Authorship Patterns in Computer Science Research in the Philippines

arXiv:1507.06369v1 [cs.DL] 23 Jul 2015

Jaderick P. Pabico[*]
Institute of Computer Science
University of the Philippines Los Baños
College 4031, Laguna
jppabico@uplb.edu.ph

ABSTRACT

We studied patterns of authorship in computer science (CS) research in the Philippines by using data mining and graph theory techniques on archives of scientific papers presented in the Philippine Computer Science Congresses from 2000 to 2010 involving 326 papers written by 605 authors. We inferred from these archives various graphs namely, a paper–author bipartite graph, a co-authorship graph, and two mixing graphs. Our results show that the scientific articles by Filipino computer scientists were generated at a rate of 33 papers per year, while the papers were written by an average of 2.64 authors (maximum=13). The frequency distribution of the number of authors per paper follows a power-law with a power of $\varphi = -2.04$ ($R^2 = 0.71$). The number of Filipino CS researchers increases at an annual rate of 60 new scientists. The researchers have written an average of 1.42 papers (maximum=20) and have collaborated with 3.70 other computer scientists (maximum=54). The frequency distribution of the number of papers per author follows a power law with $\varphi = -1.88$ ($R^2 = 0.83$). This distribution closely agrees with Lotka's *law of scientific productivity* having $\varphi \approx -2$. The number of co-authors per author also follows a power-law with $\varphi = -1.65$ ($R^2 = 0.80$). These results suggest that most CS papers in the country were written by scientists who prefer to work alone or at most in small groups. These also suggest that few papers were written by scientists who were involved in large collaboration efforts. The productivity of the Philippines' CS researchers, as measured by their number of papers, is positively correlated with their participation in collaborative research efforts, as measured by their number of co-authors (Pearson $r = 0.7425$). The Filipino CS scientists follow a low dissortative mixing when choosing a collaborator either in terms of the collaborator's number of papers ($r = -0.1015$), or its

number of co-authors ($r = -0.0398$). This means that a Filipino CS researcher with high numbers of papers and co-authors chooses a collaborator whose numbers of papers and co-authors are low.

Keywords
Authorship patterns, collaboration graph, computer science research, Philippines

1. INTRODUCTION
The patterns of authorship of scientific research articles reflect how the volume of knowledge was generated by the scientists in a country. The number of quality papers that a nation's researchers write within a time period reflects the scientific productivity of that nation's scientists. The number of authors who wrote a particular research article, on the other hand, mirrors the number of manpower needed to generate the knowledge embodied in the paper. The number of co-authors that a scientist has tells the participation of that scientist in collaborative research efforts, as well as that scientist's dependency with other researchers to generate knowledge. This paper presents the authorship patterns of computer science (CS) research in the Philippines as induced from the archives of scientific papers presented in the Philippine Computer Science Congresses (PCSC) from 2000 to 2010.[1] Although the subject of this paper falls under the CS subdisciplines of graph theory, data structures, information retrieval and mining, visualization, and pattern discovery, the subject matter will be of more interest to the whole computing science community in the Philippines for just one reason: *it is all about the Filipino computer scientists*. We hope that with this paper, we can understand several factors in CS research that are unique in the Philippine setting. For example, we can quantify the bounds of the amount of scientific knowledge that the Filipino computer scientists generated, as well as the bounds of the number of Filipino computing scientists who conducted research in the past years. We can also identify who are the most prolific computer scientists, as well as those who are with the most number of research collaborators. In general, understanding the patterns on how the Filipino computer scientists generate knowledge may provide discernment on information breakdowns, bottlenecks, and

---

[*]http://www.ics.uplb.edu.ph/jppabico

[1]The PCSC started in 2000 but the 2001 papers are not accessible to the author. There was no PCSC conducted in 2002 [1].

structural holes in the scientific community of CS in the Philippines.

In recent years, the advent of advanced computer-based archiving technologies made most scientific works in the last 10 to 50 years easily accessible via any digital media by virtually anyone from anywhere. Examples of such archives are the Los Alamos e-Print Archive (LAePA) [15], the Medline Database (Medline) [37], the Standford Public Information Retrieval System (SPIRES) [35], the Network of Computer Science Technical Reference (NCSTRL) [17], the DBLP Computer Science Bibliography (DBLP) [14], the Samahang Pisika ng Pilipinas (SPP) [34], the Transactions of the National Academy of Science and Technology-Annual Scientific Meetings (NAST-ASM), and the Proceedings of the Philippine Society of Agricultural and Biosystems Engineering (BAE) [25–27] (Table 1). These archives compile scientific papers that were recently used by some researchers [7, 8, 18, 19, 22–24, 38] who conducted data mining techniques to understand the complex nature in scientific research in various fields. Inferred from these archives are results that show that the average papers per author ranges from 2 to 7, while the papers were written by an average of 2 to 9 authors. Depending on the scientific discipline, a given author has an average of 3 to 173 collaborators [7, 18, 22–24].

Table 1: List of example scientific paper archives as used by various researchers [7, 8, 18, 19, 22–24, 38]. The number of papers and of authors are the numbers when the studies were conducted by the respective authors.

Archive Name	Year Started	Number of Papers	Number of Authors
International Archives			
LAePA	1992	$> 161,000$	$> 94,000$
Medline	1961	$> 216,000$	$> 152,000$
SPIRES	1990	$\approx 66,000$	$> 56,000$
NCSTRL	1974	$\approx 13,000$	$\approx 12,000$
DBLP	1960	$\approx 84,000$	$\approx 95,000$
Philippine-based Archives			
SPP	2001	699	840
Agriculture	2006	235	645
NAST-ASM	2006	720	1,780
ABE	2007	90	171

In the Philippines, we have previously utilized the archives of scientific posters presented at the recent NAST-ASM in an initial attempt to understand the authorship patterns of Filipino agricultural scientists [22]. Although Philippine-based scientific journals and proceedings in agricultural science abound, we assumed that the papers compiled in the NAST-ASM archives represent the majority of scientific knowledge discovered by Filipino agricultural scientists, not only because of the sheer volume of knowledge it contains, but also because of the quality of knowledge presented having been reviewed, and often times authored, by no less than the nation's Academicians and National Scientists. The papers used in this study were categorized under the Agricultural Sciences Division (ASD) and

involved 235 poster abstracts written by 645 authors spanning the recent four years from 2006 to 2009. In this study, we have found that the Filipino agricultural scientists have written an average of 1.39 papers (maximum=13), while they have collaborated with an average of 2.70 scientists (maximum=28). Their research papers were written by an average of 3.81 Filipino authors (maximum=15).

Using the same NAST-ASM archives, we recently expanded [24] the above study to involve all six NAST scientific divisions encompassing various scientific disciplines: The ASD; The biological sciences (BSD); The chemical, mathematical and physical sciences (CMPSD); The engineering sciences and technology (ESTD); The health sciences (HSD); And the social sciences (SSD). This expanded study involved 720 papers written by 1,780 authors. Because of the sheer volume of scientific discoveries contained in the archive, we assumed that the papers represent the major scientific work of Filipino scientists in various disciplines in the past four years from 2006 to 2009. Again, our previous assumption holds that the archives not only contain high quantity of scientific discoveries in the Philippines, but more importantly, high quality research results for the same reason as mentioned above. The results of our analysis show that the Filipino scientists have written an average of 1.52 papers (maximum=40), while they have collaborated with an average of 2.82 scientists (maximum=66). The scientific papers have been written by an average of 3.70 Filipino authors (maximum=22).

