UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Svetlana Vorotnikova

Effects of Homophily on Citation Patterns in Scientific Communities

Master's thesis (30 ECTS)

Supervisor: Luciano García Bañuelos, Ph.D.

Author: ..... "....." June 2011

Supervisor: ..... "....." June 2011

Professor: ..... "....." June 2011

TARTU 2011

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Luciano García Bañuelos for his support and guidance throughout the research. I am also grateful to all LiquidPub research team for their support in the beginning of the research process.

I would like to extend my gratitude to my family and especially to my husband, Sergey Omelkov, who always encouraged me during the project.

Abstract

Individuals tend to establish ties in higher rate with individuals that exhibit some kind of affinity than with dissimilar ones. This principle, referred to as homophily, has a important impact in the shape of the social interactions, i.e., topologies of the underlying social networks. In the context of bibliometrics, the effects of homophily can be observed in patterns of citation: references often include a non-negligeable number of self-citations and citations from close collaborators. In light of the above, we aim at designing bibliometric indicators that allow us to modulate the effect of homophily in the ranking induced by metrics. Clearly, homophily-aware metrics would favor communities where citations involve a broader participation. In this thesis, we present homophily-trimmed and homophily-weighted versions of the citation count and report on the patterns of citation uncovered by such metrics over the citation network for three different communities.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Commonly research results are published in scientific journals or presented on the conferences. After publication, research results are used by other researchers in their subsequent articles. The citation of one article by another is may be taken as a way to recognize the impact of the cited paper. Therefor, this data can be used to statistically and mathematically measure the quality of scientific product or even the scientist himself. The methods utilized in this measurement are known as *bibliometrics*.

In recent years the usage of bibliometric indicators in evaluation of research impact has become mainstream. Various rankings and metrics have been developed and adopted, such as citation count, h-index and Impact Factor. Such metrics are easy to compute and they provide a meaningful estimation of the impact and productivity of the scientist or his/her product. Oftentimes the simplicity of formulation makes them sensitive to noise and even vulnerable to attacks. For instance the simple metric such as citation count might be enough when someone is interested in having an idea about productivity and impact of the results of the researcher. However, this metric would clearly be unfair with young researchers. Lalo and Mosseri [1] suggested that having a paper with 5 or 6 co-authors would positively affect metrics, as this practice would benefit from a larger number of publications and potentially increased number of cross-references.

Citation of scientific literature is intrinsically a social process. Through the principles of social systems citations refer to *homophily*. The principle of homophily implies the fact that individuals are tend to establish ties with individuals who behave similarly than with dissimilar ones. Moreover, homophily has an important impact into the shape of the network that models social interactions. Homophily can be understood as being inversely proportional to the distance in the social graph.

In light of the above, we define homophily-aware metrics as the attempt to understand this social phenomenon in bibliometrics. Using public datasets, we analyze the patterns

of citation observed in major conferences and journals in different scientific communities.

The remaining of the document is organized as follows. In Chapter 2 we review the concept of homophily, then briefly summarize the field of bibliometric indicators, their advantages and problems. Then we examine different aspects of social network analysis applicable to citation and coauthorship graphs in details. In Chapter 3 we propose a family of bibliometric indicators that takes into account the effects of homophily. In Chapter 4 we discuss three different datasets that we used for assessing the usefulness of the proposed metric. Thus we present an analysis of citation patterns observed in those datasets, which corresponds to three different research communities, namely Computer Science, High-Energy Physics and Computational Linguistics.

Chapter 2

Background and state of the art

In this chapter we introduce the concept,that provide a background to our work. Firstly, we define the concept of homophily. Secondly, we provide short overview of existing bibliometric indicators. Thirdly, we discuss the social network analysis techniques which used in present work. Finally, existing work on relevant fields is covered.

2.1 Homophily

People with different characteristics (gender, ethnicities, age, education) appear to have different qualities. We tend to consider this qualities as essential aspects of their category membership. For instance, women are emotional, educated people are tolerant, and gang members are violent [2]. By interacting only with people who are like ourselves, anything that we experience gets reinforced.

Homophily is the principle that each individual tend to establish ties with individuals who exhibit kind of affinity than with dissimilar ones. It is expressed in proverb “Birds of a feather flock together”. The fact of homophily means that any kind of information (cultural, behavioral, genetic) tend to be localized in the scope of sociodemographic space. Homophily implies the distance in social terms as a distance in network.

Probably the most influential source of homophily is space, people more likely to have contact with those who are closer in geographic location than those who are distant. The advent of new technologies like telephone and Internet may have loosened the bounds of geography by decreasing the effort involved in contact, but in fact actually not eliminated old patterns.

Other research in this area studied how ties in co-membership are affected by similarity to other members of a group. Both strong and weak ties to others in the group,

which are also likely to be among similar others, tend to increase the duration of memberships [3].

2.2 Bibliometric indicators

Bibliometric indicators seek to measure the quality and scientific impact of books, articles and other forms of publications. Indicators usually based on the number of scientific papers and citations they received.

There are three types of bliometric indicators [4]: *Quantity indicators* measure the productivity of the particular researcher or research group. *Performance indicators* measure the quality of a journal, researcher or group. *Structural indicators* measure connections between researchers, journals or research fields.

In the following Section we describe some example metrics of each type.

2.2.1 Quantity indicators

Number of publications. The most simple and straightforward method is to count the number of articles published by particular author or research group. It is generally not possible to compare authors or groups since the number of publication does not reflect the quality of the article. When comparing groups, one has to keep in mind the fact that number of publications depends on the group's size[5].

To count *number of publications in Top-ranked journals is* the way to overcome this limitations. However, this method do not cope with the effect of the group size.

2.2.2 Performance indicators

Performance indicators usually intended to measure performance along the quality dimension and can be used to measure the impact of the research into scientific community. How often an article or a journal is cited by others is an indication of performance. Impact factor and h-index, two highly discussed bibliometric indicators in this field, are briefly discussed in this section.

ISI impact factor(IF) was first proposed by Gerfild in 1955 [6]. The IF of a journal is established each year on the basis of previous two-year period. It is determined as follows:

C_x- citations in year x to articles published in the two preceding years,

P_{x-1-2} - number of articles published in year $x-1$ and $x-2$,

$I_{ISIx} = C_x/P_{x-1-2}$ - impact factor in year x .

Strength of the indicator includes its independence from the size of the journal, stability and high comprehensibility. However, IF has several limitations. Firstly, it does not reflect the quality of each article published in journal. Thus it is not clear whether the high degree is a result of high quality of all articles or only some of them. Secondly, multidisciplinary journals have a higher IF than specialized journals [7]. The possible explanation of this fact could be that multidisciplinary journals have large relationships from variety of fields and thus cited more frequently. Third, there are difference between research fields. Highly ranked journals in each specialized field can have significantly different IF, so it becomes impossible to compare interdisciplinary journals. Fourth, the type of articles also influence journal IF. For example, some technical reports and review articles are more likely to receive citations than original research articles and case reports[8]. And finally, editors of journals can influence to increase journal IF, by recommending for authors to cite articles published within the journal. The conclusion was given by Weingart in 2005 [9]: impact factors in their undifferentiated form are outdated and should not be used as measures in any evaluative context at all. Yet, they are probably the most popular bibliometric measure of all.

H-index [10] attempts to measure productivity and the impact into research area of the scientist. H-index is based on the scientist's top cited papers and number of citations they received. The Figure 2.1 shows the curve giving the number of citations versus paper number, the intersection with 45 degree line gives h-index.

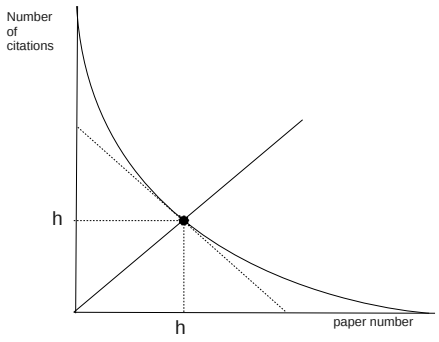


Figure 2.1: H-index visualization graph

But such metrics have a bunch of weaknesses. So Laloe and Mosseri [1] found that according to this metrics the contribution of an author is exactly the same whether he/she has 10 co-authors. Authors also emphasized that in some scientific areas after technical breakthroughs appears flurry of publications and most of them becomes forgotten after

some time. As a consequence, bibliometric indicators are vary sensitive to fashion. The more popular scientific field, the easier to obtain higher metric. This research shows that such bibliometric indicators are very sensitive to noise and vulnerable to attacks.

