

Journal of STI Policy and Management

Publication details, including instructions for authors and subscription information: <http://www.stipmjournal.org/>

Predicting Potential Co-Authorship using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences

Rizka Rahmaida^{1,2}, Asep Saefuddin¹, Bagus Sartono¹

¹ Department of Statistics, Kampus IPB Dramaga, IPB University, Indonesia

² Indonesian Institute of Sciences, Jakarta, Indonesia

Version of record first published: 15 December 2019

To cite this article: Rahmaida, R., Saefuddin, A., and Sartono, B. (2019). Predicting Potential Co-Authorship using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences. *Journal of STI Policy and Management*, 4(2), 143–152

To link to this article: <http://dx.doi.org/10.14203/STIPM.2019.170>

ISSN 2540-9786 (Print); ISSN 2502-5996 (online)




Accreditation Number: 21/E/KPT/2018

Full terms and conditions of use: <https://creativecommons.org/licenses/by-nc-sa/4.0/>

You are free to:

- Share : copy and redistribute the material in any medium or format
- Adapt : remix, transform, and build upon the material
- The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

-  Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
-  NonCommercial — You may not use the material for commercial purposes.
-  ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

- You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.
- No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.
- If you copy the dataset merely to extract the uncopyrightable data elements would not need permission to do so. However, if you republish the full dataset or using the copyrightable data layers require a permission from Research Center for STIPM, Indonesian Institute of Sciences.

**JOURNAL OF SCIENCE, TECHNOLOGY AND INNOVATION
POLICY AND MANAGEMENT (STIPM JOURNAL),
Volume 04, Issue 02, December 2019**

FOREWORD by EDITOR-in-CHIEF

We are pleased to present the STIPM Journal Vol 4, No. 2, December, 2019. This issue brings together research findings related to science, technology, and innovation policy and management from Japan and Indonesia.

First article was written by **Djisman Simanjuntak *et al.*** entitled *Exploring the Transition to Eudaimonic Tourism: A Case Study of Bali*. This article discusses innovation in tourism focus on the dynamics of tourism grows. As tourism grows, carrying capacity is stretched or even overstretched in some places and industries. A shift toward more eudaimonic tourism is needed, and the innovative elements of eudaimonia include geographical treasure, biodiversity, and local deep culture.

Taeko Suehiro and Kumiko Miyazaki present an article entitled *Accumulation of Knowledge by Strategic Public Procurement through Public-Private-Partnership for Service Innovation in Japan*. This study focuses on how governments strategically procure public service through Public-Private Partnership (PPP)—or more specifically, Private Finance Initiative (PFI) arrangements.

Erman Aminullah presents *E-Cigarette as Disruptive Innovation: Forecasting of Conventional Cigarette Substitution in Indonesia*. This article intends to forecast conventional cigarette substitution by e-cigarette in the context of disruptive innovation. E-cigarette as disruptive innovation has been driven by technology innovation to create e-cigarette products for global market. The advancement of e-cigarette technology innovation would continue to create smart and less harmful e-cigarette as alternative tobacco products in future.

Kumiko Miyazaki, Santiago Ruiz Navas, and Ryusuke Sato present the fourth article entitled *Evolutionary Path of Development of AI and Patterns of Knowledge Convergence over the Second and Third AI Boom*. AI has been through several booms and we have currently reached the 3rd AI boom which followed the 2nd AI boom centering mainly on expert systems. The current AI boom started around 2013 and AI is beginning to affect corporate management and operations. AI has been evolving over six decades but it seems that the current boom is different from the previous booms.

The fifth article entitled *Predicting Potential Co-Authorship using Random Forest: Case of Scientific Publication in Indonesian Institute of Sciences* by **Rizka Rahmida, Asep Saefudin, and Bagus Sartono**. Co-authorship network is one of the proxies to evaluate the emerging research collaborations. Co-authorship that happens for the first time among a pair of author plays an important role as the key of success for their co-authorship in the future.

Finally, **Hiroki Idota *et al.***, present an article entitled *Conducting Product Innovation by Using Social Media among Japanese Firms*. This article based on a study that attempts to conduct an empirical

analysis of how social media use promotes product innovation in Japanese firms by collaboration with consumers based on survey data from Japanese firms using probit analysis. This study finds that collaboration with consumers by using social media is important for innovation, particularly in developing concepts and devising methods of use.

The STIPM Journal is indexed by Google Scholar, ISJD, IPI, DOAJ, BASE, and OCLC World Cat. This make the journal dissemination wider. We would like to thank all the reviewers for their excellent work and the authors who kindly contributed their papers for this issue. We are also indebted to the *STIPM Journal* editorial office at P2KMI-LIPI and the publishing and production teams at LIPI Press for their assistance in preparation and publication of this issue.