Using the NAST-ASM archives to infer the authorship patterns of scientists from specific disciplines proved to be difficult to do, even though works of scientists in a specific field might already be included in the archives. Examples of such disciplines are the Physics, the Agricultural and Biosystems Engineering (ABE), and the CS disciplines. The reason for this is that the NAST-ASM archives did not label both the scientists and the research works as belonging to either the Physics, the ABE or the CS discipline. In fact, Physics papers are classified only under CMPSD, while ABE papers may be classified within two out of six NAST divisions namely, ASD and ESTD. Both ASD and ESTD involved papers from various other fields that are not ABE in nature, such as entomolgy, biochemistry, forestry, information technology and all other engineering fields. Researches from the CS discipline, on the other hand, maybe classified under CMPSD and ESTD, which also involved various other fields that are not CS in nature. Thus, to analyze the authorship patterns of Filipino scientists and researchers in specific disciplines, separate archives must be used to better reflect the works and workers in the said discipline. In the case of Physics and ABE, their respective archives actually exist as the Proceedings of the Samahang Pisika ng Pilipinas [34], and the Proceedings of the Joint International Agricultural Engineering Conference and Exhibition of the Philippine Society of Agricultural Engineers (PSAE) [25–27]. Both proceedings are archived in digital format. We have analyzed the authorship patterns of Filipino physicists from 2001 to 2005 involving 699 papers written by 840 authors [38],

as well as those of the ABE scientists over the recent 3-year period from 2007 to 2009 involving 90 papers written by 171 authors [23]. Our results in these studies are summarized in Table 2, together with the summary of the previous works discussed above for comparison purposes.

In this current effort, to understand the authorship patterns of Filipino CS researchers, we have applied data mining and graph theory techniques on archives of papers presented in the PCSC [2–4, 6, 28–32] from 2000 to 2010. The 9–year archive has accumulated 326 papers written by 605 authors. We have found out, on the average, that the CS research papers were authored by 2.64 Filipino scientists, while the CS researchers themselves have written 1.42 papers and have collaborated with 3.70 other scientists. Aside from computing these fundamental quantities to compare the CS community with other scientific disciplines in the country, we also computed the respective frequency distributions of these quantities. The power law nature of these distributions suggest that most CS papers were authored by those who have a few collaborators, while few of the papers were authored by those who have a large list of collaborators. We have found a statistical evidence suggesting that the productivity of computer scientists in the country is positively correlated with the scientists' participation in a number of collaborative research endeavors. We have also observed low assortative mixing among authors when choosing a collaborators in terms of the collaborator's scientific productivity, as well as the collaborator's number of collaborators. We hope that the results contributed by this paper could later be used to aid the various stakeholders (e.g., funding agencies and professional organizations) in providing opportunities to accelerate knowledge generation in the field of CS in the country, as well as in strengthening the efficiency and effectiveness of existing formal research and technical communication channels.

2. MATERIALS AND METHODS

2.1 Archive of Scientific Papers

We have utilized the author information from 326 peer-reviewed papers presented during the 2000 to 2010 PCSC [2–4, 28–32]. The papers presented each year are archived electronically in CDROM format, which is distributed to PCSC participants and paper presentors during the conference. The CDROM contains papers that are usually in portable document format (PDF) and comes with a table of contents that is also in PDF. An easily parseable hypertext markup language (HTML) format of the archive is also accessible from the website of the Computing Society of the Philippines [9].

Table 3 summarizes the particulars of various PCSC such as their respective proceedings, the number of papers presented, and the number of authors who wrote the papers during each year. The number of papers and authors during the 2000 PCSC were closed to the annual average, respectively. Both counts increase steadily in the earlier 4–year span from 2003 to 2006. When the PCSC was held in Boracay in 2007, both the

number of authors and papers dropped considerably. However, both counts gain momentum and increase considerably in the recent 4–year span from 2007 to 2010. The 2010 PCSC has received a record number of paper submissions, and thus reflects the record-breaking number of papers accepted and presented, as well as the number of authors who wrote the papers. The 9–year PCSC has attracted an annual average of 36 papers and 83 authors.

In this study, we considered a scientific paper as either a keynote paper, a plenary (invited) paper, a tutorial paper, or a contributed paper. These paper types are present in all PCSC with the exception of the first year and the latest two years. In 2000 PCSC, a poster paper session (POSTERS) was included and the 2000 archive includes these paper type. In PCSC 2009, the research-in-progress session (RIPS) was instituted. RIPS allows the oral presentation of papers that are usually authored by undergraduate students and are categorized by the paper review panelists as incomplete or *in progress* but are already worthy of oral presentation. The PCSC 2009 archive, however, did not label whether the paper was RIPS or not. Thus, we assumed here that the 2009 PCSC archive does not include the RIPS. In PCSC 2010, POSTERS was reinstituted. Both RIPS and POSTERS papers are included in the 2010 PCSC archive. However, we did not include these papers in our study because as of this writing, the author information is incomplete for papers with more than one author.

In our analysis of the co-authorship patterns, we considered an archive $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ of N scientific papers, with each paper P_i having a list $\mathcal{A}_i = \{A_j | A_j \in \mathcal{A}, |\mathcal{A}_i| = M_i\}$ of M_i authors. From the author information in $\mathcal{P}$, we created a database of distinct authors $\mathcal{A} = \bigcup_{i=1}^N \mathcal{A}_i$, such that $M = |\mathcal{A}| \leq \sum_{i=1}^N M_i$. We note here that $M = \sum_{i=1}^N M_i$ implies $\bigcap_{i=1}^N \mathcal{A}_i = \emptyset$, which means that all authors have written exactly one paper. Our results show that this is not the case in Philippine CS research.

2.2 Building the Paper-Author Bipartite Graph

Given $\mathcal{P}$ and $\mathcal{A}$, we built the paper–author bipartite graph **PAG** $= (\mathcal{P} \bigcup \mathcal{A}, \mathcal{E})$, where $\mathcal{E} = \{(i, j) | P_i \in \mathcal{P}, A_j \in \mathcal{A}\}$. For each paper P_i, we created a bipartite subgraph (sub-bigraph) PAG_i composed of a type–P vertex labeled P_i, and M_i type–A vertices with the respective labels as in $\mathcal{A}_i$. We then created edges in PAG_i by connecting the type–P vertex with all the M_i type–A vertices. The ith sub-bigraph induced by P_i represents the one-to-many relationship between the ith paper and its M_i authors. We then connected all N sub-bigraphs via each sub-bigraph's common type–A vertices. The resulting graph $\bigcup_{i=1}^N \text{PAG}_i$ is the paper–author bipartite graph **PAG**. Intuitively, **PAG** may be built with a time complexity of $O(N \times M)$ but we reduced this to $O(N \times \log M)$ by using a balanced binary tree structure for $\mathcal{A}$.

Figure 1(a–c) shows how the **PAG** was created for a hypothetical paper archive $\mathcal{P}$ composed of two papers P_1

Table 2: Fundamental statistics of various different scientific collaboration networks: Average number of authors per paper ($A_{\mathcal{P}\text{AVG}}$), average number of papers per author ($P_{\mathcal{A}\text{AVG}}$), and average number of co-authors per author ($C_{\mathcal{A}\text{AVG}}$).

