Strength of Co-authorship Ties in Clusters: a Comparative Analysis

Michele A. Brandão and Mirella M. Moro

Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
micheleabrandao@dcc.ufmg.br, mirella@dcc.ufmg.br

Abstract. We analyze the strength of ties through three different clustering algorithms applied to co-authorship social networks from three different research areas. This study reveals if tie strength metrics can be used to evaluate clusters quality. We obtain different results for each algorithm and observe that Markov cluster algorithm provides the best results for co-authorship social networks. Also, researchers in overlapped communities detected by clique percolation method work as bridges.

Keywords: Social Networks, Tie Strength, Clustering Algorithms

1 Introduction

Clustering algorithms represent a classical problem of data mining and has many applications over a plethora of domains. Then, identifying which algorithm is proper to one such domain is a challenge per se. Likewise, evaluating the quality of the created clusters is hard due to its problem-driven nature, as a good clustering algorithm for a problem may not be as good for another [1].

In the context of social networks (SN), clustering algorithms are useful for detecting (finding) communities. Examples of studies include to explore regional innovation systems, clustering effect in scientific communities and concentration of developers in a country [5]. Specially in academic SN, detecting clusters helps to discovery patterns that may increase the researchers' productivity, reveal the impact in research policy and understand group formation [11]. However, once again, the problem is how to verify the quality of the created clusters.

Here, we apply clustering techniques in co-authorship SN, a type of academic SN in which the nodes are researchers and there are edges between those who have published together. By definition, a cluster in SN is a collection of individuals with dense interactions patterns internally and sparse interactions externally [13]. Therefore, to evaluate the quality of the created clusters, we use existing metrics to assess the strength of co-authorship ties intra and inter clusters.

In summary, when the strength of ties is measured by metrics that consider the neighborhood of nodes, the strength of ties intra cluster should be higher than inter clusters. Hence, we measure tie strength using two metrics that provide such information [3,7,15]: *Neighborhood Overlap* (NO), the collaboration between two nodes regarding their neighbors; and *co-authorship frequency* (W), the absolute number of publications between two persons.

Brief Related Work. There are many clustering techniques and they are applied to different types of networks, for example, similarity graphs [9], directed networks [12], social professional networks [5] and mobile SN [10]. From these techniques, we have chosen three that are commonly applied to undirected graphs and represent good strategies commonly used to detect communities in SN [13].

Also, there are different ways to measure clustering quality, such as BetaCV, C-index and modularity [5]. However, identifying whether such metrics give the expected answer for a graph is very difficult [1]. Moreover, most of these metrics are biased and unreliable in larger real graphs. Indeed, in this work, we investigate whether tie strength metrics can be used to evaluate clustering quality. This study represents a new direction in the evaluation of clustering algorithms and may help to fill this gap in the state-of-the-art.

Contributions. Overall, our contributions are the analyses of: the distribution of strong and weak ties intra and inter clusters and the dynamism of the strength of ties through different clustering algorithms. Such analyses reveal whether tie strength metrics can be used to evaluate clusters quality based on the definition that ties intra a community should be strong and inter should be weak.

Next, we present the analysis setup that includes creating reals SNs (Section 2). Then, we analyze three clustering methods: Louvain method (Section 3.1), clique percolation method (Section 3.2) and Markov cluster algorithm (Section 3.3), and compare their results (Section 4). We have chosen such algorithms because they are important to detect core groups (a.k.a. clusters or communities) on SN [10]. Thus, we use both terms interchangeably and maintain the nomenclature of the clustering methods' authors.

2 Analyses Setup

A co-authorship social network can be modeled as a weighted graph $\mathcal{G}^w = (\mathcal{V}, \mathcal{E}^w)$, with $\mathcal{V}$ the set of nodes and $\mathcal{E}^w$ the set of non-directed weighted links. Nodes are researchers (or authors), a tie between any two researchers exists if they have published together, and the tie weight represents the absolute number of publications between them, called as *co-authorship frequency* or simply W, which has been applied to measure the strength of ties [15].