We are expecting that STIPM will always provide a higher scientific platform for the authors and the readers, with a comprehensive overview of the most recent STI Policy and Management research and development at the national, regional dan international level.

Happy New Year 2020 to all of you...

Jakarta, December 2019

Editor-In-Chief

JOURNAL OF STI POLICY AND MANAGEMENT

Volume 4, Number 2, December 2019

LIST OF CONTENTS

Exploring the Transition to Eudaimonic Tourism: A Case Study of Bali Djisman Simanjuntak, Alvin Desfiandi, Erica Lukas, Isti Setiawati, Nakita Sabrina, and Stanley Makalew	77–99
Accumulation of Knowledge by Strategic Public Procurement through Public-Private Partnerships for Service Innovation in Japan Taeko Suehiro and Kumiko Miyazaki	101–112
E-Cigarette as Disruptive Innovation: Forecasting of Conventional Cigarette Substitution in Indonesia Erman Aminullah	113–124
Evolutionary Path of Development of Artificial Intelligent (AI) and Patterns of Knowledge Convergence over the Second and Third AI Booms Kumiko Miyazaki, Santiago Ruiz Navas, and Ryusuke Sato	125–142
Predicting Potential Co-Authorship using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences Rizka Rahmaida, Asep Saefuddin, and Bagus Sartono	143–152
Conducting Product Innovation by Using Social Media among Japanese Firms Hiroki Idota, Sheikh Abu Taher, Teruyuki Bunno, and Masatsugu Tsuji	153–166



Predicting Potential Co-Authorship using Random Forest: Case of Scientific Publications in Indonesian Institute of Sciences

Rizka Rahmaida^{1*}, Asep Saefuddin¹, Bagus Sartono¹

rizkarahmaida@gmail.com, asaefuddin@gmail.com, bagusco@apps.ipb.ac.id

¹ Department of Statistics, Kampus IPB Dramaga, IPB University, Indonesia

ARTICLE INFO

Article History:

Received : 05 August 2019

Revised : 30 September 2019

Accepted : 28 November 2019

Available online : 15 December 2019

Keywords:

Research collaboration,
Co-authorship prediction,
R&D management,
Random forest classifier

ABSTRACT

Research collaboration is one of the strengths in research management due to its advantages in quantity and quality of the research. Co-authorship network is one of the proxies to evaluate the emerging research collaborations. Co-authorship that happens for the first time among a pair of author plays an important role as the key of success for their co-authorship in the future. Therefore, the research aims to build a model predicting new co-authorship as potential co-authorship. This research used scientific articles in Indonesian biodiversity research published in Scopus during 2006–2015. New co-authorship of between 4,628 pair of authors were analyzed in terms of their similarity in co-authorship network, research interest, and community to predict whether a pair of author will have a new co-authorship in future. Random forest classifier was used to build the model after applying 10-fold cross validation in various parameter and random undersampling technique as preprocessing procedures. The result shows that the similarity in network, community network, and research interest and becomes good features to predict the potential co-authorship among a pair of author. Furthermore, paired authors that predicted to be co-authored and involving authors from Indonesian Institute of Sciences are identified as the potential partners recommended for development of research teams.

©2019 PAPPITEK-LIPI All rights reserved

I. INTRODUCTION

Research collaboration is defined as working together between researchers to produce new knowledge by research activities (Katz & Martin, 1997). Researchers have some purpose

in performing collaboration: satisfy intellectual interest, share the excitement of an area with other people, keep themselves more focussed on research, and create a network with other people (Beaver, 2001). Collaboration has some benefits such as: enhanced productivity, research progressed more rapidly (Beaver, 2001), better quality research (Ibáñez, Bielza, & Larrañaga,

* Corresponding Author.: +62-812-9808-0360
E-mail: rizkarahmaida@gmail.com

2013), and generate new ideas and learn new skills (Bammer, 2008). Therefore, collaboration is an important aspect in research management.

Process in building collaboration for the first time is not an easy task because researchers should find suitable partners who potentially will succeed in the future (Pavlov & Ichise, 2007). Under this condition, researchers faced uncertainty about the suitable person to collaborate with. This problem can be overcome if researchers have access to information related to other researchers, for example information on research interests and activities that are being carried out by other researchers (Yu, et al., 2014). This kind of information is generally not available in Indonesia. For example, Center for Development, Education, and Training, Indonesian Institute of Sciences (LIPI) has been developing the National Credit Score Assessment System since the end of 2015. The main purpose of this system is to assist researchers in submitting credit numbers that contain information on their research output needed to occupy a functional position of researcher. Unfortunately, this application cannot be used to build collaboration prediction model between researchers due to limited range and structure of data. Therefore, another approach in analyzing collaboration is a necessity.

