Research Collaboration and Topic Trends in Computer Science – An Analysis Based on UCP Authors

Yan Wu
wy011@ie.cuhk.edu.hk

Srinivasan Venkatramanan
sriniv.venkat@gmail.com

Dah Ming Chiu
dmchiu@ie.cuhk.edu.hk

Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong

ABSTRACT

Academic publication metadata can be used to analyze the collaboration, productivity and hot topic trends of a research community. Recently, it is shown that authors with uninterrupted and continuous presence (UCP) over a time window, though small in number (about 1%), amass the majority of significant and high-influence academic output. We adopt the UCP metric to retrieve the most active authors in the Computer Science (CS) community over different time windows in the past 50 years, and use them to analyze collaboration, productivity and topic trends. We show that the UCP authors are representative of the overall population; the community is increasingly moving in the direction of Team Research (as opposed to Soloist or Mentor-mentee research), with increased level and degree of collaboration; and the research topics become increasingly inter-related. By focusing on the UCP authors, we can more easily visualize these trends.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrival**]: Information Search and Retrieval—*Information filtering*; J.4 [**Social and Behavioural Sciences**]: Sociology

General Terms

Measurement

Keywords

UCP author; research collaboration; topic trend

1. INTRODUCTION

As a research field established in the 1960s [1], Computer Science has gone through rapid development and becomes a mature field. Much can be learned about the developments and trends in Computer Science by analyzing the publication metadata. In this study, we take a particular approach, by focusing on analyzing the top 1% authors with uninterrupted and continuous presence, referred to as the UCP authors. The idea is that these authors are the core of the community and are representative of what the whole community is doing. By analyzing their activities, we can visualize the major trends of the whole community.

The term "UCP author" was used in [5] to describe the major finding of their study of publication metadata obtained from Scopus in a specific time window of 16 years. That is, during the period from 1996 to 2011, the number of authors who published papers every year without stop amounts to about 1% of all authors; and these UCP authors co-authored a much larger percentage of papers and amassed a high percentage of total citations, compared to the average researchers.

In our study, we further explore the nature and extent of collaborative efforts by these UCP authors in comparison with average authors. Recently, it was pointed out [12] that "Team Science" is an important trend in how research is done. The phenomenon is manifested in steady increase in the number of co-authors for publications over time. Since this trend exists not only in science but also in other research fields, we can refer to it as "Team Research". The "team" in Team Research may correspond to an organized group within an organization, or collaboration partnership between researchers in different organizations and countries. From the collaboration patterns of UCP authors, we can get more insight about Team Research, in particular its correlation to research productivity.

Since our metadata comes with classification of each paper into a research subdomain in Computer Science (e.g. "Databases" , "Machine learning and Pattern Recognition" or "Networks and Communications" etc), it is possible to tell the subdomains each UCP author works in. Given the moderate size for UCP authors, it is possible to apply graph clustering algorithms to find the collaboration clusters for UCP authors in Computer Science over time, and characterize these clusters in terms of the major subdomains they are working on. This analysis shows a trend of a continuous convergence towards fewer large clusters of inter-related inter-disciplinary research.

The rest of the paper is organized as follows. Section 2 briefly reviews related works. Section 3 gives a description of the data set for our analysis. Section 4 analyzes UCP authors and the team research phenomenon in detail. Section 5 shows the research topic trends based on the clustering of

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742015.

UCP authors. Section 6 concludes the paper and provides some directions for future work.

2. RELATED WORK

One of the classic works on research collaboration is by Newman [8]. He treated the coauthor network as a special kind of social network and showed some structural properties of such a network. Subsequent research focused mainly in two aspects. One aspect is the statistical change of the amount of collaboration. For example, [9] showed the overall increase in the number of coauthored articles in the literature; [12] coined the term "team science" and showed the increasing dominance of team science in the production of knowledge; and [11] studied the network effects on authors' collaboration behaviors. The other aspect is on the structural evolution of the collaboration network. For example, [6] analyzed the eigenvector evolution of the coauthor network and proposed a spectral evolution model to show the change of coauthor structures; [4] proposed a stochastic Poisson model with optimization tree, which can efficiently predict the increment of collaboration based on local neighborhood structure.