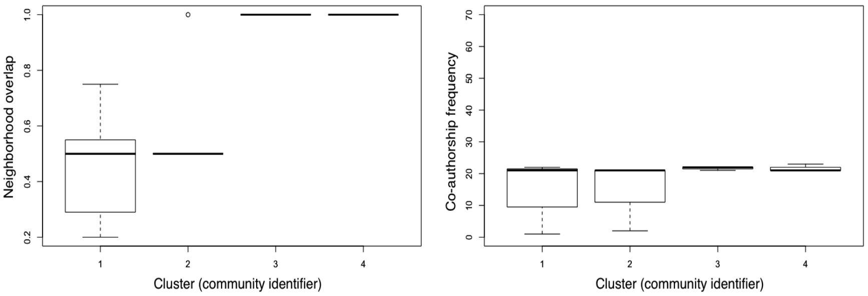
Another topological property to measure the strength of ties is *Neighborhood Overlap – NO* [3,7]. The NO of an edge connecting researchers v_i and v_j is given by the equation: $\frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)| - \{v_i, v_j\}}$, where $\mathcal{N}(v_i)$ represents the co-authors of researcher v_i, and $\mathcal{N}(v_j)$ the co-authors of v_j.

Here, we build three co-authorship SNs using the *CiênciaBrasil* datasets[1]. The publications available in *CiênciaBrasil* are from Brazilian researchers and have been collected from Lattes, an online platform for archiving researchers' curriculum vitae, in November 2013. Each network represents the co-authorships among researchers from three areas: computer science, medicine and sociology. Table 1 has the datasets statistics: number of authors (researchers), number of

[1] Datasets available at `http://www.dcc.ufmg.br/∼mirella/projs/apoena`

Table 1: Dataset statistics per research area.

Area	# authors	# publications	AvgPubA	# pairs (# dist)
Computer Science	543	48,706	89.69	16,312 (1,563)
Medicine	368	75,553	205.30	16,089 (778)
Sociology	96	7,195	74.95	322 (39)



(a) *Neighborhood Overlap* (NO)

(b) *Co-authorship Frequency* (W)

Fig. 1: The strength of ties intra-communities in a perfect clustering. In each box plot, the central rectangle spans the first to the third quartiles, the segment inside is the median, traits above and below the rectangle represent the minimum and maximum values. The clusters' identifiers order the box plots.

publications, average number of publications per author and number of pairs of co-authors (and number of distinct pairs of co-authors).

Considering these datasets, we apply three clustering algorithms: Louvain method (LM), clique percolation method (CPM) and Markov cluster algorithm (MCL). Then, we measure the strength of ties for each pair of researchers in each cluster detected by the algorithms. Such strength is measured by using NO and W. Following Brandão and Moro [3], we consider that a tie is weak when NO is in the range $[0; 0.2]$ and strong otherwise. Likewise, a tie is weak when W is in the range $[1; 5]$ and strong otherwise.

Overall, analyses provide insights whether the strength of ties metrics can be used to evaluate clustering quality. By clusters definition [2,14,16], ties intra-clusters should be strong and ties inter-clusters should be weak. Therefore, a cluster should have most pairs of researchers (ties) classified as strong and most ties that connect different clusters as weak.

One of the problems in evaluating clustering quality is the absence of a ground truth for comparison [1]. Thus, we now verify the strength of ties in a synthetic data that represents a situation with perfect clustering. According to Harman et al. [8], a perfect clustering has a perfect modularization, i.e., all modules in a cluster are connected to all other modules and there are no inter-cluster connections. Thus, we build a graph with 17 nodes and 23 edges (two randomly chosen prime numbers). We link the nodes in a way to form four clusters and there are no connections among nodes from different clusters. Figures 1a and 1b present NO and W of a perfect clustering, respectively. Cluster #1 is the largest one (7 nodes and 12 edges), cluster #2 is the second largest (4 nodes and 5 edges), clusters #3 and #4 have the same size (3 nodes and 3 edges).

