



Network analysis of temporal trends in scholarly research productivity

Hyoungshick Kim^a, Ji Won Yoon^{b,*}, Jon Crowcroft^a

^a Computer Laboratory, University of Cambridge, Cambridge, UK

^b Statistics Department, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Article history:

Received 18 December 2010

Received in revised form 28 May 2011

Accepted 31 May 2011

Keywords:

Publication analysis

Research trend

Comparative analysis

Network analysis

Security research

Korean research

ABSTRACT

We propose a method to identify the journals or proceedings that are most highly esteemed by a research group over some time frame. Using open publication databases, we identify the experts in the community, and analyse their publication pattern, and then use this as a guideline for evaluating scientific outputs of other groups of researchers publishing in the same domain. To illustrate the practicality of our method, we analyse the scientific output of Korean researchers in the security subject domain from 2004 to 2009, and comparing this groups' output with that of well-known researchers. Our empirical analysis *demonstrates* that there is a persistent gap between these two research groups' publications impact over this period, although the absolute number of journal publications greatly increased over recent years.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Over the last decade, the South Korean government invested a great deal of money hoping to make Korean research universities and institutes globally competitive and to produce more high-quality research outputs. In particular, a research funding program called Brain Korea 21 (BK 21) was launched by the Korea Ministry of Education (Shin, 2009), especially designed to strengthen the competitiveness of universities in Korea. This provides fellowship funding to graduate students and professors in research groups at top Korean universities. As a result, research publication outputs from Korean research universities and laboratories has grown perceptibly: for example, the number of research publications from Seoul National University (SNU) was over 4000 in 2004, comparable with world-class universities such as Stanford, Johns Hopkins, and UC Berkeley. In 2004, article publications from SNU approached around 50% of those from Harvard university which has the highest number of total Science Citation Index (SCI) publications worldwide that year. Comparing this with 1995, when the production of published articles from SNU was about 18% of Harvard's, we can consider as an outstanding accomplishment (Shin, 2009). However, questions about research quality and direction are continuously posed, despite this quantitative growth.

Bibliometric indicators such as the number of publications seem to be simple and practical metrics to evaluate scientific contributions at the macroscopic level. However, it is not desirable to use this metric for all purposes. Bibliometric indicators such as the number of publications, journal impact factors, number of citations, and citation index are often readily available, and provide further meaningful information about the level of research productivity and scientific impact. Not surprisingly, it is really important to use a bibliometric database which is suitable for this purpose since these indicators can vary greatly depending on the source used.

* Corresponding author.

E-mail addresses: hk331@cl.cam.ac.uk (H. Kim), yoonyj@tcd.ie (J.W. Yoon), Jon.Crowcroft@cl.cam.ac.uk (J. Crowcroft).

The Thomson Reuters (formerly ISI) bibliographic database, which includes the Arts and Humanities Citation Index (A & HCI), Science Citation Index (SCI) and Social Sciences Citation Index (SSCI), has been used for decades as de facto standard databases for conducting publication and citation analyses (Meho, 2007). However, this is not a universally applicable resource. First of all, the coverage of the database is by no means complete across all subjects. Different research fields are covered unevenly, and few of conference proceedings and books are included, despite being important in some scientific areas. Unlike fields such as natural sciences and life sciences, prestigious conferences hosted by professional computer science societies such as ACM/IEEE are preferred to journals as a place to present original and important results (Fortnow, 2009; Meyer, Choppy, Staunstrup, & Leeuwen, 2009). Although we can try to use a database for proceedings (e.g. Thomson Reuters' a databases for proceedings), prestige proceedings would still need to be chosen manually. Moreover, some national journals, important in the social sciences and humanities, may not be covered because the databases have an English language bias (Seglen, 1998). Lastly, although the database attempts to include the most important scientific literature for some specific subject, it is difficult to isolate scientific contributions relevant to that specific subject, since unrelated literature is also sometimes included. Our study is partly motivated by problems caused by naive use of such problems with the bibliographic database.

Our key insight is to extract experts' publication pattern for a specific subject and then use this as a guideline or reference point to help evaluate the scientific contribution of a second target research group. We can identify a set of prestige journals and proceedings for a specific subject by analysing experts' scientific contributions over a given time window. Much faith is already placed in peer review – the idea that scientific contribution can be effectively evaluated by experts in the same field. We make the following two contributions, inspired by recent advances in complex network analysis.

- We propose a method to analyse the scientific production for a research group (Section 3). From open databases, we construct the researchers' publication graphs, and analyse the relationships and topological positions between journals and proceedings. From these graphs, we identify a set of prestige journals and proceedings for a specific subject over a given period of time. We extend this to a new research evaluation method by comparing the publication graphs of different research groups. For example, we can compare the publication graph for an intended research group with that of the most well-known researchers. We suggest reasonable metrics to compare the similarity, connectedness and gaps between the research groups' publication outputs.
- We analyse publications on information security from South Korea, and evaluate them using our proposed method (Section 4). Korea, China and Japan are distant, both geographically and linguistically, from established research groups in North America and Western Europe. There has been constant concern that research groups may become isolated and introverted, working on topics in which research leaders have lost interest. We show quantitative evidence of the problem – effectively providing an early warning system for possible diminishing research impact. These effects were measured by comparing the relationship graphs constructed for Korean researchers' publication data with those defined by the most well-known researchers' publication data during the period 2004–2009. The experimental result shows that there is still a gap between the two research groups' publication data in 2009 at similar level to 2004 (Section 5).

2. Related work

The use of bibliometric indicators in research evaluation emerged in the 1960s and 1970s (Leydesdorff, 2005), and is in widespread use today due to the proliferation of relevant databases. These indicators provide useful output measures of activity and performance in scientific research and have become standard tools for research evaluation (Almeida, Pais, & Formosinho, 2009). However, these indicators can vary widely depending on the bibliometric database used. Our work is mainly motivated by the limitation of the bibliometric database constructed for general-purpose since this database has some limitations. There have been studies of scientific output across various subject areas, however, we are mainly interested here in those related to computer science (Meyer, Choppy, Staunstrup, & Leeuwen, 2009; Wainer, Xavier, & Bezerra, 2009). In particular, Meyer et al. (2009) introduced several criteria and principles to evaluate research quality in computer science. Interestingly, for us, they warned that simple publication counts, weighted or not, must not be used as the indicators of research value.

An alternative approach is to analyse researchers' social networks such as citation networks and co-authorship networks. A citation network consists of a set of nodes representing papers and a set of edges, where an edge from node x to node y represents a citation from paper x to paper y . The citation graph can be used to evaluate authors, papers and journal/proceedings. For this purpose, the PageRank algorithm (Brin & Page, 1998) has been most popularly used to rank the nodes in citation graphs (Sidiropoulos & Manolopoulos, 2005) since a citation graph could be defined as a directed graph. In addition, citation graphs provide useful statistical information. For example, a set of journal/proceedings (or authors) can be grouped by using clustering algorithms (Biryukov & Dong, 2010; Zaane, Chen, & Goebel, 2009; Zhou, Ji, Zha, & Giles, 2006). Another branch of previous work focuses on finding the most important journal/proceedings (or authors) in the sense of citation (Nerur, Sikora, Mangalaraj, & Balijepally, 2005; Yan & Lee, 2007). Zhuang, Elmacioglu, Lee, and Giles (2007) proposed some heuristics to assess the quality of conferences from analysing the characteristics of program committee members in the conferences, based on the assumption that the quality of conferences are highly correlated with the program committee members of the conferences.

