

The Artificial Immune Systems Domain: Identifying Progress and Main Contributors Using Publication and Co-Authorship Analyses

Alaa ABI HAIDAR^{1,2,3,4}, Adrien SIX^{1,2}, Veronique THOMAS-VASLIN^{1,2}, and
Jean-Gabriel GANASCIA^{3,4}

¹ CNRS, UMR 7211, Immunology, Immunopathology, Immunotherapy, 75013 Paris

² UPMC Univ Paris 06, UMR 7211, Integrative Immunology team, 75013 Paris

³ CNRS, UMR 7606, ACASA, LIP6, 4 place Jussieu, 75005 Paris

⁴ UPMC Univ Paris 06, UMR 7606, ACASA . LIP6 . 4 place Jussieu, 75005 Paris

{alaa.abi_haidar, adrien.six, veronique.thomas-vaslin, jean-gabriel.
ganascia}@upmc.fr

<http://www.immunocomplexit.net>

Abstract. Much can be learned about the progress, fathers and future of a scientific domain from the analysis of a collection of relevant articles and their corresponding authors. Here, we study the highly interdisciplinary domain of Artificial Immune System (AIS) since its birth, a couple of decades ago. We apply Social Network Analysis to the co-authorship network of the most comprehensive publicly accessible AIS bibliography. We automatically extract publication dates and author names from the bibliography and evaluate authors with the highest degree (unique collaborations) and centrality (influence).

Our results highlight the relative growth of publication volume and identify significant contributors in the AIS field. Furthermore, our findings are not only encouraging for the AIS community but may be useful for analyses of other scientific communities and leading contributors therein.

Keywords: Artificial Immune Systems, Social Network Analysis, Co-authorship, Information Extraction, Text Mining

1 Introduction

Artificial Immune Systems (AIS) are adaptive systems, inspired by theories and observed principles of the immune system, and applied towards solving computational problems [1]. Common AIS techniques are based on specific theoretical models explaining the behavior of the vertebrate adaptive immune system such as negative selection, clonal selection, immune networks and dendritic cells [2]. AIS are mainly classified into two categories: The first one aims at mathematically modeling the immune system to better understand its behavior. The second one uses the immune system as a metaphorical inspiration to engineer algorithms that are capable of learning and solving a huge variety of machine learning problems such as classification, clustering and regression analysis.

AIS is a relatively new field which began in the mid 80's with the modeling and refinement of Jerne's immune network theory [3] by Farmer et al. [4], and later by Bersini and Varela [5, 6]. However, it wasn't until negative selection was used by Forrest [7] for protecting computer networks from viruses that the AIS domain was established. Cooke and Hunt [8, 9] adapted immune networks for classification and Timmis [10] further improved it while De Castro et al [11, 12] worked on aiNet for multimodal function optimization and data analysis. The first book on AIS was edited by Dasgupta in 1998 [13].

In the past few years, several review papers have discussed the slow advances in AIS and proposed improvement strategies through novel and simpler AIS models (inspired by the vertebrate innate immune systems and immune systems of plants) as well as the development of a unified architecture for integration of existing models [14, 2, 15, 16]. Timmis argues that AIS has reached an impasse [2] and Timmis et al pointed out a dearth of theory to justify the use and continuity of AIS [17].

The domain of AIS has existed for a couple of decades but has never been analyzed quantitatively. Furthermore, the fear that the domain is stagnating was only based on qualitative studies targeting specific AIS models and frameworks, such as negative selection, clonal selection and more recently, the danger theory.

In this paper we use techniques from co-authorship network analysis and statistical methods in order to investigate the current state and future of the domain of AIS. Moreover, we identify leading contributors to the field using co-authorship network analysis and discuss our results. In the following section, we discuss the methods used for information extraction and statistical analysis, in addition to social network analysis of the co-authorship network. In section 3, we illustrate and discuss the growth of the AIS domain and compare it to the general growth of Medline indexed articles in general. Finally, in section 3.3, we discuss our results and the future directions of the field.

