, which is a clique of size 3. This shows that in this group, authors tend to publish in highly connected communities. The most rare models of collaborations in this group are motifs of (3-1), (4-2) and (5-9), showing that it is uncommon that two co-authors of an author do not collaborate. In other words, we can say that in these fields, co-authors of an author are tipically also co-authors.

The second group of scientific fields includes Computer Science, Engineering, Materials Science and Pharmacology & Toxicology. In opposition to the first group, the clique motif of size 5 is not the most significant collaboration pattern here, but we still have a dense subgraph of size 5, motif (5-20), as the main pattern. This means that researchers in this group also publish in highly connected communities, but not so dense as in the first group.

The most different field of study is Economics & Business, where other types of motifs are also important models of collaboration. Motifs of (5-7), (5-8) are even more common

than the clique patterns of 4 and motifs of (4-5), (4-4) are more important than clique of size 3. In this field, the co-authors are indirectly connected. We should point out this field exhibits an uncharacteristic paper with 36 authors, that creates a 36 node clique that functions as an "outlier" that can somehow skew the results.

Finally, the last group includes Mathematics and, surprisingly to us, Microbiology, where the building blocks of the co-authorship network are motifs of (4-2) and (5-9). In this group, the co-authors are not necessarily directly related as co-authors.

III. Conclusion

The ability to compare scientific production across different fields is important, specially in the view of a research assessment by science funding organizations. Most of the studies in this domain rely mainly on statistical properties of co-authorship networks either at the level of individual authors or at the level of the entire network. In this paper, we presented an approach for comparison of co-authorship networks at the subgraph level. Network motifs, defined as overrepresented subgraphs, are here used to build a fingerprint for the comparison of collaboration patterns in different scientific fields.

Based on a large set of publications authored by researchers at the University of Porto, we built co-authorship networks for each of the 22 ISI scientific fields. With our approach, we were able to determine four distinct groups of fields, holding different collaboration patterns with different significance levels. Whilst in Chemistry and Physics, researchers publish in large collaboration groups, in the form of motif (5-21), a clique of size 5, in other fields, such as Mathematics, researchers exhibit a less dense collaboration pattern, in the form of motifs (4-2) and (5-9), with co-authors rarely publishing between themselves.

Although being limited to the University of Porto in terms of raw data used, the derived results are comprehensive in terms of analyzed fields and they are consistent with the results of previous research studies on different data sets, such as the one made by Newman [6], which used more typical global statistical properties. We were able to identify a list of frequently occurring subgraphs in the network of each field, and this can be used as a fingerprint for the purpose of network comparison.

The motif based comparison can be further enhanced by incorporating weight for edges or nodes. Social networks represent human interactions and these are more complex than simple binary links. In real networks, a connection holds more information if the weight is also considered. In future work, we intend to consider more information about the connections, such as the weight of edges, which can in our case be defined as the number of co-authored publications, or citation rate of co-authored papers. In addition to adding weight to motifs to improve comparison precision, we believe there is also information to be gained by also involving more traditional statistical properties of networks such as degree distribution, in an hybrid approach. We will also try to increase the size of the considered motifs, by resorting to high performance computing to achieve reasonable computation times.

Acknowledgments

This work is in part funded by the HORUS project (PTDC/EIA-EIA/100897/2008), the ERDF/COMPETE Programme and by FCT within project FCOMP-01-0124-FEDER-022701. Sarvenaz Choobdar is funded by an FCT Research Grant (SFRH/BD/72697/2010). Pedro Ribeiro is funded by an FCT Research Grant (SFRH/BPD/81695/2011). Sylwia Bugla is also funded by an FCT Research Grant (SFRH/BI/51051/2010).

References

[1] P. Batista, M. Campiteli, and O. Kinouchi, "Is it possible to compare researchers with different scientific interests?" *Scientometrics*, vol. 68, no. 1, pp. 179–189, Jul. 2006.

[2] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215 – 239, 1979.

[3] B. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. Stockwell, and T. Ideker, "Pathblast: a tool for alignment of protein interaction networks," *Nucleic Acids Research*, vol. 32, no. suppl 2, p. W83, 2004.

[4] J. Berg and M. Lässig, "Cross-species analysis of biological networks by bayesian alignment," *Proceedings of the National Academy of Sciences*, vol. 103, no. 29, p. 10967, 2006.

[5] R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nature Biotechnology*, vol. 24, no. 4, pp. 427–433, 2006.

[6] M. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, p. 5200, 2004.

[7] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, p. e177, 2007.

[8] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *Journal of the Royal Society Interface*, vol. 7, no. 50, p. 1341, 2010.

[9] T. Milenković, W. Ng, W. Hayes, and N. Pržulj, "Optimal network alignment with graphlet degree vectors," *Cancer Informatics*, vol. 9, p. 121, 2010.

[10] G. Mills-Tettey, A. Stentz, and M. Dias, "The dynamic hungarian algorithm for the assignment problem with changing costs," *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 2007.

[11] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[12] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, p. 1538, 2004.

[13] L. Krumov, C. Fretter, M. Mller-Hannemann, K. Weihe, and M. Htt, "Motifs in co-authorship networks and their relation to the impact of scientific publications," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 84, pp. 535–540, 2011.

[14] S. T. Bugla, "Name Identification in Scientific Publications," Master's thesis, MSc in Networking and Informatics Systems Engineering, Faculty of Science, University of Porto, December 2009.

[15] P. Ribeiro and F. Silva, "G-tries: an efficient data structure for discovering network motifs," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, 2010, pp. 1559–1566.

[16] I. Jolliffe and MyiLibrary, *Principal component analysis*. Springer, 2nd ed., 2002.

[17] G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.