Co-authorship networks are an important class of social networks and have been studied extensively to analyse uncover patterns, motivation, and structure of scientific collaboration. Morris (2005) proposed a model to monitor the birth and development of a scientific speciality using a collection of journal papers. Kandylas, Upham, and Ungar (2008) intensively analyzed how communities of researchers are evolved over time and designed the model for community growth. Zhou, Councill, Zha, and Giles (2007) also proposed a method to discover the temporal trends of researchers by constructing their co-authorship graphs over time and comparing their communities. Lee (2008) analysed the research trends in the information security field using “co-word analysis”.

Our work is an extension and modification of these approaches, focusing on the relationship between, and topological positions of researchers’ publications. Furthermore we develop a model and metrics to analyse the research group’s connectedness to the research mainstream.

3. The proposed method

3.1. Construction of publication graph

Given a set of researchers \mathbf{R} , we construct the publication graph $G_{\mathbf{R}}^{\mathbf{J}}$ by taking the following steps:

- 1 For each researcher $a \in \mathbf{R}$, collect the a s publication outputs within a time window (e.g. within 2009).
- 2 Generate the bipartite graph $G_{\mathbf{R}}$ with these collected publication data, whose nodes are divided into a set of authors \mathbf{A} and a set of journal/proceedings \mathbf{J} and an edge (a, j) means that the author a published a paper in the journal (or proceeding) j for $a \in \mathbf{A}$ and $j \in \mathbf{J}$.
- 3 Construct the \mathbf{J} -projected graph $G_{\mathbf{R}}^{\mathbf{J}}$ compressed by \mathbf{J} -projection, which is a well-known technique so-called one-mode projecting to show the relations among a particular set of nodes only (Guillaume, 2005; Zhou, Ren, Medo, & Zhang, 2007). The \mathbf{J} -projection means a network containing only nodes in \mathbf{J} , where two nodes are connected when they have at least one common author. We construct $G_{\mathbf{R}}^{\mathbf{J}}$ as a *weighted* graph. The weight of an edge (x, y) is assigned to be inversely proportional to the number of the shared authors between two journal/proceedings. We note that the edge weight in the \mathbf{J} -projected graph represents a distance measure between journal/proceedings rather than similarity. We use $w(x, y)$ to denote the weight of the edge (x, y) .

A \mathbf{J} -projected graph represents the relationship between journal/proceedings that have mainly been published in a research group. We name it by publication graph. In fact, this graph gives the information about not only a set of popular journal/proceedings for the research group but also the relative importance of them by computing their centrality metric values such as *degree*, *closeness* (Newman, 2001) and *betweenness* (Brandes, 2001). A research group can be generally applied to a set of researchers in a department (e.g. Computer Laboratory in University of Cambridge), university, the program committee of a conference, and a research field (e.g. information security). We do not restrict our attention to specific applications.

For an example, suppose that we have two researchers: $\mathbf{R} = \{a_1, a_2\}$. When the researchers a_1 and a_2 published their papers in the journals $\{j_1, j_2, j_3\}$ and $\{j_2, j_3, j_4\}$, respectively, then we have a bipartite graph as shown in Fig. 1(a). From the bipartite graph, we can construct the \mathbf{J} -projected graph $G_{\mathbf{R}}^{\mathbf{J}}$ as in Fig. 1(b). In this example, we can observe that the degree values of j_2 and j_3 are greater than those of j_1 and j_4 . Also, the weight of the edge (j_2, j_3) is 0.5 since two researchers a_1 and a_2 commonly published their papers in both journals j_2 and j_3 . We will discuss how to analyse the constructed publication graphs in the next section.

3.2. Comparison of publication graphs

Given several publication graphs, we can analyse the similarity, connectedness and gaps between the publication graphs which imply the scientific contributions of the research groups. In order to measure them quantitatively, we propose the three functions as: the “fraction of the overlapping nodes or interactions” for the similarity; the “network centrality of the overlapping nodes” for the connectedness; and the “distance between the graphs” for the gap.

We first define the proposed metrics in this section and will test the effectiveness of these metrics through an empirical analysis based on real datasets in Section 4.

3.2.1. Fraction of overlapping nodes/interactions

To measure the similarity between the publication graphs, we can count the number of the shared nodes or edges: it is simple to implement and computationally efficient.

The first metric is computed as the “ratio of the overlapping nodes” between publication graphs. Formally, given k \mathbf{J} -projected graphs $G_1^{\mathbf{J}} = (V_1, E_1), \dots, G_k^{\mathbf{J}} = (V_k, E_k)$, we use the symbols V_U and V_A to represent the superset of nodes in the graphs $(\bigcup_{i=1}^k V_i)$ and the set of the shared nodes between the graphs $(\bigcap_{i=1}^k V_i)$, respectively. With these symbols, we define the “ratio of the overlapping nodes” between publication graphs as follows:

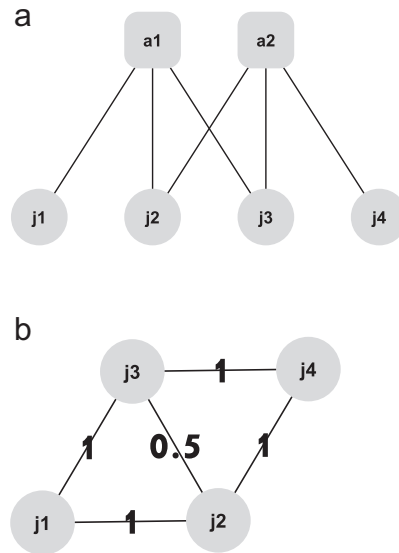


Fig. 1. An example of journal/proceeding graph construction. Given the publication results $\{j_1, j_2, j_3\}$ and $\{j_2, j_3, j_4\}$, respectively, by two researchers a_1 and a_2 , we have (a) the bipartite graph and construct and (b) the \mathbf{J} -projected graph G_k^J . In (b) an edge (x, y) is labelled with its weight $w(x, y)$. For example, the weight of the edge (j_2, j_3) is 0.5 which indicates that the two journals j_2 and j_3 have two common authors, a_1 and a_2 .

Definition ratio of the overlapping nodes – Given k \mathbf{J} -projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we then define the “ratio of the overlapping nodes” as

$$\text{Ratio}_V(G_1^J, \dots, G_k^J) = \frac{|V_A|}{|V_U|} \tag{1}$$

The “ratio of the overlapping nodes” between publication graphs can be interpreted as a measurement of the size of the set of journal/proceedings commonly published by the research groups used in constructing the given \mathbf{J} -projected graphs. This means that a set of research groups with a high value of this metric have many shared journals/proceeding for publication. However, this metric does not take into account the topological positions of nodes and the weights of edges. In other words, with this naive metric only, we cannot measure the relative importance of a common journal/proceeding node in a publication graph compared with another common journal/proceeding node since its relations with the other journal/proceeding nodes are not considered. An example in Fig. 2 shows a limitation of this metric. Suppose that we have two different subgraphs denoted by G_1^J (subgraph with grey coloured nodes) and G_2^J (subgraph with square shaped nodes). In this case, grey and square nodes are represented as the overlapping nodes of two subgraphs $G_1^J \cap G_2^J$, which is located in the centers of both examples. The “ratio of the overlapping nodes” cannot explain the difference between Fig. 2(a) and (b). The relationships between the shared node (grey and square) and the other nodes are ignored in computing the metric value. This is not satisfactory; we expect that the \mathbf{J} -projected graphs, G_1^J and G_2^J in Fig. 2(b) is more highly correlated since the shared node in Fig. 2(b) plays a more important role than the shared node in Fig. 2(a).