Another group of papers related to our work studied the factors that may influence productivity or authors' research behavior. For example, [3] showed the effects of aging on researchers' publication patterns and described researchers' publication style during different stages of career. The study in [10] found the existence of the Matthew effect in academic publishing, which may favor senior and experienced researchers. Finally, [5] was the first to introduce the notion of UCP as a way to identify a set of core authors in a research community, and showed the dominance of these authors in the production of academic outputs.

Most of the previous works (except [5]) analyzed the entire population of a community. By focusing on UCP authors, which is a much smaller, but important and representative subset of the overall population, we are able to find more results about trends in research collaboration (team research), its relationship to research productivity, and the evolution of research topics and focus as well.

3. DATASET

Our data is collected from Microsoft Academic Search (MAS)[1]. MAS has maintained a huge amount of publication data for a wide range of (15) research fields. For each field, it further categorizes the papers to belong to different subdomains in that field. For example, in the Computer Science field, it includes 24 subdomains such as "Databases", "Machine learning and Pattern Recognition", "Networks and Communications" and so on[2]. Each paper is labelled with a unique numerical ID; its metadata includes paper title, author list, publication year, publication venue and a reference list. Likewise, authors are maintained as another type of object. Each author is labelled with a unique numerical ID as well; its metadata includes current affiliation and publication history. An author's research field and research subdomains in that field can be obtained from his publication history. We choose the Computer Science field, which seems most complete (and we are most familiar with) for a case study in this paper. The same methodology can be applied to data in other research fields.

Considering the fact that the data for the earlier and the most recent years are less complete, we take the data in the 50-year window $[1960, 2009]$ for our analysis in this paper[3]. We filter out paper records without publication year or author information for this study. Table 1 presents a general description of our dataset.

Table 1: Dataset Description

Field coverage	Computer Science
Time coverage (year)	$1960 - 2009$
#papaers	2698044
#authors	1393143
#publishing links	6643575

4. UCP AUTHORS AND TEAM RESEARCH

In this section, we analyze and compare the collaboration levels and patterns of UCP authors versus average authors, as well as their productivity. We end with a discussion how this analysis helps us further understand the trend of team research, and the role of UCP authors in it.

4.1 UCP Author Slightly Redefined

In [5], authors defined UCP authors by considering a specific window of years, from 1996 to 2011, and observed that there are about 1% such authors. For our purposes, we make a small twist in the definition. For each year to be used as the start of a window, we find the top 1% authors in terms of the length of uninterrupted and continuous presence from that starting year. This gives rise of a window size for UCP authors for each year. For example, starting from year 1988, the UCP window size needs to be set as 8 (which means the ending year of that UCP window is 1995), in order to make the percentage of UCP authors among authors with at least one publication in that UCP window around 1%. Smaller window size will lead to a higher percentage than 1% while larger window size will lead to a lower percentage.

With this modified definition, it is observed that the window size required to be counted as a UCP author is different for each starting year. In fact, this UCP window size is growing steadily over the years, as shown in Fig. 1. This certainly correlates well with our impression that the top authors are becoming more and more active. The number of years required to become a UCP author starting from 1996 is around 11, which is a little less than that found in [5]. This is not too surprising considering the research field and dataset studied are both different. But the result is in the same ball park.

4.2 Comparison on Collaboration

We first compare the collaboration patterns of UCP authors with that of average authors. UCP authors is the author set including all UCP authors in a UCP window, while "average authors" is the author set including any author with at least one publication in a UCP window. So average authors is a superset of UCP authors. We take the UCP year window $[1988, 1995]$ for comparison. We compare the nature and extent of collaboration, such as coauthor size, collaboration strength and team connectivity, and then the

[1]http://academic.research.microsoft.com/

[2]Some papers are only classified as Computer Science papers, but not categorized into any subdomain.

