On Good and Fair Paper-Reviewer Assignment

Cheng Long, Raymond Chi-Wing Wong, Yu Peng, Liangliang Ye
The Hong Kong University of Science and Technology
Kowloon, Hong Kong
{clong,raywong,gracepy,llye}@cse.ust.hk

***Abstract*—Peer review has become the most common practice for judging papers submitted to a conference for decades. An extremely important task involved in peer review is to assign submitted papers to reviewers with appropriate expertise which is referred to as *paper-reviewer assignment*. In this paper, we study the paper-reviewer assignment problem from both the goodness aspect and the fairness aspect. For the goodness aspect, we propose to maximize the *topic coverage* of the paper-reviewer assignment. This objective is new and the problem based on this objective is shown to be NP-hard. To solve this problem efficiently, we design an approximate algorithm which gives a $\frac{1}{3}$-approximation. For the fairness aspect, we perform a detailed study on conflict-of-interest (COI) types and discuss several issues related to using COI, which, we hope, can raise some open discussions among researchers on the COI study. Finally, we conducted experiments on real datasets which verified the effectiveness of our algorithm and also revealed some interesting results of COI.**

I. Introduction

Peer review has become the most common practice for judging papers submitted to a conference for decades [1]. One important task involved in peer review is to assign reviewers from the program committee to submitted papers. This task is referred to as *paper-reviewer assignment* (PRA). Since the number of submitted papers and the size of the program committee in a conference is very large (e.g., in ICDM 2012, there were 756 submissions and 234 PCs [1]), *manual* PRA is not quite feasible. Instead, people resort to *automatic* PRA methods.

The first work about automatic PRA was due to Dumais and Nielsen [2], where PRA is regarded as an *information retrieval* problem. Specifically, a paper is used as a *query* and each reviewer is represented by a text document (e.g., the expertise statement provided by the reviewer or simply the publications of the reviewer), The problem is to retrieve a certain number of reviewers who are the most *relevant* to the paper. Following the same idea of [2], some other methods for automatic PRA have been proposed [3, 1, 4-7]. These retrieval-based methods differ from each other by using different information retrieval techniques: Latent Semantic Indexing (LSI) [2, 3], Vector Space Model [1, 4], Topic Model [5], Mixture Language Model [7] and other models [6].

A common drawback of these retrieval-based methods is that the retrieval process has to be done for each paper *independently* such that an assignment between papers and reviewers can be constructed, which, however, introduces several problems. First, it might happen that some popular reviewers are assigned with excessive papers while some other reviewers with few or even no papers. Second, even though they can avoid the first problem by incorporating a hard constraint on the reviewers' workload, they still suffer a lot. One example is that the constructed assignment is sensitive to the order of papers processed since the papers processed earlier can be assigned with relevant reviewers but the papers processed later may be assigned with irrelevant reviewers due to the hard constraint introduced. Another example is that the process is heuristic-based without any optimized objectives.

To avoid the above drawback of the retrieval-based methods, more recent studies regard PRA as a *matching* problem between a paper set and a reviewer set [8-15]. In this way, all papers *as a batch* are assigned to the reviewers. In general, matching-based methods has the following two-phase framework. In the first phase, a *weighted* bipartite graph between a paper set and a reviewer set is constructed. The weight of an edge between a paper and a reviewer is usually set to a value denoting the *relevance* between the paper and the reviewer. In the second phase, based on the constructed bipartite graph, a matching, which is used to construct the final paper-reviewer assignment, is found such that several constraints (e.g., each paper is assigned to a certain number of reviewers and each reviewer is not assigned with excessive papers) are satisfied and one appropriate objective function is optimized. Under this framework, most of the matching-based methods [16, 17, 1, 14] share the following ideas. Firstly, the relevance used in the first phase is defined based on *topics*. Specifically, each paper (and also each reviewer) is associated with a set of *topics* from a topic domain (pre-specified or learned). The relevance between a paper and a reviewer is defined based on the number of *common* topics shared by the paper and the reviewer. Secondly, the matching described in the second phase usually corresponds to the *maximum weight matching* (or its variant) based on the bipartite graph. With these ideas, these methods assign papers to reviewers so the *total weight* of a matching is maximized. This idea is intuitive, but it does not directly maximize the *coverage* of the topics of the papers *covered* by the assigned reviewers, which could be quite problematic in some cases. To illustrate, let us work through a toy example in detail.

Example 1 (Motivating Example): In Figure 1(a), we have 1 paper p_1, a topic domain containing 5 topics, namely $t_1, ..., t_5$, and 4 reviewers, namely $r_1, ..., r_4$. A link between a paper and a topic denotes that the paper covers the topic,

[1] http://www.cs.uvm.edu/~icdm/Slides/ICDM12-CommunityMtng.pdf

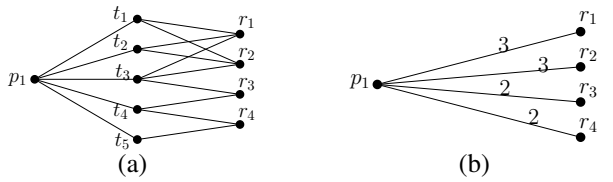


Fig. 1. A motivating example

and a link between a reviewer and a topic denotes that the reviewer has his/her expertise on the topic. Assume that we want to assign 2 reviewers to paper p_1. Existing methods such as [17] would first construct a bipartite graph as shown in Figure 1(b). The weight of edge (p_1, r_1) in the graph is equal to 3 since paper p_1 and reviewer r_1 share 3 topics (i.e., t_1, t_2 and t_3). Similarly, we have the weights of other edges as shown in the graph. Then, the assignment $A_1 = \{(p_1, r_1), (p_1, r_2)\}$ (which means that paper p_1 is assigned to reviewers r_1 and r_2) is computed since it corresponds to the *maximum weight matching* subject to the constraint that p_1 is assigned to 2 reviewers. Unfortunately, assignment A_1 is not good enough since all the assigned reviewers (i.e., r_1 and r_2) have their expertise on only 3 *distinct* topics of p_1(i.e., t_1, t_2 and t_3) while there are still 2 topics of p_1 (i.e., t_4 and t_5) left uncovered. This might result in a fairly poor judgement on p_1. A better choice is to assign r_1 (or r_2) and r_4 to p_1 since they together have their expertise in more distinct topics of p_1 (because the number of distinct topics covered is 5) and thus the resulting judgement is probably more qualitative. □

Motivated by the above observation, in this paper, we propose a new problem called ***Max**imum **T**opic **C**overage **P**aper-**R**eviewer **A**ssignment* (MaxTC-PRA), which assigns papers to reviewers such that the total number of *distinct* topics of papers that are covered by the assigned reviewers is maximized and the following three constraints are satisfied: (1) *Paper Demand Constraint:* Each paper is reviewed by a certain number of reviewers, (2) *Reviewer Workload Constraint:* Each reviewer reviews at most a certain number of papers, and (3) *COI Constraint:* There exists no conflict-of-interest (COI) between the authors of each paper and the assigned reviewers.

Compared with the existing matching-based methods such as those in [14, 16, 17], our MaxTC-PRA gives a *broader* coverage of the topics of the papers by only counting the *distinct* topics covered by the assigned papers. This makes a big difference since it is desirable that the topics covered by the assigned reviewers together should be as *broad* as possible so that multiple aspects of the paper can be judged qualitatively.