2.2.3 Structural indicators

There are some structural indicators that can be computed. For example, one can analyze the field where the unit was published and field in which it is cited. Further one can make description of the cognitive structure of the research field, or of their co-authors and co-authors affiliations[4]. Since this part straightly related to the present research, it will be discussed in section 2.4 more precisely.

2.3 Social network analysis

Social network analysis is the key technique in modern sociology. It is the mapping and measuring relationships and flows between the connected entities. In our case the nodes in the network are people, while links show collaborations or citation relationships between the nodes. The shape of the network can determine the network usefulness to its individuals. More open networks with weaker connections are more likely to introduce new opportunities to their members. A group of individuals with connections to other social networks is likely to have access to wider range of information. Also individuals can play a role of connector by bridging two networks that are not directly connected. Social network analysis produces an alternative view, where attributes of individuals are less important than their relationships within the network.

Variety of metrics has been proposed to measure the social networks. The concept were taken from [11].

Let $G = (V, E)$ be an undirected, simple (no self-loops, no multiple edges) graph (network) with a set of nodes (vertexes) V and a set of edges E.

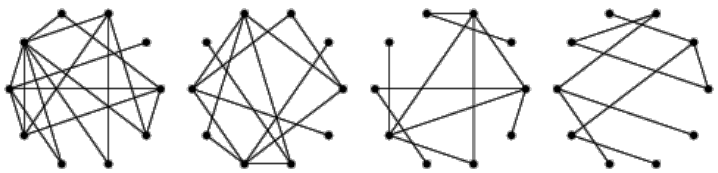


Figure 2.2: Random graphs of 10 vertexes have diameter 3,4,5 and 7, respectively.

Graph's diameter is the largest number of vertexes which must be traversed in order to travel from one vertex to another when paths which backtrack, detour, or loop are ex-

cluded from consideration. In other words, the length $max_{u,v}d(u,v)$ of the longest shortest path between any two graph vertexes (u,v) of a graph, where $d(u,v)$ is a graph distance (Figure 2.2).

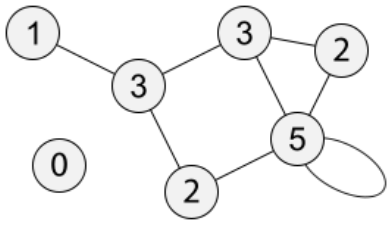


Figure 2.3: Graph with vertexes labeled by degree

The *degree* $d(v)$ of node v is defined to be the number of nodes in V that are adjacent to v. On figure 2.3 minimum degree is 0 and maximum is 5.

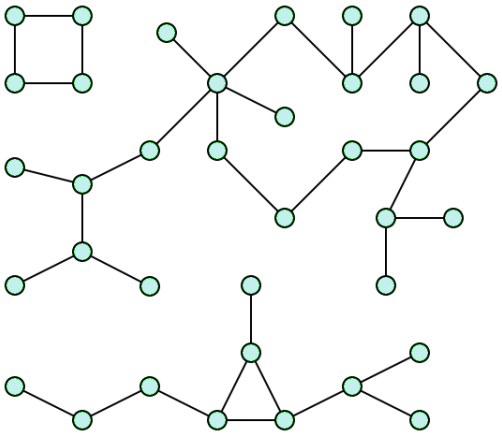


Figure 2.4: Connected components of the graph

The connected component of undirected graph is the subgraph where each two nodes are connected directly or indirectly. On the graph shown on figure 2.4 there are three connected components. *Largest connected component* in this case is 22, which make 61% from whole graph.

Average distance between nodes is defined as average number of "hops" that required to move from one any given node in the network to another.

2.4 State of the art

The usage of bibliometric indicators in the assessment of research outcomes is commonplace. A considerable number of similarity measures have been studied in the field of

bibliometric. For example, in [12] authors estimated communities similarity in Orkut social network. They proposed 6 different measures, which are based on the number of mutual users of communities, that was normalized in a different manner. Lehmann et al [13, 14] tried to estimate authors quality on the basis of citation data. They examined citation distribution in the High Energy Physics dataset. This approach emphasizes the popularity of publications rather than their prestige. To address this limitation, variants of PageRank have been proposed. The notion of PageRank firstly was proposed by Page et al. [15] in 1998 in the context of relative web page importance based solely on the page location in the Web's graph structure. Lately, in 2006, Bollen et. al [16] proposed a weighted version of PageRank algorithm to obtain the metric that reflects prestige in the domain of scholarly assessment. These measures however still work on the basis of the citation network only. As a result, they do not distinguish between intra-community and inner-community citations. Liu et al. [17] propose an extension of PageRank that takes into account citation networks and co-authorship networks. Meanwhile, Zhou et al. [18] introduced Co-Rank, which makes use of three networks: the citation network, the co-authorship network, and author's social network. Despite of the plenty of networks, it still does not differentiate between inter-community and intra-community citations.

In [19, 20] Newman characterized the patterns of collaboration among scholars of several disciplines by analyzing the characteristics of the co-authorship graph. Newman observed important differences in the collaboration patterns in several disciplines: mathematicians and physicists seem to work mostly individually or with a few coauthors, whereas biologists reveals highly collaborative settings.

In [21] authors studied information diffusion in the citation network. They proposed a method to identify citations between communities and concluded that citations that occur within communities lead to the slightly higher number of direct citations; and also citing more recent papers corresponds to receiving more citations in turn.

Chapter 3

Homophily aware metrics

The work presented in this chapter is related to the homophily aware metrics, which allow to modulate the effect of homophily in the ranking induced by metrics.

In this work we study network of scientists in which two scientists considered connected if they have at least one mutual paper. We did not considered the fact that some scientists could know one another but have never collaborated. The other measure of similarity is related to citation ties.

3.1 Similarity measures

As it was mentioned before, the concept of homophily describes relationships that are based on some measures of similarity. Despite the fact that no single definition exists for similarity, in the scope of this work, we consider two types of relationships to define similarity: *citations and co-authorship*.

Using these relationships, we can measure similarity to be inversely proportional to the distance that separates researchers in the co-authorship network and papers in the citations network. Following this basic intuition, the rationale behind our homophily aware metrics consists on using the distance on these graphs to ponderate traditional metrics, which in this particular work is applied to citation count. Is therefore important to start these section by defining both of these networks.

Definition 1 (Co-authorship Graph). A *Co-authorship Graph* is a tuple $G_a = (A, Co, W, Ta)$, where

- A is the set of vertexes on the graph, each vertex representing an author
- $Co \subseteq A \times A$ is a set of undirected edges (i.e. unordered pairs), each edge representing a co-authorship relation between two authors

- $W : Co \to N$ is a weight function that maps each edge $\{a_1, a_2\} \in C$ to the number of papers that a_1 and a_2 have co-authored
- $T_a : Co \to N$ is a function that maps each edge $\{a_1, a_2\} \in Co$ to the year of publication of the first paper co-authored by a_1 and a_2.

Definition 2 (Citation Graph, Citation Count). A *Citation Graph* is a tuple $G_c = (P, Ct, T_p)$, where:

- P is the vertexes on the graph, each vertex representing a paper
- $Ct \subseteq P \times P$ is a set of directed edges (i.e. ordered pair), each edge $(p_1, p_2) \in Ct$ denoting the fact that paper p_1 cites paper p_2
- $T_p : P \to N$ s a function that maps a paper $p \in Ct$ to the year of publication of paper p.

Let $p \in P$ be a paper. The Citation Count of paper p is defined as follows:

$$CC(p) = |\{q \mid (q, p) \in Ct\}|$$

Given these two graphs, the homophily ponderation is done by using the concept of *neighborhood* of a paper within the Co-authorship Graph and up to a certain distance. The next element to define is therefore a function that given a paper return the set of authors on its neighborhood Γ.