Scientific Discipline	No. of Years	Fundamental Statistics			Literature Reference
		$A_{\mathcal{P}\text{AVG}}$	$P_{\mathcal{A}\text{AVG}}$	$C_{\mathcal{A}\text{AVG}}$	
International Research					
Biomedical Research	40	3.75	6.40	18.10	Newman [18]
High-energy Physics	27	8.96	11.60	173.00	Newman [18]
CS	10	2.22	2.55	3.59	Newman [18]
Philippine-based Research					
Physics	5	3.16	-	10.80	Villanueva and Pabico [38]
Agriculture	4	3.81	1.39	2.70	Pabico [22]
Various Fields	4	3.70	1.52	2.82	Pabico and Micor [24]
ABE	3	3.02	1.59	2.35	Pabico [23]
CS	**9**	**2.68**	**1.42**	**3.58**	

Table 3: Basic information about the 2000 to 2010 Philippine Computing Science Congress: Year and site each held, number of papers presented, number of authors, and proceedings reference.

Year	PCSC Site	Number of Papers	Number of Authors	Proceedings Reference	Remarks
2000	De La Salle University	35	78	Azcarraga [6]	POSTERS
2001	MSU-IIT	-	-	-	Data not available
2002	-	-	-	-	Not held
2003	Philippine Science HS	15	31	Saldaña and Caro [31]	
2004	UP Los Baños	29	61	Albacea et al. [4]	
2005	University of Cebu	33	80	Saldaña and Chua [32]	
2006	Ateneo de Manila	38	101	Saldaña [28]	
2007	Boracay Island	33	74	Saldaña [29]	
2008	UP Diliman	37	76	Saldaña [30]	
2009	Silliman University	41	93	Adorna and Saldaña [3]	RIPS
2010	Ateneo de Davao	61	148	Adorna [2]	RIPS and POSTERS
	Average	36	83		

and P_2 written by authors A_1, A_2, A_3, and A_4. In this scenario, P_1 was co-authored by A_1 and A_2, while P_2 was jointly written by A_2, A_3, and A_4. In both papers, A_2 was the common author. Separately, the sub-bigraph induced by P_1 is $\text{PAG}_1 = (\{P_1, A_1, A_2\}, \{(1,1), (1,2)\})$, while the sub-bigraph induced by P_2 is $\text{PAG}_2 = (\{P_2, A_2, A_3, A_4\}, \{(2,2), (2,3), (2,4)\})$. The sub-bigraphs PAG_1 and PAG_2 are connected through the common vertex A_2 to create $\textbf{PAG} = (\mathcal{P}\bigcup\mathcal{A}, \mathcal{E})$, where $\mathcal{P} = \{P_1, P_2\}$, $\mathcal{A} = \{A_1, A_2, A_3, A_4\}$, and $\mathcal{E} = \{(1,1), (1,2), (2,2), (2,3), (2,4)\}$.

2.3 Building the Co-authorship Graph

We built the co-authorship graph $\textbf{CAG}$ from $\textbf{PAG}$ as follows. For each vertex $P_i \in \textbf{PAG}$, we deleted all incedent edges to (or from) P_i, as well as P_i itself, and created in its instead a complete subgraph $\text{CAG}_i = (\mathcal{A}_i, \mathcal{E}_i)$, where $\mathcal{E}_i = \{(j,k)|A_j, A_k \in \mathcal{A}_i, j \neq k\}$ and $|\mathcal{E}_i| = M_i(M_i - 1)/2$ connecting all pairwise combinations of $A_j, A_k \in \mathcal{A}_i$, $j \neq k$. The fully-connected subgraph CAG_i represents the co-authorship graph of authors who co-wrote the ith paper P_i. The resulting graph $\textbf{CAG} = \bigcup_{i=1}^{N} \text{CAG}_i$ is the co-authorship graph of CS researchers in the Philippines. Because some authors have not collaborated, some vertices $A_i \in \textbf{CAG}$ are not connected to any of the other vertices $A_j \in \textbf{CAG}$.

Figure 1(e–g) shows the flow diagram of the procedure on how $\textbf{CAG}$ was created from the hypothetical example mentioned above. The co-authorship subgraph induced by P_1 is $\text{CAG}_1 = (\{A_1, A_2\}, \{(1,2)\})$, while the co-authorship subgraph induced by P_2 is $\text{CAG}_2 = (\{A_2, A_3, A_4\}, \{(2,3), (2,4), (3,4)\})$. The subgraphs CAG_1 and CAG_2 are connected through the common vertex A_2 to create the co-authorship graph $\textbf{CAG}(\mathcal{A}, \mathcal{E}_c)$, where $\mathcal{A} = \{A_1, A_2, A_3, A_4\}$ and $\mathcal{E}_c = \{(1,2), (2,3), (2,4), (3,4)\}$.

In building $\textbf{CAG}$, we adopted the same assumptions made by Newman [18]: (1) That all pairs of authors A_i and A_j, $\forall i \neq j$, who have written a paper together are genuinely acquainted with one another; and (2) That the co-authorship graph $\textbf{CAG}$ reflects a genuine professional interaction between Filipino computer scientists.

2.4 Computing for node degrees

From $\textbf{PAG}$, we can infer an $N \times M$ matrix $\textbf{PAM}$ that mathematically represents the adjacency of $\textbf{PAG}$. Each matrix element $\textbf{PAM}_{i,j} = 1$ if the ith paper is written or co-written by the jth author. Otherwise, $\textbf{PAM}_{i,j} = 0$. The $\textbf{PAM}$ of the hypothetical $\textbf{PAG}$ discussed above is shown in Figure 1(d). It is interesting to note that $\textbf{PAM}_{i,j} \ngtr 1$ as no distinct author name appears more than once in the author line of a paper.
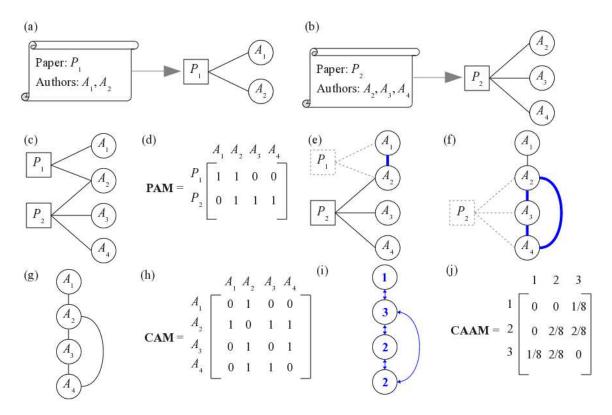
Figure 1: The process flow for building the paper–author bipartite graph PAG and the paper–author matrix PAM using an archive with two hypothetical papers P_1 and P_2: (a) The hypothetical paper P_1 and its corresponding sub-bigraph; (b) The hypothetical paper P_2 and the sub-bigraph induced by it; (c) The resulting $\mathbf{PAG} = \mathrm{PAG}_1 \bigcup \mathrm{PAG}_2$; and (d) The equivalent PAM. The process flow for building the co-authorship graph CAG and the co-authorship matrix CAM: (e) Deleting vertex P_1 and edges $(1,1)$ and $(1,2)$, and creating the completely connected subgraph CAG_1; (f) Deleting vertex P_2 and edges $(2,2)$, $(2,3)$, and $(2,4)$, and creating the fully-connected subgraph CAG_2; (g) The resulting $\mathbf{CAG} = \mathrm{CAG}_1 \bigcup \mathrm{CAG}_2$; and (h) The equivalent CAM. The process flow for transforming CAG into CAAG and the corresponding mixing matrix CAAM: (i) The mixing network when $\tau = \Delta$; and (j) The resulting CAAM. In the visualization of the different graphs, square vertices represent papers while circle vertices represent authors.

