Conflict of Interest Declaration and Detection System in Heterogeneous Networks

Siyuan Wu Leong Hou U
University of Macau
Macau SAR
mb55408,ryanlhu@umac.mo

Sourav S Bhowmick
Nanyang Technological University
Singapore, Singapore
assourav@ntu.edu.sg

Wolfgang Gatterbauer
Northeastern University
Boston, USA
wolfgang@ccs.neu.edu

ABSTRACT

Peer review is the most critical process in evaluating an article to be accepted for publication in an academic venue. When assigning a reviewer to evaluate an article, the assignment should be aware of conflicts of interest (COIs) such that the reviews are fair to everyone. However, existing conference management systems simply ask reviewers and authors to declare their explicit COIs through a plain search user interface guided by some simple conflict rules. We argue that such declaration system is not enough to discover all latent COI cases. In this work, we study a graphical declaration system that visualizes the relationships of authors and reviewers based on a heterogeneous co-authorship network. With the help of the declarations, we attempt to detect the latent COIs automatically based on the meta-paths of a heterogeneous network.

KEYWORDS

Conflict of Interest, Peer Review Process, Heterogeneous Network

1 INTRODUCTION

In academic peer review, a single conflict of interest (COI) case may be enough to turn a decision around. From an author's point of view, the fairness of the review process is crucial as she may spend months (or even years) to prepare an article. To avoid COI cases, a common practice is to ask reviewers and authors to declare their COIs guided by a set of pre-defined COI rules (e.g., advisor-advisee relation and collaboration(s) in past 3 years). We argue that such self-declaration approach is insufficient since (1) reviewers and authors may not exhaustively check the conflict list (especially when there are thousands candidates in the list) and (2) the declaration rules are too strict so some latent conflicts are not required to be declared, e.g., an "academic sibling": two researchers with the same advisor if they never published together. Even though program committee chairs (in conferences) or editors (in journals) may check suspicious COIs on their own [3], the checking is very time consuming and most likely incomplete. All of these motivate us to re-visit the COI declaration system and study how to detect COIs more intelligently and develop an approach that can automatically detect COIs.

CIKM'17, November 6–10, 2017, Singapore.
 ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: 10.1145/3132847.3133134

In this work, we share the intuition with [2, 7] that author relationships could be extracted from their publication records. In [2, 7], the publication records are modeled as a heterogeneous network, that consists of author nodes and various types of edges (e.g., co-authorship, time of publication, publication venue, author affiliation, etc). The richness of these relationships paves the way to study various analytical tasks, such as link prediction [7], recommendation [5], and similarity measure [4]. However, we found very limited work on the COI detection problem. Cheng et al. [3] classified some COI cases based on the authorship records and discussed the effect of COIs in the paper assignment process. Aleman-Meza et al. [1] introduced a rule-based COI detection method based on a friend-of-a-friend model, where the weight of a rule is designed subjectively. We are aware of the difficulties of the COI detection problem, including (1) unavailability of ground-truth data and (2) no model for measuring how a conflict pair affects the decision of an article. All of these difficulties retard the development of the detection system.

To overcome these difficulties, we propose a mutual reinforced approach that consists of two processes, *manual declaration* and *auto-detection*. We first study an interactive declaration system that attempts to reduce the search space of latent conflicted cases. Specifically, the declaration system visualizes latent COI cases on a meta-path graph [2] so that latent conflict cases can be identified by navigating the meta-paths. As an example, an author can find a conflicted reviewer who was under the same supervisor using meta-path "*advisor-advisee*". This is particularly helpful for those fresh PhD graduates who submit their first few studies after graduation.

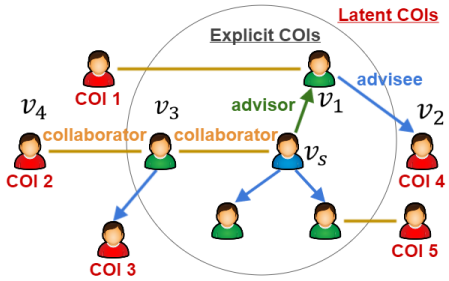


Figure 1: Meta-Path based Conflict of Interest Declaration

Figure 1 shows a heterogeneous network composed of user nodes and three types of edges, including co-authorship (orange lines), advisee (blue arrows), and advisor (green arrow). In this example, we assume that all one-hop neighbors have conflict with the author v_s and all red nodes are in the program committee of a conference. The author may find some latent COI cases by navigating these

paths. For instance, v_2 is likely a COI case as v_2 is the advisee of the author's advisor. However, v_4 is less likely a COI case of the authors since she may not know the collaborator of v_3.