[3]The data from MAS was lastly collected on July 31, 2012.

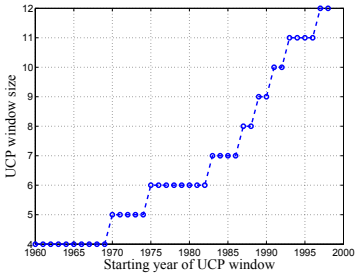


Figure 1: Change of UCP window size.

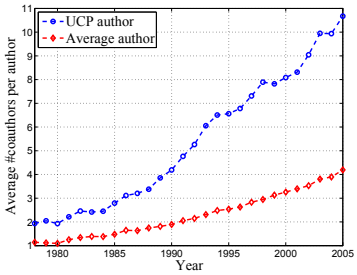


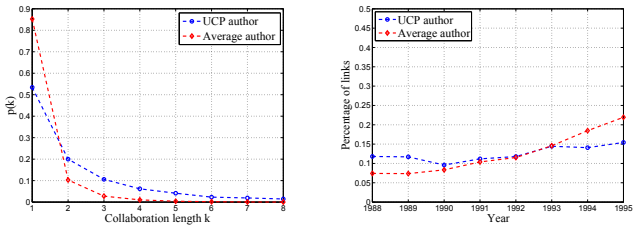
Figure 2: Comparison of coauthor size between UCP authors and average authors.

productivity. Note that although we only present the comparison for one UCP window here, the results for other UCP windows are similar, and are not presented due to space limitation.

One important measure for the extent of collaboration of an author, obviously, is the size of the coauthors set. Fig. 2 shows the average number of coauthors per author for UCP authors and average authors on an annual basis. Here the coauthors include UCP coauthors and non-UCP coauthors. The figure shows that UCP authors generally have more coauthors than average authors. Moreover, the UCP authors also have a significantly higher increase rate of coauthors, although more collaboration is the trend for the whole community [9, 12].

A likely reason for the much higher number of coauthors for UCP authors is directly due to "team science" and the growth in team size. As the collaboration pattern in a team is often hierarchical, and the UCP authors are more likely at the root of the hierarchy, they would naturally have more collaborators and benefit from growth of team sizes. If we assumed the coauthor network is built by preferential attachment [7], we would reach the same conclusion for UCP authors. To further understand the collaboration pattern by UCP authors, we also show the coauthor size of the same set of UCP authors during their pre UCP period and post UCP period for ten years in Fig. 2. It is clear that even before and after their UCP period, UCP authors tend to have more coauthors. The difference between UCP authors and average authors in the ten years before their UCP period is not so much as that in later periods. But there still exists slight advantage to UCP authors. This indicates that in order to become UCP authors, it is important for authors to build and expand their research teams in the very beginning. The further growth of the number of coauthors in the post UCP window is likely due to the reputation and connections they accumulated during their UCP period.

In a social network, while the number of friends may show the size of one's social connections, the tie strength can better reflect the extent of one's influence in his social network. Similarly, in the study of research collaboration, we can also use tie strength to represent the collaboration strength between each pair of coauthors. We define the collaboration strength as the number of years one collaborates with another in a time window, i.e., their collaboration length. Again we compare UCP authors with average authors using the UCP window [1988, 1995]. Fig. 3(a) shows the distribution of collaboration length in the 8-year UCP window for UCP authors and average authors. We observe that for both UCP authors and average authors, more than half of the collaboration links exist for only one year, which shows the dominance of short-term research teams. However, UCP authors are more likely to have longer collaboration relationship with others. This indicates that although UCP authors have a rapid expansion rate of coauthors, there still exist some stable collaboration links. For the transient links which last for only one year, the distribution of the year in which the one-year collaboration happens is plotted in Fig. 3(b). While it is almost uniformly distributed in the 8 years for UCP authors, it is left skewed for average authors. UCP authors keep a regular proportion of transient collaboration links, while the average authors have more short-term collaboration links in recent years, which may be the result of rapid increase of paper publishing over years.