The degree $\Delta_i^{\mathcal{P}}$ of the ith P–type vertex P_i represents the number of authors that wrote paper P_i. Conversely, the degree $\Delta_j^{\mathcal{A}}$ of the jth A–type vertex A_j represents the number of papers that author A_j wrote. The vertex degrees $\Delta_i^{\mathcal{P}}$ and $\Delta_i^{\mathcal{A}}$ can be computed using **PAM** as shown in Equations 1 and 2, respectively. We can use the vertex degrees to compute for the minimum ($A_{\mathcal{P}\mathrm{MIN}}$), average ($A_{\mathcal{P}\mathrm{AVG}}$) and maximum ($A_{\mathcal{P}\mathrm{MAX}}$) number of authors per paper (Equations 4 to 6), as well as the minimum ($P_{\mathcal{A}\mathrm{MIN}}$), average ($P_{\mathcal{A}\mathrm{AVG}}$) and maximum ($P_{\mathcal{A}\mathrm{MAX}}$) number of papers per author (Equations 7 to 9). The degree $\Delta_i^{\mathcal{P}}$ of the ith P–type vertex P_i represents the number of authors that wrote paper P_i. Conversely, the degree $\Delta_j^{\mathcal{A}}$ of the jth A–type vertex A_j represents the number of papers that author A_j wrote. The vertex degrees $\Delta_i^{\mathcal{P}}$ and $\Delta_i^{\mathcal{A}}$ can be computed using the matrix **PAM** as shown in Equations 1 and 2, respectively. We can use the vertex degrees to compute for the minimum ($A_{\mathcal{P}\mathrm{MIN}}$), average ($A_{\mathcal{P}\mathrm{AVG}}$) and maximum ($A_{\mathcal{P}\mathrm{MAX}}$) number of authors per paper (Equations 4 to 6), as well as the minimum ($P_{\mathcal{A}\mathrm{MIN}}$),

average ($P_{\mathcal{A}\mathrm{AVG}}$) and maximum ($P_{\mathcal{A}\mathrm{MAX}}$) number of papers per author (Equations 7 to 9).

In the hypothetical archive discussed above, $\Delta_1^{\mathcal{P}} = 2$ while $\Delta_2^{\mathcal{P}} = 3$. Conversely, $\Delta_1^{\mathcal{A}} = 1$, $\Delta_2^{\mathcal{A}} = 2$, $\Delta_3^{\mathcal{A}} = 1$, and $\Delta_4^{\mathcal{A}} = 1$. $A_{\mathcal{P}\mathrm{MIN}} = 2$, $A_{\mathcal{P}\mathrm{AVG}} = 2.5$, and $A_{\mathcal{P}\mathrm{MAX}} = 3$. Similarly, $P_{\mathcal{A}\mathrm{MIN}} = 1$, $P_{\mathcal{A}\mathrm{AVG}} = 1.25$, and $P_{\mathcal{A}\mathrm{MAX}} = 2$.

From **CAG**, we can infer an $M \times M$ diagonally symmetric co-authorship matrix **CAM** that mathematically represents ties between the M scientists. Each matrix element $\mathbf{CAM}_{j,k} = \mathbf{CAM}_{k,j} = 1$ if and only if author A_j has collaborated with author A_k with at least one paper. Since collaboration is a symmetric relation, $\mathbf{CAM}_{j,k} = 1$ implies $\mathbf{CAM}_{k,j} = 1$, which means that author A_k collaborates with author A_j in response. Without losing generality, we set all diagonal elements $\mathbf{CAM}_{j,j} = 0$. If A_j has not collaborated with A_k, then $\mathbf{CAM}_{j,k} = \mathbf{CAM}_{k,j} = 0$. Figure 1(h) shows the computed **CAM** of the hypothetical **CAG**. Using **CAM**, the vertex degree Δ_i^C of the ith author,

which reflects the number of co-authors A_i has, is computed as shown in Equation 3, while the minimum $C_{\mathcal{A}\text{MIN}}$, average $C_{\mathcal{A}\text{AVG}}$, and maximum $C_{\mathcal{A}\text{MAX}}$ number of co-authors are respectively computed as in Equations 10 to 12.

$$\Delta_i^{\mathcal{P}} = \sum_{j=1}^{M} \mathbf{PAM}_{i,j} \tag{1}$$

$$\Delta_j^{\mathcal{A}} = \sum_{i=1}^{N} \mathbf{PAM}_{i,j} \tag{2}$$

$$\Delta_i^{C} = \sum_{j=1}^{M} \mathbf{CAM}_{j,i} \tag{3}$$

$$A_{\mathcal{P}\text{MIN}} = \min_{i=1}^{N} \Delta_i^{\mathcal{P}} \tag{4}$$

$$A_{\mathcal{P}\text{AVG}} = \frac{\sum_{i=1}^{N} \Delta_i^{\mathcal{P}}}{N} \tag{5}$$

$$A_{\mathcal{P}\text{MAX}} = \max_{i=1}^{N} \Delta_i^{\mathcal{P}} \tag{6}$$

$$P_{\mathcal{A}\text{MIN}} = \min_{j=1}^{M} \Delta_j^{\mathcal{A}} \tag{7}$$

$$P_{\mathcal{A}\text{AVG}} = \frac{\sum_{j=1}^{M} \Delta_j^{\mathcal{A}}}{M} \tag{8}$$

$$P_{\mathcal{A}\text{MAX}} = \max_{j=1}^{M} \Delta_j^{\mathcal{A}} \tag{9}$$

$$C_{\mathcal{A}\text{MIN}} = \min_{i=1}^{M} \Delta_i^{C} \tag{10}$$

$$C_{\mathcal{A}\text{AVG}} = \frac{\sum_{i=1}^{M} \Delta_i^{C}}{M} \tag{11}$$

$$C_{\mathcal{A}\text{MAX}} = \max_{i=1}^{M} \Delta_i^{C} \tag{12}$$

2.5 Degree Distributions in PAG and CAG

The frequency distribution $\rho(\Delta)$ of a vertex degree Δ is a graph–based quantity that has been much studied and applied recently for various co-authorship graphs [8, 18, 19] and social networks [5, 21]. It provides the frequency that a randomly selected vertex has Δ edges (or degrees). Graphs with high-degree yet low cardinality vertices have long-tailed $\rho(\Delta)$ and are called *scale-free* graphs. Such graphs follow the power law distribution (Equation 13) and oftentimes model the relationships of naturally occuring entities, such as that of proteins and their interactions [33]. We hypothesized that **PAG** and **CAG** are scale-free and thus their respective ρ follow a power-law. To test this hypothesis, we fitted a power law line each on $\rho(\Delta^{\mathcal{P}})$, $\rho(\Delta^{\mathcal{A}})$, and $\rho(\Delta^{C})$ and statistically tested the power to be significantly different from zero at $\alpha = 0.05$ (where α is taken as the probability of the two–tailed alternative greater than the test statistics). The power law distribution is statistically estimated by the frequency y in Equation 14 and involves the vertex degree Δ, a constant c, and the power φ, which is also known as the fractal dimension [13]. We estimated the values of c and φ by using a linear regression analysis in the power law's linear form (Equation 14).

$$y = c\Delta^{\varphi} \tag{13}$$

$$\log y = \log c + \varphi \log \Delta \tag{14}$$

2.6 Productivity and collaboration

An author $A_i \in \mathcal{A}$ has an inherent vector of valued attributes (τ_1, τ_2), wherein in this research we set $\tau_1 = P_{\mathcal{A}}$ and $\tau_2 = C_{\mathcal{A}}$. we hypothesized that $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$ have a high positive correlation such that authors who are productive, as measured by their high $P_{\mathcal{A}}$, are also those who have high number of memberships in various collaboration efforts, as measured by their high $C_{\mathcal{A}}$. High positive correlation would also mean that authors who are less productive (i.e., low $P_{\mathcal{A}}$) are those who write alone or their number of collaborators is relatively small (i.e., low $C_{\mathcal{A}}$). We tested the hypothesis by estimating the Pearson correlation r and statistically testing it against zero (i.e., we hypothesize that $r \neq 0$). We utilized the Pearson statistics because the causality relation between $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$ was not established (i.e., we do not know whether $P_{\mathcal{A}}$ causes $C_{\mathcal{A}}$, or *vice versa*, or whether such relation exists at all).