2 Methods

We adopt data mining techniques in order to extract author names and publication dates from the most comprehensive AIS bibliography [18] comprising 1044 articles and 994 authors. We only had access to a binary PDF format of the bibliography and therefore we had to convert it into an ASCII textual format. The converted text was far from the desired clean structured text and therefore a significant amount of manual curation was involved in order to systematically mark the end of the authors' list before automating the extraction of author names.

Publication dates are limited to the year of publication consisting of a 4-digit number $YYYY$ where $1980 < YYYY < 2010$ in order to avoid confusion with other numbers such as digital object identifiers (doi), volume numbers and serial numbers. More challenging was the task of author name extraction with some authors having their first names abbreviated others having full first names and others not respecting the first-name-last-name order. We manually restructured

the bibliography, ensuring that last names were preceded by abbreviated or full names. We prepended shared last names with the abbreviated first names to avoid the risk of agglomerating authors with a common last name. For example “X Lee” and “Y Lee” are considered two distinct authors and represented as: X-Lee and Y-Lee, respectively. Still, we face the possibility that two different authors would be counted as one if they share both lastnames and firstnames (at least initial). However, that is outside the scope of our analysis.

2.1 Social Network Analysis

Social Network Analysis (SNA) has played major roles in many disciplines in the past few years [19, 20]. Co-authorship Networks (CN) are social networks consisting of scientific collaborations and collaborators [21]. In CN, the authors are represented as nodes (or vertices) and collaborations as undirected edges. CN are similar to citation networks in the scientific literature [22]. However, CN have better social and collaborative implications [23]. CN Analysis has already been studied and applied to a couple of fields but never to AIS [23–25]. Indeed, studies on co-authorship analysis have shown how visible and influential can article be [26]. Other studies have focused on examining academic research performance based on a co-authorship network of centrality and gender [27].

We use several existing methods to analyze a CN of AIS as follows:

We used the R package [28] to calculate the degree, closeness, and betweenness centrality for the binary undirected co-authorship network. In the following sections, we illustrate and discuss the 20 highest ranking authors for each of the following metrics:

Degree is a measure of the unique number of collaborators an author has.

Closeness, that is only applied to the largest (connected) component, is a measure of how authors are directly connected to a well-connected author.

Betweenness is a measure of a node’s influence for information flow in the network. Betweenness measures how many times a node is visited when two random nodes are connected through a path of nodes.

For more information about centrality measures and SNA, please refer to [20].

2.2 PageRank

PageRank [29], or eigenvector centrality, is used by the Google search engine to determine a page’s relevance or importance. Important pages receive a higher PageRank and are more likely to appear at the top of the search results. PageRank is based on backlinks. The more quality backlinks the higher google pagerank.

Liu et al [23] have applied PageRank to a co-authorship network of the Digital Science domain in order to identify prestigious authors. They transformed each undirectional edge into a set of bi-directional, symmetrical edges.

2.3 AuthorRank

Liu et al [23] also define a modification of PageRank, that they call AuthorRank. PageRank assumes that when a node A connects to n other nodes, it receives a weight based on an equal fraction $\frac{1}{n}$, whereas AuthorRank attributes different weights for each author based on the number of their publications in common *ibid.*

Both PageRank and AuthorRank are measures of prestige that we use to identify leading authors in the domain of AIS as discussed in the following section.

We used the R package [28] to implement PageRank and AuthorRank to rank the top ranking authors and to visualize the co-authorship network using a Fruchterman Reingold Layout.

3 Results and Discussion

3.1 Publication Distribution Analysis

In order to answer the numerous doubts about a decline in the AIS field, we have measured the number of publications that are relevant to the field over time. Furthermore, we have fitted our observation using exponential functions to study the growth of publication size over time and to predict it over the years to come. We have used the coefficient of determination R^2 in order to validate the fitness function and its prediction as shown in figure 1.