Note the minimum value of NO is 0.2, i.e., most communities are composed by strong ties. The smallest clusters have NO equal to 1 (i.e., all ties are strongly connected), because all nodes are connected to each other, but in a real social network this hardly happens. We emphasize that a high NO indicates that pairs of researchers are more connected to each other intra a cluster. Also, W of all clusters has the median higher than 20. This is a property strictly related to the frequency of nodes interactions – not always found in real networks. However, co-authorship SN with a high degree of collaboration tend to have a high W [3]. Hence, most detected clusters should have more strong ties than weak ones.

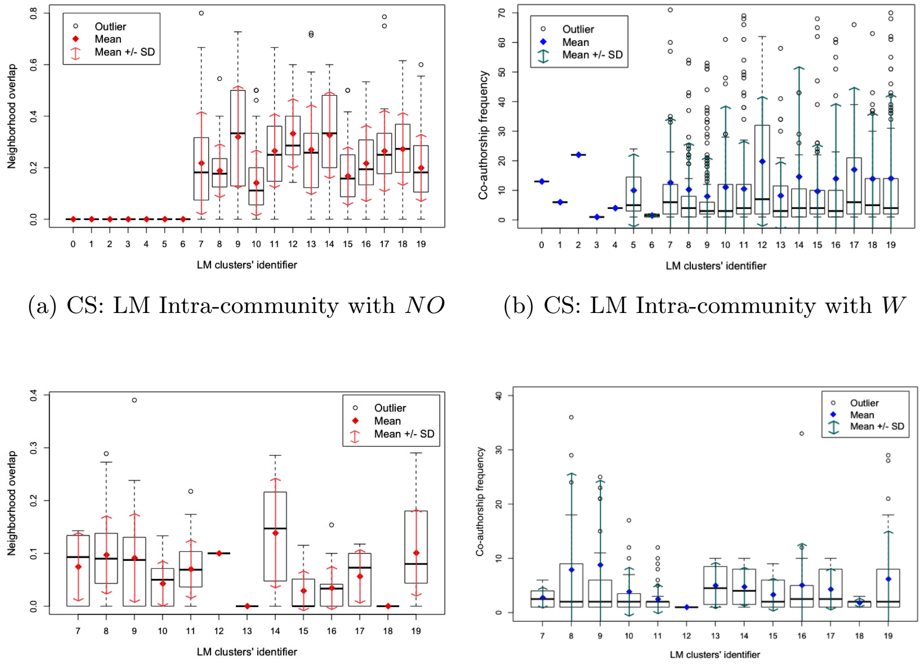
3 Evaluated Clustering Techniques

In this section, considering three real co-authorship networks, we analyze three clustering techniques: Louvain method (Section 3.1), Clique Percolation method (Section 3.2) and Markov Cluster algorithm (Section 3.3). For space constraint, all graphs are presented in [4] and we discuss only the main findings next.

3.1 Community Detection Using Louvain Method

The Louvain method (LM) [2] is a simple, efficient and a very common method for detecting communities in large networks. It makes greedy seeks to optimize the modularity of a partition of the network, where modularity is a topological property and designed to measure the density of links intra communities [2]. As it works over unweighted networks, it allows to study the links between researchers in clusters that are formed by the modularity and the network topology.

Considering only computer science (CS), Figure 2 presents the results for measuring intra and inter-communities created by LM using both NO and W. Overall, medicine has more communities with smaller mean and median NO values than computer science, but W of such communities are higher than computer science. Also, in both areas, the communities with highest NO do not indicate communities with highest W. In sociology, NO and W of researchers in each community is small, but communities with the highest NO do not have the highest W. Such aspects suggest that the strength of the intensity of co-authorships among researchers measured by W does not always correspond to the strength of the interactions among researchers' neighbors measured by NO. Moreover, there are communities with NO equal to zero, few edges compose all such communities in the three SN, and all edges have one node in common. These smaller communities are detected by LM because the researchers are not



(a) CS: LM Intra-community with NO (b) CS: LM Intra-community with W

(c) CS: LM Inter-community with NO (d) CS: LM Inter-community with W

Fig. 2: LM intra/inter-communities with NO and W for CS.