The new objective of MaxTC-PRA is more natural and desirable than those of the existing methods but it makes the problem harder. Firstly, the existing matching-based methods cannot be adapted to our problem. This is because our MaxTC-PRA problem cannot be fit in the framework of matching-based methods since the weight between a paper and a reviewer is correlated with the weight between the same paper and another reviewer (in terms of the topic coverage). Secondly, we prove that MaxTC-PRA is NP-hard. Although solving MaxTC-PRA optimally is difficult, we propose a greedy algorithm which iteratively assigns a paper to a reviewer such that the *marginal gain* of the objective is the greatest. Interestingly, this greedy algorithm provides a $\frac{1}{3}$-factor approximation for MaxTC-PRA.

In addition, we discuss in detail three issues of using COI in PRA, namely, (1) what types of COI (e.g., the co-author relationship is a type of COI) should be used in PRA; (2) whether it is reliable to use the COIs specified by the authors and/or the reviewers only; and (3) whether it is always good to use as many COIs as possible. We hope these issues could open some more thorough and useful discussions among researchers on the COI study.

We summarize our main contributions as follows.

- To the best of our knowledge, we are the *first* to propose the MaxTC-PRA problem, which assigns reviewers to papers for maximizing the total number of *distinct* topics of papers covered by the assigned reviewers. MaxTC-PRA is superior over existing methods since it gives a *broader* coverage of the topics of the papers.
- We prove that MaxTC-PRA is NP-hard and design an approximate algorithm for MaxTC-PRA, which provides a $\frac{1}{3}$-factor approximation.
- We discuss three issues related to using COI in PRA, which we hope, could serve as the initiative efforts on the COI study.
- We conducted a comprehensive empirical study which verifies the superiority of our MaxTC-PRA and the effectiveness of our approximate algorithm. Besides, we observed some interesting findings of the effects of different COI types on the paper-reviewer assignment.

The remainder of this paper is organized as follows. We define and solve the MaxTC-PRA problem in Section II and in Section III, respectively. Then, we discuss three issues related to COI in Section IV and provide the empirical study in Section V. We review the related work in Section VI and conclude the paper in Section VII.

II. Problem Definition

Let $P = \{p_1, p_2, ..., p_n\}$ be a set of n papers and $R = \{r_1, r_2, ..., r_m\}$ be a set of m reviewers. Let $T = \{t_1, t_2, ..., t_k\}$ be the topic domain whose collection will be described in Section V-A. Each paper p_i (reviewer r_j) is captured by a set of topics from the topic domain T, denoted by $T(p_i)$ ($T(r_j)$) for each $i \in [1, n]$ ($j \in [1, m]$). For example, in Example 1, $P = \{p_1\}$, $R = \{r_1, r_2, r_3, r_4\}$, $T = \{t_1, t_2, ..., t_5\}$, $T(p_1) = \{t_1, t_2, ..., t_5\}$ and $T(r_1) = \{t_1, t_2, t_3\}$.

Given a paper $p_i \in P$ and a reviewer $r_j \in R$, (p_i, r_j) is said to be a *match* if p_i is assigned to r_j. Given a match $M = (p_i, r_j)$, we denote the paper (reviewer) involved in M by $M.p$ ($M.r$) which is p_i (r_j).

Any possible set of matches corresponds to an *assignment* between the paper set P and the reviewer set R. Given an assignment $A \subseteq P \times R$, we denote by $A(p_i)$ the set of matches in A that involve paper $p_i \in P$, i.e., $A(p_i) = \{M | M \in A, M.p = p_i\}$. Similarly, we define $A(r_j) = \{M | M \in A, M.r = r_j\}$ for each $r_j \in R$. For

example, in Example 1, $A = \{(p_1, r_1), (p_1, r_4)\}$ denotes an assignment. Besides, we have $A(p_1) = \{(p_1, r_1), (p_1, r_4)\}$ and $A(r_1) = \{(p_1, r_1)\}$.

In a typical PRA scenario, each paper should be assigned to a certain number τ_p of reviewers for cross-verification consideration (e.g., in ICDM, each paper is assigned to 3 reviewers), which we call the *Paper Demand Constraint*, and each reviewer should not be assigned with more than a certain number τ_r of papers for workload consideration, which we call the *Reviewer Workload Constraint*. Here, τ_p and τ_r are two positive integers at least 1.

Besides, we have to take COIs into consideration in PRA, which we call the *COI Constraint*. Specifically, a paper should not be assigned to a reviewer if the authors of the paper have COI with the reviewer. We can always capture the COIs by a set $\mathcal{C}$ which contains all pairs of papers and reviewers in the form of (p_i, r_j) such that p_i cannot be assigned to r_j. We will discuss the issues related to COI in Section IV in detail.

Let A be an assignment between the paper set P and the reviewer set R. Let $\sigma(A)$ denote the total number of *distinct* topics of all papers covered by the assigned reviewers wrt assignment A. For instance, in Example 1, consider the assignment $A_1 = \{(p_1, r_1), (p_1, r_2)\}$. We have $\sigma(A_1) = 3$ since 3 distinct topics of p_1 are covered by the assigned reviewers in A_1. Similarly, for the assignment $A_2 = \{(p_1, r_1), (p_1, r_4)\}$, we have $\sigma(A_2) = 5$. Then, the MaxTC-PRA problem is to find an assignment A with the greatest value of $\sigma(A)$ among all the assignments satisfying the Paper Demand Constraint, the Reviewer Workload Constraint and the COI Constraint.

Problem Statement. Formally, we present the MaxTC-PRA problem as follows.

Problem 1: Given a set P of n papers, namely $p_1, p_2, ..., p_n$, a set R of m reviewers, namely $r_1, r_2, ..., r_m$, two integers τ_p and τ_r, and a COI set $\mathcal{C}$, **the MaxTC-PRA problem** is to find the assignment A that satisfies

- **Paper Demand Constraint**: Each paper is assigned to τ_p reviewers, i.e., for each $p_i \in P$, $|A(p_i)| = \tau_p$,
- **Reviewer Workload Constraint**: Each reviewer reviews at most τ_r papers, i.e., for each $r_j \in R$, $|A(r_j)| \leq \tau_r$,
- **COI Constraint**: $A \cap \mathcal{C} = \emptyset$,

and maximizes $\sigma(A)$. □

Consider Example 1. Suppose $\tau_p = 2$, $\tau_r = 1$ and $\mathcal{C} = \{(p_1, r_1)\}$. Then, the solution of the MaxTC-PRA problem is $A = \{(p_1, r_2), (p_1, r_4)\}$ with the greatest value of $\sigma(A) = 5$.

Note that the definition of MaxTC-PRA in Definition 1 does not guarantee that there always exists a solution. For example, when $n \cdot \tau_p > m \cdot \tau_r$, the Paper Demand Constraint and the Reviewer Workload Constraint cannot be satisfied simultaneously. Even when $n \cdot \tau_p \leq m \cdot \tau_r$, if the COI set is large (an extreme case is that all pairs of papers and reviewers cannot be made into matches), it is possible that there exist no assignments that satisfy all the three constraints. Motivated by these considerations, we replace the Paper Demand Constraint with a *relaxed* one: we only require that each paper is reviewed *at most* τ_p reviewers. The MaxTC-PRA problem adopting the relaxed Paper Demand Constraint enjoys two benefits. First, it always guarantee an existence of a solution. Second, in the case that there are enough reviewers and/or the COI set is not that large (which is the common case in practice), the optimal solution of the MaxTC-PRA problem with the relaxed constraint could be made exactly the same as that of the original version of MaxTC-PRA either *automatically* (because of the objective function $\sigma(\cdot)$ since including more matches in the assignment usually increases the objective function) or *manually* by post-assigning those papers that have not been assigned to exactly τ_p reviewers to the reviewers who are assigned with fewer than τ_r papers. Thus, in the following, we focus on the MaxTC-PRA problem with the relaxed Paper Demand Constraint only.