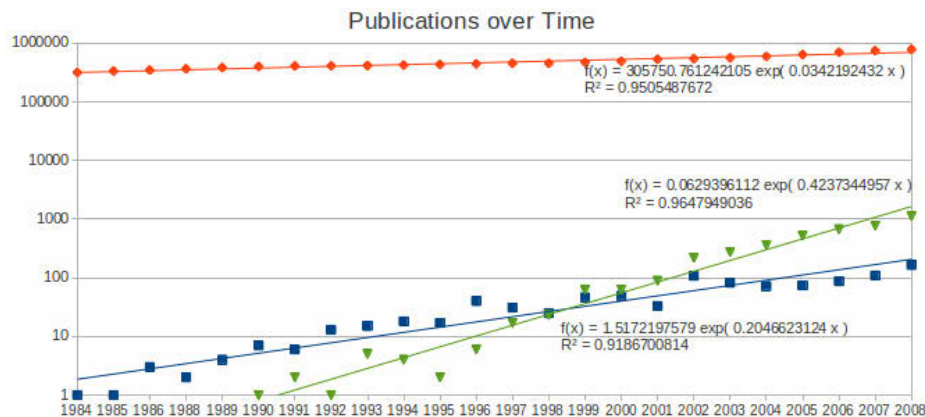


Fig. 1. Number of Publications per year 1) according to the AIS bibliography [blue], 2) according to google scholar results for “Artificial Immune Systems” [green] and 3) for all indexed Pubmed articles [red] per year. The results show a relative growth in the domain based on both AIS bibliography and Google Scholar results. The exponential fits are validated using the R^2 coefficient.

Contrary to previous fears of a stagnating AIS field, we have shown that the field of AIS is ever growing by measuring the number of publications over time. Indeed, we have used exponential fitness functions in the form of $f(x) = c.e^{(x.b)}$, where c is a constant, x is the year index and b is the growth factor. We have compared the growth in the volume of AIS articles according to our AIS bibliography ($b = 0.2$) to that of indexed Pubmed articles ($b = 0.03$) between the years of 1984 and 2008¹ using Corlan's Medline trend (<http://dan.corlan.net/medline-trend.html>) to cement our conclusion regarding the relative expansion of the AIS domain. Furthermore, we add to our perspective the number of articles returned by Google Scholar for the query "Artificial Immune Systems" which results in a faster growing trend ($b = 0.4$) as shown in green in figure 1. Dasgupta's Bibliography includes conference proceedings, as can be found in Google Scholar, whereas PubMed indexes journal articles only. However, we argue that in the fields of informatics and engineering, conference and workshop publications typically have higher impact

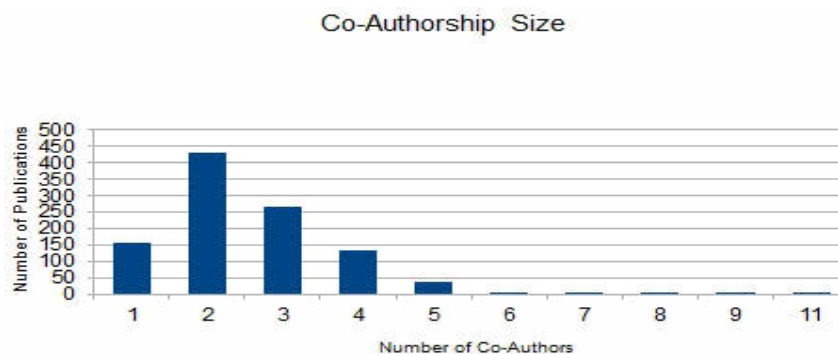


Fig. 2. The distribution of publications per authors forms a gaussian distribution around two authors.

In addition, we have analyzed the number of authors per publication that is an indication of collaboration strength. As shown in figure 2, most collaborations include two authors. Moreover, there are more 3 co-authored publications than (and almost as many 4 co-authored publications as) single authored ones. We presume that this may be as a result of the field of AIS being a very collaborative one. Indeed, a highly interdisciplinary field such as AIS invites collaborations from the fields of immunology, systems biology, artificial intelligence, machine learning, data mining...etc. Similar studies have been conducted on the field of the Digital Library Research Community [23] yielding similar results with 28.5% of papers authored by 2 authors, 23.6% by 3 authors, 19.6% by a single author and 9.4% by 4 authors.