2.7 Assortativity in CAG

Given an attribute τ of a vertex, the assortativity r of a graph is the tendency of vertices to be connected to like vertices [20], such that there are more edges between vertices with high τ values than between a high–τ vertex and a low–τ vertex. We started its computation by relabeling each vertex $A_i \in \mathbf{CAG}(\mathcal{A}, \mathcal{E}_c)$ by its τ, and converting all undirected edges in $\mathcal{E}$ to bidirectional edges to create $\mathcal{E}_d$. The resulting graph $\mathbf{CAAG}(\mathcal{A}', \mathcal{E}_d)$, where $\mathcal{A}'$ is just the relabeled vertices in $\mathcal{A}$, and $|\mathcal{E}_d| = 2 \times |\mathcal{E}_c|$. We used a *mixing matrix* $\mathbf{CAAM}$, where each matrix element $\mathbf{CAAM}_{i,j}$ represents the fraction of all edges in $\mathbf{CAAG}$ that start at a_i and end at a_j, such that $\sum_{i,j} \mathbf{CAAM}_{i,j} = 1$. Let f_i be the fraction of all edges in $\mathbf{CAAG}$ that are incident to a_i, thus $f_i = \sum_j \mathbf{CAAM}_{i,j}$. The assortativity r can be approximated by the Pearson correlation coefficient discussed by Newman [20] and subsequently used by Bird et al. [8]. *Assortativity* is when all vertices in $\mathbf{CAG}$ are connected only to vertices with similar τ (i.e., $r > 0$). *Dissortativity* (or negative assortativity $r < 0$) is when high–τ vertices are only connected to low–τ ones. Using the degree Δ as τ, Figure 1(i–j) shows how the $\mathbf{CAG}$ of the hypothetical example discussed above was transformed into $\mathbf{CAAG}$, as well as how the $\mathbf{CAAM}_{i,j}$ was computed. In this paper, we independently used $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$ as τ to separately discover the general preference of CS researchers in choosing a collaborator in terms of the collaborator's $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$, respectively.

3. RESULTS AND DISCUSSION
3.1 The PCSC Paper Archive

For this study, we utilized the archive $\mathcal{P}$ of papers presented during the 2000 to 2010 PCSC to infer the authorship patterns of Filipino computer scientists. The total number of papers presented in these conventions is $N = 326$, while the number of authors is $M = 605$. As pointed out by Newman [18], one particular issue that we were concerned about was the

number of names L that appear in $\mathcal{P}$, which clearly identifies distinct authors. This is because L is not necessarily the same as M. For example, author A_i may format his name differently on different papers, such that the names *Juan dela Cruz, Dela Cruz, Juan,* and *J. dela Cruz* could all belong to him. This scenario would mean that $M = 3$, but in fact $L = 1$. On the contrary, two distinct authors A_i and A_j may have the same name, such that the name *Maria Maquiling* could belong to both. This means that $M = 1$, while in fact $L = 2$. This apparent name ambiguity problem has already been given approximate solutions by various techniques [10, 11, 36, 39] that use additional information found in the papers, such as the names of the authors' respective home institutions and their subdisciplines. However, we could not use these additional information because there are authors who belong to more than one institution, and due to multi-specialty research collaborations, they could publish in other subdisciplines. Further, the author information in $\mathcal{P}$ rarely includes the subdisciplines. In order to solve these issues, we performed our analysis using the author's surname and first and second names' initials, knowing full well that we may be overestimating the true value of M. In this regard, having $L \geq M$ in this research may give us a guarantee that our results provide the respective upper bounds of the patterns.

Figure 2 shows the annual trend of cumulative number of authors and papers presented in the 9–year PCSC. Based on simple regression analysis, we found out that PCSC has attracted about 60 new authors per year who helped co-write about 33 new papers annually. After extrapolating these lines to 5 years into the future, we can see that in 2015 the number of distinct authors that will be contributing to PCSC will reach to 843 while the number of papers that will be contributed will reach to 458.

3.2 Inferences from PAG

Table 4 summarizes the values inferred from **PAG**. On the average, the CS authors in the Philippines have writen about 1.42 papers, while papers were written by an average of 2.64 authors. The Filipino authors have collaborated, on the average, with 3.70 other authors. We have shown the comparison of these simple statistics with other various national and international research co-authorship graphs (Table 2). As inferred also from **PAG**, we have identified the top five researchers with the most number of papers in the archive: PC Naval (20 papers), RP Salaña (19), HN Adorna (16), RC Sison (15), and REO Roxas and JDL Caro (10 each). We have annotated the vertices in Figures 3 and 4 to visualize the respective relative positions of these authors in **PAG** and **CAG**.

3.3 Number of authors per paper

The Filipino CS research papers have been written on the average by 2.64 authors, which is lower compared to that of the ABE ($A_{\mathcal{P}\text{AVG}} = 3.02$), the agricultural science ($A_{\mathcal{P}\text{AVG}} = 3.81$) and NAST sciences ($A_{\mathcal{P}\text{AVG}} = 3.70$) in the country. This means that in the Philippines, creating new scientific information requires less
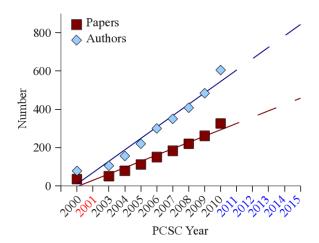

Figure 2: Annual trend of the cumulative number of papers (red square) and authors (blue diamond). The red solid and dashed lines, respectively, represent the regression and the 5-year extrapolation line of the yearly cumulative number of papers (slope= 32.99, $R^2 = 0.95$). The blue solid and dashed lines, respectively, represent the regression and the 5-year extrapolation line of the yearly cumulative number of authors (slope= 59.54, $R^2 = 0.96$). (This figure is in color in the digital format of this paper.)

number of authors in CS than in other disciplines. In the international co-authorship graphs, more authors are needed to write new information in the field of biomedical research ($A_{\mathcal{P}\text{AVG}} = 3.75$), and significantly more authors in the high-energy physics ($A_{\mathcal{P}\text{AVG}} = 8.96$). However, the Filipino CS research papers needed more authors on the average compared to that in the international CS's ($A_{\mathcal{P}\text{AVG}} = 2.22$).