Even though some COI cases can be detected based on the meta-paths, it is also interesting to explore if the conflict degree can be estimated automatically. In this work, we attempt to estimate the weight of these path patterns by a logistic regression model. The regression result is then used to estimate the degree of conflict between an author and a reviewer.

The key contribution of this paper is a proposal solve the cold-start problem of automatic COI detection by a novel integrated manual and automatic detection approach. This approach simultaneously (1) helps users find possible COIs with an intuitive interface ("manual declaration") and (2) suggests possible COIs to authors ("auto-detection") based on a trained model from past manually declared conflicts.

2 SYSTEM OVERVIEW

Currently, there is no ground truth set of COIs that would allow to reliably train and evaluate automatic detection systems. Our approach is thus to (1) facilitate users finding possible COIs ("manual declaration"), (2) then use this ground truth data set to train a model, which (3) suggests possible COIs to future authors ("auto-detection").

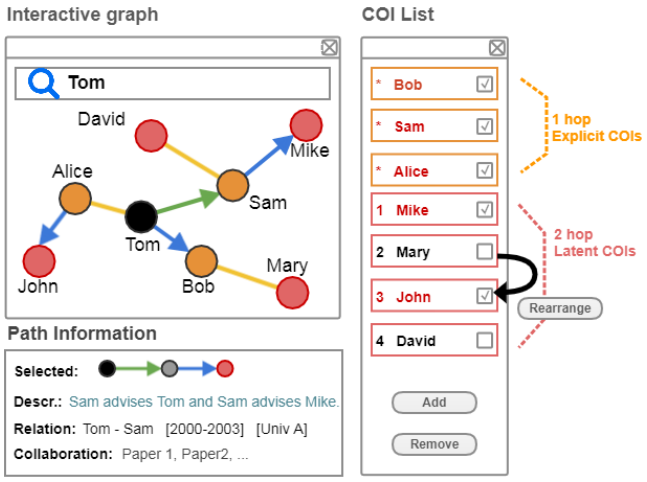


Figure 2: Manual Declaration System

Manual Declaration. We provide a user-friendly interface to visualize the heterogeneous collaboration network between authors, where the network contains different types of edges (e.g., co-authorship and advisor-advisee[1]). Figure 2 shows an example on how our declaration system is used in a conference. When author *Tom* searches his name, the system shows the collaboration network of *Tom* and the program committee members (highlighted by red color) on the left hand side of the system. To avoid being overwhelmed by information, the system only shows two-hops neighbors initially and allows a user to expand a branch by clicking on it. Given the collaboration network, *Tom* can add any author node into his COI list (on the right hand side of the system) and find the semantic meaning of an edge by simply clicking on it. *Tom* can also prioritize the conflict cases where these inputs will be considered into the *auto detection* process.

Auto Detection. We attempt to study an auto-detection method based on the meta-path of the heterogeneous network. Specifically, our method extracts different meta-paths from the heterogeneous network and tries to train the weight vector of the meta-paths by a logistic regression. However, as mentioned in Section 1, there is no ground-truth data available so that the training process is hard to proceed and evaluate. In this work, we suggest to treat the manual declaration data as the ground truth to train the model. To reinforce these two sub-processes, the auto-detection result is then used as the initial COI cases of the declaration system.

Our reinforced COI system can substantially reduce the workload of reviewers and authors since (1) the latent COI cases are now visualized on a sub-graph and (2) the semi-automatic process already suggest possible COIs. While the manual declaration process secures the reliability of the conflict cases, the auto detection and the graph visualization enhances the effectiveness of the COI declaration. As these two steps are mutually reinforced, we believe that the auto detection can provide more reliable COI cases when the system collects more declaration data from the users.

3 COI DETECTION PROBLEM

Given a collaboration network $G(V, E)$, V indicates author set and E indicates their heterogeneous relationship. Every edge $e \in E$ represents a collaboration record between two author nodes. Given a source author $s \in V$, our problem is to return a latent top-k COI list, $COI(s) = (v_1, v_2, \ldots, v_k)$ where the order is decided by their latent COI scores $score(v_i)$ and $\forall v_i, v_j \in V : i < j \Rightarrow score(v_i) \geq score(v_j)$.