The second metric is computed as the ratio of the shared interactions between publication graphs. This metric illustrates how similar two network structures are (e.g. interaction patterns). At first glance, it seems that we need to consider inter-actions of non-existing edges in publication graphs since non-existing edges also give some information about relationship between nodes in general. However, we would not consider non-existing edges for our purpose since the publication graphs are generally too sparse. That is, non-existing edges will dominate the information about existing edges.

Formally, given k \mathbf{J} -projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we use the symbols E_U and E_A to represent the superset of nodes in the graphs $(\bigcup_{i=1}^k E_i)$ and the set of the shared nodes between the graphs $(\bigcap_{i=1}^k E_i)$, respectively. With these symbols, we define the “ratio of the overlapping interactions” between publication graphs as follows:

Definition ratio of the overlapping interactions – Given k \mathbf{J} -projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we then define the “ratio of the overlapping nodes” as

$$\text{Ratio}_E(G_1^J, \dots, G_k^J) = \frac{|E_A|}{|E_U|} \tag{2}$$

The “ratio of the overlapping interactions” between \mathbf{J} -projected graphs can be interpreted as a measurement of the size of the set of authors publishing their work in the same pair of journal/proceedings. Among a set of research groups with a high value of this metric, there are many researchers who used the same pair of journal/proceedings. However, the metric value is not proper if the fraction of overlapping edges is small. We would not recommend using this metric to explain the similarity between \mathbf{J} -projected graphs in such situations.

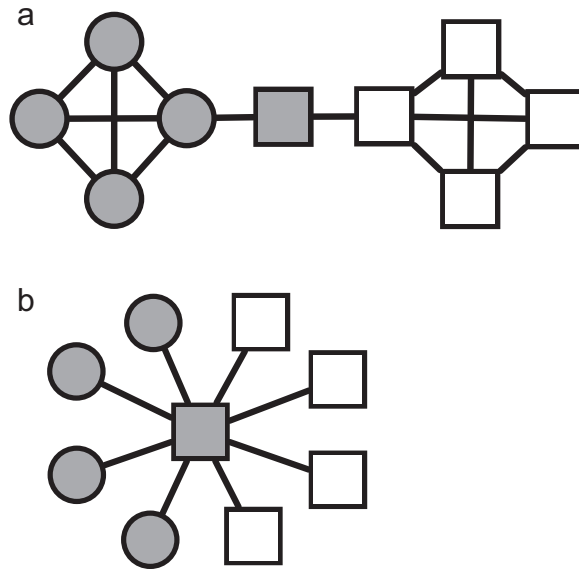


Fig. 2. Two examples of journal/proceeding graphs to explain the limitation of the “ratio of the overlapping nodes”. In each graph, grey and square nodes represent G_1^j and G_2^j , respectively. The “ratio of the overlapping nodes” values of both examples are the same ($|G_1^j \cap G_2^j| = 1$) although their structures are totally different. (a) When an outlier is shared and (b) when a hub is shared.

3.2.2. Network centrality of the overlapping nodes

A reasonable metric should measure not only the number of the shared journal/proceedings between **J**-projected graphs but also their relative importance or relationships with the other journal/proceedings. For this purpose, we extend the naive metrics of the “fraction of the overlapping nodes” in Section 3.2.1 to consider the network position of the shared nodes. We find it intuitive to calculate the network centrality values of the shared journal/proceedings between publication graphs. Network centrality can provide relative measures that can be used to compare nodes against each other based on network topology. Formally, given a graph $G=(V, E)$, we use the standard definition of the *degree*, *closeness* and *betweenness* centrality values of a node u as follows (Wasserman & Faust, 1994):

- **Degree centrality** measures the number of direct connections to other nodes. This is calculated for a node u as the ratio of the number of node u 's neighbour to the total number of all other nodes in the network:

$$C_{deg}(u) = \frac{deg(u)}{|V| - 1} \tag{3}$$

where $deg(u)$ is the number of edges of node u .

In a **J**-projected graph, the *degree* centrality can be interpreted as a measure of how many the other journal/proceedings in which the corresponding authors' work is published exist.

- **Closeness centrality** measures how near nodes are to each other or in practical terms how quickly a node can communicate with all other nodes in a network. This is calculated for a node u as the average shortest path length to all other nodes in the network:

$$C_{clo}(u) = \frac{1}{|V| - 1} \sum_{v \neq u \in V} dist(u, v) \tag{4}$$

where $dist(u, v)$ is the length of the shortest path from node u to node v .

We here use the conventional definition of the length of path $p = \langle v_0, v_1, \dots, v_k \rangle$ is the sum of the weights of its constituent edges:

$$length(p) = \sum_{i=1}^k w(v_{i-1}, v_i) \tag{5}$$

where $p = \langle v_0, v_1, \dots, v_k \rangle$.

In a **J**-projected graph, the *closeness* centrality can be interpreted as a measure of how close a journal/proceeding is on average to all other journal/proceedings. Then, journal/proceedings with high *closeness* values could be viewed as those which can be more highly selected by researchers than others.

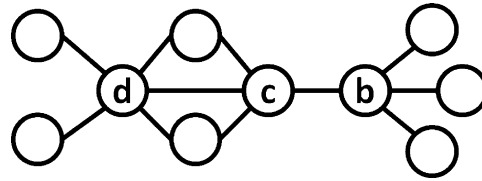


Fig. 3. The characteristics of network centrality. In this network, *d* has higher *degree* centrality than *c* since *d* has five neighbours while *c* has higher *closeness* centrality than *d*. We note that *d* is located at the periphery of the network compared to *c*. Interestingly, *v* has the highest *betweenness* centrality. We can see that *v* plays a ‘bridge’ role for the rightmost nodes.

• **Betweenness centrality** measures the paths that pass through a node and can be thought of as the proportional flow of data through each node. This is calculated for a node *u* as the proportional number of shortest paths between all node pairs in the network, that pass through *u*:

$$C_{bet}(u) = \sum_{s \neq u, t \neq u \in V} \frac{\sigma_{s,t}(u)}{\sigma_{s,t}} \tag{6}$$

where $\sigma_{s,t}$ is the total number of shortest paths from source node *s* to destination node *t*, and $\sigma_{s,t}(u)$ is the number of shortest paths from source node *s* to destination node *t*, which actually pass through node *u*.

In a **J**-projected graph, the *betweenness* centrality can be interpreted as a measure of how many a journal/proceeding is involved between other journal/proceedings in the network. This measure favours nodes that join communities (dense sub-networks), rather than nodes that lie inside a community. So we can identify a bridge journal/proceeding between the different fields in the sense of topic or quality.