Intractability. Unfortunately, it turns out that the MaxTC-PRA problem is NP-hard.

Lemma 1: The MaxTC-PRA problem is NP-hard. □

Proof: We prove by transforming a *Maximum Coverage* [18] problem instance, which is NP-hard, to a MaxTC-PRA problem instance.

The Maximum Coverage problem is described as follows. Given a universe set U of n elements, namely $e_1, e_2, ..., e_n$, a collection C of some subsets of U, where $C = \{S_1, S_2, ..., S_m\}$, and a positive integer m', it finds a set C' of at most m' subsets in C such that the (distinct) number of elements in U that are covered by the subsets in C', i.e., $|\cup_{S \in C'} S|$, is maximized.

The transformation is described as follows. We regard the universe set U as a paper and each subset S_l in C as a reviewer for $1 \leq l \leq m$. We regard each element as a topic and thus the paper corresponding to U has the set of topics as $\{e_1, e_2, ..., e_n\}$. Similarly, the reviewer corresponding to S_l has the set of topics equal to S_l. Besides, we set τ_p to be m', τ_r to be 1 and $\mathcal{C}$ to be $\emptyset$.

Clearly, the transformed MaxTC-PRA problem instance is equivalent to the Maximum Coverage problem instance, which finishes the proof. ■

III. Solution of MaxTC-PRA

A. A Greedy Algorithm

In this section, we develop an approximate algorithm called *Greedy* for MaxTC-PRA.

Let A be the assignment to be returned by the algorithm. *Greedy* initializes A to be $\emptyset$ and thus all the three constraints of the MaxTC-PRA problem are satisfied at the right beginning (note that we consider the relaxed Paper Demand Constraint in the problem). Let U be a set of all possible matches to be used in the algorithm. It initializes U to $P \times R - \mathcal{C}$, which excludes all pairs of papers and reviewers in $\mathcal{C}$ for the COI Constraint. During the execution, it maintains U such that for each match $M \in U$, the relaxed Paper Demand Constraint and the Reviewer Workload Constraint are satisfied by $A \cup \{M\}$ (where A is being updated during the execution) by removing all those matches from U that violate this condition. Note that any match in U that violates this condition cannot be

Algorithm 1 The *Greedy* algorithm

Input: A paper set $P = \{p_1, p_2, ..., p_n\}$; A reviewer set $R = \{r_1, r_2, ..., r_k\}$; Parameters τ_p, τ_r and a COI set $\mathcal{C}$

Output: An approximate solution of the MaxTC-PRA problem

1: $A \leftarrow \emptyset$
2: $U \leftarrow P \times R - \mathcal{C}$
3: **while** U is non-empty **do**
4: pick the match $M \in U$ such that $\sigma(A \cup \{M\}) - \sigma(A)$ is maximized
5: $A \leftarrow A \cup \{M\}$; $U \leftarrow U - \{M\}$
6: remove each match $M' \in U$ such that the relaxed Paper Demand Constraint or the Reviewer Workload Constraint is not satisfied by $A \cup \{M'\}$
7: **return** A

included in A (or any super-set of A) subject to the constraints of MaxTC-PRA. Besides, note that there is no need to do any removal operation on U initially since the initial content of U (i.e., $P \times R - \mathcal{C}$) contains all possible matches to be used in the algorithm. Then, it iteratively augments A with the match $M \in U$ such that the marginal gain (in terms of the objective function) of adding M into A is maximized. Note that all the three constraints are satisfied if one of the matches in U is inserted into A. The process stops when U is empty. Since A satisfies all the constraints at the beginning and augmenting A in each iteration does not break these constraints, we know that the final assignment A satisfies all the constraints of MaxTC-PRA. We present the *Greedy* algorithm in Algorithm 1.

Detailed Steps and Complexity Analysis. In the implementation of the Greedy algorithm which relies on the concept of the marginal gain of adding a match M into an assignment A (i.e., $\sigma(A \cup \{M\}) - \sigma(A)$), we maintain a priority-queue to store all matches in U based on their marginal gain values.

Before the while-loop, we need to construct U, which takes $O(|P \times R - \mathcal{C}|)$ time. Besides, we compute the marginal gain of each match in U, which could be finished in $O(|P \times R - \mathcal{C}| \cdot |T|^2)$ time (since we have $|P \times R - \mathcal{C}|$ matches in U and the cost of computing the marginal gain of each match is simply $O(|T|^2)$).

Within each iteration of the while-loop, the first step is to obtain the match M in U with the greatest marginal gain in $O(1)$ time from the priority-queue. The second step is to update the marginal gain of each remaining match in $U - \{M\}$ in $O(|R| \cdot |T|^2)$ time (since at most $|R|$ matches' marginal gains are affected after a new match is included in A). The third step is to re-insert the matches with the updated marginal gains into the priority-queue, which can be done in $O(|R| \log |P \times R - \mathcal{C}|)$. The fourth step is to remove some matches in U due to the relaxed Paper Demand Constraint and the Reviewer Workload Constraint, which takes $O(|P| + |R|)$ time (since at most $|P| + |R|$ matches are affected).

Since we have at most $\min\{|P| \cdot \tau_p, |P \times R - \mathcal{C}|\}$ iterations in total, we know that the time complexity of *Greedy* is $O(|P \times R - \mathcal{C}| + |P \times R - \mathcal{C}| \cdot |T|^2 + \min\{|P| \cdot \tau_p, |P \times R - \mathcal{C}|\} \cdot (1 + |R| \cdot |T|^2 + |R| \log |P \times R - \mathcal{C}| + |P| + |R|)) = O(|P \times R - \mathcal{C}| \cdot |T|^2 + \min\{|P| \cdot \tau_p, |P \times R - \mathcal{C}|\} \cdot (|R| \cdot |T|^2 + |R| \log |P \times R - \mathcal{C}| + |P|))$.

B. Approximation Quality Analysis

In this part, we prove that the *Greedy* algorithm in Algorithm 1 gives a $\frac{1}{3}$-factor approximation of MaxTC-PRA by using the theory of *submodularity* [19] and *p-system* [20].

Given an assignment $A \subseteq P \times R$, we say that A is *feasible* if it satisfies all the three constraints, namely, the relaxed Paper Demand Constraint, the Reviewer Workload Constraint and the COI Constraint. Let $\mathcal{F}$ be the set containing all feasible assignments between P and R. Note that $\mathcal{F} \subseteq 2^{P \times R - \mathcal{C}}$. Here, 2^S corresponds to the power set of a set S. Then, the optimal solution of the MaxTC-PRA problem is the assignment in $\mathcal{F}$ with the greatest value of $\sigma(\cdot)$.

Submodularity. First, we show that $\sigma(\cdot)$ is a *submodular* function.

Definition 1 (Submodularity [19]): Let U be a universe set. Function $f : 2^U \rightarrow \mathbb{R}$ is said to be *submodular* iff given any two sets X and Y where $Y \subseteq X \subseteq U$, $\forall e \in U - Y, f(X \cup \{e\}) - f(X) \leq f(Y \cup \{e\}) - f(Y)$. □

Given a set X and an element e, the marginal gain of adding an element e in the set X is denoted by $f(X \cup \{e\}) - f(X)$. For a submodular function, the marginal gain of adding an element in the set is non-increasing when the set becomes larger. As will be shown later, a submodular function enjoys some nice properties.