The COI score $score(v_i)$ indicates the relative conflict degree of v_i to the source author s. In this work, we propose to estimate the degree of a COI case by a weighted count of all meta-paths between an author and a PC member. In general, the relationship between two nodes is large in a heterogenous graph when they are connected by many meta-paths. However, as we discussed in Section 1, we should not equally weigh these meta-paths as some are more important than the others. Thereby, we will discuss our ranking model in the next section.

4 COI DETECTION ON HETEROGENEOUS GRAPH

We first introduce how to construct the heterogeneous graph from the co-authorship information in Section 4.1. Next, we discuss the meta-path patterns being considered in this work in Section 4.2. Lastly, we introduce our training and ranking model in Section 4.3.

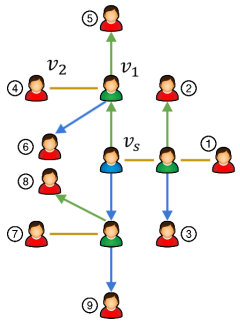
4.1 Heterogeneous Graph Construction

The heterogeneous graph $G(V, E)$ is constructed as follows. Initially, we set all authors as the graph nodes V and we add an edge $e = (a, b)$ into E if author a collaborated with author b in the past. To design the type of the *simple* edges E, we utilize the advisor-advisee relationship detection algorithm proposed in [8], which uses a

[1] Advisor-advisee relationship can be extracted by some learning approaches, e.g., [8].

time-constrained probabilistic factor graph model to estimate the advisor and advisee of the authors based on their collaboration periods, affiliation changes, and collaboration probabilities. The result of [8] is a set of advisor-advisee pairs denoted as A. For every edge $e \in A$, we replace e by two directed edges of types "advisor" and "advisee". For all other edges $e' \in E \setminus A$, we simply set their type to "collaborator". In summary, after the construction, we have three types of edges in the heterogeneous graph, $\varphi = \{advisor, advisee, collaborator\}$.

4.2 Meta-Path Patterns



(a) Heterogeneous Graph

No.	Meta Paths
θ_1	$\overrightarrow{collaborator}\ \overrightarrow{collaborator}$
θ_2	$\overrightarrow{collaborator}\ \overrightarrow{advisor}$
θ_3	$\overrightarrow{collaborator}\ \overrightarrow{advisee}$
θ_4	$\overrightarrow{advisor}\ \overrightarrow{collaborator}$
θ_5	$\overrightarrow{advisor}\ \overrightarrow{advisor}$
θ_6	$\overrightarrow{advisor}\ \overrightarrow{advisee}$
θ_7	$\overrightarrow{advisee}\ \overrightarrow{collaborator}$
θ_8	$\overrightarrow{advisee}\ \overrightarrow{advisor}$
θ_9	$\overrightarrow{advisee}\ \overrightarrow{advisee}$

(b) Meta Paths

Figure 3: Meta-Path Patterns

In the heterogeneous network, a meta-path is a path pattern that describes a sequence of node types and edge types [6]. In this work, we generate the meta-paths by enumerating all possible edge type combinations and then train their weights by a logistic regression model (being discussed shortly in the next subsection). For instance, $\overrightarrow{collaborator}\ \overrightarrow{collaborator}$ is a 2-hop meta-path. And $\overrightarrow{advisor}\ \overrightarrow{collaborator}\ \overrightarrow{advisee}$ is a 3-hop meta-path. We consider both 2-hop and 3-hop patterns in the experiment section. We could consider other relationships, such as working in the same project and the same affiliation, which may further enhances the reliability. Figure 3 shows all meta-path patterns of 2-hops and its corresponding heterogeneous network.

4.3 COI Score Estimation

We first define the proximity of an edge $e = (i, j)$ as follows.

$$prox(e) = \frac{\sum_{p \in e} 1/|authors(p)|}{papers(i)} \tag{1}$$

where p indicates a co-author paper between i and j, $|authors(p)|$ indicates the number of authors in paper p, and $papers(i)$ indicates the number of papers written by i which can be viewed as a normalized factor. Accordingly, the proximity of a 2-hops path, $e \rightarrow e'$, can be defined as

$$prox(e) \cdot prox(e') \tag{2}$$

Given a meta-path θ_i, a source node s, and a target node t, we can define the proximity of the meta-path between s and t as follows.

$$prox(s, t, \theta_i) = \sum_{(s,v)\rightarrow(v,t)\in\theta_i} prox((s, v)) \cdot prox((v, t)) \tag{3}$$

where $(s, v) \rightarrow (v, t) \in \theta_i$ indicates the pattern of these two edges matches θ_i.