3.4 Number of papers per author

On the average, the Filipino CS researchers have written less papers ($P_{\mathcal{A}\text{AVG}} = 1.42$) than their ABE ($P_{\mathcal{A}\text{AVG}} = 1.59$) and NAST ($P_{\mathcal{A}\text{AVG}} = 1.52$) counterparts, but more than the agricultural ($P_{\mathcal{A}\text{AVG}} = 1.39$) scientists in the country. However, the average scientific productivity of Filipino computer scientists, measured by the number papers written per author, still falls behind the international averages. The international biomedical researchers, high-energy physicists, and computer scientists have respectively written an average of 6.4, 11.6, and 2.55 papers.

3.5 Inferences from CAG

Figure 4 presents a visualization of the co-authorship graph **CAG** created from the papers in $\mathcal{P}$. In this visualization, it can easily be seen that the graph of CS research co-authorship in the Philippines is composed of disconnected subgraphs. We have found out that authors in each of the subgraphs belong to the same institution. This means that CS authors collaborate only to authors who belong to the same institution, and that cross–institution collaborations do not exist yet in the Philippines setting. It is understandable,
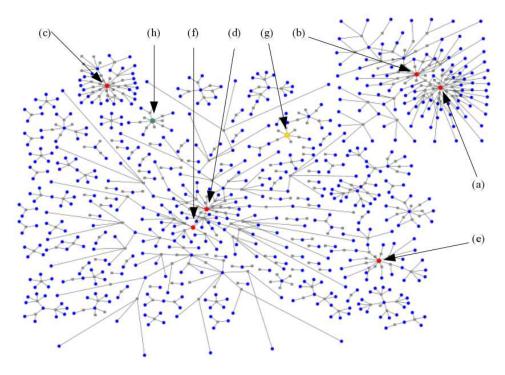
Figure 3: The paper–author bigraph PAG drawn with the graph visualization algorithm by Kamada and Kawai [12]. In this visualization, colored circles represent authors while gray squares represent papers. The labels correspond to some identified authors with the most number of papers: (a) PC Naval, (b) HN Adorna, (c) RP Saldaña, (d) R Sison, (e) REO Roxas, and (f) D Cheng. In addition, (g) is this journal's editor-in-chief EA Albacea, while (h) is this paper's author. (This figure is in color in the digital format of this paper.)

Table 4: Values of inferred statistics from PAG and CAG: Minimum, average and maximum number of authors per paper ($A_\mathcal{P}$), number of papers per author ($P_\mathcal{A}$), and number of collaborators per author ($C_\mathcal{A}$); As well as the respective degree distribution's power law coefficients (φ) and the corresponding statistics (R^2).

Statistics	Mininum	Average	Maximum	Degree Distribution	
				φ	R^2
Number of authors per paper	$A_{\mathcal{P}\text{MIN}} = 1$	$A_{\mathcal{P}\text{AVG}} = 2.64$	$A_{\mathcal{P}\text{MAX}} = 13$	-2.04	0.71
Number of papers per author	$P_{\mathcal{A}\text{MIN}} = 1$	$P_{\mathcal{A}\text{AVG}} = 1.42$	$P_{\mathcal{A}\text{MAX}} = 20$	-1.88	0.83
Number of collaborators per author	$C_{\mathcal{A}\text{MIN}} = 0$	$C_{\mathcal{A}\text{AVG}} = 3.70$	$C_{\mathcal{A}\text{MAX}} = 54$	-1.65	0.80

however, that not much nationally important computational problems exist, or have been identified, nowadays to bring researchers from several institutions together to solve a common problem. We have also identified and labeled some *central* authors in some of the subgraphs. We have identified the top five scientists with the most number of collaborators namely, PC Naval with 54 collaborators, RC Sison with 30, D Cheng with 29, RP Saldaña with 25, and HN Adorna with 20. We believed that these authors, together with those whom we identified with the most number of papers, are the *central* scientists in their respective subgraphs. By central we mean the most influential person among the connected authors in the subgraph.

3.6 Number of collaborators per author

In the area of collaborative research, the Filipino CS researchers have collaborated with more other researchers ($C_{\mathcal{A}\text{AVG}} = 3.70$) compared to that of their ABE ($C_{\mathcal{A}\text{AVG}} = 2.35$), agriculture ($C_{\mathcal{A}\text{AVG}} = 2.70$) and NAST ($C_{\mathcal{A}\text{AVG}} = 2.82$) counterparts in the country. The Filipino physicists, however, have more collaborators ($C_{\mathcal{A}\text{AVG}} = 10.80$) than the computer scientists. Similarly, the international scientists have collaborated significantly more compared to the Filipino computer scientists, with the biomedical researchers and high-energy physicists having an average collaborators of 18.1 and 173, respectively. This seemingly high number of collaborators in high-energy physics is actually achievable, as pointed out by Newman [18], because of the significantly higher average number of authors per paper in their community ($C_{\mathcal{A}\text{AVG}} = 8.96$). Thus, the *mega–collaboration* average of 173 is actually just a product of their high $A_\mathcal{P}$. The Filipino CS researchers, on the other hand, have collaborated with almost the
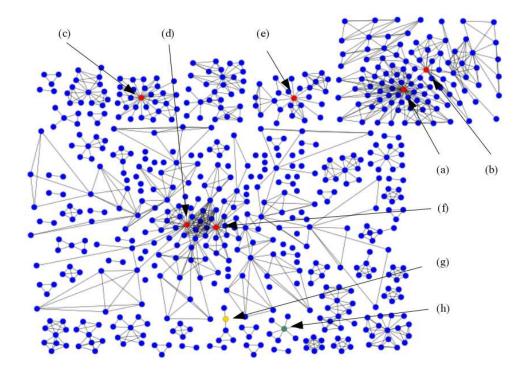
Figure 4: The co-authorshop graph CAG of Filipino computer scientists is a *sociogram* that shows the professional relations between scientists involved in scientific research. The sociogram was drawn using the two-dimensional force-directed algorithm by Kamada and Kawai [12]. The labels correspond to some identified authors believed to be central persons in their respective subgraphs. (a) P Naval, (b) HN Adorna, (c) RP Saldaña, (d) R Sison, (e) REO Roxas, and (f) D Cheng are the top researchers with the most number of collaborators. (g) This journal's editor-in-chief EA Albacea with his own collaboration subgraph. (h) This paper's author, who is also a central person in his own, although small, subgraph. (This figure is in color in the digital format of this paper.)

same number of collaborators as that of the international counterparts ($C_{\mathcal{A}\text{AVG}} = 3.59$).

3.7 Degree Distributions

Figures 5 shows the respective degree ($\Delta^{\mathcal{P}}$, $\Delta^{\mathcal{A}}$, and Δ^C) frequencies of the vertices in **PAG** and **CAG**, each plotted in scatter (for raw data) and line (predicted) plots. Figure 5(a) shows the scatter and predicted line plots of the frequency distribution of the number of authors per paper in normal and log-log scales. Here we see that the predicted line plots follow a power law form. The power law line that we we found has the form $y = 269.15(\Delta^{\mathcal{P}})^{-2.04}$ with $R^2 = 0.71$. Both coefficients $c = 269.15$ and $\varphi = -2.04$ are significantly different from zero at 1% statistics, respectively, confirming our hypothesis that $\rho(\Delta^{\mathcal{P}})$ obeys a power law distribution. We did not include the distribution for $\Delta^{\mathcal{A}} = 0$ because no paper could have been written by zero authors (i.e., no paper has a missing author information).

