Discovering Scientific Communities using Conference Network

Alejandro Mussi, Fabio Casati, Aliaksandr Birukou, Luca Cernuzzi
DISI, University of Trento - Italy
{mussi, casati, birukou}@disi.unitn.it, lcernuzzi@uca.edu.py

ABSTRACT

This paper presents an algorithm and a tool for discovering scientific communities. Several approaches have been proposed to discover community structure applying clustering methods over different networks, such as co-authorship and citation networks. However, most existing approaches do not allow for overlapping of communities, which is instead natural when we consider communities of scientists. The approach presented in this paper combines different clustering algorithms for detecting overlapping scientific communities, based on conference publication data. The Community Engine Tool (CET)[1] implements the algorithm and was evaluated using the DBLP dataset, which contains information on more than 12 thousand conferences. The results show that using our approach it is possible to automatically produce community structure close to human-defined classification of conferences. The approach is part of a larger research effort aimed at studying how scientific communities are born, evolve, remain healthy or become unhealthy (e.g., self-referential), and eventually vanish.

1. INTRODUCTION

The increase in the number of scientific publications has made the search of digital scientific literature a difficult task, which is highly dependent of the researcher ability to search, filter and classify content. Most used scientific literature search engines and portals, such as Google Scholar [6], Citeseer [11] and ACM [2], use only simple text-based and citation-based score to rank the query result, and the rank is barely useful [10].

The world of science has many fields and sub-fields such as Biology, Mathematics, Computer Science, and so on. Each of them has different structures and publication dynamics. An example is the number of citations in the top-20 most cited journals in Computer Science is 4 times higher than the top-20 most cited journals in Social Science [12]. Therefore, good contributions that belongs to a community with lower productivity may be overlooked because of those which are in a community with a higher productivity. The same problem happens when we rank researchers: it is unfair to compare researchers using citation-based metrics without a context, in other words, the community they belong to. Also, because of different sizes of communities, it is hard to measure the productivity or impact of researchers from different communities fairly, using traditional citation-based metrics, such as H-index, since researchers from communities with higher productivity are likely to produce more citations than those from communities with lower productivity.

In this paper we present a model and a tool for discovering and evaluation of scientific communities. The use of discovered communities will improve two important activities in scientific research: the *search* of scientific contributions and the *assessment* of people (researchers), as explained in the following.

By using community-aware search mechanisms it would be possible to narrow down the domain of the queries to specific communities, or, vice versa, extend it to different communities to obtain diversity of content. Moreover, having a framework that supports discovering scientific communities will provide the means for a better understanding of the social behavior in the scope of scientific research, enabling us the possibility to identify patterns in developments of projects, research trends, successful research profiles, etc. in different communities.

Regarding the assessment of people, in [1] it is suggested that numerical indicators must not be used to compare papers or researchers across different disciplines. Since nowadays the boarders between disciplines are blurring, it is hard to define a priori the disciplines to which a paper or a researcher belongs. Ad-hoc and evolving communities can provide a better way for such comparison.

The approach presented in this paper combines different techniques for detecting scientific communities, based on conference publication data. The Community Engine Tool (CET) has implemented the algorithm and has been evaluated using the DBLP dataset, which contains information on more than 12 thousand conferences. The results showed that using our approach it is possible to automatically produce community structure close to human-defined classification of conferences. The approach is part of a larger research effort aimed at studying how scientific communities are born, evolve, remain healthy or become unhealthy (e.g., self-referential), and eventually vanish.

2. BACKGROUND

This section introduces the model of scientific communities and techniques used for detecting communities.

2.1 Scientific Community Model

[1] `http://project.liquidpub.org/research-areas/scientific-community`

The concept of community can be defined in different terms, in the highest level can be considered as a set of related people. The type of relations we consider will determine the type of community we are capturing. Therefore, at the end we aim at detecting sets of people that are strongly related by some pre-defined type of relation within the set and less related among other groups. In this field, Newman has proposed a property for a graph called **community structure** which is focused on capturing these groups of nodes that are densely related and not as densely between other groups.

In this paper we aim at capturing scientific communities which are composed of people and also of scientific entities, where a scientific entity refers to an abstract representation of all scientific content, such as journals, papers, conferences, among others. We define a scientific community, identified by a name, as a set of scientists and other scientific entities that are densely connected within the community and sparsely connected among other communities.