¹ The years covered in the AIS Bibliography

3.2 Co-Authorship Network Analysis

The co-authorship network is summarized in table 1 and visualized in figure 3.

| Network Statistics | |
|------------------------|--------|
| Articles | 1044 |
| Number of Vertices | 994 |
| Number of Edges | 1768 |
| Network Density | 0.0036 |
| Number of Clusters | 159 |
| Largest Component Size | 55% |

Table 1. Co-Authorship network summary

The number of publications per author may be a good indicator of an author's contribution to the field however that can be biased in favor of non-collaborative authors with many published articles. The author's degree, or number of unique collaborators, can be a better measure of an author's collaborative efforts in a field. Degree centrality measures authors' connectivity with immediate neighbors or collaborators. Some authors may, however, be locally well connected but not globally with the entire network. Closeness centrality expands on degree centrality and favors authors that are connected (directly or indirectly) to as many authors in the network. Betweenness is another measure of centrality that measures how often a node is on a shortest path between any two random nodes in the network. Betweenness conveys the role of an author as an information spreader or a hub. The authors with the highest number of publications, degree, betweenness and closeness are listed in table 2. Authors with the highest degree, betweenness and closeness measures are illustrated in figure 4.

Several studies have analysed Co-authorship Network (CN) components for various scientific domains. Nascimento [24] reports the largest component in SIGMODs² CN having about 60% of all authors. Newman [25] has studied several CN with the smallest "largest component", containing 57.2% of all authors. Liu et al [23] report in the JCDL³ CN the largest component of only 38% (599 authors) of all authors for the years between 1994 and 2003. The low percentage may be due to a relative immaturity of the Digital Library (DL) field. The AIS CN has the largest component of 55% (550 authors) of all authors, thus showing a relative maturity of the field. We suspect that the maturity of AIS is related to 10 years of the dedicated conference, ICARIS, since 2002 and 20 years since the beginning of AIS, i.e. twice as longer than the DL field. The largest component of 550 authors is visualized in black at the center of figure 3, whereas the remaining components alternate in various colors around it. Moreover, table 3

² SIGMOD is a Special Interest Group on Management Of Data under the Association of Machinery, ACM