Figure 5(b) shows the scatter and predicted line plots of the frequency distribution of the number of papers per author. Here we see that the line plot follows a power law of the form $y = 138(\Delta^{\mathcal{A}})^{-1.88}$ with $R^2 = 0.83$. The coefficients $c = 138$ and $\varphi = -1.88$ are significantly different from zero at 1% statistics. Here, we did not include the distribution for $\Delta^{\mathcal{P}} = 0$ because the CS researchers who have written zero papers are

not included in the archive. The distribution generally characterizes a high number of authors who wrote a small number of papers, and a small number of authors who wrote a very large number of papers. Thus, in CS research the Philippines, the number of highly productive researchers is a relatively small fraction of all Filipino CS scientists. The power $\varphi = -1.88$ is in close agreement with Lotka's *law of scientific productivity* found in an experiment in 1926 to be ≈ -2 [16], while the coefficient $c = 138$ uniquely characterizes the scientific productivity of CS researchers in the Philippines.

Figure 5(c) shows the degree frequency of the vertices in **CAG** with a power law line fit of the form $y = 251.03(\Delta^C)^{-1.65}$ having a $R^2 = 0.80$. We accept that the power law is the best model for $\rho(\Delta^C)$ because we have found both $c = 251.033$ and $\varphi = -1.65$ to be significantly different from zero at 1% statistics. A power law fit suggests that:

1. Only a few number of authors have the most number of co-authors in **CAG**. These authors act as information hubs in the co-authorship graph, and therefore has the potential control of information flow through the network. We deemed such authors as influential or *central*. We have already identified these *central* persons in § 3.2 and § 3.5.
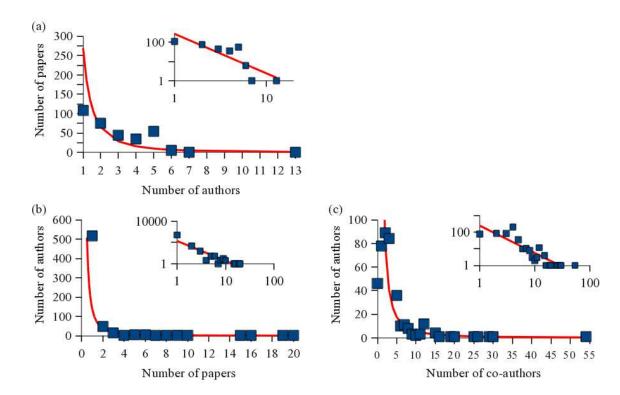
Figure 5: The vertex degree distributions follow the power law. Blue squares mean frequency of the vertex degree while the red dashed line is the power law fit. INSET: The same scatter and line plots in log-log scale. (a) $\rho(\Delta^{\mathcal{P}}) : y = 269.15^*(\Delta^{\mathcal{P}})^{-2.04^*}, R^2 = 0.71;$ **(b)** $\rho(\Delta^{\mathcal{A}}) : y = 138^*(\Delta^{\mathcal{A}})^{-1.88^*}, R^2 = 0.83;$ **and (c)** $\rho(\Delta^C) : y = 251.03^*(\Delta^C)^{-1.65^*}, R^2 = 0.80.$ ***The estimated coefficients are significantly different from zero at** $\alpha = 0.01$ **statistics. (This figure is in color in the digital format of this paper.)**

2. The co-authorship of CS research in the country is scale invariant. This means that the properties of **CAG** that we observed in this study, as well as the patterns of co-authorship and publication, will not change as much when the number of authors M increases. This makes **CAG** a *scale-free* graph.

3.8 Correlation Between $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$

Figure 6(a) shows the scatter plot between $P_{\mathcal{A}}$ and $C_{\mathcal{A}}$. The scatter plot shows that they are positively correlated with $r = 0.7425$. This suggests that the scientific productivity of the country's CS researchers, as measured by their number of papers, is correlated with the researchers' participation in collaborative research efforts, as measured by their number of co-authors. A highly productive scientist is most likely to have a high number of collaborators, and *vice versa*. This observation is particularly true in scientific publications because a large group of scientists has more manpower available for writing papers.

3.9 Assortative mixing in CAG

Figure 6(b–c) shows the mixing plots for correlating the $P_{\mathcal{A}}$ and the $C_{\mathcal{A}}$ of each researcher. These correlations quantify how a computer scientist chooses his collaborator based on the similarity or dissimilarity of his and the collaborator's attributes. Based on the Pearson correlation analysis, a computer scientist chooses a collaborator who has a dissimilar $P_{\mathcal{A}}$ (-0.1015) or $C_{\mathcal{A}}$ ($r = -0.0398$) as he has. We expect that a computer scientist with a low $P_{\mathcal{A}}$ will most likely chooses a collaborator whose $P_{\mathcal{A}}$ is high.

4. SUMMARY AND CONCLUSION

In this paper, we inferred two graphs **PAG** and **CAG** from the author information of CS papers in the country using various computational techniques. The graphs were based on publication data in various PCSC with 326 papers written by 605 authors. A large number of calculations were performed on the graphs, including the fundamental averages $A_{\mathcal{P}\text{AVG}} = 2.64$, $P_{\mathcal{A}\text{AVG}} = 1.42$, and $C_{\mathcal{A}\text{AVG}} = 3.70$. The respective frequency distributions of these quantities follow a power law which suggests that most papers were written by scientists with a small number of collaborators, while few papers were authored by those with large number of collaborators. Specifically, the power $\varphi = -1.88$ of the frequency distribution for $P_{\mathcal{A}}$ closely agrees with Lotka's *law of scientific productivity*. The productivity of the scientists, as measured by $P_{\mathcal{A}}$, is positively correlated with the scientist's participation in a number of collaborative research efforts, as measured by $C_{\mathcal{A}}$, suggesting that highly productive scientists are more likely to have a high number of collaborators, and scientists with high number of collaborators are more likely to be highly
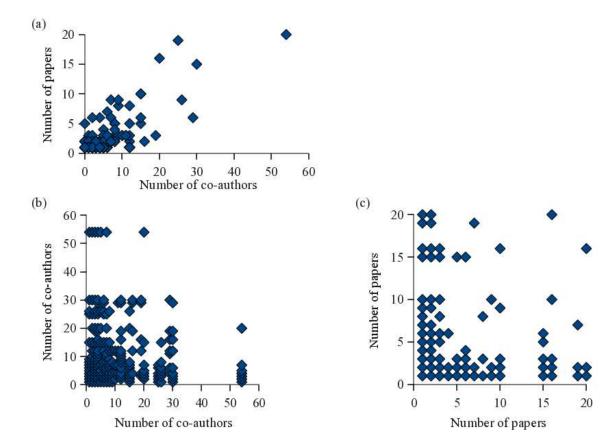
Figure 6: (a) Scatter plot between the number of papers and the number of collaborators of each author shows a positive correlation of $r = 0.7425$; (b) Mixing plot between the number of papers per author shows a low negative correlation of $r = -0.1015$; and (c) Mixing plot between the number of co-authors per author shows a low negative correlation of $r = -0.0398$.

productive. The assortativity tests show that scientists prefer to conduct collaborative research endeavors with scientists whose number of papers and collaborators are different from theirs. It is therefore reasonable to suppose that the scientific enterprise in the CS field in the Philippines will be significantly be given a boost if collaboration among scientists will be promoted (e.g., maybe through governmental policies and other programs).

The following efforts are already underway as extensions to this research endeavor:

1. Time study to measure the dynamics and evolution of **CAG**. The current effort did not measure how the current **CAG** has evolved to what it is today. Thus, the extended study tests several hypotheses regarding the nature of the development of the **CAG**, including the social phenomenon called *preferential attachment*. Preferential attachment, also known as the âĂIJrich gets richerâĂİ adage, is the tendency of new scientists to build collaborations with prolific scientists, and then later on seek more collaborations with other prolific ones. These tendencies make scientists with high number of papers to write more papers in a given time than others.