Co-authored publication has been used as a basic counting unit to measure collaborative activity (Katz & Martin, 1997). Therefore, co-authorship can be approached to solve the problem. Co-authorship data contains information about scientific articles including all of the authors. Furthermore, analysis on co-authorship prediction can be used to predict authors that potentially to collaborate in conducting research (Guns & Rousseau, 2014).

Previous research on co-authorship prediction used random forest model to predict co-authorship in malaria and TBC (Guns & Rousseau, 2014), computer sciences (Asil & Gurgun, 2017), and physics studies (Aouay, Jamoussi, & Gargouri, 2014). Generally, research on co-authorship prediction only used network similarity in co-authorship network (Aouay, et al., 2014; Asil & Gurgun, 2017; Guns & Rousseau, 2014; Roopashree & Umadevi, 2014; Yu, et

al., 2014). In those research, the similarity were calculated based on number of similar co-authors without concerning research interest among a pair of authors. Meanwhile, Chuan, et al. (2017) used Latent Dirichlet Allocation (LDA) to investigate topic similarity to measure research interest by analyzing title and abstract in document article. Curiskis, Osborn, dan Kennedy (2015) used social network feature and community feature in building model prediction. He detected the community by clustering co-authorship networks. Both LDA and clustering a large co-authorship network had disadvantages because they required a high performance of computing time and resources.

The contribution of this research is highlighted as follows. We used simple approaches to measure knowledge similarity and community network in building prediction model. We used research object, title, and journal name to measure research interest similarity. We also used journal, institution, and country to detect community among the pairs. Those approaches can provide a good prediction model with average computation resources.

Research using co-authorship data in Indonesia is still restricted to descriptive analysis on chemical (Nadhiroh, 2015), biodiversity (Handayani, Amelia, Rahmida, Hardiyati, & Nadhiroh, 2016) and statistics and mathematics studies (Nadhiroh, Hardiyati, Amelia, & Handayani, 2018). Research on co-authorship prediction using data from Indonesia is limited yet.

Indonesia has the highest national biodiversity index in the world (UNEP, 2001). Its natural resources attracted many foreign researchers to conduct their research in Indonesia (Handayani, et al., 2016). The existence of various research objects in Indonesia and high interest of foreign researchers provides a large potential in research collaboration. As mentioned before, research collaboration is an important aspect in research management but so far there is no research related to determining potential collaboration partners in Indonesia. Therefore, this research used publications on Indonesian biodiversity research as a case study. At the other side, Indonesian Institute of Sciences (LIPI), as the biggest research institution in Indonesia, has

to improve its performance and should carry out good research management. In order to satisfy the necessity, co-authorship prediction analysis can give input to generate policy concerning research collaboration management.

This study aimed to build an acceptable model using random forest classifier in order to predict potential co-authorship. Random forest is an advanced classifier that theoretically and empirically has good prediction results. The application of this classifier is easy because it doesn't need any preprocessing steps and, has ability to handle numerical and categorical predictors. Then, the prediction model is used to predict co-authorship involving authors from LIPI. Prediction determined by the model yield a recommendation for research management in LIPI.

II. ANALYTICAL FRAMEWORK

Scientific collaboration can be defined as interaction among scientists to complete research tasks. Task within a collaboration often have a high degree of uncertainty. In a research collaboration, it is not clear whether the goal can be achieved or what is the best way to achieve it (Sonnenwald, 2005). Two authors who never collaborated will consider some factor to initiate new collaboration.

A. Co-authorship network

According to Owusu-nimo and Boshoff (2016), existing personal or working relationship and a mutual acquaintance are factors that influenced the initiation of collaboration among researchers. Collaboration, as represented in co-authorship, is also more likely to be characterized with (1) a pattern of history among co-authors, (2) frequent communication, (3) some level of mutual trust, and (4) shared socialization or educational history (Ponomariov & Boardman, 2016).

Personal factors also play important roles in order to initiate the collaboration. For example, similar approaches to science, trust, and the ability to get along with each other are also used to identify and select collaborators (Maglaughlin & Sonnenwald, 2005).

Collaboration of a large number of authors, represented in co-authorship relationship, formed a social network. It becomes a foundation for collaboration (Sonnenwald, 2005). In a social network, two scientists are more likely to collaborate and co-author a paper if they have a co-author in common (Newman, 2001).

Two authors are connected if they had co-authored at least one article in common. In terms of neighborhood, we said that author A is the neighbor of author B if they had at least one article in common. The similarity of two authors are measured by Common Neighbour (CN), Jaccard Coefficient (JC), and Adamic Adar (AA) (Chuan et al., 2017). The calculation of these features are based by a number of co-author of an author that called neighbors $\Gamma(u)$ and obtained three numerical variables. In this research, co-authorship network is assumed as a binary network.