2.2 Community Detection

The detection of community structure on complex networks has become an interesting focus of investigation in different disciplines such as physic, social sciences, computer science, among others. Girvan and Newman were the first to introduce the property of community structure of a network [5], and an index to measure the quality of the structure called *modularity* Q [9]. Many algorithms have been developed in the sake of detecting community structure of complex networks [14][9][3][7], but the vast majority of these algorithms do not take into account the overlapping of these communities [13].

The analysis of the modularity Q of a graph opens a variety of algorithms to detect community structure based on the optimization of Q. However, exact modularity optimization is a problem that is computationally hard [4]. Hence, efficient algorithms must deal with some heuristic in order to get result in polynomial-time.

The combination of community detection clustering algorithms and an index to measure the community structure will help us to evaluate the algorithm and select the best classification on a hierarchical output.

3. DISCOVERING SCIENTIFIC COMMUNITIES

In this section the complete process of the detection of scientific communities is summarized. We start from the description of the problems we need to face in order to achieve the goals, followed by the proposed model, the algorithm used for the detection of communities, and the creation of the community network.

3.1 Discovering Scientific Communities: Problem and Scope

The problem of modeling, managing and analyzing scientific communities contains a wide range of different aspects that need to be confronted (Figure 1). Each of these subproblems has its own complexity and challenges.

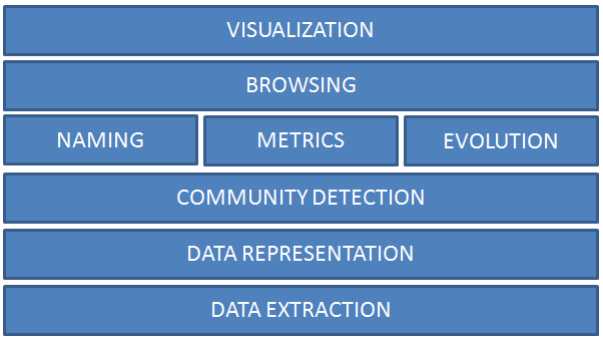

Figure 1: Scientific Communities problem stack

- **Problem 1** - Scientific Data Extraction: the first step of the process is to provide the data for the framework. This problem is focused on extraction of data from different sources.
- **Problem 2** - Data Representation: the way of representing connections between entities will define the shape of the communities. This issue is about establishing a model for communities and extracting the data for their detection.
- **Problem 3** - Community Detection: the problem consists in developing algorithms capable of detecting community structure.
- **Problem 4** - Naming: once communities are detected, each of them should be identified by a name that characterizes the community.
- **Problem 5** - Metrics: this problem is about proposing new community-aware metrics for the sake of improving the evaluation of scientific content and researchers.
- **Problem 6** - Evolution and Trends: as scientific communities are not static, methods for managing the evolution of communities along the time are necessary. This problem deals with the design and implementation of business logics to support evolution of communities.
- **Problem 7** - Browsing: once information about communities is available, methods to query and navigate through the communities are required. Thus, the design and implementation of a browsing interface for communities is also an important problem.
- **Problem 8** - Visualization: this problem is about designing and implementing a visual model for communities that enable users to interact with communities.

The listed problems provide an overview of different aspects of community discovery to be considered.

3.2 Conference Network

Different scientific networks can be build by combining information about scientific entities. The detection of communities on these network will provide different community structures and meanings. For example, the citation network

will tend to provide topic-related communities, while the authorship network will highlight social relations inside the community since co-authors often know each other.

One of the main problems is that the vast majority of the clustering algorithms used to detect communities do graph partition on the network [13]. This means that after the clustering process a node only belongs to a particular community. This is a problem if we seek for the communities with overlapped members, such as communities of authors or conferences[2].

In this paper a new type of network *Conference Network* is proposed, that will support overlapping communities of scientific entities, such as authors, reviewers, and scientific publications.

A conference network is defined as a weighed graph where nodes represent conferences, and the weight of the edge between any two different nodes, A and B, is defined as the number of authors that have published in both conferences (A and B).

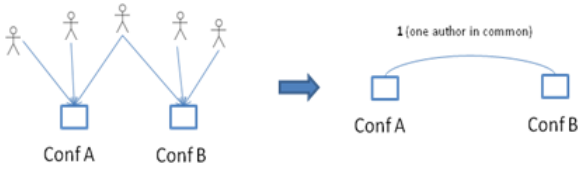

Figure 2: Graphical representation of a Conference Network

The conference network groups authors into conferences reducing the size of the network to be computed without loosing information on connection points, because the relation between members are represented in the weights of edges. Figure 2 provides a simple example of how an author connects two conferences.