To estimate the COI score between s and t, a naïve idea is to aggregate the proximity of all meta-paths based on Equation 3. However, it is obvious that the conflict degree of different meta-paths should not be the same. In the following, we study a training model to estimate the conflict degree of the meta-paths.

In the training process, we need to categorize some author pairs as positive pairs P and negative pairs N. Positive pairs P are the set of co-author pairs and self-declared pairs (e.g., from the self declaration system). Negative pairs N are sampled by a random walk on the heterogeneous graph starting from the source node s. Specifically, s and t can be treated as a negative case if their random walk probability is very low. In addition, we should also consider those COI cases being removed from the declaration system as the negative cases (cf. Figure 2).

For a training pair $x = (s, t)$, we define the weight and the proximity of all meta-paths as two vectors $W = (w_1, \ldots, w_9)$ and $L = (prox(s, t, \theta_1), \ldots, prox(s, t, \theta_9))$, respectively. We can calculate the probability of the positive cases and negative cases by a logistic regression model as follows.

$$p_x = p(y \mid x) = \begin{cases} \frac{e^{W^T L+b}}{e^{W^T L+b}+1} & \text{if } y = 1 \cap x \in P \\ \frac{1}{e^{W^T L+b}+1} & \text{if } y = 0 \cap x \in N \end{cases}$$

where b is a constant in the logistic regression model.

Our objective is to maximize the probability of the positive cases so the objective function can be written as

$$\prod_{x \in P \cup N} p_x^{y_x} (1 - p_x)^{1-y_x} \tag{4}$$

where this objective function can be solved by maximum likelihood estimation. After the training, the COI score between s and t can be calculated by

$$score(s, t) = \frac{e^{W^T L+b}}{e^{W^T L+b} + 1} \tag{5}$$

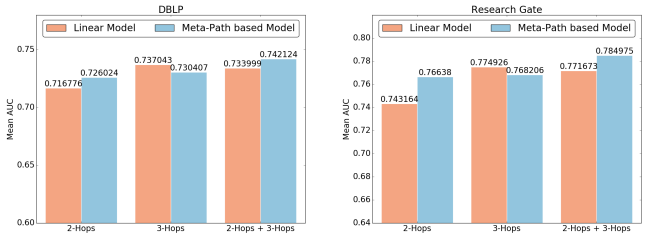
5 EXPERIMENTS

Our publication records are extracted from DBLP, that contain 28 top-tier conferences and 31 leading journals in different research areas from 1970 to 2016. Given the publication records, we construct a heterogeneous graph of 82,713 author nodes and 673,280 co-authorship edges, among which ~16% of edges are extracted as advisee-advisor relationships by the method proposed in [8].

DBLP. We make an assumption that two authors might have conflict of interest if they will collaborate in the short term future (e.g., next 3 years). To evaluate the effectiveness, we simply partition the DBLP dataset into two parts, $D_1(1970 - 2013)$ and $D_2(2014 - 2016)$. We use both positive and negative cases from D_1 to train the model and evaluate the model using the positive pairs in D_2.

ResearchGate. As we do not have the ground truth of real COI cases, we follow the suggestion of [9] who evaluate social media research by external sources when there is no ground truth data available. The external source adopted in this work is ResearchGate, which is an academic social networking platform for sharing academic activities. We set a pair (a, b) is positive if a is *following* b on ResearchGate.

5.1 Effectiveness Evaluation



(a) DBLP **(b) Research Gate**

Figure 4: Average AUC(area under the ROC curve)

We compare our auto detection method with a simple linear aggregation approach (i.e., with unified weight meta-paths). Figure 4 shows the average AUC (Area Under the ROC Curve) of two methods on the meta-paths of 2-hops (i.e., 9 patterns) and 3-hops (i.e., 27 patterns). The result indicates that the logistic regression model is effective on 2-hops but less effective on 3-hops. This is caused by the sparsity of the inputs, i.e., some 3-hops meta-path patterns are very rare. The result of the combined method (2-hops + 3-hops) demonstrates the effectiveness of the logistic regression technique. However, it is hard to draw any conclusion that we find better COIs than other methods as ground truth data is not available.