Definition 3 (Paper k-Neighborhood). Given a paper p, the 0-neighborhood of p, written $\Gamma(p, 0)$, is the set of authors of the paper. The 1-neighborhood of p, written $\Gamma(p, 1)$, is the set of authors who have coauthored at least one paper with one or more co-authors of p, prior to the year of publication of p. Recursively, for a given integer $k > 0$, the *k-Neighborhood* of p, written $\Gamma(p, k)$, is the union of the $\Gamma(p, k-1)$ and all authors who have co-authored a paper with at least one author in $\Gamma(p, k-1)$ prior to the year of publication of p. Formally:

$$\Gamma(p, 0) = AuthorsOf(p)$$

$$\Gamma(p, k) = \Gamma(p, k-1) \cup \left\{a \mid \exists \{a, b\} \in Co : T_a(\{a, b\}) \leq T_p(p) \wedge b \in \Gamma(p, k-1)\right\}$$

Where:

- $AuthorsOf : P \rightarrow \mathscr{P}(A)$ is a function to represent the set of authors of a single given paper.

With the concept of neighborhood, the citation count can be redefined to include only citations coming from the neighborhood, which is equivalent to count citations coming from similar researchers, where similarity is based on the closeness between authors.

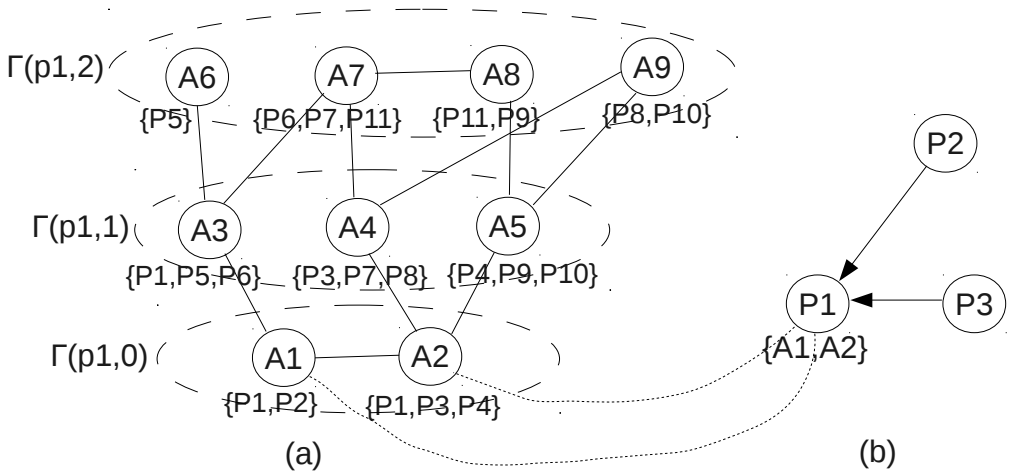


Figure 3.1: k-neighborhood for the paper P1

To illustrate the concept lets consider co-authorship (a) and citation (b) graphs in figure 3.1. Assume that paper P1 coauthored by authors A1 and A2, then to the 0-neighborhood for paper P1 corresponds authors A1 and A2 themselves, which is on the bottom of the picture 3.1(a). The 1-neighborhood of paper P1 consists of authors {A3,A4,A5}, they are on the middle of picture 3.1(a). The 2-neighborhood of paper P1 is on the top of this picture and contains authors {A6,A7,A8,A9}.

Definition 4 (k-Neighborhood Citation Count). Let $q, p \in P$ be papers in the citation graph and k the paper neighborhood threshold. The *k-Neighborhood Citation Count* of paper p is the set of papers citing p by authors up to distance k in the co-authorship graph and is formally expressed as:

$$HCC(p,k) = \{q \mid (q,p) \in Ct \wedge \Gamma(p,k) \cap \Gamma(q,k) \neq \emptyset\}$$

The structure of collaboration networks usually shows clusters around collaboration between authors and the classical *six degrees of separation* rule holds also for this kind of networks[19]. This concept includes the fact that each node in social graph can be

reached from any other node with at most 6 hops, in other words any two people in the Earth can be connected in six steps or fewer.

The rational behind homophily-based citation count consist on penalize citations coming from nearby collaborators and give higher relevance to those that are further away in the graph. In this way, the measure can better represent the impact of an author in terms of how far his/her contribution reach.

Based on this basic rationale and using definition 4, citation counts can be trimmed by not counting citations coming from the neighborhood. The basic case is the citation count without considering self-citations, which, according to definition 3, would be the case of $k = 0$. Definition 5 formalizes this idea.

Definition 5 (Homophily k-Trimmed Citation Count). Let $q, p \in P$ be papers in the citation graph and k the paper neighborhood trimming threshold. The *Homophily Trimmed Citation Count* of paper p is:

$$HkTCC(p,k) = CC(p) - HCC(p,k)$$

Having defined and clearly separated citations coming from the neighborhood and citations coming from further away, it is possible to calculate a citation sum where closest citations are weighted. The weighting function in this case is a simple linear function in the range of $[0,1)$. The analysis of different weighting functions is left for future research.

Definition 6 (Homophily k-Weighted Citation Count). Let $q, p \in P$ be papers in the citation graph and k the paper neighborhood trimming threshold. The *Homophily k-Weighted Citation Count* of paper p is:

$$HkWCC(p,k) = \sum_{l=0}^{k} \frac{HCC(p,l) \cdot l}{k+1} + HkTCC(p,k)$$

Figure 3.2 presents an example of k-neighborhood citation count. Note that the citation graph corresponds to the co-authorship graph on Figure 3.1(a). Citation count for papers $p1$ is 11 - the total number of incoming citations, $CC(P1) = 11$. The 0-neighborhood of paper 1 is the set $\{P2, P3, P4\}$, that corresponds to the set of direct coauthors. 1-neighborhood of paper 1 is the set of papers $\{P5, P6, P7, P8, P9, P10\}$ since their authors have coauthored at least one paper with co-authors of P1. Similarly, 2-neighborhood of P1 corresponds to: $HCC(p1,2) = \{P11\}$.

Thus, $HkTCC(P1,0) = CC(P1) - HCC(P1,0) = 11 - 3 = 8$,

$HkTCC(P1,1) = CC(P1) - HCC(P1,1) = 11 - 9 = 2$

$HkTCC(P1,2) = CC(P1) - HCC(P1,2) = 11 - 10 = 1$

and $HkWCC(P1,1) = \frac{HCC(P1,1)}{2} + HkTCC(P1,1) = \frac{3}{2} + 2 = 3.5$

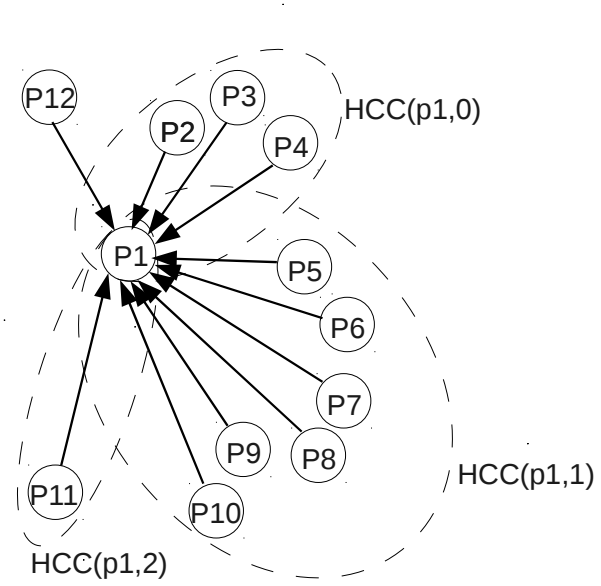


Figure 3.2: An example of k-neighborhood Citation Count

$HkWCC(P1,2) = \frac{HCC(P1,1)}{3} + \frac{HCC(P1,2)*2}{3} + HkTCC(P1,2) = \frac{3}{3} + \frac{1*2}{3} + 1 = 2,66$

So far, we have defined two variants of homophily-aware metrics, namely homophily-trimmed citation count (cf. Definition 5) and homophily-weighted citation count (cf. Definition 6). Both metrics are computed on individual publications. However, they can be adapted to provide an insight about the impact of individual researchers as described in the following Definition 7:

Definition 7(Single Author Homophily -aware citation counts). Let $a \in A$ be an author and $P_a \subseteq \mathscr{P}(P)$ be the set of publications of author a. The author homophily-aware citation counts up to neighborhood k are:

$$HkTCS(a,k) = \sum_{p \in P_a} HkTCS(p,k)$$

and

$$HkWCS(a,k) = \sum_{p \in P_a} HkWCS(p,k)$$

The metrics can also be extended to evaluate the aggregate production of research groups (e.g. laboratories, Universities, research centers, etc.), journals, venues (e.g. Conferences, Symposia, etc.) and so forth. Hereinafter, we refer to this as aggregate authorship entity. Thus to calculate any homophily-aware metric, we only need to determine the set of publications that are associated to a given aggregate authorship entity and proceed

in a similar way as for individual researchers.