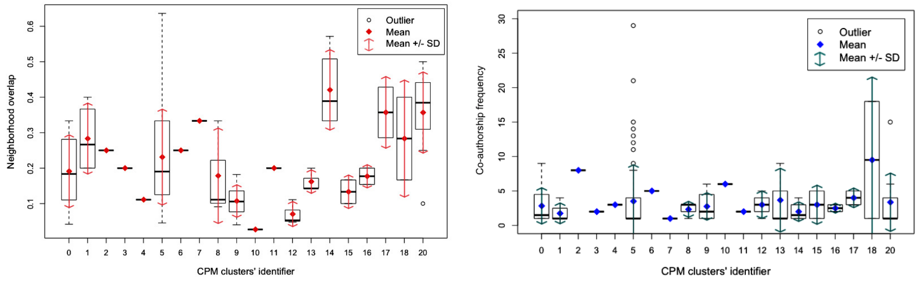
densely connected to other communities. Also, considering the outliers, the communities have less outliers to NO than to W. For future work, the study of such outliers might reveal interesting properties about co-authorships.

3.2 Uncovering Communities with Clique Percolation Method

The clique percolation method (CPM) locates the k-clique communities of networks and considers that a typical node in a community is linked to many others, but not necessarily to all other nodes in the community [14]. Overall, a community is a union of smaller fully connected subgraphs that share nodes. Such complete subgraphs are called *k-cliques*, where k refers to the number of nodes in the subgraph. Then, *k-clique-community* is defined as the union of all k-cliques that can be reached from each other through adjacent k-cliques [14].

We apply this method using the algorithm implemented in CFinder[2] and k=3. By definition, a community is actually a connected graph when k=2 and

[2] CFinder: `http://www.cfinder.org`



(a) Med: CPM Intra-community with NO (b) Med: CPM Intra-community with W

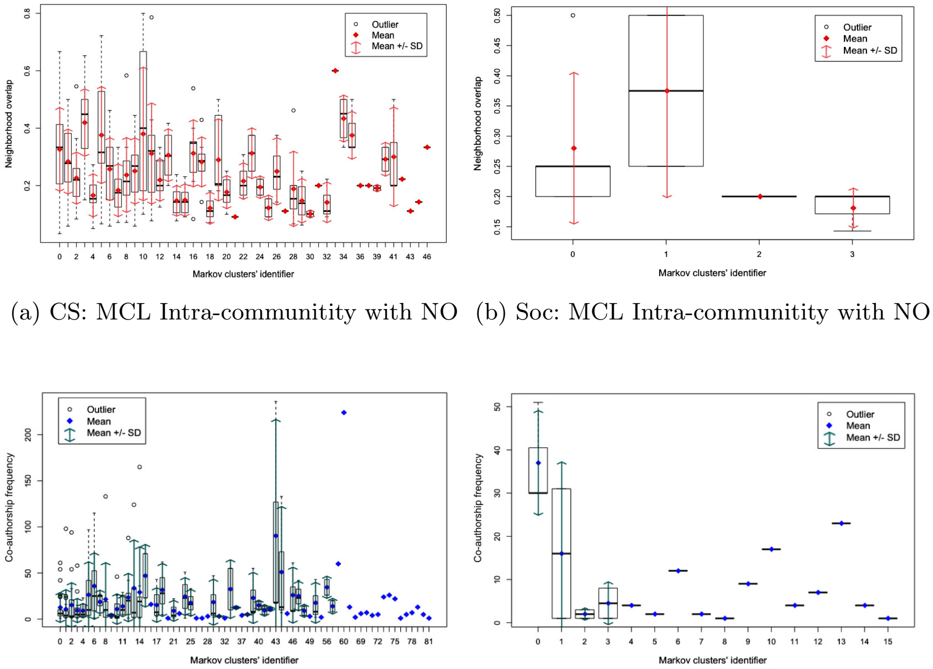
Fig. 3: CPM intra/inter-communities with NO and W for medicine.