3.3 Community Detection Clustering Algorithm

The algorithm used for detecting communities is based on *Edge Betweenness* (EB), which has been proposed by Girvan and Newman [5]. The algorithm captures those edges that connects most communities in order to remove it. The Edge Betweenness of an edge is defined as the number of shortest paths between all combinations of two different nodes that pass through the edge. We adapted the algorithm to use it in a weighted graph by considering also the weights in the computation of the shortest path. The highest EB value corresponds to an edge that has the maximal value of EB (see Figure 3). Hence, if we remove this edge it will separate different communities.

The described method can iterate until no connection/edge remains. This process is known as divisive clustering algorithms. The output produces a dendrogram which represents the entire hierarchy of possible community division of the graph. For each graph we calculate the value of modularity Q and then select the graph with the highest modularity, in other words the one corresponding to the best community structure.

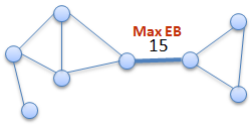

Figure 3: Calculation of the Edge Betweenness value of two nodes

The *modularity* Q of a graph was proposed in [8] and defined as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \quad (1)$$

where c_i is the community to which vertex i is assigned, $\delta - function$ $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, A_{ij} is the weight of edge from i to j, and $m = \frac{1}{2} \sum_{ij} [A_{ij}]$ is the number of edges in the graph. If we preserve the degrees of vertices in our network but otherwise connect vertices together at random, then the probability of an edge existing between vertices i and j is $\frac{(k_i k_j)}{2m}$, where k_i is the degree of vertex i.

The modularity measures the fraction of the edges in the graph that connect vertices of the same type minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices. Q values near 0 indicates randomness, while higher values mean strong community structure.

In summary, the algorithm performs the following steps:

1. Calculate the betweenness for all edges considering weights.
2. Divide the EB values of all edges by their weights.
3. Remove the edge with the highest betweenness.
4. Recalculate betweenness value for all edges.

Finally, the division with the highest modularity Q is selected in order to create the Community Network.

The betweenness score for all m edges in the graph of n vertices can be calculated in $O(mn)$ time using the fast algorithm of Newman [4]. Since, this calculation has to be repeated per each removal of edges, the entire algorithm runs in worse-case time $O(m^2 n)$.

3.4 Building the Community Network

Once communities are detected, we create a *Community Network*. This network will allow us to visualize and analyze scientists and scientific content. We start by formally defining a scientific community in the community network.

A scientific community is a labeled set of scientific entities defined by the membership function.

[2]Nowadays, both researchers and conferences belong to several communities, due to interdisciplinary nature of the modern research

$$C_i = (L, (e^{[w]}, t))$$

Where:

- C_i is the Community
- L is the label that identifies the community
- e is a scientific entity that can be any of the following: a scientific contribution, a person, an event or a collection (of scientific entities)
- w is the relatedness coefficient that represents the degree to which an entity is part of the community
- t is the time relation between the entity and the community that represents the period of time when the entity is part of the community.

The community network is built as a directed graph where nodes represent communities and edges represent the overlap of members between them. More formally:

$$CN = (\{(C_i, C_j, O_{ij})\}),\ \forall ij, i \neq j \quad (2)$$

Where:

- C_i : community i
- O_{ij} : the overlap from Community i to j.

The *overlap* (connection) between communities is defined as the percentage of elements two communities share. If two communities share entities, an edge between communities is created, and the weight is proportional to the number of entities the community has.

3.5 Naming Communities

Communities should be identified by a certain name which has to characterize the community. In this work we adopt two different approaches. The first method proposed for creating the name of the community is based on using the names conferences which are part of the community. The algorithm selects the two conferences in the community that have more members and use the acronym name to label the community. The names of the conferences help researchers to roughly understand the topic of the community if they know the conferences. The second approach is based on adding extra information about the topics of the community (tags). The algorithm for tagging communities checks the classification of conference from DBLP and tags the communities by matching the conferences found in the community with respect to the conference found in the DBLP classification.

3.6 Community-based Metrics

The community network provides a different view and organization of all scientific information, offering alternatives ways for searching and assessing people and scientific contents. In this section we propose some community-based metrics for the analysis of the discovered communities.