3.2 Chapter summary

In this chapter we presented a family of bibliometric indicators that take into account the effect of homophily. We defined two variants of homophily-aware citation counts: homophily and weighted versions. The above metric rely on the notion of paper neighborhood, which has also been introduced. Finally, we indicate the way to move from single paper Homophily-aware metric to aggregate authorship that corresponds to collective research entities such as laboratory or research group.

Chapter 4

Homophily and citation patterns

In this chapter we present an analysis of the impact of homophily in the patterns of citation on three different scientific communities. The analysis was conducted on the citation and coauthorship networks, which were constructed from real world datasets. Foremost, we build the set of experiments on the distinct venues, and on the top ranked authors afterwards.

4.1 Datasets

For evaluation we considered three datasets that are publicly available in the web. The analysis of datasets given in Subsection 4.1.1.

DBLP is based on data from the DBLP Computer Science Bibliography. The dataset was consolidated in the context of the academic search system ArnetMiner[1] , as described in [22]. The snapshot covers publications in the period from 1947 to 2010. Data was integrated from different sources, such as ACM digital library, CitrSeer and others, cleaned and free from name ambiguity problem.

HEP-TH is a dataset containing publications on to high-energy physics theory. This data initially was set for a competition held in conjunction with the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) in August 2003 [23]. The data covers papers in the period from January 1993 to April 2003. Data appears sufficiently cleaned and without name ambiguity.

ACL is a dataset containing publications on Bibliometric and Network Analysis of the field of Computational Linguistics. The ACL Anthology Network was built from

[1] http://www.arnetminer.org/citation

the original pdf files available from the ACL Anthology. Publication period is from 1965 till 2009. The ACL Network was originally created by Mark Thomas Joseph and is currently being maintained by Pradeep Muthukrishnan under the supervision of Professor Dragomir R. Radev [24]. The dataset was cleaned and did not required additional cleaning.

For the work in hand, we parsed the available datasets and installed them into a local MySQL database. The database schema is described in appendix A.

4.1.1 Dataset analysis

Short summary about datasets is presented in Table 4.1.

	DBLP	HEP-TH	ACL
Number of papers	1397237	29554	15160
Total citation count	3021489	352807	62825
Citation per paper	2.16	11.9	4.14
Number of authors (vertexes)	1063048	14939	12470
Number of collaborations(edges)	2763240	24760	34911
Collaborations per author (n of edges/n of vertexes)	2.6	1.65	2.79
Papers per author	1.314	1.98	1.22
Average collaborators per paper	2.23	1.869	2.285
Largest component	69%	62%	72%
Average distance	5.78	7.3	4.99

Table 4.1: Short summary about studied datasets

The DBLP dataset if significantly wider than two other communities, it has about one million of distinct authors, more then three million of citations and 1.3 million of papers versus 14 thousands and 12 thousands of author in HEP-TH and ACL respectively. Meantime, an average citation count in DBLP is 2.16 citation per paper. This dataset teems with

inaccuracies such as missing information (53% of papers do not have any citation ties at all). This limitation is an effect of gathering information from different sources and some citation ties can be missed. Meanwhile High-energy physics and Computational Linguistics datasets have 11.9 citation/paper and 4.14 citation/paper respectively.

Number of collaborations per author is similar for ACL and DBLP datasets, about 2.6 in both cases and 1.8 for HEP-TH dataset. This fact may indicate more collaborative character of computer scientists and linguists than physicists. This fact also confirmed by the average number of collaborators: physicists are more used to work alone or with 1.8 collaborators per paper in average, while in rest datasets average paper is written by 2.2 collaborators. We will return to this question in Section 4.1.2.

The number of papers per author is similar among three studied fields, beween 1 and 2 in each case. However, despite the fact that physicists covers smaller time period they produced slightly more papers than their more practically minded colleagues from ACL and DBLP datasets.

Table 4.1 gives the size of largest connected component in each of the networks. For each of the studied networks the size of largest connected component vary from 62% for physicists to 72% for linguists. Overall, this picture tells us that some of the nodes are in kind of intellectual isolation from the mainstream. Most scientists who do not belong to the largest component are members of small disconnected components. As we can see, physicists dataset has the smallest connected component, which means that bigger part of scientists have never collaborated with others.

We computed the distance between 25% randomly chosen pairs of individuals in a network using a breadth-first search algorithm and then take the average to give the numbers shown in Table 4.1. For each network this number is quite small, at least compared with the size of the network.

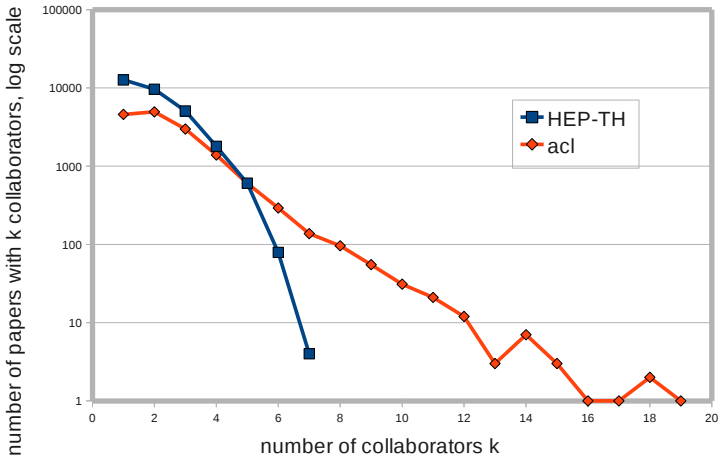
4.1.2 Patterns of scientific collaboration

In Figure 4.1 we show the distribution of the number of coauthors that scientists have for each of the three datasets. The distribution is quite similar, however physicists (squares) are more likely to work alone than computer scientists. A little less than the half of the papers were written alone (43%), 0.2% of papers was written in group of 6 and only 4 articles has 7 authors.

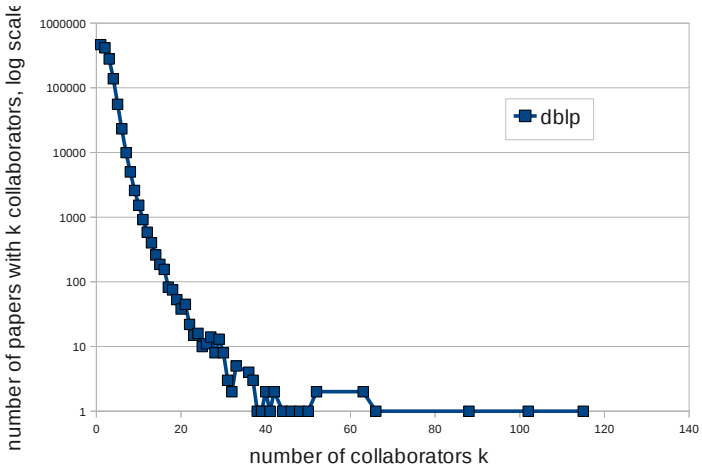
Meantime, about one third of computational linguists papers (rhombus) were written alone. The distribution for linguists has a longer tail, reflecting the higher mean number of collaborators that articles have in this field. At most this network has 19 authors per

paper. Linguists more frequently write articles in pair than alone. High collaborative nature of linguists confirms the average distance between authors, in this case it is 4.99 (Table 4.1), which means that in average 5 “hops” required along links if one needs to move from one given node to another, it comparable with DBLP average distance - 5.78. The average distance for physicists is 7.3.

