Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

Profile Matching in Heterogeneous Academic Social Networks using Knowledge Graphs*

Research Article

Sahar Rezazadeh Saatlou[1], Behshid Behkamal[2], Havva Alizadeh Noughabi[3], Davood Rafiei[4]

DOI: 10.22067/cke.2023.84559.1104

Abstract: With the increasing popularity of academic social networks, many users join more than one network to benefit from their unique features. However, matching the profiles of a user, despite being crucial for data verification and update synchronization, is challenging due to the differences in profile structures across different networks. In this paper, we propose an academic profile-matching approach that utilizes an Academic Knowledge Graph (AKG) to overcome the diversity problem in profile structures. Our approach includes three components: (1) candidate profile generation, which retrieves related profiles from the target network based on name similarity to the source profile; (2) profile enrichment, which uses AKG to discover relations between the attributes of the source and target profiles; and (3) profile matching, which selects one candidate as a matched profile. Through experiments on real-world datasets, we demonstrate that the proposed approach is effective in matching academic profiles across different networks, outperforming state-of-the-art baselines.

Keywords: Entity Matching, Heterogeneity, Academic Social Networks, Knowledge Graph.

1. Introduction

Academic Social Networks (ASN) such as ResearchGate, LinkedIn, and Google Scholar offer researchers a platform to connect and collaborate with other researchers and share their publications and research interests. With a large number of entities (e.g. publications and scholars) and relationships (e.g. citations and co-authorships), ASNs are considered complex heterogeneous networks [1], [2].

A problem associated with using academic social networks is that researchers often have accounts on multiple platforms, which can result in their social behaviors being fragmented across different networks [1], [3]. Furthermore, the data is highly unstructured, and noisy. [4], [5]. Therefore, finding the same user on various platforms is a challenging task due to the inconsistency and diversity of the data. In general, User alignment refers to matching user profiles from different academic profiles that are associated with a real person [6]. As the number of accounts continues to grow, there has been increasing interest in matching user profiles across various academic social networks. Matching user profiles has the potential to synthesize users' behaviors, which can be beneficial for a large number of applications, including recommendation systems and marketing strategies [7], [8]. Accurately linking user identities across ASNs is a critical step in collecting disparate but related data about them [5].

Previous research [4], [8] primarily focused on aligning networks based on profile features or network structures. However, this posed a limitation of matching profiles when confronted with networks having disparate profile feature names and structures. In other words, since the data associated with each user profile may be different in two networks, it is difficult to accurately compare them, especially when a source profile matches more than one profile in the target network [1], [3]. In response to this gap, we introduce a semantic approach that leverages a knowledge graph. This novel method overcomes the challenge of aligning networks with different feature names and structures. For example, consider a researcher whose field of study is listed as Computer Science in the source network, but the profile of the same researcher in the target network only mentions research keywords and not the field of study. As shown in Figure 1, finding a match between the two profiles hinged on detecting the field of study of the researcher based on the research keywords. To address this, we rely on the Academic Knowledge Graph (AKG) [9]. The idea is to leverage AKG and extract a field of study attribute for the target profile, making it more comparable with the source profile. Our hypothesis is that this can improve the performance of profile matching and help to establish a stronger relationship between the two distinct fields. In general, AKG enables the discovery of connections between corresponding attributes of profiles in different networks.

Our proposed profile-matching approach is based on the hypothesis that leveraging the Academic Knowledge Graph (AKG) can enhance profile-matching performance

* Manuscript received: 2023 September 21, Revised, 2023 October 27, Accepted, 2023 December 10.