In Fig. 3, for example, the nodes *d*, *c*, and *b* illustrate the characteristics of these network centrality metrics. These nodes have the highest *degree*, *closeness* and *betweenness* centrality, respectively.

The network centrality value of a journal/proceeding illustrates its relative importance in a publication graph. Since our main interest here is analysing the relative importance of the shared journal/proceedings compared with the overall publication graphs, we extend the analysis of a node’s centrality into the analysis of the shared nodes’ centrality. When the graphs are strongly correlated, these values will be significantly greater than those of weakly correlated graphs. We define these as follows:

Network definition centrality of the overlapping nodes – Given *k* **J**-projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we define the “network centrality of the overlapping nodes” as follows:

$$C_{type}(G_1^J, \dots, G_k^J) = \frac{\sum_{v \in V_A} C_{type}(v)}{\sum_{u \in V_U} C_{type}(u)} \tag{7}$$

The notation, C_{type} , is used to denote the network centrality values such as *degree* (C_{deg}), *closeness* (Newman, 2001) (C_{clo}), and *betweenness* (Brandes, 2001) (C_{bet}).

3.2.3. Network distance between publication graphs

To measure the diversity of graphs, we also consider the network distance between the nodes in the **J**-projected graphs. For publication networks in our study, the network distance between two journal/proceedings means the differences of their quality or topic for publication under the assumption that an author publishes her work in journal/proceedings at a similar level of quality or topic.

The first metric is computed as the total sum of distances from the overlapping nodes among the **J**-projected graphs and the other remaining nodes in the graphs. We define this as follows:

Minimum definition distance to the overlapping nodes – Given *k* **J**-projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we define the “the minimum distance to the overlapping nodes” as follows:

$$Dist(G_1^J, \dots, G_k^J) = \begin{cases} \frac{1}{|V_U|} \cdot \sum_{u \in V_U} \min_{v \in V_A} \{dist(u, v)\}, & \text{if } V_A \neq \emptyset, \\ \infty, & \text{Otherwise.} \end{cases} \tag{8}$$

This metric measures how close all journal/proceedings in the network are to their shared journal/proceedings. We can explain how much closer a node in each graph is, to the overlapping nodes among the graphs on average, using this value. This value will be exactly 0 if and only if V_A is the same as V_U . With this metric, we can explain how much similar the journal/proceedings are, which are selected by research groups, respectively.

Table 1

Summary of publication data for Korean researchers and well-known researchers: R_K and R_W represent a set of Korean and well-known researchers, respectively. Also, n_r , n_p and n_j represent the number of the authors, their publications and the ratio of journals in the publications, respectively. Interestingly, in R_K , n_j increases over time in spite of the decrease in n_p . It is clearly distinguished from the publication trend of R_W .

Group		2004	2005	2006	2007	2008	2009
R_K	n_r	17	20	25	25	30	32
	n_p	51	96	97	78	71	63
	n_j	1	10	14	23	27	33
R_W	n_r	88	78	95	109	135	119
	n_p	368	364	486	504	651	433
	n_j	89	92	131	106	132	115

For some applications, it is also important to analyse how far unrelated journal/proceedings are included in their publication graph. In other words, we need to measure the distance between most unrelated journal/proceedings. Basically, this property is closely related to the *network diameter* of a graph. *Network diameter* is the maximum distance between nodes in the network (Per and Frank, 1995). Therefore we need to measure how many the *network diameter* of the union graph $G_U^J = (V_U, E_U)$ is increased after combining all the **J**-projected graphs where $E_U = \bigcup_{i=1}^k E_i$.

We compute the increase in the diameter of the union graph G_U^J as follows:

Increase definition in the diameter – Given k **J**-projected graphs $G_1^J = (V_1, E_1), \dots, G_k^J = (V_k, E_k)$, we define the “increase in the diameter” of the union graph as follows:

$$\Delta \text{Diam}(G_1^J, \dots, G_k^J) = \frac{1}{k} \cdot \sum_{i=1}^k d_i \tag{9}$$

where $d_i = \max\{\text{diameter}(G_U^J) - \text{diameter}(G_i^J), 0\}$ for $1 \leq i \leq k$.

4. The case study of Korean security research

We demonstrate the practicality of our method by comparing the scientific contributions by Korean researchers with those by the most well-known researchers for information security in 2004–2009.

We use sample sets since it is infeasible to collect all publications relating to security. To obtain a reasonable sample for researchers each year from 2004 to 2009, we select all Korean researchers from the program committee members of prestigious conferences relating to security that were held in South Korea. Two conferences, “International Workshop on Information Security Applications” (WISA)¹ and “International Conference on Information Security and Cryptology” (ICISC),² are selected by considering their relatively large scales and long history compared to other conferences. We assume that for each year the program committee members of these conferences are representative researchers to show the research trend of information security in Korea. Let R_K be a set of all Korean program committee members of these conferences. In a similar manner, we obtain a reasonable sample set of well-known researchers by using the top international conferences for security, “IEEE Symposium on Security and Privacy” (SP), “ACM Conference on Computer and Communications Security” (CCS) and “USENIX Security Symposium” (USENIX). These conferences are selected under the conference ranking of well-known web sites (Computer Science Conference Ranking, 2009; Computer Security Conference Ranking and Statistic, 2009; Security Conference Ranking, 2007). Let R_W be a set of all program committee members of these conferences. We collected R_K s and R_W s publication results from 2004 to 2009, respectively. For simplicity, we only consider the bibliographic information indexed by the publication information indexed by the Digital Bibliography & Library Project (DBLP) (Ley, 2002) under the assumption that this database provides the most bibliographic information on major computer science journals and proceedings. Table 1 shows the number of the authors and their publications, respectively, in the collected datasets.

We compare the number of journal/proceedings between Korean researchers and well-known researchers. The results are shown in Fig. 4. Fig. 4(a) shows the average number of publications per researcher and Fig. 4(b) shows the ratio of journals only in the total publications. Interestingly, the ratio of journals for the Korean researchers has dramatically increased from 2004 to 2009 while the total number of publications has steadily decreased from 2005 to 2009. We note that the ratio of journal articles for well-known researchers is relatively stable compared with that of the Korean researchers. We are curious whether this dramatic increase of journal publication in the last few years means a significant reduction of the gap between the Korean research and the research mainstream. Our work is originally motivated by this question.

With the collected publication data, we construct the **J**-projected graphs for each research group per year from 2004 to 2009. We use G_K^J and G_W^J to denote the Korean and the well-known researchers’ graphs, respectively. Without loss of generality, in the case of a disconnected graph, we eliminate the small disconnected components and focus only on the

¹ <http://www.wisa.or.kr/>.

² <http://www.icisc.org/>.