The DBLP community (Figure 4.1(b)) has a bit different distribution, the tail is significantly longer then in previous network. The interesting exception is that community even has an article with 115 and another with 102 collaborators.



(a) HEP-TH and ACL



(b)DBLP

Figure 4.1: Distribution of numbers of collaborators for scientists in each dataset (note that y-axis is logarithmic)

4.2 Community citation patterns evaluation

We analyzed the citation patterns observed in 3 foregoing datasets. As we mentioned before, about half of DBLP papers do not have any citations. To overcome this limitation in community analysis we have chosen only venues with at least 8 citations per paper. Figure 4.2 presents the homophily-trimmed citation count of neighborhoods 0 up to 3 for the conferences in the analysis. The values of the metrics have been normalized to the total number of citations to allow a direct comparison of the citation patterns. The trimmed version of the metric was chosen to emphasize the importance of citations coming from close collaborators.

Without any additional experiments from the Figure 4.2 one could notice the fact that the shape of trend lines in DBLP and ACL datasets are steeper, while the lines of HEP-TH are smoother. The reason of such behavior is explained in section 4.1.2. The three lower lines in Figure 4.2(b) are the exception, which correspond yo three highly ranked journals in the High-energy physics field and that will be further discussed in subsection 4.2.1.

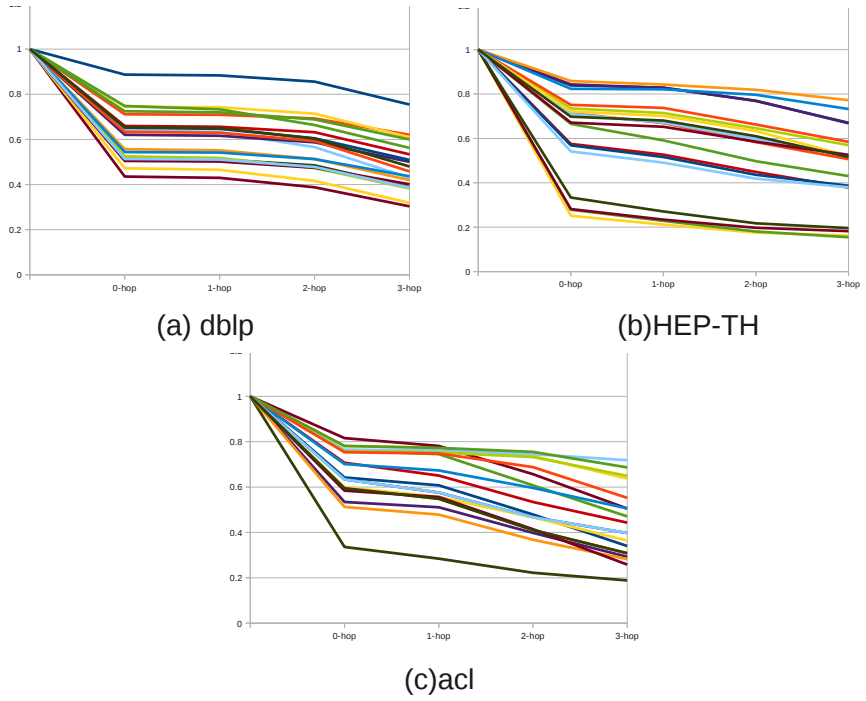


Figure 4.2: Variations of homophily trimmed citation count

4.2.1 Information diffusion in citation networks

Information diffusion is the communication of knowledge among members of a social system. In order to analyze information diffusion, one needs to study the overall information flow and individual information cascades in the networks [21]. In this subsection we examine information flow between different journals in studied networks and consider correlation between resulting clusters and homophily aware metrics.

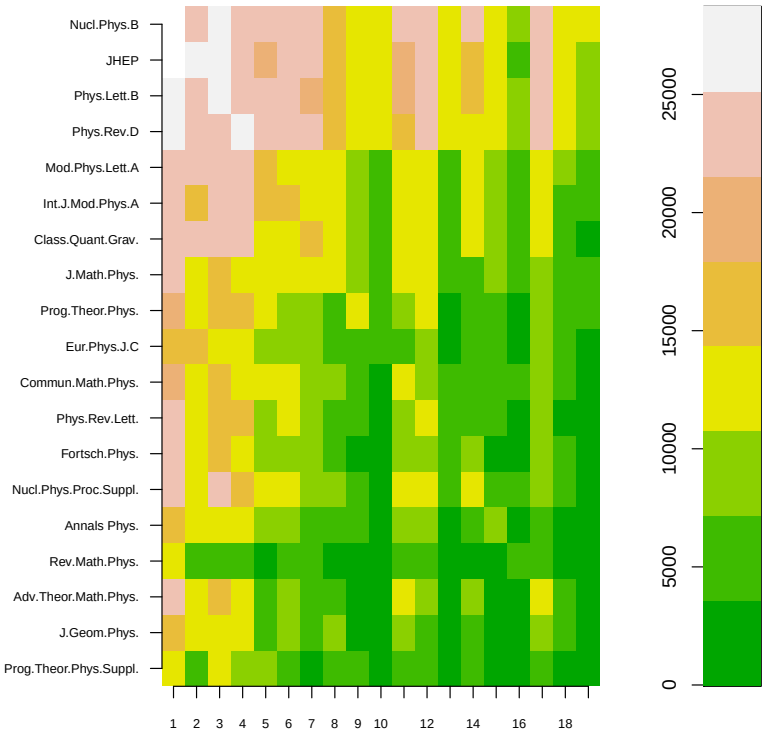


Figure 4.3: Visualization of inter-citation matrix for HEP-TH dataset. White and green cells represents the highest and lowest amount of citations respectively.

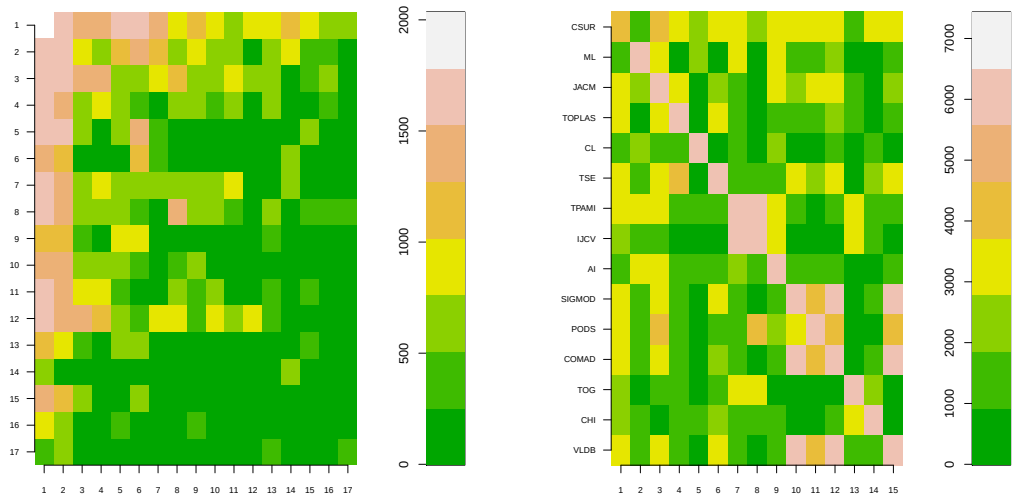
In order to quantify the density of information flow from journal to journal, we count the number of citations between every pair of journals. The results for HEP-TH dataset presented in Figure 4.3. Citing journals are on the y-axis and cited journals are on x-axis. Journals are ordered with respect to increasing of Homophily Weighted citation count with 0-hops.