Common neighbors of author u and v is defined as the number of common neighbors shared by author u and author v . Newman (2001) verified a correlation between CN of author u and v at the current time, and the probability that they will collaborate in the future. Common Neighbors of authors $(u, v)(u, v)$ is calculated by:

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Where $CN(u, v)$ is Common neighbors of authors (u, v) , $\Gamma(u)$ is set of neighbors of author u , and $\Gamma(v)$ is set of neighbors of author v .

Jaccard's coefficient is a normalized measure of common neighbors. It computes the ratio of common neighbors out of all neighbors, and can be used for comparing the similarity and diversity of neighbor set. The calculation of Jaccard's coefficient of authors $(u, v)(u, v)$ follows this formula:

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

Where $JC(u, v)$ is Jaccard's coefficient of authors (u, v) , $\Gamma(u)$ is set of neighbors of author u , and $\Gamma(v)$ is set of neighbors of author v .

Adamic Adar, a weighted version of common neighbors, using greater weight to common neighbors w of author u and author v which

themselves have fewer neighbors. This means the contribution of a common neighbor to the score is weighted in proportion to the rarity of the neighbor. This measurement is calculated by:

$$AA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\ln(|\Gamma(z)|)}$$

Where $AA(u, v)$ is Adamic Adar of of authors (u, v) , $|\Gamma(z)|$ is number of $\Gamma(z)$ member, and $\Gamma(v)$ is set of neighbors of author v .

B. Community network

Researchers who joined the same community tends to have a social network (Beaver, 2001). Therefore, a community can be another form of social network. Cusriskis, et al. (2015) predicted co-authorship using community network. They defined community based on the co-authorship network by clustering large co-authorship network. Katz and Martin (1997) defined different levels of collaboration in terms of intra and inter at institution and country level. Those levels can be viewed as communities due to its administration boundaries.

Technical computation becomes consideration for proposing new measurements. Technically, process of clustering large network requires high performance computing resources. In this research, we propose two measurements of community network: institution network and country network. Two authors are in the same community if they are from same institution (or country). Pairs of authors from the same institution (or country) was coded '1' while others was coded '0' for their corresponding measurement.

C. Knowledge similarity

Scientists identify ideas for new projects and select collaborators in their social network (Beaver, 2001; Katz & Martin, 1997). Chuan, et al. (2017) predicted co-authorship by measurement of content similarity that reflect the knowledge of each author. He used topic modelling algorithm in a whole text of the document. This method had a disadvantage that the proposed method has

high computational time in comparison with the relevant algorithms.

A journal usually reflects a specific research area. Therefore, authors who had published articles in that journal, usually have similar knowledge. The calculation of journal network is analogue with co-authorship network. Pairs of authors who have published article in at least one journal in common were considered to have relationship in journal network and were coded '1'.

Glanzel (2003) propose co-word analysis to map knowledge in a research area. He also mentioned that co-word analysis is based on frequency analysis of co-occurrence of keywords extracted from titles, abstracts or text, in general. In this research, knowledge similarity between two authors can be measured by calculating their word similarity in tittle of their articles, and their journal names. Because the case chosen in this study was the publication of biodiversity, this study also considered the object of research as keywords that can be extracted. The categorization of research object is based on seven categories of kingdom in taxonomy: Plantae, Animalia, Protozoa, Fungi, Bacteria, Archaea, and Chromista. Based on expert judgment, those categories could be used in the process of collecting articles on Indonesian biodiversity research. Each article collected must contain at least one of seven kingdom names in its title, keywords, or abstract. Therefore, this step guaranteed that every article belonged to at least one kingdom category. This also enabled us to measure similarity of every pair of authors using co-word analysis in terms of co-occurrence of kingdom category (as research object) extracted from their articles. These measurements use a simple natural language processing called Jaccard Index (JI). JI calculates the ratio of the number of similar words in common out of all words. Calculations of JI were carried out after deleting stopwords as preprocessing technique. JI of two sets of words A and B were obtained using:

$$JI(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where $|A \cap B|$ is number of intersection of A and B, and $|A \cup B|$ is number of union of A and B.

III. METHODOLOGY

Collaboration can be analyzed using co-authorship network. In this network, co-authorships between authors were described by a network in which a node represented an author and an edge represented a co-authorship relation. Co-authorship networks during some periods have dynamic forms in terms of which authors published articles and had co-authorships.