Interestingly, we found that the objective function $\sigma(\cdot)$ of the MaxTC-PRA problem is submodular.

Lemma 2: Function $\sigma : 2^{P \times R} \rightarrow \mathbb{R}$ is submodular. □

Proof: Given an assignment A, we denote by $T(p_h, A)$ the set of distinct topics of paper p_h where $1 \leq h \leq n$ that are covered by the assigned reviewers in A.

Let X and Y be any two subsets of $P \times R$ with $Y \subseteq X$ and $e = (p_i, r_j)$ be an element in $P \times R - Y$. Note that each of X and Y corresponds to an assignment between P and R.

Consider $\sigma(X)$. We have $\sigma(X) = \sum_{h=1}^{n} |T(p_h, X)|$.

Consider $\sigma(X \cup \{e\})$. Since adding $e = (p_i, r_j)$ in X covers no topics of other papers than p_i, we know that $T(p_h, X \cup \{e\}) = T(p_h, X)$ for $h = 1, 2, ..., i-1, i+1, ..., n$. Therefore, $\sigma(X \cup \{e\}) = \sum_{h=1}^{n} |T(p_h, X \cup \{e\})| = \sum_{h=1}^{i-1} |T(p_h, X)| + |T(p_i, X \cup \{e\})| + \sum_{h=i+1}^{n} |T(p_h, X)|$. As a result, we have

$$\sigma(X \cup \{e\}) - \sigma(X) = |T(p_i, X \cup \{e\})| - |T(p_i, X)| \quad (1)$$

Similarly, we have

$$\sigma(Y \cup \{e\}) - \sigma(Y) = |T(p_i, Y \cup \{e\})| - |T(p_i, Y)| \quad (2)$$

Since $T(p_i, X \cup \{e\}) = T(p_i, X) \cup (T(p_i) \cap T(r_j))$, we have

$$|T(p_i, X \cup \{e\})| - |T(p_i, X)| = |(T(p_i) \cap T(r_j)) - T(p_i, X)| \quad (3)$$

Similarly, we know that

$$|T(p_i, Y \cup \{e\})| - |T(p_i, Y)| = |(T(p_i) \cap T(r_j)) - T(p_i, Y)| \quad (4)$$

Since $Y \subseteq X$, we know that

$$T(p_i, Y) \subseteq T(p_i, X) \tag{5}$$

By combining Equations (1), (2), (3), (4) and (5), we deduce that $\sigma(X \cup \{e\}) - \sigma(X) \leq \sigma(Y \cup \{e\}) - \sigma(Y)$. ■

For illustration, consider Example 1. Consider two subsets of $2^{P \times R}$: $X = \{(p_1, r_1), (p_1, r_4)\}$ and $Y = \{(p_1, r_4)\}$. Note that $Y \subset X$. Suppose that we add $e = (p_1, r_2)$ to both X and Y. The marginal gain of adding e into X, which is equal to $\sigma(X \cup \{e\}) - \sigma(X) = 5 - 5 = 0$, is smaller than that of adding e into Y, which is equal to $\sigma(Y \cup \{e\}) - \sigma(Y) = 5 - 2 = 3$. That is, the marginal gain of adding a match into a larger assignment is smaller.

p-system. Next, we show that the set $\mathcal{F}$ containing all feasible assignments forms a *2-system* [20].

Before we give the definition of p-system, we introduce the concept of *independent system*.

Definition 2 (Independent system [20]): Let U be a universe set. $I \subseteq 2^U$ is said to be an *independent system* wrt U if (1) $\emptyset \in I$ and (2) for each $X \in I$ and any $Y \subseteq X$, $Y \in I$. □

It is easy to verify that $\mathcal{F}$ is an independent system wrt $P \times R - \mathcal{C}$ since $\emptyset \in \mathcal{F}$ (an empty assignment is feasible) and for any feasible assignment $A \in \mathcal{F}$, we know that any $A' \subseteq A$ is also feasible (since dropping any match in a feasible set does not break any of the three constraints).

Definition 3 (p-system [20]): Given a universe set U and a positive integer p, an independent system I wrt U is said to be a *p-system* wrt U if for any $S \subseteq U$ and any $X, Y \subseteq S$ such that X and Y are *maximal* wrt (S, I) [2] with the maximum and minimum sizes, respectively, $|X|/|Y| \leq p$. □

A p-system defines a constrained independent system I, where the size ratio between the largest and smallest sets that are maximal wrt (S, I) is bounded by a constant p for any subset S of the universe set. A *matroid* [21] forms a p-system with $p = 1$. Interestingly, we find that $\mathcal{F}$ forms a p-system with $p = 2$ (i.e., 2-system).

Lemma 3: $\mathcal{F}$ which corresponds to the set containing all feasible assignments is a 2-system wrt $P \times R - \mathcal{C}$. □

Proof: Let S be a subset of $P \times R - \mathcal{C}$. Let X (Y) be the subset of S which is maximal wrt $(S, \mathcal{F})$ with the greatest (smallest) size. Let P_S (R_S) be the set of papers (reviewers) that are involved in the matches of S.

We say that each match in $X - Y$ is a *white* match and each match in $Y - X$ is a *black* match. Note that $(X - Y) \cap (Y - X) = \emptyset$.

For each paper $p \in P_S$, we define a variable $v(p)$ to be the difference between the number of white matches involving p and the number of black matches involving p. Specifically, $v(p) = |X(p) - Y(p)| - |Y(p) - X(p)|$. Similarly, for each reviewer $r \in R_S$, we define $v(r) = |X(r) - Y(r)| - |Y(r) - X(r)|$.

[2] A subset X of set $S \subseteq U$ is maximal wrt (S, I) if $X \in I$ and $\forall e \in S - X$, $X \cup \{e\} \notin I$.

First, we introduce a property which will be used for proving our lemma.

Property 1: For each white pair $(p, r) \in X - Y$, we have $v(p) \cdot v(r) \leq 0$. □

Proof: We prove this property by contradiction. Assume that $(p, r) \in X - Y$ is a white pair with $v(p) \cdot v(r) > 0$. We have two cases.

Case 1: $v(p) > 0$ and $v(r) > 0$. In this case, we have $|Y(p)| - |X(p)| = |Y(p) - X(p)| - |X(p) - Y(p)| = -v(p) < 0$ which implies that $|Y(p)| < |X(p)| \leq \tau_p$. Similarly, we have $|Y(r)| - |X(r)| = |Y(r) - X(r)| - |X(r) - Y(r)| = -v(r) < 0$ which implies that $|Y(r)| < |X(r)| \leq \tau_r$. As a result, we know that $Y' = Y \cup \{(p, r)\}$ is feasible, which, however, leads to a contradiction that Y is maximal wrt $(S, P \times R - \mathcal{C})$.

Case 2: $v(p) < 0$ and $v(r) < 0$. In this case, we can verify the property by using the fact that X is maximal wrt $(S, P \times R - \mathcal{C})$ in a similar way as in Case 1. ■

Let $P^+ \subseteq P_S$ be the set containing all those papers p with $v(p) > 0$.

Consider a paper p in P^+ and a white match $(p, r) \in X(p) - Y(p)$. Since $v(p) > 0$, we know that $X(p) - Y(p) \neq \emptyset$. Besides, according to Property 1, we know that $v(r) \leq 0$ which implies that $Y(r) - X(r) \neq \emptyset$ (since $(p, r) \in X(r) - Y(r)$ and $|Y(r) - X(r)| - |X(r) - Y(r)| = -v(r) \geq 0$).