³ Joint Conference on Digital Library

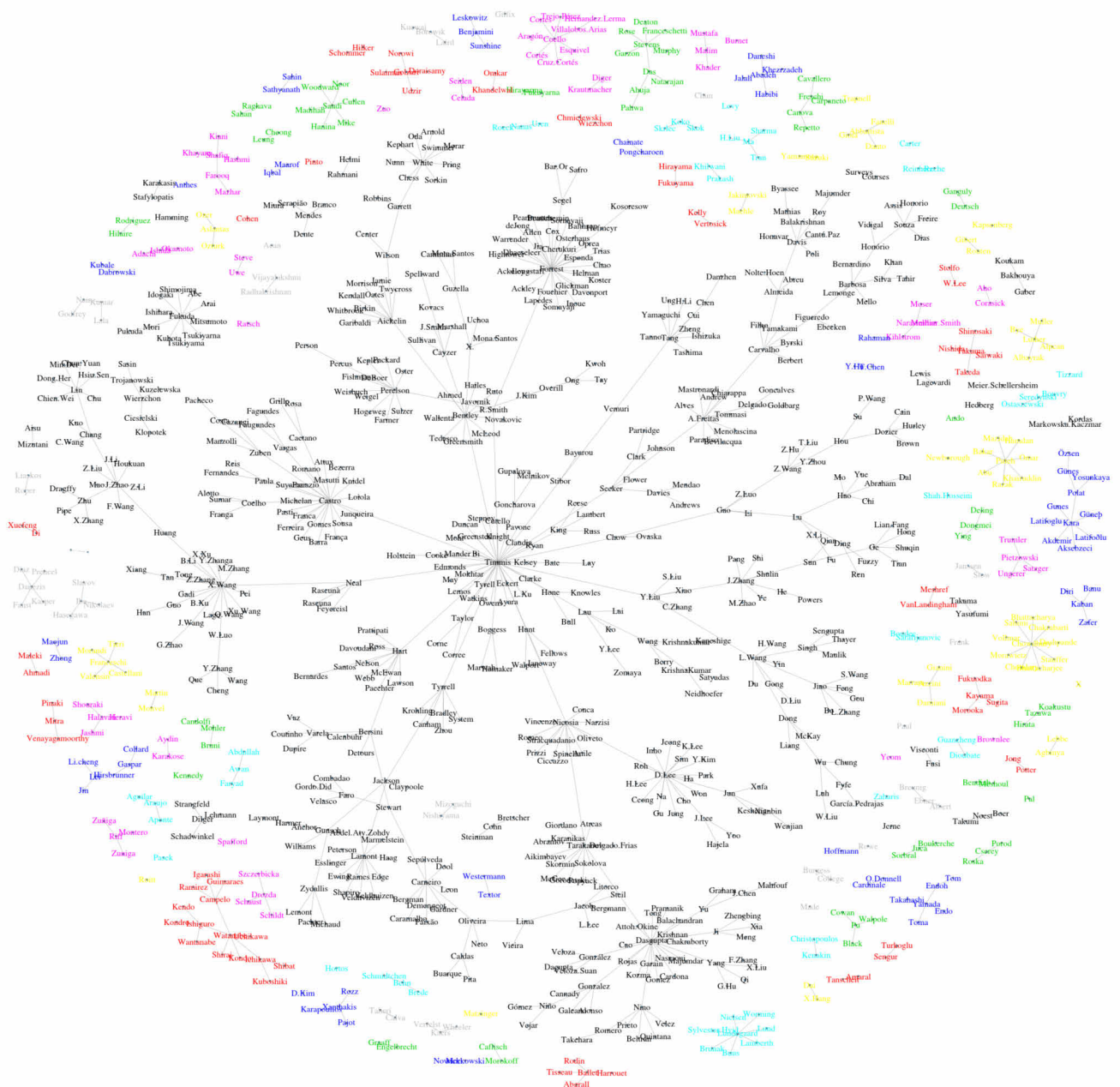


Fig. 3. A visualization of the AIS co-authorship network with nodes representing authors and edges representing collaborations (with at least one co-authored article). Each component is represented in a different color. In particular, the largest component (in black) contains 55% of all authors.

| Rank | Author | Publications | Author | Degree | Author | Betweenness | Author | Closeness |
|------|------------|--------------|-----------|--------|-----------|-------------|------------|-----------|
| 1 | Timmis | 83 | Timmis | 50 | Timmis | 86260 | Timmis | 0.22405 |
| 2 | Dasgupta | 71 | Forrest | 35 | Dasgupta | 35531 | R-Smith | 0.22391 |
| 3 | Forrest | 68 | Castro | 29 | Forrest | 20535 | Bentley | 0.22390 |
| 4 | Castro | 58 | Dasgupta | 28 | Castro | 18859 | Nicosia | 0.22389 |
| 5 | Zuben | 42 | Perelson | 28 | X-Wang | 17737 | Tyrrell | 0.22389 |
| 6 | Aickelin | 38 | X-Wang | 24 | A-Freitas | 17628 | Chow | 0.22386 |
| 7 | Perelson | 35 | D-Lee | 22 | Chow | 16984 | Y-Liu | 0.22386 |
| 8 | Hart | 20 | Zuben | 20 | Hart | 16183 | Clark | 0.22385 |
| 9 | Gonzalez | 19 | J-Kim | 18 | R-Smith | 15714 | Neal | 0.22385 |
| 10 | Stibor | 18 | Lamont | 18 | Tyrrell | 15451 | Lau | 0.22384 |
| 11 | Neal | 17 | Aickelin | 17 | Y-Liu | 14936 | Forrest | 0.22384 |
| 12 | Hunt | 17 | Tarakanov | 16 | Bentley | 14827 | Cutello | 0.22384 |
| 13 | Bersini | 17 | A-Freitas | 15 | Perelson | 13960 | Pavone | 0.22384 |
| 14 | Lamont | 16 | Nicosia | 15 | Stewart | 13312 | Goncharova | 0.22384 |
| 15 | Lau | 16 | Faro | 15 | Aickelin | 12671 | Knight | 0.22384 |
| 16 | Esponda | 15 | Bentley | 13 | Gao | 12462 | Stepney | 0.22384 |
| 17 | J-Kim | 15 | Oliveira | 13 | Carvalho | 11618 | Dasgupta | 0.22383 |
| 18 | Bentley | 15 | Hart | 13 | Jackson | 11594 | Castro | 0.22383 |
| 19 | Greensmith | 15 | Clark | 12 | Huang | 11481 | Hart | 0.22383 |
| 20 | Tarakanov | 15 | Gonzalez | 12 | Bersini | 11119 | A-Freitas | 0.22382 |