Two authors who haven't been connected yet until current period have two possibility in future period: being connected or not being connected. This event can be viewed as a binary variable having two possibilities of outcome. Their characteristics can also be observed based on information in current period. The information can be generated from co-authorship network, community network, and knowledge similarity.

We built the relationship prediction models that model the probability of co-authorship between two authors as a function of topological features between them. We chose the Random Forest (RF) as our prediction model. RF is one of the most widely used classification methods. It has good prediction performance both theoretically and empirically. RF is a robust machine learning technique for classification and regression (Breiman, 1999). It is an ensemble of many decision trees where each tree is built starting from a bootstrap sample of the input data. Each node (i.e., each decision) in a tree is based on a random subset of the available features (variables). Randomness at the data and model level yield accurate and robust results. RF can automatically predict the probability of an item belongs to a certain class. In this case, RF can predict the probability of a pair to have co-authorship.

This research used secondary data that contained information about scientific articles in Indonesian biodiversity research published in Scopus during 2006–2015 (Handayani, et al., 2016). Data collection is based on focus group discussion with experts in biodiversity research to ensure that data can truly describe the research.

Collection process resulted in 3,563 scientific articles covering article titles, journal names, author names, and objects of research. Objects of research are determined by seven categories of kingdom in taxonomy.

Two networks were built based on publications in T1=[2006-2010] and T2=[2011-2015]. Social network features were extracted from co-authorship network of T1, while corresponding labels (whether there was a new co-authorship relation in T2 between a pair of authors or not) were extracted from co-authorship network of T2. Some authors only published their papers in one of the two periods, T1 or T2. This condition implies that the feature and corresponding label couldn't be extracted completely. To solve this problem, we restricted authors to those who published article in both periods, resulting 1,267 authors. This research uses 9 variables based on co-authorship network, community network, and scientific interest (table 1).

Table 1.

Variables used in this research

Type of Feature	Variables	Type of variables
Co-authorship Network	Common Neighbors	Numeric
	Jaccard Coefficient	Numeric
	Adamic Adar	Numeric
Community network	Institution network	Categoric
	Country network	Categoric
Knowledge similarity	Journal network	Categoric
	Research object similarity	Numeric
	Article title similarity	Numeric
	Journal names similarity	Numeric

We also restricted author pairs to those who did not have relationship in T1 but had a new relationship in T2. We only took those pairs that have at least one common neighbor, called 2-hop authors. Based on these constraints, we found 4,628 pairs to observe, of which 346 pairs (7.5%) had a new relationship in T2.

To ensure that the model had good performance in all parts of the data, we use 10-fold cross validation techniques in model

building. This step enabled each observation contributed both training and testing. Therefore, it guarantees that the quality of random forest generated is fitted to a whole data. We also built models using 99, 199, 299, and 399 trees then we tuned the number of features used (from 2 to 8) for each number of trees to choose the best parameter used in random forest classifier

In a classification problem, models were evaluated by comparing its sensitivity and AUC value. Sensitivity shows how good the model can predict collaboration in future time while AUC measures quality of a probabilistic classifier. In this co-authorship prediction context, it can be used to quantify the overall ability of the model to discriminate between those author pairs (Yu, et al., 2014). Generally, a model that obtained 80% of AUC is said to have a good discrimination ability.

IV. RESULTS AND DISCUSSION

A. Data Exploration

Exploration of a numerical variable can be viewed from boxplot. This kind of plot describe distribution of variables and show whether outliers exist. Figure 1 compares distribution of CN, JC and AA (as co-authorship network variables) between pairs of authors who initiated co-authorship and those who didn't. Figure 1 shows that there are many outliers in those variables represented by separated dots outside the boxplots. By comparing two boxplots on each variable, we know that the distribution of those variables in pairs of authors, who initiated co-authorship, are relatively higher than those of authors who didn't.

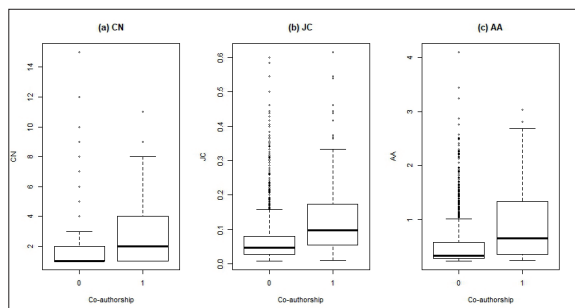


Figure 1. Boxplot of co-authorship network variables by co-authorship

CN is a variable that measures how close a pair of authors were in a co-authorship network. These results are in accordance with Newman (2001) who revealed that two authors are co-author a paper if they have a co-author in the collaboration. This indicates that the process of scientists introducing their collaborators to one another is an important one in the development of scientific communities. This also applied in JC and AA as similar closeness measurement but with standardizing and weighting.