Next, we maintain a set of black matches for each paper $p \in P^+$, denoted by $B(p)$, as follows.

We first initialize all the black matches in $Y - X$ to be *un-tagged*. Let p be a paper in P^+. For each white match (p, r) that involves p, we pick an *un-tagged* black edge from $Y(r) - X(r)$ and include it in $B(p)$. Note that this un-tagged black edge always exists since (1) the black edges in $Y(r) - X(r)$ are only retrieved when processing the white matches involving r; (2) each white match involving r is processed at most once; (3) $|Y(r) - X(r)| - |X(r) - Y(r)| \geq 0$ (since $v(r) \leq 0$). Then, we tag this black edge in $Y(r) - X(r)$.

Now, we are ready to deduce that $|X|/|Y| \leq 2$ as follows.

$$\begin{aligned} |X| - |Y| &= |X - Y| - |Y - X| \\ &= \sum_{p \in P_S} |X(p) - Y(p)| - \sum_{p \in P_S} |Y(p) - X(p)| \\ &= \sum_{p \in P_S} (|X(p) - Y(p)| - |Y(p) - X(p)|) \\ &= \sum_{p \in P_S} v(p) \leq \sum_{p \in P^+} v(p) \leq \sum_{p \in P^+} |B(p)| \\ &\leq |Y - X| \leq |Y| \end{aligned} \tag{6}$$

Equation (6) implies that $|X|/|Y| \leq 2$. ■

To illustrate Lemma 3, consider Example 1. Let $\tau_p = 2$ and $\tau_r = 1$. Suppose $\mathcal{C} = \{(p_1, r_3)\}$ and thus $P \times R - \mathcal{C} = \{(p_1, r_1), (p_1, r_2), (p_1, r_4))\}$. Let $U = P \times R - \mathcal{C}$ and $I = \mathcal{F}$ (recall that $\mathcal{F} \subseteq 2^{P \times R - \mathcal{C}} = 2^U$). Let $S \subseteq U$ be $\{(p_1, r_1), (p_1, r_2)\}$. It could be verified that there exists exactly one set that is maximal wrt (S, I), which is S itself. Therefore, we know that X, which is maximal wrt (S, I) with the greatest

size, is S and Y, which is maximal wrt (S, I) with the smallest size, is also S. Thus, $|X|/|Y| = 1 \leq 2$.

Approximation Factor. Now, we are ready to introduce the approximation quality of the *Greedy* algorithm, which is presented in the following Lemma 4.

Lemma 4: The *Greedy* algorithm in Algorithm 1 gives a $\frac{1}{3}$-factor approximation of MaxTC-PRA. □

Proof: Note that the MaxTC-PRA problem is equivalent to the problem of maximizing $\sigma(A)$ (which is submodular by Lemma 2) over $\mathcal{F}$ (which is a 2-system by Lemma 3). According to [20], a greedy algorithm for the problem of maximizing a submodular function over a p-system achieves a $\frac{1}{p+1}$-factor approximation. By using this result, we know the *Greedy* algorithm in Algorithm 1, which is the greedy algorithm for maximizing $\sigma(A)$ over $\mathcal{F}$, achieves a $\frac{1}{3}$-factor approximation. ■

As will be shown in our experiments, the practical approximation factor (which is greater than 0.9 in most experiments) is much better than this theoretical bound.

IV. Conflict-of-Interest

A *conflict-of-interest* (COI) between an author of a paper and a reviewer means that the reviewer has an a-priori bias for/against the paper. Therefore, avoiding different types of COI when assigning papers to reviewers is critical for un-biased judgement of papers. In this part, we discuss 3 issues related to the problem of using COI for a PRA task.

Issue 1: What types of author-reviewer relationship should be considered as COI types?

We do a literature study first. Some author-reviewer relationships that are well-recognized as COI types by popular conference management systems (e.g., Microsoft Research's CMT tool (http://msrcmt.research.microsoft.com/cmt/), EasyChair (http://www.easychair.org/) and CyberChair (http://www.borbala.com/cyberchair/)) include the *co-author relationship* (the author and the reviewer have co-authored some papers), the *colleague relationship* (the author and the reviewer have worked/collaborated at the same affiliations), and the *advisor-advisee* relationship (the author was/is the thesis advisor of the reviewer or vice versa).

We note that all these COI types share a common feature that a reviewer who has a COI with an author will be biased *for* the paper of the author. In other words, they are used to avoid *false positives* of paper judgements. For example, a reviewer who was the thesis advisor of an author would probably be biased for the paper submitted by the author. In contrast, no COI types have been proposed for the case that a reviewer is biased *against* the paper of the author. In this paper, for the first time, we identify such a COI type called the *competitor relationship*.

An author who submitted a paper and a reviewer are said to be in a *competitor relationship* with each other if the reviewer has also submitted a paper in the conference and the fraction of the common topics of these two papers is above a pre-specified threshold δ. Note that it is a common case that the program committee members (i.e., reviewers) of a conference would usually submit papers to the conference as well. As a result, it is not uncommon that a reviewer is assigned with a paper which has quite similar topics as his/her own paper submitted to the same conference (e.g., the two papers appear in the same area track or even study the same problem). In this case, it is hard to assume that the reviewer would judge the paper objectively since s/he has his/her own paper which is competitive with the paper being considered. Thus, in order to avoid this potential un-biased factor for safety, it is reasonable to consider the competitor relationship as a COI type. Clearly, the competitor relationship could be used to avoid *false negatives* of paper judgements. This competitor relationship can also be supported by the reviewing policy in some top-tier conferences like SODA, one of the top-tier conferences in the theory field, where program committee members are disallowed to submit papers within the same conference in order to avoid false negatives. Note that indicating the competitor relationship as a COI type looks less demanding compared with the reviewing policy used in SODA because reviewers can also submit papers when the competitor relationship can be specified as a COI type. However, it is arguable whether this competitor relationship should be used or not similar to the reviewing policy used in SODA. Thus, in this paper, whether this relationship is considered as a COI type can be regarded as a user input. If the user would like to include it, then it can be involved as a COI type. Otherwise, it can also be disregarded from being a COI type. Our purpose here is to raise some issues about the COI type for peer review. Whether we finally adopt this relationship is left as a discussion among all researchers.

In this paper, we study 4 types of author-reviewer relationships as COI types, namely, the co-author relationship, the colleague relationship, the advisor-advisee relationship and the competitor relationship. The first three are existing COI types and the last one is newly-proposed in this paper.

Issue 2: Is it reliable to use the COI information specified by the authors and/or the reviewers only?

With the growing of the number of paper submissions, the number of reviewers (i.e., program committee members) in a conference could simply be several hundreds (e.g., ICDM 2012 involves 234 reviewers), in which case, the task of specifying COIs manually by authors and/or reviewers becomes time-consuming and it could happen often that some COIs are left un-specified simply because of the tedious aspect of this task. In addition, what should not be ignored is that there might exist some bad-citizens who would cheat the system by either leaving some COIs un-specified or specifying some fake COIs intentionally. The above two considerations trigger us to come up with the idea of specifying the COIs *automatically* which are then used as a complementary source.

In this paper, we mine multiple sources from the Web for detecting the aforementioned 4 types of COIs automatically. More details will be discussed in Section V-C.

Issue 3: Is it always good to use as many COIs as possible?