3.6.1 *Community Impact* C_{IMP}

This index aims at assessing the scientific productivity or possible impact of a Scientific Community, by analyzing the h-index of the community members.

A community has a scientific impact $\mathbf{n}$ *(*$C_{IMP} = n$*) if* $\mathbf{n}$ *of their authors have h-index equal to at least* $\mathbf{n}$*, and the other authors have at most* $\mathbf{n}$ *h-index each.*

This metric is an extension of the h-index definition to a community context.

3.6.2 *Community Health* C_{HT}

The Health of a community C_{HT} *is defined as the number of communities that share authors in common with this community (overlapping).* Communities which are not well connected with others communities (known as **closed communities**) do not help to the transference of knowledge, nor the dynamic of the community. In the opposite, a community that shares members in many other communities will tend to have a good transference of knowledge, and will help to the dynamic of the members (new members coming). this type of community is defined as a **healthy community**.

3.6.3 *Author Membership Degree* A_{MD}

It is important, when talking about the members of the community, to analyze the membership degree of authors. For example, if an author has published in the Community A 10 papers, and only 1 paper in the Community B, it is unfair to consider the same degree of membership, especially when analyzing the impact of the community. The metric is defined as follows:

Let $|C_{A_i}|$ *be the number of contributions of author A in the community i, and* $|C_A|$ *the total amount of contributions of author A. Hence, the authorship degree of author A in community i is defined as:* $A_{MD}(A_i) = \frac{|C_{A_i}|}{|C_A|}$

The value is the total number of publications an author has in the community with respect to his total number publications. With this metric, a threshold can be defined for computing metrics. For example we can consider for computing the C_IMP only authors with the membership degree greater than 0.3.

4. COMMUNITY ENGINE TOOL

The Community Engine Tool (CET) is a desktop application that was designed and developed in order to support the requirements for all the process, previously described, of the detection and evaluation of scientific communities. This tool is part of the **Community Discovery Module**, which is one of the components of the LiquidPub[3] architecture.

[3] `http://project.liquidpub.org`

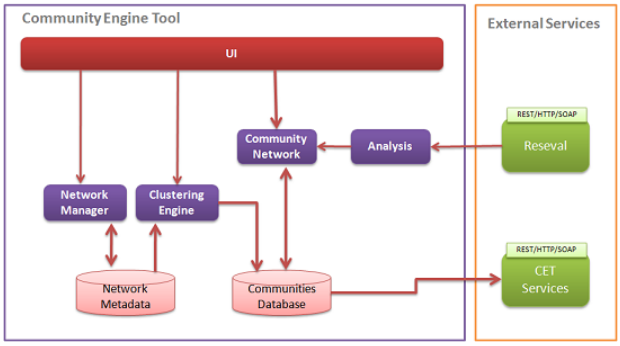

Figure 4: Community Engine Tool Architecture

The architecture of the Community Engine Tool is composed of five main components:

1. **Network Manger (NM)**: this module manages the transformation of the source data into a network of conferences. All the pre-processing steps are done in this module.

2. **Clustering Engine (CE)**: all the community detection clustering algorithms are built in this component. The network of conference is received as input, and user defined cluster algorithms are applied in order to finally obtain cluster of conferences.

3. **Community Network (CN)**: this component manages the complete creation of the CN, the members of each community, and the overlapping between them based on the obtained cluster of conferences. Figure 5 shows the community network and the members of a selected community.

4. **Analysis**: this module analyzes the CN, it interfaces with the ResEval tool[4] by calling its REST services in order to get author metrics such as h-index, g-index, and total citation count. The communities and people are analyzed in this component.

The Community Engine Tool is still in development phase, and more functionalities are intended to add in the next beta version such as the evolution analysis of communities and authors. We are also planning to make the source code publicly available once the tool reaches stability.

5. EVALUATION

The data set used for validation purpose, was a DBLP dump[5], which contained conference and workshop proceedings (12.227), papers in these proceedings (747.752), and authors (533.334) as of 08/03/2009.

In order to test the algorithm, we carried out two experiments that run the community discovery algorithm on the DBLP data set and compared the community structure obtained with the manual topic classification of conferences done by DBLP. The first experiment was performed on the conferences that DBLP has classified in Artificial Intelligent and Cryptology (AI/CRYPTO) research area, while the second experiment was performed on the conferences of Hypertext and Information Retrieval (HT/IR) area.