(a) Collaboration length (b) Distribution of one-year link

Figure 3: Comparison of collaboration strength between UCP authors and average authors.

Next, we analyze the structure of the coauthor networks built in the 8-year window by UCP authors and average authors respectively. Here the UCP coauthor network contains the collaboration between UCP authors only and the average author coauthor network consists of all the authors with publications in the specific time window. For simplicity, we have removed authors with no coauthors (single nodes only) in the two networks. The result is shown in Table 2, where we focus on the analysis of connected components in the two networks. The number of connected components is a lower bound to the estimated number of clusters in the coauthor network. It reflects the connectivity in the network as a whole. As shown by Table 2, UCP authors are more connected with each other while for average authors, small teams are more popular.

4.3 Comparison on Productivity

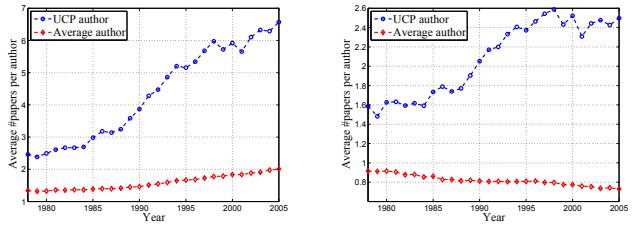
Besides the collaboration patterns, a more direct assessment of an author's activity in the community is produc-

Table 2: Analysis on Connected Components

Statistical measurement	UCP author	Average author
#nodes in network	2317	212527
#links in network	5677	410606
#connected components	61	25201
#nodes in giant component	2170	135391

tivity, which is often reflected by the annual publication rate of an author. Before going to the detailed discussion of productivity, we define two notations first: individual productivity (IP) and community productivity (CP). IP is the annual number of claimed papers per author. Thus IP is incremented for an author every time his name appears in a paper. CP, on the other hand, is based on the fractional contribution of each coauthors towards a paper (equal division assumed). CP counts each paper only once, while IP counts each paper n times when there are n coauthors. Fig. 4 shows the comparison of IP and CP for UCP authors and average authors.

Similar to the comparison of coauthors, we also include production behaviors in the ten years before and after the UCP period. For the comparison of IP, we can see that IP almost doubles in the 28 years for average authors, while it is much more than doubled for UCP authors. This can be partially explained by the different coauthor sizes of UCP authors and average authors. Different from CP, the contribution of coauthors can help increase one's IP. We can see in Fig. 2 that while the annual number of coauthors for average authors increases from 1 to 4 in the 28 years, it increases from 2 to 11 for UCP authors in the same period. Such a rapid expansion of collaboration thus inevitably leads to more productivity for UCP authors. For CP, there is a slight decreasing trend for average authors, whereas for UCP authors, the trend is increasing over the 28 years. This shows that although the UCP authors are consistently increasing their productivity, whether measure by IP or CP, the productivity (CP) of the average authors are actually decreasing. This phenomenon was also observed and discussed in the context of team science [12].



(a) Comparison of IP (b) Comparison of CP

Figure 4: Comparison of productivity for UCP authors and average authors.

4.4 Discussion

The analysis and comparison of UCP authors and average authors, on both collaboration and productivity, give us further insights into the trend of team research. UCP authors are able to achieve more sustained research activity, much higher level of research output, and accelerated growth in research output, because of their ability to build research teams to help them as well as extensive research collaboration with a broader range of other authors including other UCP authors. In this regard, the UCP authors are serving as the core of the research community.