Table 2. AIS authors ranked according to their number of publications, degree (or number of unique collaborators), betweenness and closeness.

| Component Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 15 | 550 |
|----------------------|----|----|----|----|---|---|---|---|---|----|----|----|----|------------|
| Number of Components | 39 | 60 | 22 | 15 | 9 | 4 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |

Table 3. AIS Co-authorship Network component sizes and frequencies with the largest component boldened

lists the component sizes and frequencies of the AIS CN showing a significant amount of smaller components. This is suggestive of many existing AIS sub-communities, that collaborate individually but can eventually collaborate with other sub-communities .

Table 2 distinguishes Jon Timmis as a top contributor and collaborator in this field, however, we are also interested in identifying other contributors in the AIS domain like the founding fathers such as Forrest and De Castro. In table 2, we identify authors with significantly high centrality (top 20 in degree and either betweenness or closeness⁴) such as Nicosia, A.Freitas in the AIS CN regardless of their relatively lower number of publications (not in the top 20 in number of publications). Conversely, authors such as Neal and Stibor rank amongst the top 20 for their number of publications but do not rank as highly

⁴ We look at the betweenness and closeness together given the minute variation in the top ranked authors according to closeness centrality shown in table 2

with the centrality measures. We assume this is due to their collaboration with a fewer selection of authors. Nevertheless, Neal and Stibor reappear in the top ranking authors according to AuthorRank in table 4 for their collaborations with few but “prestigious” authors (in terms of AuthorRank) such as Timmis, Hunt and Eckert.

PageRank and AuthorRank are measures of prestige. An author ranks highly if he or she collaborate with many authors that, in turn, have many collaborators – *ad libitum*. AuthorRank gives more weight to repeated collaborations (in a weighted network), whereas PageRank distributes weights evenly (in a boolean network). The top 20 authors according to PageRank and AuthorRank are listed in table 4.

| Rank | PageRank | AuthorRank |
|------|-----------|------------|
| 1 | Timmis | Timmis |
| 2 | Forrest | Castro |
| 3 | Dasgupta | Neal |
| 4 | Perelson | Stibor |
| 5 | Castro | Hunt |
| 6 | X-Wang | Knight |
| 7 | D-Lee | Eckert |
| 8 | Zuben | Zuben |
| 9 | Lamont | A-Freitas |
| 10 | Aickelin | Watkins |
| 11 | J-Kim | Andrews |
| 12 | Coello | Clark |
| 13 | Tarakanov | Hart |
| 14 | Hart | Ayara |
| 15 | Nicosia | Lemos |
| 16 | Chowdhury | Mohr |
| 17 | Fukuda | Secker |
| 18 | Yang | Owens |
| 19 | Gonzalez | Hone |
| 20 | J-Zhang | Kelsey |

Table 4. Authors ranked according to the metrics of PageRank and AuthorRank.