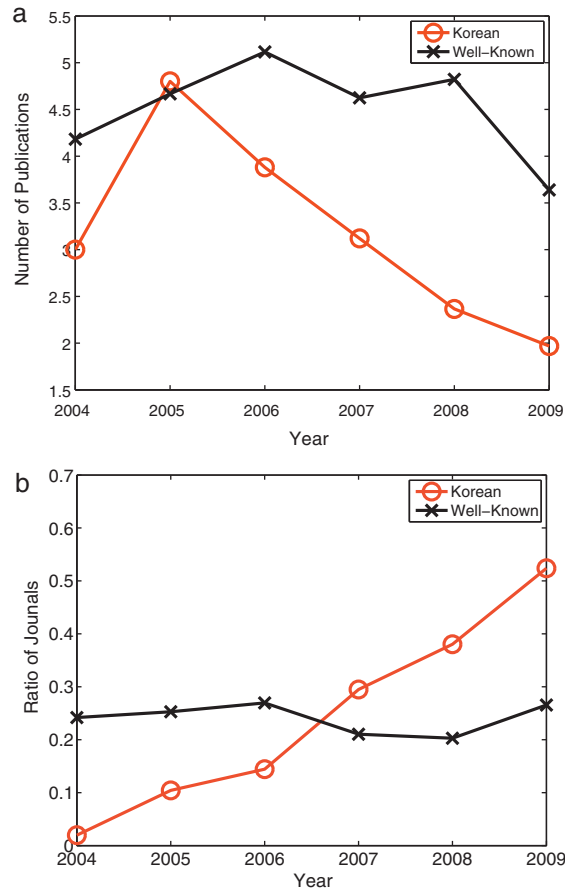


Fig. 4. The publication trends for Korean researchers (red circle) and well-known researchers (black cross) between 2004 and 2009. For Korean researchers, the average number of publications decreases over time from 2005 while the ratio of journal articles in the total publications significantly increases. The publications of well-known researchers resulted in only relatively small changes compared with those of Korean researchers. (a) Average number of publications. (b) Ratio of journals in publications. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

giant component since a small disconnected component can be interpreted as an outlier. We merge G_K^I and G_W^I per year to represent their relationship visually as shown in Fig. 5. The layout was produced and scaled by Cytoscape (Shannon et al., 2003), using an edge-weighted spring-embedded model, meaning that journal/proceedings sharing more neighbours are closer on the display.

Basically, it is not straightforward to interpret the meaning of the proposed metric values; to evaluate the relative gap, we additionally analysed the publication graphs of two research groups, which consist of the program committee members in SP and CCS, respectively. We expect that these publication graphs are more likely to be highly connected to each other under the assumption the committee members in these conferences are sufficiently qualified have similar research interest. We use G_S^I and G_C^I to denote the SP and the CCS researchers' graphs, respectively. We merge G_S^I and G_C^I per year to represent their relationship visually as shown in Fig. 6.

From the graphical presentation of publication graphs in Figs. 5 and 6, we can observe that G_K^I and G_W^I share only a small number of common nodes and edges compared with G_S^I and G_C^I . In order to analyse the relationships of these publication graphs quantitatively, we compute the proposed six metrics in Section 3.2. The results are shown in Table 2.

From these results, we first discuss the results of the "fraction of the overlapping nodes" (**Ratio_v**) and the "fraction of the overlapping interactions" (**Ratio_E**). For the "ratio of the overlapping nodes", we can see that the values between G_K^I and G_W^I are under 0.1 from 2004 to 2009 while the values between G_S^I and G_C^I are over 0.27. In other words, the nodes in G_S^I and G_C^I are highly overlapped compared with those in G_K^I and G_W^I . For the "ratio of the overlapping interactions", the computed metric values show significant differences between both cases. In particular, in the case of G_K^I and G_W^I , the "fraction of the overlapping interactions" is extremely small. In 2008 and 2009, the computed metric values are getting close to zero.

For the "network centrality of the overlapping nodes", we can see that C_{deg} , C_{clo} , and C_{bet} values between G_S^I and G_C^I are significantly greater than those between G_K^I and G_W^I ; the values are below 0.2 in C_{deg} , below 0.1 in C_{clo} , and below 0.3 in

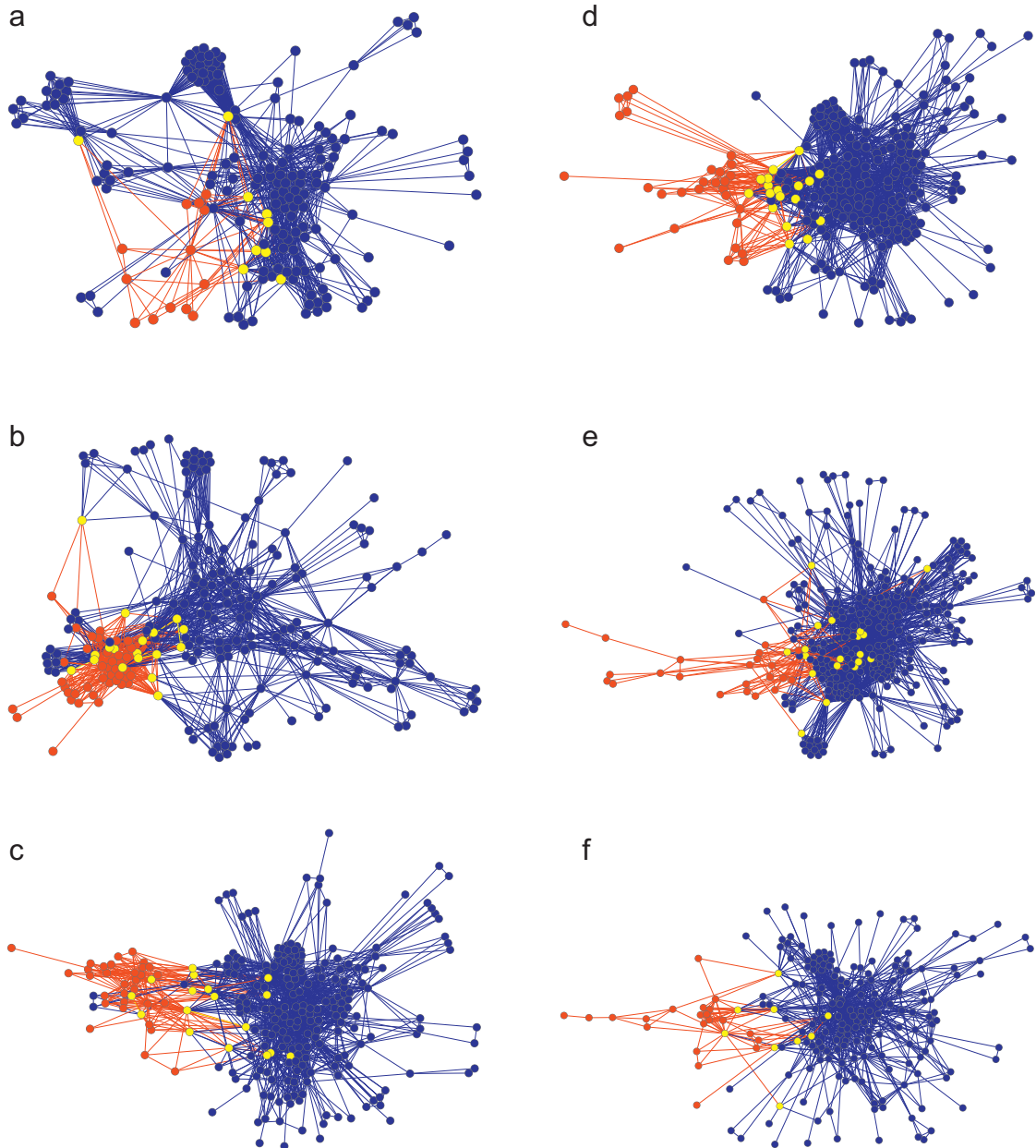


Fig. 5. G_K^J and G_W^J per year from 2004 to 2009. The red and blue nodes/edges represent G_K^J and by G_W^J , respectively. The yellow nodes/edges represent the commonly shared nodes/edges between them. The graphical presentation of their publication graphs show that the structure of G_K^J is easily identified from that of G_W^J . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

C_{bet} for the comparison of G_K^J and G_W^J while the values are above 0.4 in C_{deg} , above 0.3 in C_{clo} , and above 0.7 in C_{bet} for the comparison of G_S^J and G_C^J . Interestingly, when the publication graphs are highly connected to each other, the values of C_{bet} are close to 1. In 2009, we can also see a meaningful difference between them.