Figure 2 compares the percentage of pairs who initiated co-authorship according to whether they connected in community network or not. The highest difference is shown in figure 2(b). The percentage of co-authorships from pairs connected in an institution network is significantly higher than the percentage of co-authorships from those who weren't connected. While Figure 2(c) shows the lower difference, Figure 2(a) shows almost no difference.

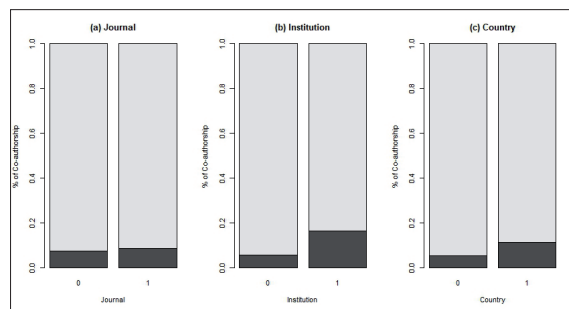


Figure 2. Percentage of co-authorship by categorical predictor variables: (a) Journal, (b) Institution and (c) Country

While past research showed that social network is an important factor influencing initiation of co-authorship (Katz & Martin, 1997) , our research shows a specific result that journal network is not a significant factor. Our research also confirmed that authors tend to choose institutional network to initiate co-authorship. This result differs from those of Ponomariov and Boardman (2016) that institutional influences may be less important than typically thought. Their result was shown by the majority of respondents' close collaborations are with individuals from outside the university, and collaborations with outside individuals tend to be more likely to result in a co-authored publication.

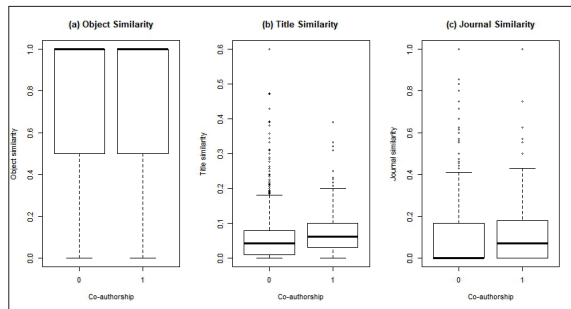


Figure 3. Boxplot of community network variables by co-authorship

Figure 3 compares distribution of object, title, and journal similarity between pairs of authors who initiated co-authorship and those who didn't. Figure 3a shows that the distributions of object similarity between two groups are almost same. Although it has some lower outliers, distributions of title similarity of pairs who initiated to collaborate is slightly higher than other's (Figure

3b). Distributions of journal similarity of them also higher than other's (Figure 3c). This results shows that title and journal similarity influences are important while object similarity influence is not.

B. Overall Accuracy

As mentioned before, our prediction networks were built based on social network and research interest. Table 2 shows that all models obtained more than 75% of sensitivity and 80% of AUC. Based on Table 2, we decided to choose random forest models using 199 tree and 5 features as the best model because it obtained the highest sensitivity among others. The reason for high prediction sensitivity rates may be explained by the fact that collaboration usually emerges from social networks.

Table 2.

Model evaluation based on various number of trees and feature

No. of Tree	No. of feature	Sensitivity	AUC
99	2	0.783	0.818
	3	0.772	0.822
	4	0.786	0.824
	5	0.777	0.824
	6	0.783	0.825
	7	0.769	0.820
	8	0.772	0.822
	199	2	0.777
3		0.783	0.825
4		0.769	0.824
5		0.795	0.824
6		0.783	0.824
7		0.783	0.821
8		0.769	0.822
299		2	0.777
	3	0.789	0.826
	4	0.783	0.826
	5	0.783	0.826
	6	0.786	0.824
	7	0.777	0.822
	8	0.780	0.823
	399	2	0.772
3		0.786	0.826
4		0.780	0.826
5		0.786	0.826
6		0.783	0.825
7		0.777	0.823
8		0.780	0.824

In this research, we added community network variables (institution and country networks) to measure social network aspect. It obtained the higher prediction performance than another model built by Yu, et al. (2014) which only used co-authorship network to measure social network aspect. Adding those variables showed that social networks can be expanded through those networks in order to predict co-authorship. Those networks can be a chance to make an information flow between two scientists.

Two scientist that have published article in at least one journal in common regularly visited the journal website in their submitting process and received online version of the journal after their articles were published. During this process, information flows between the scientists as the authors of a journal. In the case of institution (or country) network, two scientists from the same institution (or country) have a higher possibility to collaborate than those from different institution (or country). This is caused by the fact that collaboration within an institution (or a country) is usually easier and faster to build than collaboration involving different institutions (or countries) in terms of regulation and administration.