We are equally interested in identifying contributors to the field with leading AIS theoretical models, such as Jerne, Matzinger, Perelson, Bersini, Varela, Carneiro and Greensmith, recurrent committee members of the International Conference of Artificial Immune Systems (ICARIS) such as Timmis, Forrest and Nicosia, as well as principle investigators or team leaders in AIS, such as Hart, Timmis, Von Zuben and D. Dasgupta. Many of these names appear repeatedly in table 2 of top ranking authors according to centrality measures, and in table 4 of top ranking authors according to PageRank and AuthorRank.

Finally, we are interested in understanding large network components that are disconnected from the largest component such as those led by Fukuda, Kara, Watanabe and Coello. We presume that separation is due to language and geographic barriers but we hope to have a more integrated network or more methodological explanations about this segregation in the near future. We are as well interested in understanding cluster properties of the largest component mainly led by team leaders, namely, Forrest, Timmis, Hart, Dasgupta and Tarakanov.

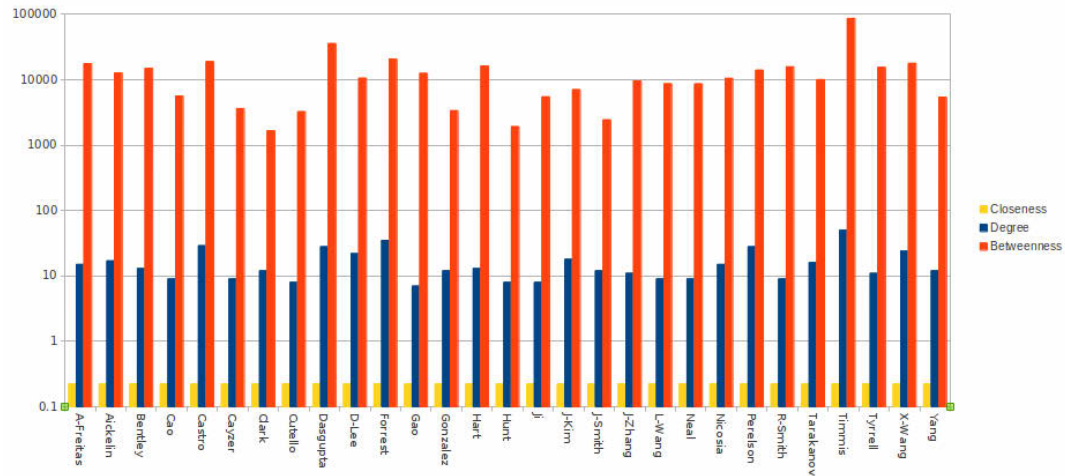


Fig. 4. Alphabetically sorted list of authors with the highest degree, betweenness and closeness

3.3 Conclusion and Future Directions

Several reviews have discussed advances in the field of Artificial Immune Systems but all from a qualitative perspective. Recent reviews have alluded to a stagnation in the field of AIS. In this work, we investigated these questions from a quantitative perspective. Our results have shown that the field has been growing ever since it was established for the past couple of decades. In addition, we have identified leading contributors by co-authorship network analysis based on AIS-relevant publications.

We acknowledge that the bibliography may be biased towards an engineering perspective as it is maintained by a computer scientist thus dismissing fundamental contribution from the theoretical immunology perspective. However, our method can be applied to any bibliography (structured or unstructured) in any scientific domain. Hence, we expect our analytical method not only to motivate the AIS community and encourage external scientists to entertain the challenges presented by AIS, but also to be a benchmark for scientific domain analyses.

Acknowledgments. This research has been supported by the DIM ISC 2011 Grant "Problématiques transversales aux systèmes complexes" Region Paris Ile-de-France through the Institute of Complex Systems (ISC-PIF). We acknowledge the help of Hugues RIPOCHE with the extraction of dates and the insights brought by Luis M. ROCHA and Chris MCEWAN.