a set of disconnected nodes without any edge when k=1. The parameter k determines the nature of the communities. Using different values for k reveals the nature of the communities [6]. We have chosen k=3 in order to discover triangles and because such a value is also used in most general cases [14]. Finally, the CPM allows overlap, i.e., a node can be a member of different communities at the same time, and communities overlap with each other by sharing nodes.

Figure 3a show the communities uncovered by the CPM (k=3) applied to the medicine SN. Note that it considers the social network as unweighted. The NO values reveal that although the communities are formed by cliques, some of them have only weak ties (i.e., pairs of researchers weakly connected regarding NO): seven in medicine, ten in computer science, and two in sociology. In other words, cliques formed by co-authorship of researchers do not have only strong interactions. Additionally, each community may have ties linking different cliques, and such ties are also weak in communities with only weak ties. Other communities have only strong ties: eight in medicine, four in computer science, and two in sociology. It is also interesting to investigate these communities in order to identify patterns in the high cooperativeness. In the remaining communities, there is a mix of strong and weak ties. Furthermore, most communities have the median and mean different from the others, meaning that researchers have distinct behavior of co-authorship in each community.

Regarding W as a measure of the strength of tie, Figure 3b shows that most communities are composed by ties between researchers with small W. In medicine and computer science, only one community has tie with W greater than 10. In sociology, W does not reach five. Although the cliques compose such communities, the high connectivity among researchers groups does not indicate a strong intensity of co-authorship.

According to CPM definition, one researcher may be in more than one community. The number of overlaps is small in the three networks: in computer science, only one researcher is in four communities; in sociology, there is no overlap; and in medicine, one researcher is in three communities. An analysis



(a) CS: MCL Intra-community with NO (b) Soc: MCL Intra-community with NO

(c) CS: MCL Intra-community with W (d) Soc: MCL Intra-community with W

Fig. 4: MCL intra-communities with NO and W for computer science (CS), sociology (soc) – clusters' identifiers in x axis are ordered by the size of communities.

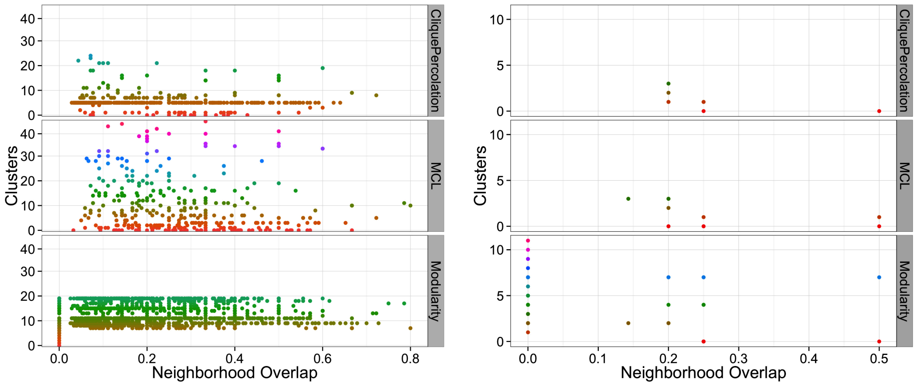
suggests that researchers in overlapped communities have weak ties with other researchers and work as a bridge (details available in [4]).

3.3 Clustering with MCL Algorithm

The Markov Cluster Algorithm (MCL) is an unsupervised clustering algorithm for graphs based on simulation of stochastic flow in graphs (known as network) [16]. MCL deterministically finds cluster structures by computing the probability of random walks though the network. We use the algorithm available in Micans[3] and keep the default values of the parameters.