For the “minimum distance to the overlapping nodes”, we can see that the values between G_K^J and G_W^J are significantly different from those between G_S^J and G_C^J ; the distance values are always above 1 for the comparison of G_K^J and G_W^J while those values are below 1 for the comparison of G_S^J and G_C^J . In 2009, the gap between them has increased considerably.

For the “increase in the diameter”, we can see that the values between G_K^J and G_W^J are different from those between G_S^J and G_C^J except 2005. In 2008 and 2009, the gap between them has not decreased but rather increased from 1 to 2.

In order to test the effectiveness of the network analysis with publication graphs, we identify a set of relatively important journal/proceedings for Korean and well-known researchers by computing the conventional *closeness* (Newman, 2001)

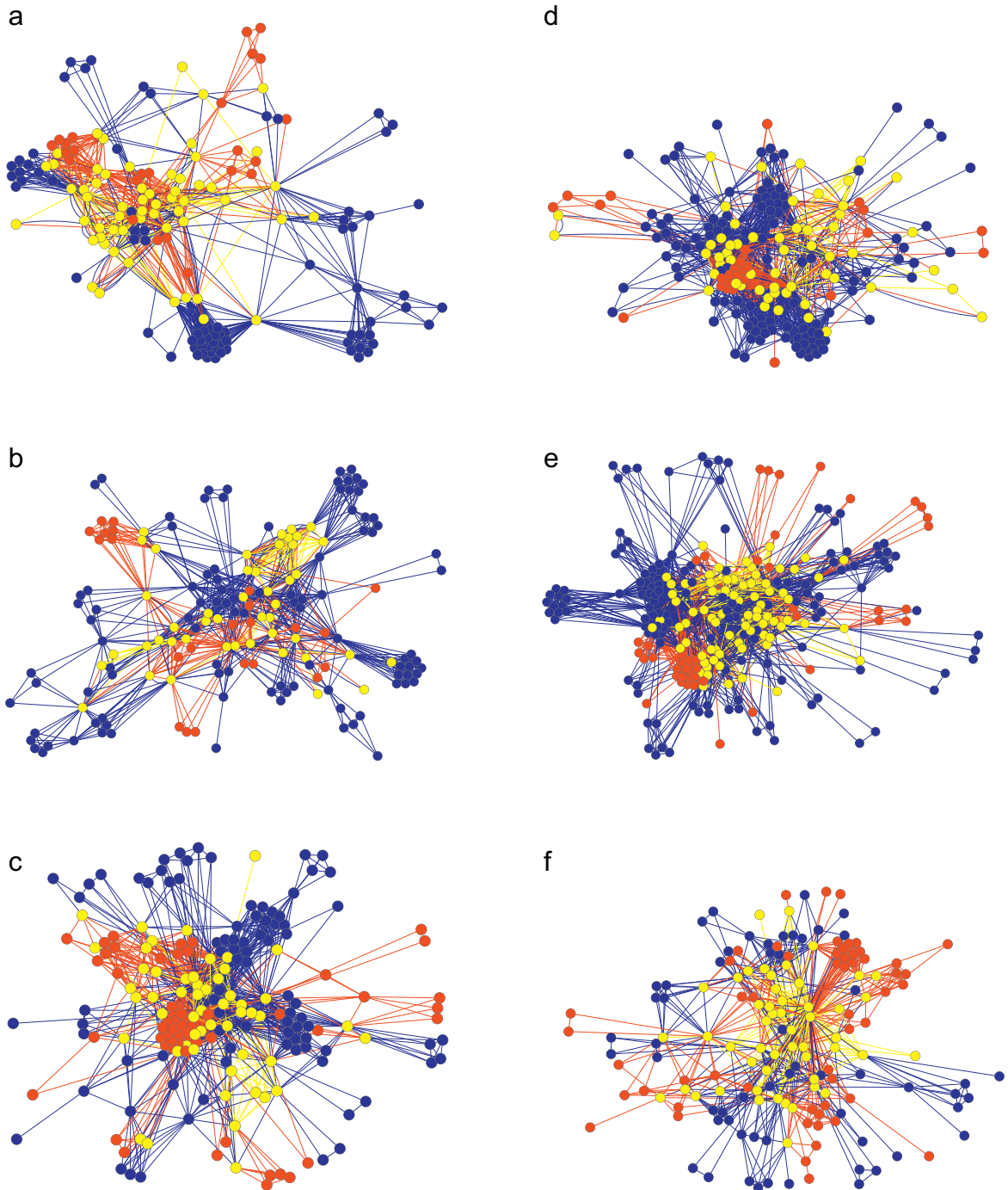


Fig. 6. G_S^J and G_C^J per year from 2004 to 2009. The red and blue nodes/edges represent G_S^J and by G_C^J , respectively. The yellow nodes/edges represent the commonly shared nodes/edges between them. Note that two communities G_S^J and G_C^J are not apparent from their graphical presentation compared to Fig. 5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

values of the nodes in G_K^J and G_W^J , respectively and compare them with the “A” ranked journal/proceedings in the ERA list. The results are shown in Table 3. As we expected, the number of highly estimated journal/proceedings in the well-known researchers’ publication graphs is greater than that in the Korean researchers’ graphs. In other words, in well-known researchers’ publication graphs, journal/proceedings with high *closeness* values could be regarded as highly estimated journal/proceedings.

Table 2

Comparison of “ G_K^J and G_W^J ” and “ G_S^J and G_C^J ” per year from 2004 to 2009. We analyse the “ratio of the overlapping nodes” (**Ratio_V**), the “ratio of the overlapping interactions” (**Ratio_E**), the “network centrality of the overlapping nodes” (**C_{deg}**, **C_{clo}**, and **C_{bet}**), the “minimum distance to the overlapping nodes” (**Dist**), and the “increase in the diameter” (**ΔDiam**), respectively (refer to Section 3.2).