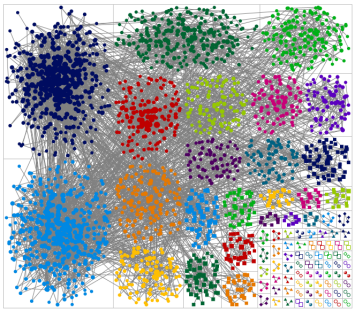
In this study, we rely on only publication and co-authorship data, so it is not possible to give any assessment of research impact. A useful future direction is to study the research impact, based on whatever reliable measures for that, of UCP authors and non-UCP authors. This will help us further understand how good research results are achieved in the era of team research.

5. FROM UCP AUTHORS TO TOPIC TREND

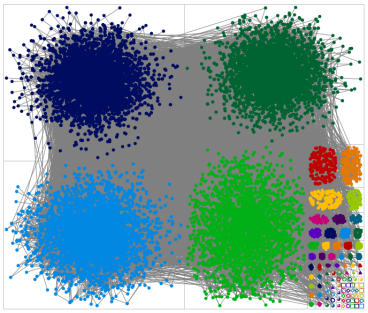
From our analysis above, UCP authors can be considered as the core of their academic community. Therefore, by observing the evolution of the research topics the UCP authors work on, we can get a sense of what have been the hot topics in the community.

For this study, we take two sets of UCP authors in the UCP windows [1988, 1995] and [1998, 2009] respectively, and compare the difference between them. As with the previous analysis, we first build the two UCP coauthor networks based on the existence of collaboration links between UCP authors in each window respectively. Note, for these two networks only the UCP authors in the respective time windows are included. The UCP authors without any coauthors (hence singleton nodes) are removed, as they will not be part of any clusters anyway. We then apply the Clauset-Newman-Moore algorithm [2] to do clustering for the two UCP coauthor networks. Fig. 5 shows the clustering result for the two windows. Different clusters are marked with different colors and put in different grids. The grey lines between different grids represent the collaboration relationships among different clusters. We can see that in the earlier window, the clusters are more fragmented. The smallest cluster contains only two authors (the minimum possible size); even the largest cluster is not so big. Besides, the connections (represented by the number of links) between different clusters are not strong. Our interpretation is that in the earlier years, researchers tended to work more in isolation or with small scale (e.g. thesis mentor-mentee type of) collaboration, with little cross teams collaboration. The period [1988, 1995] is also before the advent of WWW, which can be attributed as an important factor of increased research collaboration. In the second window [1998, 2009], however, four largest clusters with similar sizes emerged and seemed to dominate all the other clusters in size. Moreover, many more collaboration links exist between different clusters. This shows that Computer Science as a research field had become more interdisciplinary (at least within its field) with much more extensive collaboration among its researchers.

Since our clustering is conducted based on the existence of collaboration links, and through the publications of each UCP author we extract their major research subdomains, we can visualize research as the mixing (or collaboration) of ideas from different research subdomains. For the years in the first window of time, a few research subdomains are the focus of research then, and many other research ideas were emerging and small research subdomains were just being formed. This is manifested by the large number of research clusters, the minimal collaboration between these clusters, and each cluster hosting a relatively homogeneous group of researchers. This is illustrated in Fig. 6(a), where each node

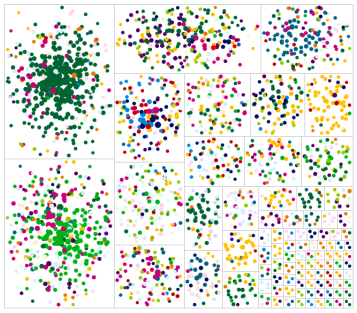


(a) 1988 – 1995 UCP window

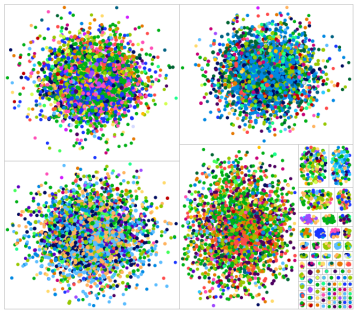


(b) 1998 – 2009 UCP window

Figure 5: Comparison of clustering results in the two UCP windows.