One input to MCL is a file describing the graph edges: the source and target nodes, and W as edges weight. The MCL interprets W of the edges as similarity to cluster the nodes. In order to understand how NO and W influence on clustering formation, we run the algorithm twice changing the value of the edge weight (one time weights equal to NO and another to W). Using NO, the MCL

[3] Micans: `http://micans.org/mcl`



(a) Computer science

(b) Sociology

Fig. 5: Results of clustering methods and measuring tie strength with NO (each cluster as a color, and each edge as a point).

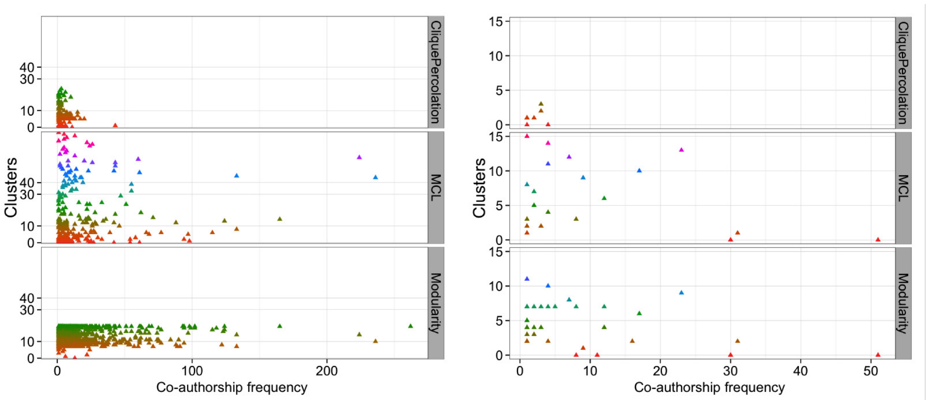
has found 140 clusters in computer science, 35 in sociology and 139 in medicine. Then, using W, the MCL has detected 82 clusters in computer science, 16 in sociology and 68 in medicine. Some clusters are composed of only one node, and they are more present in clusters formed with NO as edge weight. This result indicates that the similarity among researchers is lower considering NO than W.

Figure 4 shows the results ordered by the size of communities when NO and W between researchers are considered as edge weight. The graphs do not include clusters with only one node for clarity. There are communities formed only by weak ties and only by strong ties, for example, the communities #25 and #34 in computer science, respectively. However, most communities include both types of tie. Considering the clusters size, the biggest communities are in the beginning of each graphic. For example, clusters #0 and #1 are the largest in sociology. The number of nodes in the largest clusters for NO and W as edges weight is respectively 30 and 27 for computer science, four nodes (two communities of the same size) and four nodes (four communities of the same size) for sociology, 22 and 17 for medicine. Figure 4 also presents that the largest clusters are more formed by strong ties than weak ties, because the first quartile of these clusters is higher or equal to 0.2. Then, the largest communities have high W, because the third quartile pass 30 in the three areas.

Lastly, MCL does not find ties connecting researchers from different communities for the three co-authorships SN. This reveals that MCL provides a good clustering result since clustering algoritms minimizes inter-cluster edges [12].

4 Comparative Analyses

We now compare the results of the three clustering methods. Figures 5 and 6 contrast the clusters of each method regarding NO and W, respectively. We ob-



(a) Computer science

(b) Sociology

Fig. 6: Results of clustering methods and measuring tie strength with W (each cluster as a color, and each edge as a triangle).

serve that LM tends to find less and larger clusters than the other two methods. Also, MCL detects a huge number of clusters, and some of them are singleton. In the co-authorship context, although CPM allows community overlaps, as a researcher may publish with researchers from others communities, MCL provides the best clusters because most of the detected ones are composed by strong ties.