		2004	2005	2006	2007	2008	2009
Ratio_V	G_K^J, G_W^J	0.050	0.082	0.069	0.076	0.065	0.045
	G_S^J, G_C^J	0.367	0.274	0.276	0.290	0.289	0.315
Ratio_E	G_K^J, G_W^J	0.002	0.006	0.001	0.004	0.000	0.000
	G_S^J, G_C^J	0.090	0.076	0.059	0.043	0.104	0.144
C_{deg}	G_K^J, G_W^J	0.097	0.198	0.099	0.122	0.133	0.074
	G_S^J, G_C^J	0.466	0.455	0.481	0.435	0.481	0.557
C_{clo}	G_K^J, G_W^J	0.056	0.094	0.071	0.079	0.070	0.047
	G_S^J, G_C^J	0.411	0.315	0.314	0.321	0.330	0.366
C_{bet}	G_K^J, G_W^J	0.114	0.239	0.134	0.176	0.165	0.125
	G_S^J, G_C^J	0.914	0.711	0.905	0.903	0.894	0.946
Dist	G_K^J, G_W^J	1.243	1.287	1.235	1.245	1.133	1.541
	G_S^J, G_C^J	0.677	0.821	0.750	0.703	0.726	0.706
ΔDiam	G_K^J, G_W^J	1.000	1.000	1.000	1.000	2.000	2.000
	G_S^J, G_C^J	0.000	1.000	0.000	0.000	0.500	0.000

Table 3

Top 10 high closeness journal/proceedings from 2004 to 2009 (in the acronyms, (J) means a journal article. We provide Appendix A to introduce the full name of journal/proceedings): G_K^J and G_W^J represent the Korean and well-known researchers’ publication graphs, respectively. The journal/proceedings are arranged as a monotonically decreasing closeness centrality values. We identified the “A” ranked journal/proceedings in the ERA list released on 9 February 2010 Excellence in Research for Australia (ERA) (2010) in a bold font. We can see the number of highly estimated journal/proceedings in the well-known researchers’ publication graphs is greater than that in the Korean researchers’ graphs. We identified the metric values of G_K^J and G_W^J in a bold font.

	2004	2005	2006	2007	2008	2009
G_K^J	ICCSA	EUC	ICCSA	HCI	ComCom(J)	IEICET(J)
	WISA	TrustBus	IEICET(J)	IEICET(J)	IEICET(J)	CT-RSA
	ICOIN	KES	APWeb	ICCSA	TKDE(J)	Crypt(J)
	ISC	HSI	WISA	ACNS	JUCS(J)	TC(J)
	ICICS	ICCSA	KES	ATC	PerCom	CSI(J)
	EuroPKI	ICOIN	EUC	AWIC	ISI	IPL(J)
	AIS	AMC(J)	OTM	CT-RSA	MUE	ACIS
	CIS	IEICET(J)	PWC	WISA	UIC	JUCS(J)
	PCM	CIS	ICCS	APWeb	FGCN	CISIS
	SenSys	ICNC	EuroPKI	PAISI	CSI(J)	ASIACCS
G_W^J	CCS	SP	CCS	ASIACCS	CCS	CCS
	USENIX	ACNS	ACSAC	CCS	ACSAC	NDSS
	SP	TDSC(J)	TISSEC(J)	ESORICS	NDSS	ESORICS
	NDSS	CCS	NDSS	SP	TISSEC(J)	ACSAC
	ACNS	NDSS	SP	WPES	USENIX	SP
	CACM(J)	ESORICS	ESORICS	IJIS(J)	SP	RAID
	TCC	WORM	DGO	FC	ESORICS	CRYPTO
	SACMAT	ACSAC	ASIACCS	VLDB	RAID	SP(J)
	ICICS	ISC	RAID	CACM(J)	WPES	SACMAT
	JCS(J)	SOSP	SACMAT	ACSAC	JCS(J)	CSF

In fact, there have been few common nodes with high closeness between G_K^J and G_W^J over time although most of top 10 high closeness nodes have been greatly changed in the Korean researchers’ publication graphs from 2004 to 2009. We note that the high closeness nodes in G_W^J are fairly stable unlike G_K^J . In particular, in G_K^J , many journals have been newly added from 2008.

5. Discussion

Our work is primarily intended to demonstrate how to compare publication patterns between different research groups. We do not claim that our method is a full replacement for widely used conventional research evaluation methods but a complementary method that could be used together. As the research fields become more specific, we imagine that the proposed research evaluation methods would be useful since it is difficult to use a database universally for research evaluation.

As a case study, we analysed the scientific production of Korean researchers in information security from 2004 to 2009. An interesting observation is that the portion of journal publication to proceeding publication has been greatly increased in Korea (see Fig. 4) although the overall volume of the publication have steadily decreased from 2005. We strongly believe that

Korean researchers' preferred journal/proceedings have been dynamically changed in response to this research evaluation policy (see Table 3) because scientific production in Korea is generally evaluated by counting SCI publication. In addition, the comparative analysis of their publication graphs provides much information about their main journals and proceedings. In Fig. 5, we found that to date the Korean and well-known research groups have shared a small fraction of journal/proceedings compared with the SP and CCS research groups which are highly connected to each other. This means that the journals and proceedings in the information security field which the Korean researchers published still deviates from the research mainstream in 2009 at similar level to 2004 (see Table 2) although the number of journal publication has steadily increased. In the information security field, the increase of journal publications does not significantly affect the relationship with the research mainstream; we note that the gap between the Korean and the well-known researchers' publication graphs is the smallest in 2005. From the characteristics of 2005 in Fig. 4, we expect that the increases in the conference proceeding publications or the total number of publications are more important than the journal publication in order to improve a research group's connectedness to the research mainstream.

The other important issue is that conference (or journal) selection is strongly related to geographical and political factors in the real world. Our analysis has some limitation since we do not consider these factors.

6. Conclusion

We propose a method to analyse the publication trends of research groups. We can construct relationship graphs from publication outputs and then analyse the properties of the graphs. Our goal is to analyse not the publication and citation counts but the group's connectedness to the research mainstream, both statically and over time.

As a case study, we analysed the scientific output for Korean researchers compared with that of the most well-known research group in information security over the years from 2004 to 2009. Our research shows that there is a gap between the two research groups' publication trends in 2009 that persists at similar level since 2004.

Our approach has potential. We can measure and visualize the similarity/gap among the research groups of interest by using the proposed network analysis. For example, we can identify the weak and strong points of each country for a given research subject by performing inter-country comparisons of the research field, and hence suggest where we need to invest more, to balance a growing research. Also, we can measure the similarity, connectedness and gaps in quality or research topics among universities or departments.

Moreover we do not restrict our attention to the publication trends of research groups but will extend it to the other applications for social network services. For example, we plan to detect onset of extremism in sub groups of bloggers or communities in a social network service such as Twitter.