(a) 1988 – 1995 UCP window



(b) 1998 – 2009 UCP window

Figure 6: Major subdomain of UCP authors in the two UCP windows.

represents a UCP author, with a color representing the major research subdomain of that author (The major subdomain is set as the one most of that author's publications during the UCP period belonging to.). By the second window of years, a large number of UCP authors belong to the four major clusters with heavy intra-cluster collaboration, with a relatively small fraction of UCP authors still working in smaller clusters. Furthermore, the four clusters are no longer so homogeneous, with a more mixed set of colors, as shown in Fig. 6(b). Again, each node corresponds to a UCP author, with a color representing the major subdomain that UCP author works in, during the respective time window.

By now, you must be curious about what the large clusters in each of these two windows are. We show the answers in Fig. 7(a) and (b) for these two time windows. For each cluster, we show its composition in terms of the distribution of its researchers from the 24 subdomains of our metadata[4]. For the earlier time window [1988, 1995], the top three clusters are made up of mostly (1) "Algorithms and Theory" people, (2) "Databases" people, and (3) "Programming Language" people respectively. By the second time window, the top four dominating clusters are each hosting a more mixed set of UCP authors, with the dominating subdomains being, respectively

(1) "Algorithms and Theory" and a set of application or technology areas, including "Networks and Communications", "Security and Privacy", "Computer vision", "Graphics" etc;

(2) "Databases" and "Artificial Intelligence" and some application or technology areas, including "Networks and Communications", "Human Computer Interactions", "Data Mining" etc;

(3) "Hardware and Architecture", "Software Engineering" and "Distributed and Parallel Computing", which may all be considered to be related to computing systems;

(4) "Artificial Intelligence", "Machine Learning and Pattern Recognition", "Multimedia", "Natural Language and Speech", and "Networks and Communications", which may all be considered to belong to multimedia technology, applications and systems.

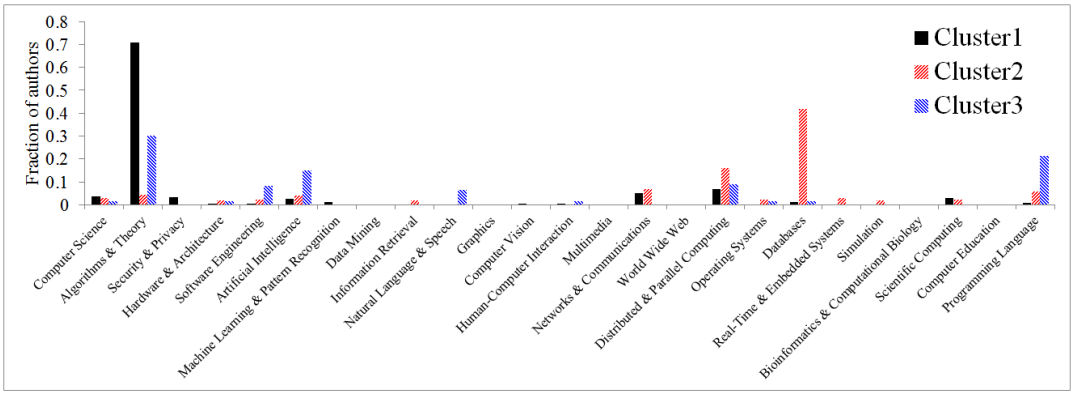
These large clusters seem to map to the hot research areas and focus in Computer Science during those time periods.

From Fig. 7(b), since there are more mixing of different subdomains in forming large clusters, we can also get a sense which subdomains tend to mix (collaborate) with others, and which subdomains tend to mix with each other. It seems "Networks and Communications", perhaps playing an infrastructure or glue role, tend to mix with others the most. "Artificial Intelligence" seems to mix mostly with "Machine Learning" and "Databases", which perhaps represent the "thinking" and "memory" aspects of artificial intelligence. Finally, it is also clear that the trend is for more and more inter-disciplinary research, rather than for people in each subdomain working alone.