The algorithm produced an entire hierarchy of possible community division of the graph, and for each partition we calculated the modularity during the process in order to select the structure which represents the best partition. Good values of modularity were obtained on Iteration 294 for HT/IR, and on Iteration 41 for AI/CRYPTO.

For AI/CRYPTO the algorithm divides the network in two communities and the members of the community match exactly with the classification of DBLP. Thus, we identified one community with all conferences of the Artificial Intelligent (on different years), and another community with all the conferences of Cryptology-Security (different years). The overlap between the communities was defined by 40 authors. Therefore, members of those communities were densely connected within the community and not as much connected between them.

As for the community structure found on HT/IR is quite different, this two groups seems to be more related, only one small community of Hypertext is not related to any community of Information Retrieval which is sigmod/2008dbtest. The number of division is the same, we have 4 communities for HT and 4 communities for IR.

Within the two topics that were not very related (CRYPTO and AI), the tool produced exactly the same human classification done by DBLP. On the other hand, for the other two topics that were more related (HT and IR) the tool outputs an equal distribution of communities with respect to the DBLP topic classification. Therefore, it has been demonstrated based on our experiments that the tool produced topic-based communities close to human defined classification.

For each community the h-index of all the members is calculated by the tool. Figure 6 shows the h-index distribution of three communities, its abscissa is ordered by authors with higher h-index first.

The chart shows high values for community *www-csa* and *sac-compsac*, while low distribution for *iceis-rcis*.

In Table 1 the healthiest discovered communities are detailed with their values. The lowest values correspond to isolated communities (See Table 2).

Table 2 lists closed communities found by the tool. These unhealthy communities have members that only published in their community and not in another. The community *iros-icra* has an important Community Impact value of 17, but it is not as healthy as others communities with similar size and impact such as *sac-compsac*, which is a little bit smaller than *iros-icra*, but it has a healthy value of 13 and a C_{HT} of 20.