C. Prediction Result

Co-authorship prediction is used to identify pairs that will have a successful co-authorship in the future. Relationship prediction is based on the independent variable in the T2 period. In that period, there were 7,667 authors who produced scientific articles. All possible pairs of authors were identified then selected pairs that had a CN > 0 and involved at least one author from LIPI were selected. The results of this selection resulted in 3,343 pairs. Of these, the authors who were connected in the T2 period were excluded remaining 2,462 pairs of authors to be predicted. The pair was formed by a total of 1,068 authors.

The pair of authors from LIPI who is predicted to have co-authorship is a potential co-authorship. This potential gives benefit to LIPI as an institution that has the main task to conduct research activities. Potential co-authorship can be seen based on the collaboration category at the institutional level, namely intra-institutional and

inter-institutional collaboration (Katz & Martin, 1997). Intra-institutional collaboration is a collaboration carried out by the authors in one institution (in this case, LIPI), while inter-institutional collaboration is a collaboration carried out by an author from LIPI with author from outside LIPI.

Table 3 shows the distribution of paired author based on the results of the predictions. The pairs predicted to have the potential co-authorship is pairs with a predicted probability ≥ 0.5 . Prediction results show that the percentage of potential pairs in the category of intra-institutional co-authorship tends to be higher compared to inter-institution co-authorship. This is due to the similarity of institutions that tend to provide convenience in terms of administration in the process of initializing co-authorship.

Table 3.

Distribution of paired author based on prediction result

Co-authorship type	Potential	Not potential
Intra-institution	87 (60.8 %)	56 (39.2 %)
Inter-institution	561 (24.2 %)	1,758 (75.8 %)

In recommending the potential of co-authorship, several studies consider the prediction of co-authorship probability generated from model to set priorities for co-authorship (Guns & Rousseau, 2014; Zhang, 2017). In this study, 100 top pairs were sorted by the highest predicted probability of co-authorship (Table 4). For each addition of 10 pairs, the pairs of potential intra-institutional co-authorship is calculated.

Table 6 shows that the potential of co-authorship intra-institutional is very low in various numbers of pairs evaluated compared with those of inter-institutional co-authorships, with the highest percentage of 15%. From the top 100 pairs examined, the random forest model only recommended three intra-institutional co-authorships. This means that high potential co-authorships are mostly in the form of inter-institutional co-authorships.

Based on research conducted by Gazni and Didegah (2011), scientific articles written by many institutions (inter-institution) tend to get higher number of citations. Then Leimu and Koricheva (2005) mentioned that the number of citations can be a measure of the quality of scientific articles. In other words, co-authorship between LIPI and non-LIPI will produce better

quality research so that LIPI have to manage them well.

Table 4.

Distribution of paired potential intra-LIPI co-authorship on various top highest potential pairs

No. of potential pairs observed	No. of co-authorship	Percentage
10	0	0.0 %
20	0	0.0 %
30	0	0.0 %
40	0	0.0 %
50	0	0.0 %
60	0	0.0 %
70	1	1.4 %
80	3	3.8 %
90	3	3.3 %
100	3	3.0 %

V. CONCLUSION

Predicting co-authorship is an important analysis in formulating recommendation for collaboration research management. This research showed that similarity in network, community network, and research interest are useful in predicting new co-authorship. Furthermore, paired authors that predicted to be co-authored and involving authors from Indonesian Institute of Sciences are identified as the potential partners recommended for development of research teams.