References

1. De Castro, L., Timmis, J.: *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer (2002)
2. Timmis, J.: Artificial immune systems today and tomorrow. *Natural Computing* **6**(1) (2007) 1–18
3. Jerne, N.: Towards a network theory of the immune system. *Ann Immunol (Paris)* **125**(1-2) (1974) 373–89
4. Farmer, J., Packard, N., Perelson, A.: The immune system, adaptation, and machine learning. *Physica* **22**(2) (1986) 187–204
5. Varela, F., Coutinho, A.: Second generation immune networks. *Immunol Today* **12**(5) (1991) 159–66
6. Bersini, H., Varela, F.: Hints for adaptive problem solving gleaned from immune networks. *Parallel problem solving from nature* (1991) 343–354
7. Forrest, S., Perelson, A., Allen, L., Cherukuri, R.: Self-nonsel self discrimination in a computer. *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy* **212** (1994)
8. Cooke, D., Hunt, J.: Recognising promoter sequences using an artificial immune system. *Proc Int Conf Intell Syst Mol Biol* **3** (1995) 89–97
9. Hunt, J., Cooke, D.: Learning using an artificial immune system. *Journal of Network and Computer Applications* **19**(2) (1996) 189–212
10. Timmis, J., Neal, M.: A resource limited artificial immune system for data analysis. *Knowledge-Based Systems* **14**(3-4) (2001) 121–130
11. de Castro, L., Timmis, J.: An artificial immune network for multimodal function optimization. *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on* **1** (2002)
12. de Castro, L., Von Zuben, F.: *aiNet: An Artificial Immune Network for Data Analysis. Data Mining: A Heuristic Approach* (2002)
13. Dasgupta, D.: *Artificial Immune Systems and Their Applications*. Springer-Verlag New York, Inc. Secaucus, NJ, USA (1998)
14. Dasgupta, D.: Advances in artificial immune systems. *Computational Intelligence Magazine, IEEE* **1**(4) (2006) 40–49
15. Twycross, J., Aickelin, U.: *Biological Inspiration for Artificial Immune Systems*. Springer Verlag, LNCS **4628** (2007) 300
16. Hart, E., Timmis, J.: Application areas of AIS: The past, the present and the future. *Applied Soft Computing Journal* **8**(1) (2008) 191–201
17. Timmis, J., Hone, A., Stibor, T., Clark, E.: Theoretical advances in artificial immune systems. *Theoretical Computer Science* **403**(1) (2008) 11–32
18. Dasgupta, D., Majumdar, N., Nino, F.: *Artificial immune systems: A bibliography*. Computer Science Division, University of Memphis, Technical Report (2007)
19. Barabasi, A., Crandall, R.: Linked: The new science of networks. *American journal of Physics* **71** (2003) 409

12 Abi-Haidar et al

20. Wassermann, S., Faust, K.: Social network analysis: Methods and applications. New York (1994)
21. Ferrara, A., Salini, S.: Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics* 1–21 10.1007/s11192-012-0810-x.
22. Garfield, E.e.a.: Citation analysis as a tool in journal evaluation. (1972)
23. Liu, X., Bollen, J., Nelson, M., Van de Sompel, H.: Co-authorship networks in the digital library research community. *Information processing & management* **41**(6) (2005) 1462–1480
24. Nascimento, M., Sander, J., Pound, J.: Analysis of sigmod’s co-authorship graph. *ACM SIGMOD Record* **32**(3) (2003) 8–10
25. Newman, M.: Scientific collaboration networks. i. network construction and fundamental results. *Physical review E* **64**(1) (2001) 016131
26. Egghe, L., Guns, R., Rousseau, R.: Measuring co-authors contribution to an articles visibility. *Scientometrics* 1–13 10.1007/s11192-012-0832-4.
27. Badar, K., Hite, J., Badir, Y.: Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in pakistan. *Scientometrics* 1–21 10.1007/s11192-012-0764-z.
28. Team, R., et al.: R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria (01/19) (2010)
29. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. (1999)