To answer this question, we should understand the effects of COIs on the PRA task. Without doubt, COIs make the PRA task fairer. But, we should also notice that a COI might prevent a paper being assigned to the reviewer who have much expertise on the topics of the paper and this might degrade the *goodness* (i.e., the topic coverage) of the resulting assignment. That is, there is a trade-off between the goodness and the fairness of the assignment between the papers and the reviewers.

In this paper, we study the effect of each of the above COI types on the goodness of the assignment returned by our *Greedy* algorithm to see to what extent this COI type degrades the goodness of the assignment. Details will be discussed in Section V-D. This result can be served as a reference to researchers to understand the significance of each COI type on the paper-reviewer assignment in order to determine whether some of the COI types should be adopted finally in the field.

V. Experiment

A. Experimental Setup

Datasets. We collected all the papers published in KDD from year 2006 to year 2010 as the submitted papers. We have 496 papers in total. We collected all the program committees (PCs) of ICDM 2010 and KDD 2010 as the reviewers (since ICDM and KDD correspond to the two major prestigious conferences in the data mining community and we expect that the program committees of these two conferences together cover the topics of data mining well). We have 550 reviewers in total. We collected all the subject areas specified in KDD 2011 for paper submission as the topic domain. We have 49 topics in total.

Topic Extraction. Following the convention [4, 3, 6], we retrieve the topics covered by a paper (reviewer) by using the *TF-IDF weighted vector space model* [22] as follows. For each paper, we combined its title, abstract and keywords (if any) as its profile. For each topic, we collected its description from Wikipedia (http://www.wikipedia.org/) as its profile. For each reviewer, we collected the title, the abstract and the keywords (if any) of each of his/her publications from CiteSeerX (http://citeseerx.ist.psu.edu/) and Google Scholar (http://scholar.google.com/) as his/her profile (Due to the incompleteness of both websites, we gathered all the information available together). In this way, we believe that the collected profile of each paper/topic/reviewer is comprehensive enough (though one can always augment the profiles with other sources if possible). Then, we computed the relevance between each pair of papers (reviewers) and topics. To retrieve the topics of a paper p (reviewer r), we introduce a parameter δ_1 (δ_2), and include all those topics t in $T(p)$ ($T(r)$) such that the relevance between p (r) and t is at least δ_1 (δ_2). In our experiments, we tuned the parameters of δ_1 and δ_2, and adopted the setting of $\delta_1 = \delta_2 = 0.04$. That is, the relevance between a reviewer's profile and a topic's profile is used to measure the expertise of the reviewer on the topic.

COI Collection. Refer to Section V-C where a case study of the COI collection is provided.

Algorithms. We considered two algorithms in our experiments, namely, *Greedy* and ILP. *Greedy* is the approximate algorithm proposed in this paper and ILP is the art-of-the-state [17] among all algorithms that consider *topic coverage* for paper-reviewer assignment. ILP is a matching-based method, which sets the weight of the edge between a paper and a reviewer as the number of common topics shared by the paper and the reviewer in Phase 1 and computes the maximum weight matching with an *integer linear program* in Phase 2.

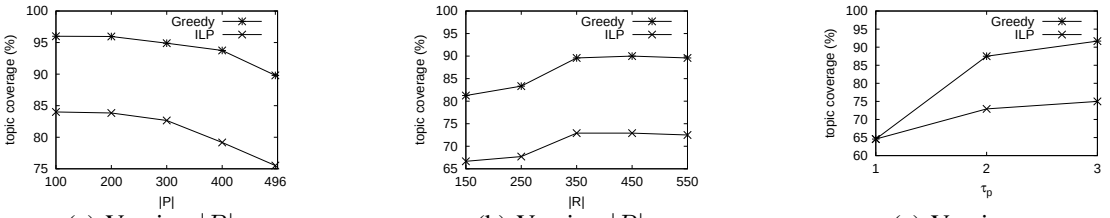
Configurations. 4 factors are studied in our experiments, namely, the number of papers (i.e., $|P|$), the number of reviewers (i.e., $|R|$), the parameter of the (relaxed) Paper Demand Constraint (i.e., τ_p) and the parameter of the Reviewer Workload Constraint (i.e., τ_r). The values used for $|P|$ are 100, 200, 300, 400 and **496**, where the value in bold font is used by default. The values used for $|R|$ are 150, 250, **350**, 450 and 550. The values used for τ_p are 1, 2 and **3** and those used for τ_r are **5**, 6, 7 and 8.

B. Verification of MaxTC-PRA & Greedy

We used a measure called *topic coverage*, which is defined to be the average percentage of the distinct topics of papers covered by the assigned reviewers. We compared the *Greedy* algorithm and the ILP algorithm. Clearly, a better algorithm gives a higher topic coverage. We used all COI types in this set of experiments. The results when varying $|P|$, $|R|$ and τ_p are shown in Figure 2(a), (b) and (c), respectively.

Varying $|P|$. We have the following observations. First, the topic coverage of the assignment returned by *Greedy* is consistently higher than that returned by ILP. In fact, the topic coverage of *Greedy* is 15% larger than that of ILP. Second, the topic coverage of *Greedy* is more than 90% in all settings, which also implies that the empirical approximation factor of *Greedy* (which is at least 0.90) is much better the theoretical one (which is $\frac{1}{3}$). Third, the topic coverage of both *Greedy* and ILP decreases slightly when $|P|$ increases. This could be explained as follows. When the number of papers increases, the total number of distinct topics of the papers is larger. Since the number of reviewers is kept unchanged, it is likely that the topic coverage would decrease.

Varying $|R|$. Again, there is a clear gap between the topic coverage of *Greedy* and that of ILP. Another observation which is different from the case of varying $|P|$ is that when $|R|$ increases, the topic coverage of both algorithms increases slightly first and then keeps stable (starting at $|P| \approx 350$). The increasing trend could be easily explained since when $|R|$ increases while $|P|$ is fixed, it is similar to the case that $|P|$ decreases while $|R|$ is fixed and thus an opposite trend of that in Figure 2(a) (i.e., a increasing tread) could be observed. The stable trend afterwords might be due to the following scenario where there exist some rare topics of the papers which are hard to cover (few reviewers have expertise on these topics) and thus simply increasing the number of reviewers does not help to increase the topic coverage significantly.



(a) Varying $|P|$ (b) Varying $|R|$ (c) Varying τ_p

Fig. 2. Topic coverage: *Greedy* vs. ILP

co-author (COI1)	colleague (COI2)	advisor-advisee (COI3)	competitor (COI4)	All COI
7,408	2,734	237	21,701	29,864

TABLE I
STATISTICS OF THE COLLECTED COIS

Varying τ_p. Still, we note that the topic coverage of *Greedy* is consistently higher than that of ILP except for the case of $\tau_p = 1$. Specifically, when $\tau_p = 1$ (i.e., each paper is reviewed by only 1 reviewer), the topic coverage of both algorithms is around 65%, which is not high enough for obtaining a qualitative paper judgement. This implies that setting $\tau_p = 1$ is not a good choice in real applications. In contrast, when $\tau_p = 3$ which is a common practice in real applications, the topic coverage of *Greedy* is more than 90%, which is about 15% higher than that of ILP.

Varying τ_r. The trends are similar to those of varying $|R|$. This is because increasing τ_r is essentially "adding" (more or less similar) reviewers into the reviewer set (with the original Reviewer Workload Constraint). Bearing this in mind, in the following, we omit the experimental results of varying τ_r.

Conclusion: Our *Greedy* algorithm works consistently better than the existing ILP algorithm in terms of topic coverage.