By examining the matrix we can observe different densities of information flow among journals:

On the left upper corner of the matrix stands out the cluster of four journals (Nucl.Phys.B, JHEP, Phys.Lett.B and Phys.Rev.D). Journals in this cluster capture most of citations in the dataset. Interestingly, papers in this set of journals also cite other journals homoge-

neously. Three of this four journals deviate from others on the figure 4.2(b). The $H0WCS$ for this journals variate from 0.25 to 0.33, which means that 75% and 67% of citations respectively comes from 0-neighborhood.

The next 4 journals also represent the cluster. The part of non-self citations among this journals is between 0.54 and 0.66. The rest of the papers can form the third cluster, which consists of journals that not grant much citations to each other.



(a) (b)

Figure 4.4: Visualization of inter-citation matrix for ACL dataset(a) and DBLP dataset(b).

In figure 4.4 we can observe the same matrix for ACL and DBLP datasets. In figure 4.4(a) the distribution of citation flow is quite uniform, however one journal stands out form the crowd - “Annual Meeting Of The Association For Computational Linguistics”. With 67% of citations from 0-neighborhood out of 14302 total citations of this venue.

DBLP dataset (Figure 4.4(b)) behave similarly. Only one venue stands out - VLDB with 12% of citation from 0-hop neighborhood. Meanwhile, main diagonal contains all big numbers, while the rest of the matrix full of low values. It is hard to emphasize clusters from this picture. Probably the reason of this fact is the limitations of dataset, which were discussed in Subsection 4.1.1. From the above matrix we can conclude that DBLP dataset suffer from lack of inter-venue citation ties and most of citations are come from the same venue.

4.2.2 Subgraph structure and Homophily-aware metrics

For every venue we computed coauthorship graph induced by the subset of authors having published at least one paper in a given venue. We computed such parameters as size of the

largest component, number of components, clustering coefficient, average distance, diameter of the subgraph. In other words we tried to understand how the subgraph structure influence Homophily-aware metrics in the scope of journals. It is important to remind that each vertex corresponds to author and each edge corresponds to the collaboration between two connected authors.

acronym	#of vertexes	# of edges	edges/vertexes	diameter	largest component	# of components	% of self citations
csur	1896	3706	1.954	22	1166	485	12%
toplas	1701	4051	2.381	17	1202	336	29%
ml	1848	3881	2.1	18	1090	403	25%
jacm	3311	7557	2.28	20	1907	931	28%
vldb	4257	18798	4.415	15	3520	402	57%
tog	2444	8147	3.333	15	1660	387	49%
sigmod	5210	18087	3.471	17	3191	1077	48%

Table 4.2: Subgraph structure summary for 7 venues in DBLP dataset

acronym	#of vertices	# of edges	edges/vert	diameter	largest component	# of components	% of self citations
Adv.Theor. Math.Phys.	221	341	1.54	6	149	46	0.17
Annals Phys.	479	470	0.98	3	28	178	0.25
Class.Quant. Grav.	1201	1678	1.39	11	585	333	0.47
Commun. Math.Phys.	845	937	1.1	6	367	237	0.29
Eur.Phys.J.C	353	347	0.98	5	93	135	0.31
Fortsch.Phys.	282	365	1.29	6	147	76	0.28
Int.J.Mod.Phys.A	1898	2565	1.35	9	956	474	0.44

Table 4.3: Subgraph structure summary for 7 venues in SNAP dataset

acronym	#of vertices	# of edges	edges/vert	diameter	largest component	# of components	% of self citations
coling	694	1729	2.49	15	516	106	37%
HLTC	309	944	3.055	12	263	22	41%
MNAAFCL	57	90	1.57	8	30	16	22%
SIGDAT	175	365	2.085	14	132	17	24%
ANLPC	184	351	1.907	16	133	79	19%
AJCL	82	147	1.792	8	46	19	24%
AMACL	1759	5486	3.118	17	1408	221	67%

Table 4.4: Subgraph structure summary for 7 venues in ACL dataset

Table 4.2 represents the sample of our results, it contains subgraph summary for 7 highly ranked venues by the number of citations per paper for DBLP dataset. The same set of experiments was conducted for 19 venues in each dataset.

It is intuitively evident that in highly collaborative communities the number of citations from close collaborators should be higher. To justify this statement we compared the different subgraph parameters between each other.

In highly collaborative communites all authors are binded with each other, which leads to decreasing of the number of disconnected components. For instance, in the community where all authors are connected would be only one component and opposite, in the isolated communities the number of components will seek to the number of nodes. This statment confirmed by the Pearson correlation coefficient in the first column in Table 4.2.

Average node degree is also a measure of graph density. The more collaborators in average has an author the more binded the network. Average degree of a node can be find by deviding number of edges in the subgraph to the number of vertexes. As shown in the second column in Table 4.5, the average degree of a graph is highly correlated with the number of citations from 0-neighborhood. Differently behave HEP-TH dataset, which, as we already know, famous by trend of authors to work alone. Lonely authors represents in a graph as disconnected nodes with degree 0. They influence the average node degree of a graph, but actually does not involved to the graph structure. Here is the reason of inconsistency.

dataset	Pearson correlation	
	# of connected components VS. 0-hop citations	edges/vertexes VS. 0-hop citations
DBLP	0.7	0.77
HEP-TH	0.73	0.1
ACL	0.62	0.81

Table 4.5: Pearson correlation coefficient between number of connected components and 0-hop neighborhood

4.3 Analysis of citation patterns for individuals

dataset	Pearson correlation		
	top-30 authors	top-60 authors	top-100 authors
DBLP	0.58	0.56	0.4
HEP-TH	0.48	0.45	0.43
ACL	0.53	0.52	0.44

Table 4.6: Pearson correlation coefficient between number of citations from 0-neighborhood and number of collaborators

We now turn our attention to citation patterns exhibited by individual researchers. It is intuitively evident that researchers are more aware of works from direct collaborators such that significant amount of references account for self-citations and citations from close collaborators. To justify above statement we compute number of collaborators for top-30, 60 and 100 most prolific authors by the number of citations per paper versus the total number of self citations received by given author (Figure 4.5). We use Pearson correlation coefficient to find out if there is any correlation between this parameters. Such coefficient can be seen from table 4.6. As we can see from the table, the correlation decreasing while threshold growing. The cause of this fact is the noise coming with authors with smaller number of citations per paper.

Above correlation is an essential principle in the definition of homophily. As a whole, the positive correlation corroborates abundance of self citations from close collaborators.