Appendix A. List of journal/proceedings

Acronym	Journals/conferences full name
ACIS	International Association for Computer and Information Science
ACNS	Applied Cryptography and Network Security
ACSAC	Annual Computer Security Applications Conference
AIS	Artificial Intelligence and Simulation
AMC(J)	Applied Mathematics and Computation
APWeb	Asia-Pacific Web Conference
ASIACCS	ACM Symposium on Information, Computer and Communications Security
ATC	Autonomic and Trusted Computing
AWIC	Atlantic Web Intelligence Conference
CACM(J)	Communications of the ACM
CCS	Conference on Computer and Communications Security
CIS	Computational Intelligence and Security
CISIS	Complex, Intelligent and Software Intensive Systems
ComCom(J)	Computer Communications
Crypt(J)	Cryptologia
CRYPTO	International Cryptology Conference
CSF	IEEE Computer Security Foundations Symposium / Workshop
CSI(J)	Computer Standards & Interfaces
CT-RSA	The Cryptographer's Track at RSA Conference (CT-RSA)
DGO	International Conference on Digital Government Research
ESORICS	European Symposium on Research in Computer Security
EUC	Embedded and Ubiquitous Computing
EuroPKI	European Public Key Infrastructure Workshop
FC	Financial Cryptography
FGCN	Future Generation Communication and Networking
HCI	Human-Computer Interaction
HSI	International Conference on Human Society at Internet
ICCS	International Conference on Computational Science
ICCSA	International Conference on Computational Science and Its Applications
ICICS	International Conference on Information and Communication Security

Acronym	Journals/conferences full name
ICNC	International Conference on Natural Computation
ICOIN	International Conference on Information Networking
IEICET(J)	IEICE Transactions
IJIS(J)	International Journal of Information Security
IPL(J)	Information Processing Letters
ISC	Information Security Conference/Workshop
ISI	Intelligence and Security Informatics
JCS(J)	Journal of Computer Security
JUCS(J)	The Journal of Universal Computer Science
KES	Knowledge-Based Intelligent Information & Engineering Systems
NDSS	Network and Distributed System Security Symposium
OTM	OTM Conferences/Workshops
PAISI	Pacific Asia Workshop on Intelligence and Security Informatics
PCM	Pacific-Rim Conference on Multimedia
PerCom	International Conference on Pervasive Computing and Communications
PWC	IFIP WG6.8 Publications
RAID	Recent Advances in Intrusion Detection
SACMAT	ACM Symposium on Access Control Models and Technologies
SenSys	International Conference on Embedded Networked Sensor Systems
SOSP	ACM Symposium on Operating Systems Principles
SP	IEEE Security & Privacy
SP(J)	IEEE Security & Privacy magazine
TC(J)	IEEE Transactions on Computers
TCC	Theory of Cryptography Conference
TDSC(J)	IEEE Transactions on Dependable and Secure Computing
TISSEC(J)	ACM Transactions on Information and System Security
TKDE(J)	IEEE Transactions on Knowledge and Data Engineering
TrustBus	Trust, Privacy and Security in Digital Business Conference
USENIX	USENIX Technical Conference & /Security Symposium
VLDB	International Conference on Very Large Data Bases
WISA	Workshop on Information Security Applications
WORM	ACM Workshop on Rapid Malcode
WPES	Workshop on Privacy in the Electronic Society

References

- Almeida, J. A. S., Pais, A. A. C. C. & Formosinho, S. J. (2009). Science indicators and science patterns in Europe. *Journal of Informetrics*, 3(2), 134–142.
- Biryukov, M. & Dong, C. (2010). Analysis of computer science communities based on dblp. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Research and advanced technology for digital libraries, volume 6273 of lecture notes in computer science* (pp. 228–235). Berlin, Heidelberg: Springer.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25, 163–177.
- Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117. Proceedings of the Seventh International World Wide Web Conference
- Computer Science Conference Ranking. <http://www.cs-conference-ranking.org/conferenc rankings/topicsi.html>, 2009.
- Computer Security Conference Ranking and Statistic <http://faculty.cs.tamu.edu/guofei>, 2009.
- Excellence in Research for Australia (ERA). <http://www.arc.gov.au/era/>, 2010.
- Fortnow, L. (2009). Viewpoint time for computer science to grow up. *Communications of the ACM*, 52(8), 33–35.
- Guillaume, J. (2005). Bipartite graphs as models of complex networks. *Lecture Notes in Computer Science*, 3405, 127–139.
- Kandylas, V., Upham, S. P. & Ungar, L. H. (2008, November). Finding cohesive clusters for analyzing knowledge communities. *Knowledge and Information Systems*, 17, 335–354.
- Lee, W. (2008). How to identify emerging research fields using scientometrics: An example in the field of Information Security. *Scientometrics*, 76(3), 503–525.
- Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *In SPIRE '02: Proceedings of the 9th international symposium on string processing and information retrieval*, pp. 1–10. London, UK: Springer-Verlag.
- Leydesdorff, L. (2005). The evaluation of research and the evolution of science indicators. *Current Science*, 89(9), 1510–1517.
- Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, 20(1), 32–36.
- Meyer, B., Choppy, C., Staunstrup, J. & Leeuwen, J. (2009). Research evaluation for computer science. *Communications of the ACM*, 52(4), 31–34.
- Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution: research articles. *Journal of the American Society for Information Science and Technology*, 56(12), 1250–1273.
- Nerur, S., Sikora, R., Mangalaraj, G. & Balijepally, V. (2005, November). Assessing the relative influence of journals in a citation network. *Communications of the ACM*, 48, 71–74.
- Newman, M. E. (2001, June). Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1), 016132.
- Per, H. & Frank, H. (1995). Eccentricity and centrality in networks. *Social Networks*, 17(1), 57–63.
- Security Conference Ranking. <http://www.doc.ic.ac.uk/cd04/ranking.html>, 2007.
- Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica*, 69(3), 224–229.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003, November). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Shin, J. (2009). Building World-class Research University: The Brain Korea 21 Project. *Higher Education*, 58(5), 669–688.
- Sidiropoulos, A. & Manolopoulos, Y. (2005). A new perspective to automatically rank scientific conferences using digital libraries. *Information Processing and Management: An International Journal*, 41(2), 289–312.
- Wainer, J., Xavier, E. C. & Bezerra, F. (2009). Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2), 535–547.
- Wasserman, S. & Faust, K. (1994). *Social networks analysis*. Cambridge University Press.

- Yan, S. & Lee, D. (2007). Toward alternative measures for ranking venues: A case of database research community. In *In Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries, JCDL '07* (pp. 235–244). New York, NY, USA: ACM.
- Zaane, O., Chen, J. & Goebel, R. (2009). Mining research communities in bibliographical data. In H. Zhang, M. Spiliopoulou, B. Mobasher, C. Giles, A. McCallum, O. Nasraoui, J. Srivastava, & J. Yen (Eds.), *Advances in web mining and web usage analysis, volume 5439 of lecture notes in computer science* (pp. 59–76). Berlin, Heidelberg: Springer.
- Zhou, D., Councill, I., Zha, H. & Giles, C. L. (2007). Discovering Temporal Communities from Social Network Documents. In *In Proceedings of the 2007 seventh IEEE international conference on data mining* (pp. 745–750). Washington, DC, USA: IEEE Computer Society.
- Zhou, D., Ji, X., Zha, H. & Giles, C. L. (2006). Topic evolution and social interactions: how authors effect research. In *In Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06* (pp. 248–257). New York, NY, USA: ACM.
- Zhou, T., Ren, J., Medo, M. & Zhang, Y. (2007, October). Bipartite network projection and personal recommendation. *Physical Review E*, 76(4), 046115.
- Zhuang, Z., Elmacioglu, E., Lee, D. & Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. In *In Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries, JCDL '07* (pp. 225–234). New York, NY, USA: ACM.