C. A Case Study of COI Collection

The COIs are collected as follows. The co-author relationship set (denoted by COI1) was collected by using the DBLP dataset (http://www.informatik.uni-trier.de/∼ley/db/). The colleague relationship set (denoted by COI2) was collected based on the researchers' homepages and LinkedIn (http://www.linkedin.com/). The advisor-advisee relationship set (denoted by COI3) was collected based on the researchers' homepages. The competitor relationship set (denoted by COI4) was collected based on P, R and T with the setting of $\delta = 0.5$.

To handle the ambiguity issue of authors' names, we append the corresponding affiliation to the end of each author name.

The sizes of the collected COI sets are shown in Table I. The size of the COI set for a COI type is the number of pairs of papers and reviewers that violate this COI type.

We show a typical example of each COI type in Table II, where the the author who is underlined and the reviewer who is shown in the same row has a corresponding COI.

D. Studies of the Effect of COI

1) On the Fairness of PRA: We evaluate the effect of each COI type on the fairness of PRA by comparing the resulting assignments with and without using this COI type. Specifically, we collected the percentage of matches in the assignment that violate the COI type being considered if this COI type is not used for PRA. The results of varying $|P|$, $|R|$ and τ_p are shown in Figure 3(a), (b) and (c), respectively.

Varying $|P|$. We have the following observations. First, surprisingly yet interestingly, we find that COI4 (i.e., the competitor relationship) has the strongest effect on PRA. Specifically, around 15-17% of the matches in the resulting assignment violate COI4 if it is not used as a COI type. This might be explained by the fact that the reviewers of a conference usually have their own papers submitted in the conference and at the same time, the papers (by others) that share some common topics with these papers would probably be assigned to these reviewers since the reviewers usually have much expertise on these common topics. This results in a situation where many author-reviewer pairs violate COI4. Second, the (decreasing) ordering of the remaining COI types by their effects on the assignment is COI1 (i.e., the co-author relationship), COI2 (i.e., the colleague relationship), COI3 (i.e., the advisor-advisee relationship). Third, the percentage of matches that violate a COI (any COI type) is more than 22%, which provides a clear argument that the assignment made without using any COI types could be fairly unfair.

Varying $|R|$ and τ_p. The results provide us similar clues and thus the discussion on these results are omitted here.

2) On the Goodness of PRA: As we have mentioned in Section IV, imposing a COI might degrade the quality of the PRA since it prevents some papers from being assigned to the reviewers who have much expertise on the topics of the papers. But, it remains unclear that *to what extent this side-effect of a COI type would be?*

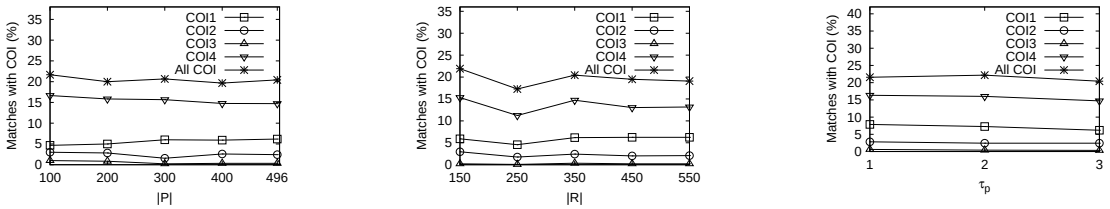
In this part, we try to answer this question by evaluating the effects of the COI types on the quality (i.e., topic coverage) of the assignment returned by our *Greedy* algorithm. The results of varying $|P|$, $|R|$ and τ_p are shown in Figure 4(a), (b) and (c), respectively.

Varying $|P|$. In Figure 4(a) where we show the effect of the whole set of 4 COI types. We observe that using the whole set of COI types degrades the topic coverage by 2%-3% only compared to the case where no COIs are used. This is a good news since it implies that the side-effect (in terms of the quality) of using all COI types studied in this paper is negligible. The results of the effect corresponding to each COI type only are not shown here since they are quite minor (e.g., ranging from 0.1% to 1%). Instead, we show the results of the overall effect of the whole set of 4 COI types.

The decreasing trends of the curves when $|P|$ increases

COI type	Reviewer	Submitted paper	Authors	Notes
Co-author	P. S. Yu	Large Human Communication Networks: Patterns and a Utility-Driven Generator	N. Du, C. Faloutsos, B. Wang, L. Akoglu	The collected COI information indicates that Yu and Faloutsos were co-authors of some papers. We manually verified that one of the papers that were co-authored by Yu and Faloutsos was published in KDD 2008.
Colleague	K. Yamanishi	Grouped Graphical Granger Modeling Methods for Temporal Causal Modeling	A. C. Lozano, N. Abe, Y. Liu, S. Rosset	The collected COI information indicates that Yamanishi and Abe have worked/collaborated at some affiliations. We manually verified that they worked in NEC research center from 1992 to 1995.
Advisor-advisee	M. Hua	Neighbor Query Friendly Compression of Social Networks	H. Maserrat, J. Pei	The collected COI information indicates that one of Hua and Pei is/was the advisor of the other. We manually verified that Hua was a PhD student supervised by Pei from 2005 to 2009.
Competitor	C. C. Aggarwal	From Frequent Itemsets to Semantically Meaningful Visual Patterns	J. Yuan, Y. Wu, M. Yang	The collected COI information indicates that Aggarwal has also submitted a paper to the same conference which is competitive with the paper being considered. We manually verified that, Aggarwal submitted a paper "Frequent pattern mining with uncertain data", which cover some similar topics, e.g., "frequent sets and patterns" and "KDD process and support tools".

TABLE II
EXAMPLES OF THE CONSIDERED COI TYPES THAT ARE MINED FROM THE WEB



(a) Varying $|P|$ (b) Varying $|R|$ (c) Varying τ_p

Fig. 3. Effects of COI on the fairness of the assignment

could be explained in the way as we did for Figure 2(a).

Varying $|R|$ and τ_p. The results are similar to the case of varying $|P|$ except for the trends of the curves in Figure 4(b) and (c) which could be explained similarly as we did for Figure 2(b) and (c), respectively.

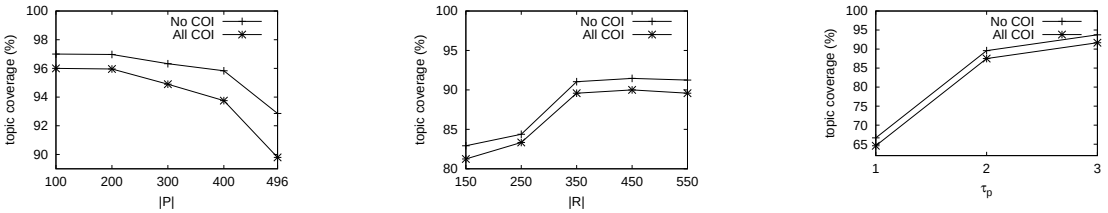
Conclusion: Among all the COI types considered in this paper, the competitor relationship (COI4), which is newly-proposed in this paper, has the strongest effect on the fairness of PRA. Besides, using the whole set of 4 COIs does not degrade the quality of PRA significantly and thus they all can be used in PRA for the fairness consideration.

VI. RELATED WORK

In the past two decades, a bulk of studies studied the problem of (automatic) paper-reviewer assignment (PRA), which we categorize into two branches, namely the *retrieval-based* methods [2, 6, 4, 1, 5, 7, 23, 24, 3] which regard each paper as a query to retrieve the *relevant* experts (i.e., reviewers) and the *matching-based* methods [8-17] which compute a *matching* based on the bipartite graph between the paper set and the reviewer set.