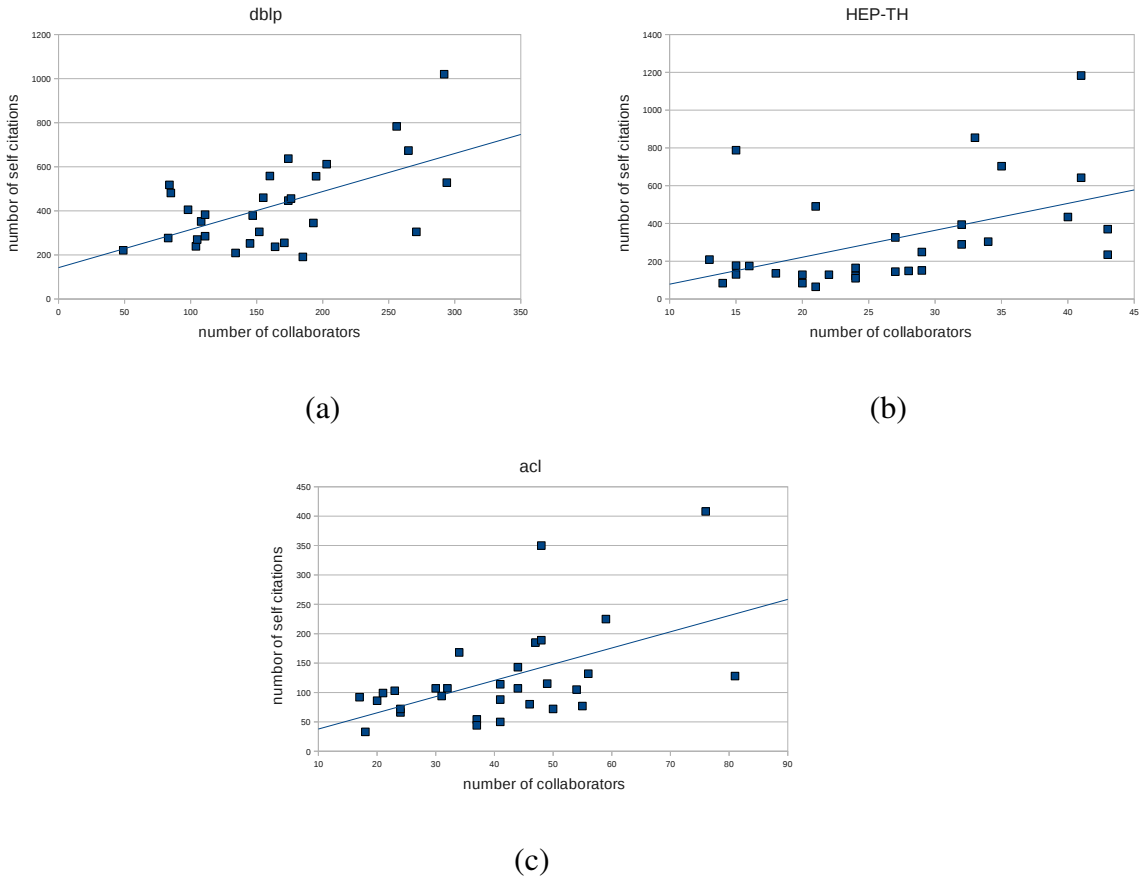


Figure 4.5: Dependency between number of collaborators and number of self citations for top-30 researchers.

4.3.1 Correlation between H-index and homophily aware metrics

H-index attempts to measure impact and productivity of scientific work. As homophily aware metrics, h-index is based on the number of citations received by researcher. This two metrics has conceptually different ideas of measuring productivity, such as h-index exclude part of publications on the basis of citation count, while homophily aware metrics get rid of citations from close collaborators considering them as unweighted.

Table 4.7 shows top-30 scientists in the field of computer science by the number of citations per paper. As expected, all scientists are highly ranked with respect to h-index (h-index varies from 40 to 105) and with high Homophily coefficient (from 0.75 to 0.93), which mean that less then 25% of citations to this authors comes from collaborators. Pearson correlation coefficient here equal 0.35. The correlation expresses the medium depth.

H-index attempts to find balance between productivity and quality and to avoid papers heavy weighted by citations. Lehmann et al [13] suppose that Hirsch establishes equality

author	h-index	H0TCC	author	h-index	H0TCC	author	h-index	H0TCC
Anil K. Jain	96	0.93	Deborah Estrin	94	0.87	Raghu Ramakrishnan	54	0.83
David A. Patterson	42	0.93	Zohar Manna	60	0.87	Michael Stonebraker	49	0.81
Jeffrey D. Ullman	87	0.93	Rajeev Motwani	75	0.87	Jiawei Han	51	0.81
Ian H. Witten	40	0.92	Scott Shenker	105	0.87	Christos Faloutsos	44	0.80
Christos H. Papadimitriou	79	0.91	Hector Garcia-Molina	89	0.87	Ken Kennedy	75	0.80
Randy H. Katz	56	0.90	Ian Foster	90	0.85	H. V. Jagadish	44	0.79
David E. Goldberg	54	0.90	W. Bruce Croft	49	0.85	Giovanni de Micheli	63	0.79
Prabhakar Raghavan	45	0.89	Rajeev Alur	49	0.88	Saul Greenberg	51	0.79
Amir Pnueli	54	0.88	Umeshwar Dayal	48	0.85	Serge Abiteboul	49	0.77
David Harel	46	0.88	David Maier	43	0.83	Oded Goldreich	44	0.75

Table 4.7: Top-30 authors in computer scince with h-index and 0-hop Homophily Trimmed Citation Count normalized to the number of citations

between incommensurable quantities, such as number of publications and number of citations. Meantime, our Homophily Aware metrics focus just on the quality of papers and do not rely on the number of papers at all. As a result this disagreement leads to discrepancy between above metrics.

H-index and Homophily-aware metrics can supplement each other. For instance, high h-index confirmed with high Homophily-aware metric indicate really productive researcher with high level of impact since high Homophily-aware metric verify that citations was gained not from close collaborators. Integration of this two metrics can be left for future work.

4.4 Homophily trimmed and weighted metrics comparison

Both trimmed and weighted versions of Homophily Citation Count metrics could be used. Recall that trimmed version of Homophily Aware metric represents the absolute number of citations excluding citations from k-neighborhood, while the weighted version of

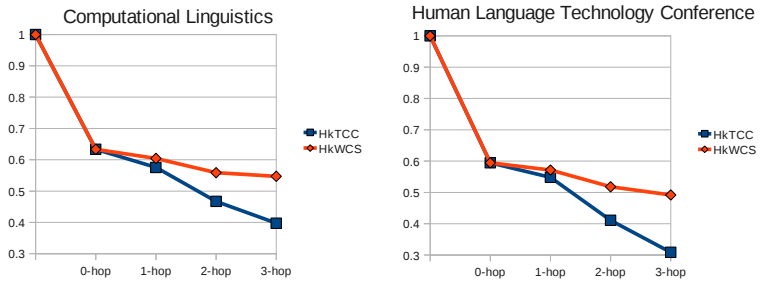


Figure 4.6: Homophily trimmed and weighted citation count comparison for two venues in ACL datasets.

metric favor citations from afar. To illustrate differences between them lets consider two venues from each dataset (Figure 4.6). The trend line in all cases behave equally: trimmed version of metric is slightly steeper and weighted version is flattened due to weighted coefficients.

4.5 Threats to validity

The experimental evaluation has been conducted using 3 foregoing datasets. The choice of those datasets was driven by the fact that they are publicly available, sufficiently prepared for scientific purposes, with wide range of publications and citations. However, it is important to note that datasets could be uncompleted or some important data could be missed, like incompleteness of DBLP dataset. This limitation has an impact on the percentage of self citations relative to the total number of citations to a paper. Due to this limitation, the absolute numbers shown in Table 4.1 need to be taken with care. Specifically, the values we have computed for the zero-trimmed citation count are likely to be lower than in reality.

Nevertheless, the biases created by DBLP dataset limitations was overcome: for analysis we choose venues with at least 5 citations per paper. This restriction affects all venues in dataset in more or less equal ways. Also the overview of datasets in Subsection 4.1.1 was based on coauthorship graph and almost didn't include citation relationships. Therefore, we suppose that the comparisons between conference citation patterns made on the basis of the above tables and figures remain valid despite this bias.

4.6 Chapter summary

In this chapter we performed various experiments on the given datasets. First, we observed three datasets in general, discussed their characteristics. We examined the tendency of authors to work in conjunction on the article. The intermediate conclusion was that physicists are more likely to work alone then other scientists under consideration. Second, we went deeper into the datasets and dwell on the level of venues. There we observed citation patterns in the scope of citation flow between different venues. Also we precisely examined subgraph structures and found some meaningful correlations between subgraph structure and homophily. Afterwords, we turned to individuals. We observed correlation between h-index and homophily, it appears that those indexes do not correlate well. And finally, we observed the essential principle of the definition of homophily, we have shown numerically the existence of correlation between the number of 0-hop citations and amount of collaborators.

During the experiments it was found that HEP-TH dataset differs from the others. Probably, this is a result of different modes of research, with computational linguists and computer scientists being applied and theoretical and high-energy physicists being primarily theoretical. The theoretical nature of the research field indirectly had an impact to the effect of homophily.

Chapter 5

Conclusion

The aim of this thesis was to define the family of homophily-aware metrics and then comprehensively examine the effect of homophily on citation patterns in scientific communities. Our metric consider two types of relationships: co-authorship and citation. This feature allows us to distinguish between citations that come from close collaborators and those which comes from afar.