5.2 Case Study

Unified weight	Logistic regression
Dario Colazzo*	Nicoleta Preda
Yannis Papakonstantinou	Stamatis Zampetakis*
Damian Bursztyn	Andrei Arion
Serge Abiteboul	Omar Benjelloun
Haris Georgiadis	Dario Colazzo*

We also conduct a case study to demonstrate the effectiveness of our method. The above table shows the top-5 COIs of *Asterios Katsifodimos* among the PC members of SIGMOD 2016. Symbol * indicates this author is in the *following* list of *Asterios Katsifodimos* on ResearchGate. In addition, the top-5 cases suggested by our method had worked in the same project team of *Asterios Katsifodimos*. This shows that our method has correctly learned a higher weight for meta-paths like "advisor-advisee" that are likely to lead to COIs.

5.3 Graph based Declaration System

We implemented a user-friendly interface for the COI declaration (see Figure 5). The left hand side is an interactive heterogeneous graph where the logged-in user is the center of the graph. The path relationship is shown when the user click on any node of the graph (i.e., potential COIs). In addition, the user can expand the graph (i.e., showing more neighbors) by clicking on a node. The right-top table shows some latent COI cases suggested by our auto-detection method (cf. Section 4). The user can also declare the conflict of interest cases by right-clicking on the graph, where the cases will be shown in the right-bottom table.

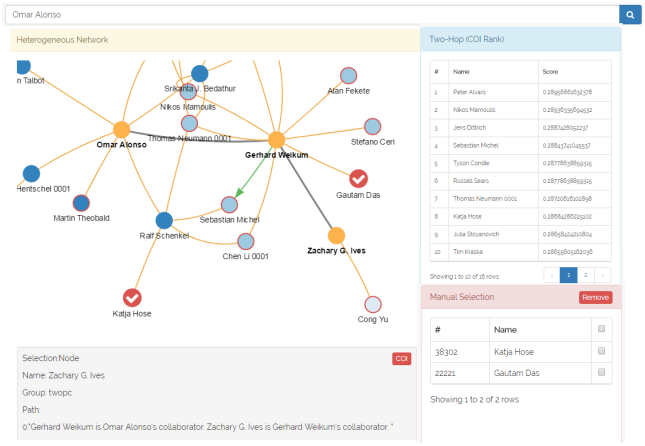


Figure 5: System Demonstration

6 CONCLUSION

In this paper, we studied a reinforcement approach that attempts to detect COIs on a heterogeneous co-authorship network in a semi-automatic manner. The reinforcement approach includes two processes, *manual declaration* and *auto detection*. The declaration system visualizes latent COI cases by an interactive graph interface that minimizes the effort of self COI declaration process. The auto detection attempts to provide an effective COI list by a logistic regression model based on meta-paths. Although we agree that it is hard to quantify the COI detection quality, this problem should draw more attention from the society since it is very important to the fairness in academic peer reviews. In the future, we aim to integrate our system into a real conference management tool.

Acknowledgements. This work has been supported in part by 61502548 from NSFC, MYRG2014-00106-FST from UMAC RC, and IIS-1553547 from NSF.

REFERENCES

[1] Boanerges Aleman-Meza, Meenakshi Nagarajan, Cartic Ramakrishnan, Li Ding, Pranam Kolari, Amit P. Sheth, Ismailcem Budak Arpinar, Anupam Joshi, and Tim Finin. [n. d.]. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In *WWW06*. 407–416.

[2] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. [n. d.]. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. In *KDD16*. 1595–1604.

[3] Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. [n. d.]. On Good and Fair Paper-Reviewer Assignment. In *ICDM13*. 1145–1150.

[4] Chuan Shi, Xiangnan Kong, Yue Huang, Philip S. Yu, and Bin Wu. 2014. HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. *IEEE Trans. Knowl. Data Eng.* 26, 10 (2014), 2479–2492.

[5] Chuan Shi, Zhiqiang Zhang, Ping Luo, Philip S. Yu, Yading Yue, and Bin Wu. [n. d.]. Semantic Path based Personalized Recommendation on Weighted Heterogeneous Information Networks. In *CIKM15*. 453–462.

[6] Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. [n. d.]. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. In *ASONAM11*. 121–128.

[7] Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla. [n. d.]. When will it happen?: relationship prediction in heterogeneous information networks. In *WSDM12*. 663–672.

[8] Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. [n. d.]. Mining advisor-advisee relationships from research publication networks. In *KDD10*. 203–212.

[9] Reza Zafarani and Huan Liu. 2015. Evaluation without ground truth in social media research. *Commun. ACM* 58, 6 (2015), 54–60.