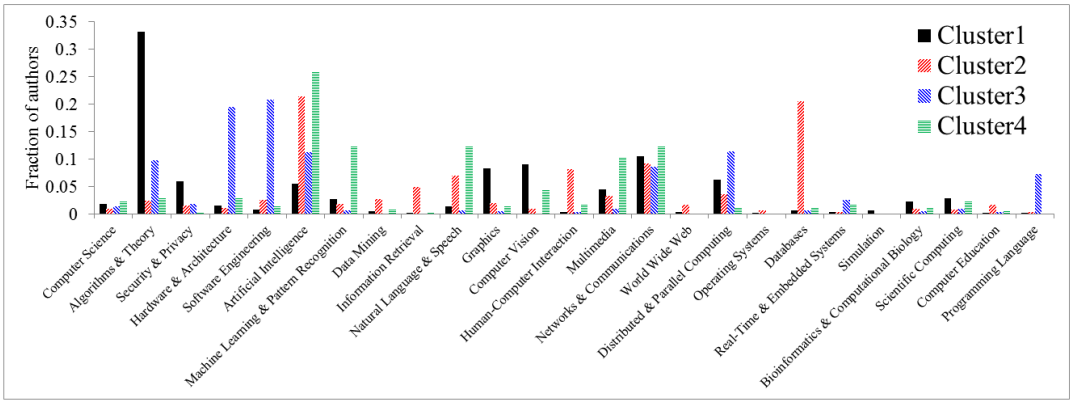
6. CONCLUSION

In this paper, we took an analysis on a new set of authors, i.e., UCP authors, who have uninterrupted and continuous

[4]We use subdomain name "Computer Science" to represent papers belonging to Computer Science, but not categorized into any subdomain.



(a) 1988 – 1995 UCP window



(b) 1998 – 2009 UCP window

Figure 7: Subdomain distribution in the two UCP windows.

presence in the scientific literature over a period. Thus UCP authors may represent the most active researchers in the community. We analyzed and compared the collaboration patterns and productivity of UCP authors versus average authors in the Computer Science field. Results show that UCP authors are serving as the core of the research community and the study of UCP authors can help us have a better understanding of the general trend of team research in the community. We also studied the research topic trends by analyzing the evolution of research topics the UCP authors work on. Results indicate that Computer Science, as a research field, is showing an increasing tendency of interdisciplinary research in the community.

Our analysis is just an initial attempt to the understanding and visualization of the general trend in the academic ecosystem. For future work, analysis on datasets in other research fields can be conducted and more measurements besides the ones we focused on in this paper can also be proposed.

7. REFERENCES

[1] J. G. Brookshear. *Computer science: an overview*. Addison Wesley, 2012.

[2] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[3] Y. Gingras, V. Lariviere, B. Macaluso, and J.-P. Robitaille. The effects of aging on researchers' publication and citation patterns. *PLOS ONE*, 3(12):e4048, 2008.

[4] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 107–116. ACM, 2008.

[5] J. P. Ioannidis, K. W. Boyack, and R. Klavans. Estimates of the continuously publishing core in the scientific workforce. *PLOS ONE*, 9(7):e101698, 2014.

[6] J. Kunegis, D. Fay, and C. Bauckhage. Network growth and the spectral evolution model. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 739–748. ACM, 2010.

[7] D. Lee, K.-I. Goh, B. Kahng, and D. Kim. Complete trails of coauthorship network evolution. *Physical Review E*, 82(2):026112, 2010.

[8] M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[9] T. L. O'Brien. Change in academic coauthorship, 1953–2003. *Science, Technology & Human Values*, page 0162243911406744, 2011.

[10] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences*, 109(14):5213–5218, 2012.

[11] S. Uddin, L. Hossain, and K. Rasmussen. Network effects on scientific collaborations. *PLOS ONE*, 8(2):e57546, 2013.

[12] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.