Communities *iscas-date* and *iros-icra* are two big closed com-

[4] `http://reseval.org`
[5] `http://dblp.uni-trier.de/xml/`

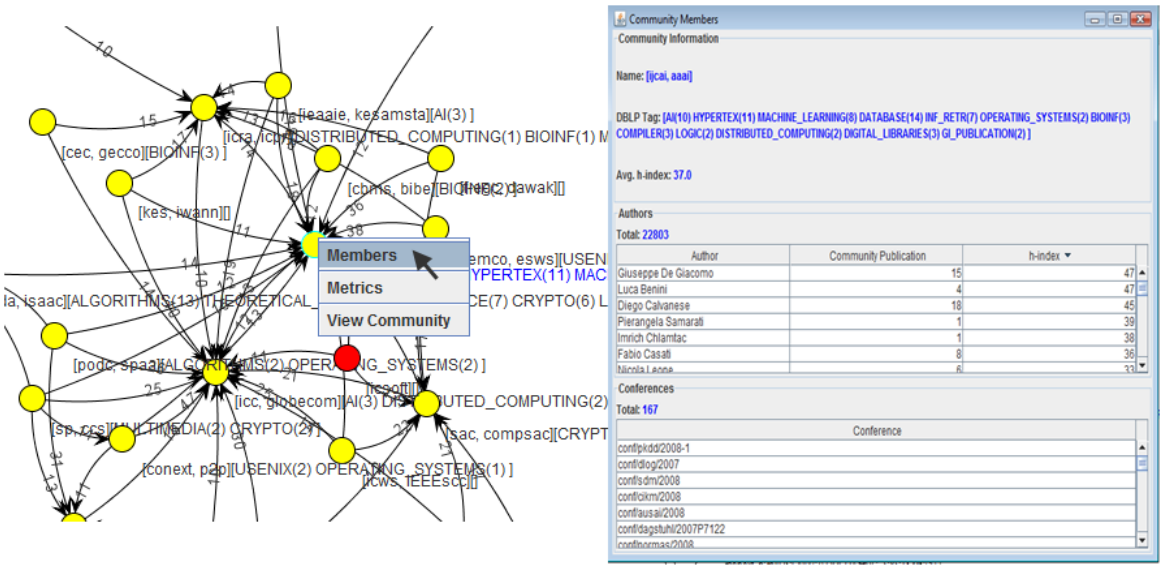

Figure 5: Member details of the selected community

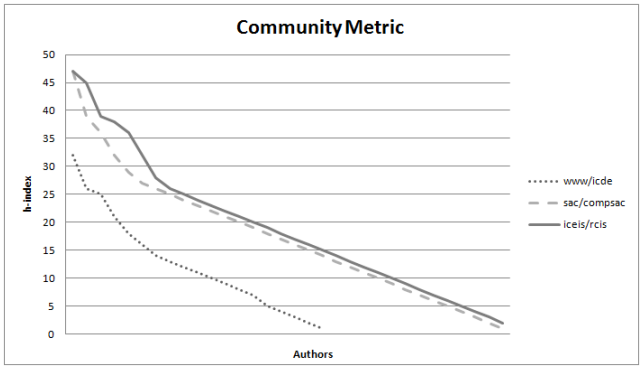

Figure 6: Community h-index distribution

Community	C_{HT}	*Authors*	$AvgH-index$	C_{IMP}
sac-compsac	13	10923	29	20
www-icde	11	15665	33	20
sc	7	74	-	-
icc-globecom	6	26517	29	18
er-bpm	6	662	22	13
icws-IEEEscc	6	2860	27	15

Table 1: Healthier communities and their scientific impact

munities. Conferences **IROS** and **ICRA** correspond to *Robotics and Automation* topic, and the conferences **ISCAS** and **DATE** correspond to *Electronic Circuits and nanotechnology*. Hence, the healthiness of these communities proof that researchers working on these topics are not interdisciplinary, they only published in their community, unlike communities with higher healthy value.

Community	C_{HT}	*Author*	$AvgH-index$	C_{IMP}
wsc-scsc	0	2348	8	6
iscas-date	0	14069	22	13
icalt-aied	0	3129	14	11
iros-icra	0	11047	24	17
biostec	0	1931	14	8
kes-iwann	0	4450	18	13

Table 2: Isolated communities and their scientific impact

In summary, the analysis showed that each community has different h-index distribution, which means that the scientific productivity differs in each community, and this affects on ranking scientific content if we do not consider the context. The healthiness of the community helps to identified closed/unhealthy and open/healthy communities. We found important difference in their healthiness between communities with similar scientific impact and size. With this metric, many search algorithms can be proposed based on these values, such as the interdisciplinary of authors, or diversity of content.

6. CONCLUSION

We have shown that scientific communities may have different productivity, and introduced community-based metrics can help to improve current search mechanism by using the power of communities. We have proposed a tool that implements the community discovery algorithm and calculates community-based metrics that seek to improve the actual way scientific content and researchers are assessed and searched.

Future work consists in providing different algorithms for discovering communities using different networks, proposing community based evaluation metrics, providing support for

evolution of the communities, and performing experiments on different datasets.

7. ACKNOWLEDGEMENTS

The LIQUIDPUB project acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 213360.

8. REFERENCES

[1] An Informatics Europe report. Research evaluation for computer science. Prepared by the Research Evaluation Committee of Informatics Europe. Version 6.0, 20 May 2008, 2008.

[2] Association for Computing Machinery. The acm digital library. Website. `http://portal.acm.org/dl.cfm`.

[3] Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008* (2008), P10008.

[4] Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. Maximizing modularity is hard.

[5] Girvan, M., and Newman, M. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*, 12 (2002), 7821.

[6] Google. Google scholar beta. Website. `http://scholar.google.com.py/`.

[7] Newman, M. Analysis of weighted networks. *Physical Review E 70*, 5 (2004), 56131.

[8] Newman, M. E. J. Analysis of weighted networks.

[9] Newman, M. E. J., and Girvan, M. Finding and evaluating community structure in networks. *Arxiv preprint cond-mat/0308217* (2003).

[10] Ratprasartporn, N., Po, J., Cakmak, A., Bani-Ahmad, S., and Ozsoyoglu, G. Context-based literature digital collection search. *The VLDB Journal 18*, 1 (2009), 277–301.

[11] The Pennsylvania State University. Citeseer: Scientific literature digital library and search engine. Website. `http://citeseerx.ist.psu.edu/`.

[12] Thomson Reuters. ISI Web of Knowledge. Website. `http://isiknowledge.com/jcr`.

[13] Wang, X., Jiao, L., and Wu, J. Adjusting from disjoint to overlapping community detection of complex networks. *Physica A: Statistical Mechanics and its Applications 388*, 24 (2009), 5045–5056.

[14] Wu, F., and Huberman, B. Finding communities in linear time: a physics approach. *The European Physical Journal B - Condensed Matter 38*, 2 (March 2004), 331–338.