REFERENCES

- Aouay, S., Jamoussi, S., & Gargouri, F. (2014). Feature based link prediction. In *ACS 11th International Conference on Computer Systems and Applications (AICCSA 2014)* (Vol. 2014, pp. 523–527). Doha: IEEE. <https://doi.org/10.1109/AICCSA.2014.7073243>
- Asil, A., & Gurgun, F. (2017). Supervised and fuzzy rule based link prediction in weighted co-authorship networks. In *2nd International Conference on Computer Science and Engineering* (pp. 1–5). IEEE.
- Bammer, G. (2008). Enhancing research collaborations: Three key management challenges. *Research Policy*, 37(5), 875–887. <https://doi.org/10.1016/j.respol.2008.03.004>
- Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, 52(3), 365–377. <https://doi.org/10.1023/A:1014254214337>
- Breiman, L. (1999). Random Forests, 5–32. Retrieved from http://machinelearning202.pbworks.com/w/file/attach/60606349/breiman_random-forests.pdf
- Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2017). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, 48(8), 2470–2486. <https://doi.org/10.1007/s10489-017-1086-x>
- Curiskis, S. A., Osborn, T. R., & Kennedy, P. J. (2015). Link prediction and topological feature importance in social networks. In M. Zahidul Islam, L. Chen, K.-L. Ong, Y. Zhao, R. Nayak, & K. Paul (Eds.), *Thirteenth Australasian Data Mining Conf.* (pp. 39–50). Sydney: Australian Comp. Soc. Inc.
- Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact : a case study of Harvard University’s publications. *Scientometrics*, 87, 251–265. <https://doi.org/10.1007/s11192-011-0343-8>
- Glanzel, W. (2003). *Bibliometrics as a research field : A course on theory and application of bibliometric indicators*. Retrieved from http://yunus.hacettepe.edu.tr/~tonta/courses/spring2011/bby704/Bib_Module_KUL.pdf
- Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2), 1461–1473. <https://doi.org/10.1007/s11192-013-1228-9>
- Handayani, T., Amelia, M., Rahmida, R., Hardiyati, R., & Nadhiroh, I. M. (2016). *Kajian Saintometrika Perkembangan Publikasi Ilmiah Keanekaragaman Hayati Indonesia Sebagai Bahan Rekomendasi Kebijakan Arah Penelitian Keanekaragaman Hayati Nasional*. Pappiptek LIPI Jakarta.
- Ibáñez, A., Bielza, C., & Larrañaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: A case study in Spanish computer science production in 2000-2009. *Scientometrics*, 95(2), 689–716. <https://doi.org/10.1007/s11192-012-0883-6>
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1–18. [https://doi.org/10.1016/S0048-7333\(96\)00917-1](https://doi.org/10.1016/S0048-7333(96)00917-1)

- Leimu, R., & Koricheva, J. (2005). Does Scientific Collaboration Increase the Impact of Ecological Articles? *BioScience*, 55(5), 438–443.
- Maglaughlin, K. L., & Sonnenwald, D. H. (2005). Factors that Impact Interdisciplinary Natural Science Research Collaboration in Academia. In *The 10th International Conference of the International Society for Scientometrics and Informetrics* (pp. 499–508). Stockholm: Karolinska University Press. Retrieved from https://www.researchgate.net/publication/236658987_Factors_that_impact_interdisciplinary_natural_science_research_collaboration_in_academia.
- Nadhiroh, I. M. (2015). *Jaringan co-authorship dan potensi kolaborasi penelitian Indonesia dengan analisis jaringan sosial [tesis]*. Bogor (ID): Institut Pertanian Bogor.
- Nadhiroh, I. M., Hardiyati, R., Amelia, M., & Handayani, T. (2018). Mathematics and statistics related studies in Indonesia using co-authorship network analysis. *International Journal of Advances in Intelligent Informatics*, 4(2), 142–153. <https://doi.org/10.26555/ijain.v4i2.120>.
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. In *Proceedings of the National Academy of Sciences* (Vol. 98, pp. 404–409). <https://doi.org/10.1073/pnas.98.2.404>.
- Owusu-nimo, F., & Boshoff, N. (2016). Research collaboration in Ghana: patterns, motives and roles. *Scientometrics*, 110(3), 1099–1121. <https://doi.org/10.1007/s11192-016-2221-x>.
- Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. In *CEUR Workshop Proceedings* (Vol. 290, pp. 42–55). Busan: ISWSA.
- Ponomariov, B., & Boardman, C. (2016). What is co-authorship? *Scientometrics*, 109(3), 1939–1963. <https://doi.org/10.1007/s11192-016-2127-7>.
- Roopashree, N., & Umadevi, V. (2014). Future Collaboration Prediction in Co-authorship Network. In *Proceedings - 2014 3rd International Conference on Eco-Friendly Computing and Communication Systems, ICECCS 2014* (pp. 183–188). Mangalore: IEEE. <https://doi.org/10.1109/Eco-friendly.2014.45>.
- Sonnenwald, D. H. (2005). Scientific Collaboration. *Annual Review of Information Science and Technology*, 41(1), 643–681.
- Yu, Q., Long, C., He, P., Shao, H., Duan, Z., Lv, Y., & Yu, Q. (2014). Predicting Co-Author Relationship in Medical Co-Authorship Networks. *PLoS ONE*, 9(7), 1–7. <https://doi.org/10.1371/journal.pone.0101214>.
- Zhang, J. (2017). Research collaboration prediction and recommendation based on network embedding in co-authorship networks. In *Proceedings of the Association for Information Science and Technology* (Vol. 54, pp. 847–849). Washington, DC. <https://doi.org/10.1002/pra2.2017.14505401182>.
- [UNEP] United Nations Environment Programme. (2001). *Global Biodiversity Outlook*. Retrieved from <https://www.cbd.int/gbo1/copyright.sht>.