Moreover, Figure 5 shows a high concentration of edges until NO reaches 0.6 in computer science and medicine. In sociology, the maximum value of NO is 0.5, and there is more concentration of edges between 0.2 to 0.3. Also, note that CPM and MCL exclude edges with NO equal to 0. Some strong ties are also removed in CPM, probably because these edges are in a 2-clique (we choose k=3 for CPM). Here, MCL better differentiates the relationships putting them in distinct clusters. On the other hand, the concentration of points in CPM and LM does not show the same for these methods. Additionally, Figure 6 shows a high concentration of edges for W less than 100 in computer science and medicine, and less than 10 in sociology. We also note that co-authorship frequencies equal to zero are not removed in any clustering method. Overall, the three algorithms form clusters with weak and strong ties. However, MCL mostly detects clusters with more strong ties than weak ones.

5 Concluding Remarks

We applied three clustering algorithms in three co-authorship SN. For the unweighted LM, its evaluation results showed it identifies less clusters than the others. When applying CPM, there was a small number of overlaps between communities and researchers in the overlaps are weak ties (they work as bridges). For MCL, we have applied it twice in each algorithm: one with NO as weight and another with W. MCL identified a larger number of clusters than the other methods. Furthermore, the tie strength inter-communities tends to be weak for LM and CPM; whereas MCL algorithm does not find edges inter-communities.

A main conclusion of using NO and W in clustering evaluation is: MCL is the best clustering algorithm to be applied in co-authorship SN when compared to LM and CPM. Nevertheless, we also conclude that considering only the strength of ties metrics is not enough to define clustering qualities. Therefore, in the next steps, we plan to apply internal measures (like BetaCV, C-index, and so on) to compare with the results generated by the tie strength metrics. Also, we plan to investigate how the network structure affects the clustering results.

Acknowledgments. Work funded by CAPES, CNPq and FAPEMIG, Brazil.

References

1. Almeida et al., H.: Is there a best quality metric for graph clusters? In: ECML-PKDD. (2011) 44–59
2. Blondel et al., V.D.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008
3. Brandão, M.A., Moro, M.M.: Analyzing the strength of co-authorship ties with neighborhood overlap. In: DEXA. (2015) 527–542
4. Brandão, M.A., Moro, M.M.: A comparative analysis of the strength of co-authorship ties in clusters. Technical Report 4, UFMG (March 2017) http://www.dcc.ufmg.br/~mirella/projs/apoena.
5. Brandão, M.A., Moro, M.M.: Social professional networks: A survey and taxonomy. Computer Communications **100** (2017) 20 – 31
6. Deb, D., Vishveshwara, S., Vishveshwara, S.: Understanding protein structure from a percolation perspective. Biophysical journal **97**(6) (2009) 1787–1794
7. Easley, D., Kleinberg, J.: Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press (2010)
8. Harman, M., Swift, S., Mahdavi, K.: An empirical study of the robustness of two module clustering fitness functions. In: GECCO. (2005) 1029–1036
9. Hassanzadeh et al., O.: Framework for evaluating clustering algorithms in duplicate detection. Proc. VLDB Endow. **2**(1) (2009) 1282–1293
10. Kim, P., Kim, S.: A detection of overlapping community in mobile social network. In: Procs. of ACM SAC. (2014)
11. Kshitij, A., Ghosh, J., Gupta, B.M.: Embedded information structures and functions of co-authorship networks: Evidence from cancer research collaboration in india. Scientometrics **102**(1) (2015) 285–306
12. Malliaros, F.D., Vazirgiannis, M.: Clustering and community detection in directed networks: A survey. Physics Reports **533**(4) (2013) 95–142
13. Mishra et al., N.: Clustering social networks. In Bonato, A., Chung, F.R.K., eds.: Algorithms and Models for the Web-Graph. Springer (2007) 56–67
14. Palla et al., G.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043) (2005) 814–818
15. Silva et al., T.H.P.: Community-based endogamy as an influence indicator. In: JCDL. (2014)
16. Van Dongen, S.M.: Graph clustering by flow simulation. PhD thesis, Utrecht University (2000)