Influence of homophily was tested on three different datasets. We numerically confirmed the existence of correlation between sociological aspect of homophily and patterns of scholarly citation ties. We have tried to address the problem from three points of view: datasets in general, at the level of venues and at the level of individual scientists. Also, we concluded that effect of homophily less noticeable in the theoretical research field.

Additionally of the work we have provided a review of existing bibliometric indicators, their advantages and limitations. We have also provided sociological point of view to the phenomenon of homophily.

Future work

The work presented in this thesis can be extended into different directions. Here are some of them:

- The analysis of different weighting functions for Homophily k-Weighted Citation Count, to more accurately encourage citations that came from afar. Provide an analysis of usage of other weighted functions.

- Propose such bibliometric indicator, that would contain advantages of h-index and also would be aware of homophily. This metric will cover the larger spectrum of aspects that can be evaluated.
- Apply utilized method to adjacent field of research, such as analysis of Web service networks.

Sarnasuse mõju viitamise mustritele teaduslikes kogukondades

Svetlana Vorotnikova

Indiviidid kipuvad looma sidemeid pigem nende isikutega, kellega neil on sarnaseid huvisid või muud ühist. Bibliomeetria kontekstis avaldub sama põhimõte viitamise mustrites - artiklites viidatakse muuhulgas ka autorite endi ning nende koostööpartnerite artiklitele. Eelneva valguses on käesoleva töö eesmärgiks projekteerida bibliomeetrilised indikaatorid, mis võimaldaks meil arvestada sarnasuse mõju teadlaste ning teadusasutuste hindamisel. Konkreetsemalt defineeritakse käesolevas töös sarnasuse alusel kärbitud ja kaalutud versioonid viidete arvu meetrikale. On ilmselge, et sarnasust arvestavad meetrikad annavad kõrgema hinnangu kogukondadele, kus on tava viidata eelkõige teiste kogukondade autorite artiklitele. Samuti kirjeldatakse ning analüüsitakse antud töös viitamise mustreid, mis tuvastati kirjeldatud meetrikate rakendamisel viidete võrgule kolme erineva teadlaste kogukonna puhul.

Bibliography

[1] Frank Laloe and Remy Mosseri. Bibliometric evaluation of individual researchers: not even right... not even wrong! *Europhysics News*, 2009.

[2] Lynn Smith-Lovin Miller McPherson and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001.

[3] P. Popielarz JM. McPherson and S. Drobnic. Social networks and organizational dynamics. *Am. Sociol. Rev.*, 1992.

[4] J. Lundberg. Bibliometrics as a research assessment tool: impact beyond the impact factor. *Stockholm, Sweden: Karolinska Institutet*, 2006.

[5] R.E. De Bruin H.F. Moed and T.N. Van Leeuw. New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. *Scientometrics*, pages 381–422, 1995.

[6] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122:108–111, 1955.

[7] R. Rousseau. Journal evaluation: technical and practical issues. *Library Trends*, 2002.

[8] PO. Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ*, pages 498–502, 1997.

[9] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, pages 27–37, 2005.

[10] J.E. HIRSCH. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences.*, 2005.

[11] F. Harary. *Graph Theory*. Reading, 1994.

[12] M. Sahami E. Spertus and O. Buyukkokten. Evaluating similarity measures: A large-scale study in the orkut social network. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[13] A.D. Jackson S. Lehmann and B. Lautrup. Measures and mismeasures of scientific quality. 2005.

[14] B. E. Lautrup S. Lehmann and A. D. Jackson. Citation networks in high energy physics. *Physical Review E*, 2003.

[15] R. Motwani L. Page, S. Brin and T. Winograd. The pagerank citation ranking: bringing order to the web. *Tech. rep., Stanford Digital Library Technologies Project*, 1998.

[16] M.A. Rodriguez J. Bollen and H.V. De Somphel. Journal status. *Scientometrics*, pages 669–687, 2006.

[17] M.L. Nelson X. Liu, J. Bollen and H. Van De Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, pages 1462–1480, 2005.

[18] H. Zha D. Zhou, S. Oshanskiy and C.L. Giles. Co-ranking authors and documents in a heterogeneous network. *In Seventh IEEE International Conference on Data Mining*, pages 739–744, 2008.

[19] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 2001.

[20] M. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, pages 5200 – 5205, 2004.

[21] L. Adamic X. Shi, B. Tseng. Information diffusion in computer science citation networks. *Proceedings of the Third International ICWSM Conference*, 2009.

[22] Limin Yao Juanzi Li Li Zhang Jie Tang, Jing Zhang and Zhong Su. Arnetminer: Extraction and mining of academic social networks. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, 2008.

[23] J. M. Kleinberg J. Gehrke, P. Ginsparg. Overview of the 2003 kdd cup. *SIGKDD Explorations*, 2003.

[24] Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The acl anthology network corpus. In *Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore, 2009.

[25] J. Cohen. *Statistical power analysis for the behavioral sciences*. 1988.

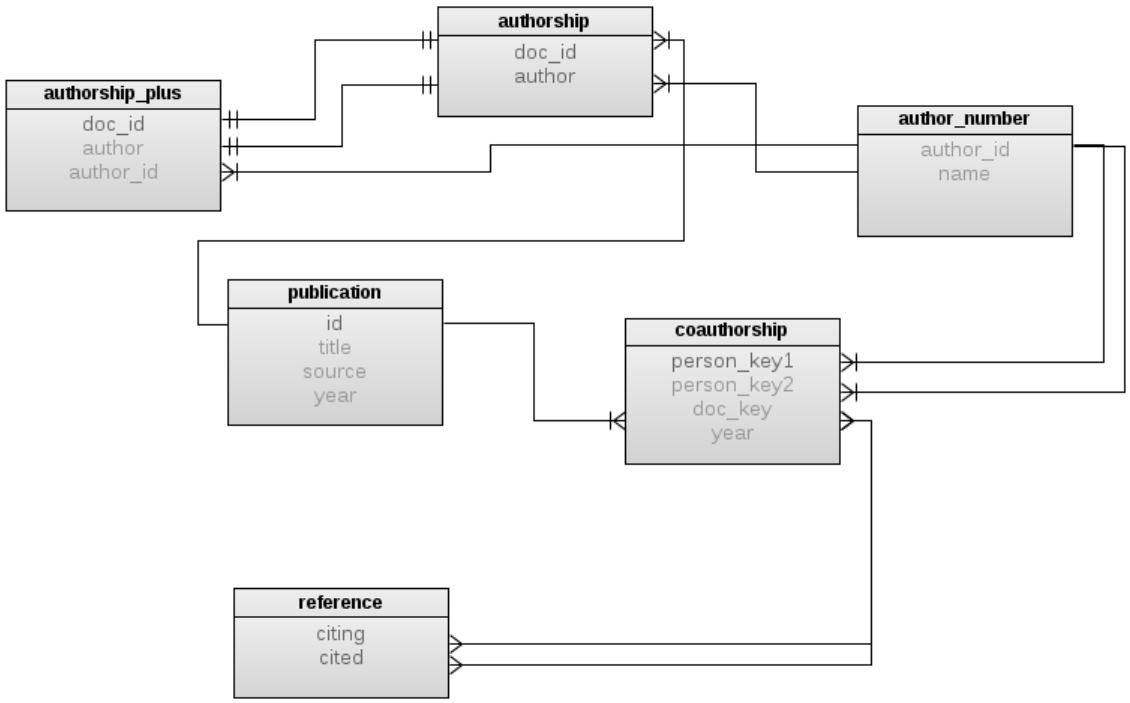
[26] L. Egghe. Theory and practice of the g-index. *Scientometrics*, 2006.

[27] RK Merton PF Lazarsfeld. Friendship as a social process: a substantive and methodological analysis. 1954.

[28] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

Appendix A

Database schema diagram



Appendix B

Supplementary materials

All related materials could be found on attached CD.

CD include:

- Datasets. Located in folder “Datasets” on the CD. Datasets represented as MySQL dump files.
- Spreadsheets with the statistics presented in the document. Located in folder “Statistics”.