2. Development of a National Researcher Database System. Due to the inherent name ambiguity encountered in the conduct of this research, it is recommended that a National Researcher Database System (NRDS) be developed. NRDS will keep track of the changes in names used by a researcher, and at the same time be a repository of scientific articles in the Philippines. The content of the repository may be used as the National Index of Scientific papers in the Philippines. With the NRDS, a citation network may also be inferred to compute the impact factor, not only of the journals and proceedings, but also of the papers themselves.

5. ACKNOWLEDGMENTS

References

[1] H.N. Adorna, July 2010. Personal Communication.

[2] H.N. Adorna, editor. *Proceedings of the 10th Philippine Computing Science Congress.* Computing

Society of the Philippines, March 2010. Ateneo de Davao University, Davao City.

[3] H.N. Adorna and R.P. Saldaña, editors. *Proceedings of the 9th Philippine Computing Science Congress.* Computing Society of the Philippines, March 2009. Silliman University, Dumaguete City.

[4] E.A. Albacea, C. Chua, and R.P. Saldaña, editors. *Proceedings of the 4th Philippine Computing Science Congress.* Computing Society of the Philippines, February 2004. University of the Philippine Los Baños.

[5] C.C.P. Arevalo and J.P. Pabico. Preferential attachment in an internet-mediated human network. *Philippine Computing Journal*, 4(1):31–35, 2009.

[6] A. Azcarraga, editor. *Proceedings of the Philippine Computing Science Congress.* Computing Society of the Philippines, November–December 2000. De La Salle University, Manila.

[7] A.L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002. DOI: 10.1016/S0378-4371(02)00736-7.

[8] C. Bird, E.T. Barr, A. Nash, P.T. Devanbu, V. Filkov, and Z. Su. Structure and dynamics of research collaboration in computer science. In *Proceedings of the 9th SIAM International Conference on Data Mining*, pages 826–837, 2009.

[9] Computing Society of the Philippines, 2010. http://www.csp.org.ph.

[10] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, 2007.

[11] J. Huang, S. Ertekin, and C.L. Giles. Efficient name disambiguation for large-scale databases. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge Discovery in Databases (PKDD 2006): 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 536–544, 2006.

[12] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, 1989.

[13] J.S. Kim, K.-I. Goh, G. Salvi, E. Oh, B. Kahng, and D. Kim. Fractality in complex networks: Critical and supercritical skeletons. *Physical Review E*, 75: 016110, 2007. DOI: 10.1103/PhysRevE.75.016110.

[14] M. Ley and P. Reuther. Maintaining an online bibliographical database: The problem of data quality. In G. Ritschard and C. Djeraba, editors, *Extraction et gestion des connaissances (EGC'2006), Actes des sixièmes journées Extraction et Gestion des Connaissances*, pages 5–10, 2006. http://www.informatik.uni-trier.de/ ley/db/.

[15] Los Alamos National Laboratory. Los Alamos e-Print Archive, 2009. http://xxx.lanl.gov/.

[16] A.J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16(12), 1926.

[17] ncstrl.org. Network Computer Science Technical Reference, 2009. www.ncstrl.org.

[18] M.E.J. Newman. Scientific collaboration networks I: Network construction and fundamental results. *Physical Review E*, 64:016131–1–016131–8, 2001.

[19] M.E.J. Newman. Scientific collaboration networks II: Shortest paths, weighted networks, and centrality. *Physical Review E*, 64:016132–1–016132–7, 2001.

[20] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126, 2006.

[21] J.P. Pabico. Inferences in a virtual community: Demography, user preferences, and network topology. *Philippine Information Technology Journal*, 1(2):2–8, 2008.

[22] J.P. Pabico. Social network analysis of research collaboration among Filipino agricultural scientists. In *Proceedings of the 9th ISSAAS-Philippines National Convention and Annual Meeting*, 2009. In CDROM.

[23] J.P. Pabico. The network structure of scientific research collaboration of agricultural engineers in the Philippines. In *Proceedings of the Joint 8th International Agricultural Engineering Conference and Exhibition, 60th Philippine Society of Agricultural Engineers Annual National Convention, and 21st Philippine Agricultural Engineering Week*, 2010. In CDROM.

[24] J.P. Pabico and J.R.L. Micor. Structural analysis of the collaboration network of Filipino scientists. In *Proceedings of the 2nd UPLB CAS Student-Faculty Research Conference*, 2009. In CDROM.

[25] PSAE, editor. *Proceedings of the 5th International Agricultural Engineering Conference and Exhibition, 57th Philippine Society of Agricultural Engineers (PSAE) Annual National Convention, and 18th Philippine Agricultural Engineering Week.* Philippine Society of Agricultural Engineers, 2007. In CDROM.

[26] PSAE, editor. *Proceedings of the 6th International Agricultural Engineering Conference and Exhibition, 58th Philippine Society of Agricultural Engineers (PSAE) Annual National Convention, and 19th Philippine Agricultural Engineering Week.* Philippine Society of Agricultural Engineers, 2008. In CDROM.

[27] PSAE, editor. *Proceedings of the 7th International Agricultural Engineering Conference and Exhibition, 59th Philippine Society of Agricultural Engineers (PSAE) Annual National Convention, and 20th Philippine Agricultural Engineering Week.*

Philippine Society of Agricultural Engineers, 2009. In CDROM.

[28] R.P. Saldaña, editor. *Proceedings of the 6th Philippine Computing Science Congress.* Computing Society of the Philippines, March 2006. Ateneo de Manila University.

[29] R.P. Saldaña, editor. *Proceedings of the 7th Philippine Computing Science Congress.* Computing Society of the Philippines, February 2007. Boracay Regency Resort and Convention Center.

[30] R.P. Saldaña, editor. *Proceedings of the 8th Philippine Computing Science Congress.* Computing Society of the Philippines, February 2008. University of the Philippine Diliman.

[31] R.P. Saldaña and J.D.L. Caro, editors. *Proceedings of the 3rd Philippine Computing Science Congress.* Computing Society of the Philippines, February 2003. Philippine Science High School-Diliman Campus.

[32] R.P. Saldaña and C. Chua, editors. *Proceedings of the 5th Philippine Computing Science Congress.* Computing Society of the Philippines, March 2005. University of Cebu-Banilad Campus, Cebu City.

[33] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32:D449–D451, 2004.

[34] Samahang Pisika ng Pilipinas. Proceedings of the Samahang Pisika ng Pilipinas, 2008. http://www.spp-online.org.

[35] Stanford University. Stanford Public Information Retrieval System, 2009. http://www.slac.stanford.edu/spires/.

[36] V.I. Torvik and N.R. Smalheiser. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3):11:1–11:29, 2009.

[37] US National Library of Medicine. Medline, 2009. http://www.nlm.nih.gov/.

[38] K.D. Villanueva and J.P. Pabico. Structural characterization and dynamics of the Samahang Pisika ng Pilipinas (SPP) collaboration network. Technical report, Special Problem in Computer Science, Institute of Computer Science, University of the Philippines Los Baños, April 2009.

[39] F. Wang, J. Li, J. Tang, J. Zhang, and K. Wang. Name disambiguation using atomic clusters. In *Proceedings of the 9th International Conference on Web-age Information Management*, pages 357–364, 2008.