The retrieval-based methods vary a lot and are different from each other in terms of the information retrieval techniques used for finding reviewers. The first retrieval-based method is due to [2], which used the Latent Semantic Indexing (LSI) for the retrieval task. In [6], Basu et al. proposed to mine the Web to capture reviewers' preferences and then to recommend papers to the reviewers based on the mined preference. In [4, 1], the authors adopted the Vector Space Model (VSM) to capture the relevance between papers and reviewers. In [5], Mimno and McCallum used the Topic Model to model the expertise of reviewers. In [7], the authors proposed three strategies for retrieving the reviewers and showed that the one based on the Mixture Language Model [25] performed the best. In [23], the authors designed a *particle-swarm* algorithm based on the co-author network for identifying reviewers. In [24], the authors adopted a *probabilistic random walk* model based on a so-called *expertise graph* containing topical documents and related persons. In [3], the authors adopted LSI for retrieving the topics of papers and reviewers, and measured the relevance between them based on their topic information. [26] gave a comprehensive survey on *expertise retrieval*. All these retrieval-based methods suffer from the drawback that to construct a paper-reviewer assignment, they have to perform the retrieval task for each paper *individually*, which does not take into consideration some constraints required by the assignment and are prone to be biased to the papers processed earlier.

The matching-based methods avoid the above drawback of the retrieval-based methods by constructing the assignment in a *collective* way. A matching-based method has two phases: Phase 1 computes the weights of the edges in the bipartite graph and Phase 2 computes an appropriate matching based on the graph. Different matching-based methods use different strategies for computing the weights at Phase 1 and/or adopt different objectives for the matching at Phase 2. In [8], the authors computed the weights based on the *expertise statements* provided by both the authors and the reviewers and computed a *bottleneck* matching. In [9], the authors proposed some desiderata for a matching-based method but left many details un-specified. In [10], the authors proposed a *heuristic* matching algorithm. In [11], it is required that the area chairs provide the weights and a maximum weight



(a) Varying $|P|$ (b) Varying $|R|$ (c) Varying τ_p

Fig. 4. Effects of COI on the goodness of the assignment

matching is computed. In [12], the authors proposed to *learn* the weights and computed the maximum weight matching. In [13], the weights are assumed to be available and a *leximin-optimal* assignment is computed. In [15], the authors explored Language Model, Regression and Collaborative Filtering for learning the weights and studied several matching objectives.

More recently, [14, 16, 17] proposed to model both a paper and a reviewer as a set of topics and computed the weights based on the topic information in Phase 1. In Phase 2, each of these studies tends to assign the papers to those reviewers who have similar topic information with them. For example, in [17], Karimzadehgan and Zhai used the number of common topics shared by a paper and a reviewer as the weight of the corresponding edge in Phase 1 and computed a *maximum weight* matching in Phase 2. Nevertheless, none of these studies aim to maximize the *topic coverage* of the papers directly, which usually results in the adverse case where each of the assigned reviewers of a paper has many topics in common with the paper, but the number of distinct topics covered by these reviewers is undesirably low. In contrast, our MaxTC-PRA favors the assignment where more distinct topics of papers are covered by the reviewers.

VII. Conclusion

In this paper, we proposed a new problem called MaxTC-PRA, which favors the diversification of the topics of the papers covered by the assigned reviewers. We showed that this problem is NP-hard and proposed a $\frac{1}{3}$-approximate algorithm for this problem whose time complexity is polynomial. We also discussed several issues related to using COI in PRA, which we hope, could serve as an initiative effort on COI study. We conducted experiments on real datasets which verified our algorithm and revealed some interesting results of COI. Some possible future work can be done. First, we would like to study the paper-reviewer assignment problem with different objectives. One possible objective is to maximize the minimum total number of distinct topics of a single paper covered by the assigned reviewers. Another possible objective is to maximize the *overall importance* of the covered topics where each topic is associated with an *importance weight*. Second, we would like to study the problem when *partial* relevance between papers and topics and *partial* relevance between reviewers and topics are considered, which is more complicated. In this paper, we consider *full* relevance between papers and topics and *full* relevance between reviewers and topics.

References

[1] H. K. Biswas and M. Hasan, "Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment," in *ICICT*, 2007.

[2] S. T. Dumais and J. Nielsen, "Automating the assignment of submitted manuscripts to reviewers," in *SIGIR*. ACM, 1992.

[3] J. Zablocki and R. Lee, "Auto-assign: An implementation to assign reviewers using topic comparison in start."

[4] S. Hettich and M. J. Pazzani, "Mining for proposal reviewers: lessons learned at the national science foundation," in *SIGKDD*. ACM, 2006.

[5] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *SIGKDD*. ACM, 2007.

[6] C. Basu, H. Hirsh, W. Cohen, and C. Nevill-Manning, "Recommending papers by mining the web," in *Proceedings of the IJCAI99 Workshop on Learning about Users*, 1999.

[7] M. Karimzadehgan, C. Zhai, and G. Belford, "Multi-aspect expertise matching for review assignment," in *CIKM*. ACM, 2008.

[8] D. Hartvigsen, J. C. Wei, and R. Czuchlewski, "The conference paper-reviewer assignment problem*," *Decision Sciences*, vol. 30, no. 3, pp. 865–876, 1999.

[9] S. Benferhat and J. Lang, "Conference paper assignment," *International Journal of Intelligent Systems*, 2001.

[10] J. Merelo-Guervós and P. Castillo-Valdivieso, "Conference paper assignment using a combined greedy/evolutionary algorithm," in *Parallel Problem Solving from Nature-PPSN VIII*. Springer, 2004, pp. 602–611.

[11] C. J. Taylor, "On the optimal assignment of conference papers to reviewers," Tech. Rep. MS-CIS-08-30, 2008.

[12] D. Conry, Y. Koren, and N. Ramakrishnan, "Recommender systems for the conference paper assignment problem," in *RecSys*. ACM, 2009, pp. 357–360.

[13] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre, "Assigning papers to referees," *Algorithmica*, 2010.

[14] W. Tang, J. Tang, and C. Tan, "Expertise matching via constraint-based optimization," in *WI-IAT*, 2010.

[15] L. Charlin, R. S. Zemel, and C. Boutilier, "A framework for optimizing paper matching," *arXiv preprint:1202.3706*, 2012.

[16] S. Ferilli, N. D. Mauro, T. Basile, F. Esposito, and M. Biba, "Automatic topics identification for reviewer assignment," *AAAI*, pp. 721–730, 2006.

[17] M. Karimzadehgan and C. Zhai, "Constrained multi-aspect expertise matching for committee review assignment," in *CIKM*. ACM, 2009, pp. 1697–1700.

[18] D. S. Hochbaum, *Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems*. PWS Publishing Co., 1997, pp. 94–143.

[19] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions-i," *Mathematical Programming*, vol. 14, no. 1, 1978.

[20] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions-ii," *Polyhedral combinatorics*, pp. 73–87, 1978.

[21] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a submodular set function subject to a matroid constraint," *Integer programming and combinatorial optimization*, pp. 182–196, 2007.

[22] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[23] M. A. Rodriguez and J. Bollen, "An algorithm to determine peer-reviewers," in *CIKM*. ACM, 2008, pp. 319–328.

[24] P. Serdyukov, H. Rode, and D. Hiemstra, "Modeling multi-step relevance propagation for expert finding," in *CIKM*. ACM, 2008, pp. 1133–1142.

[25] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999.

[26] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise retrieval," *Found. Trends Inf. Retr.*, vol. 6, pp. 127–256, 2012.