[1] MSc Student, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
[2] Corresponding author. Associate Professor, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
Email: behkamal@um.ac.ir
[3] PhD Candidate, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
[4] Professor, Department of Computer Science, University of Alberta, Edmonton, Canada.

and strengthen the relationship between two distinct fields. The approach comprises three main components: (1) candidate profile generation, which involves retrieving related profiles from the target network based on name similarity with the source profile; (2) profile enrichment, which uses a domain-specific knowledge graph to discover the connections between the attributes of the source and target profiles; and (3) profile matching, which selects a candidate as the matched profile. The key contributions of our proposed method are:

1. Introducing an academic profile-matching method that overcomes the diversity and heterogeneity problem in profile structures by exploiting the knowledge graph.
2. Leveraging a domain-specific Knowledge Graph to discover semantic relationships between two profiles.
3. Demonstrating a significantly high accuracy in matching academic profiles across heterogeneous academic social networks based on experiments conducted on real-world datasets and compared to the state-of-the-art baselines.

The rest of this paper is organized as follows. Related work is discussed in the next section, our approach is introduced in Section 3. Experimental evaluation is discussed in Section 4, and the discussion is explained in Section 5. Finally, concluding remarks are given in Section 5.

2. Related works

Identifying users over cross-social networks plays an important role in many research areas, including cyber security, and information retrieval [10]. The latest developments in the area [11]–[13] can be categorized into (1) user profile-based methods, (2) user-generated content-based methods, and (3) network structure-based methods, which are illustrated in Figure 3.

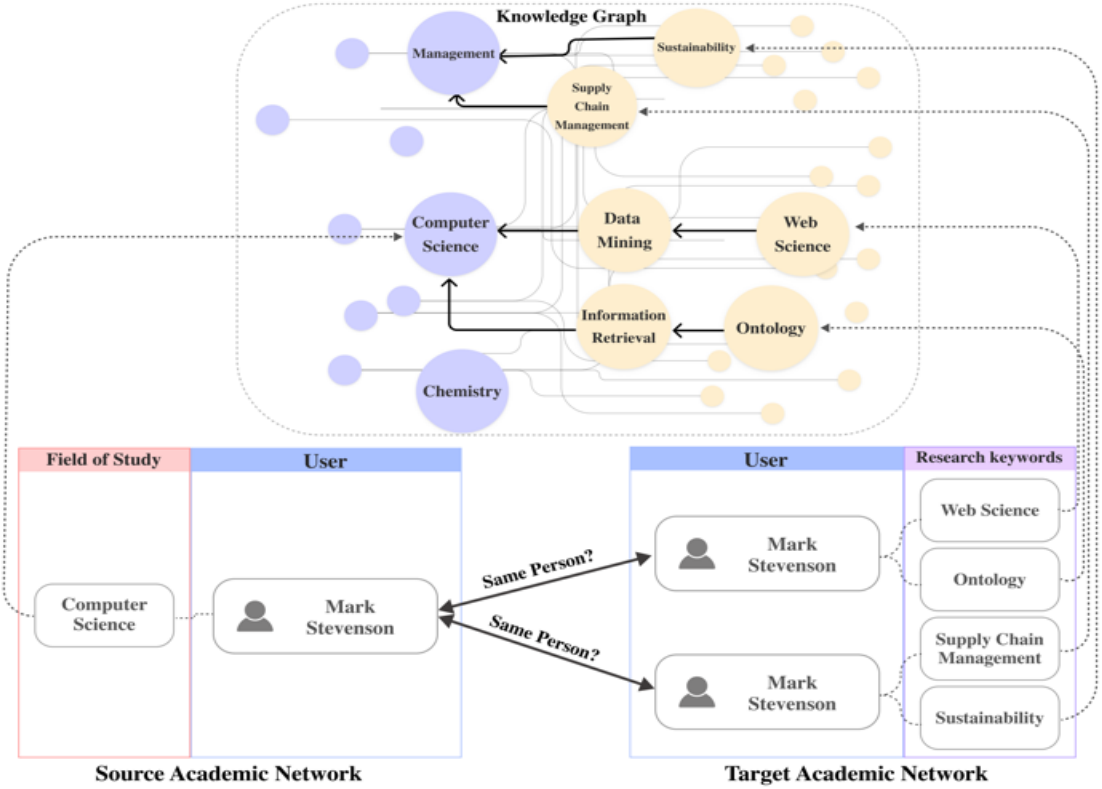

Figure 1. An academic profile matching scenario, showing the problem of matching heterogeneous profiles

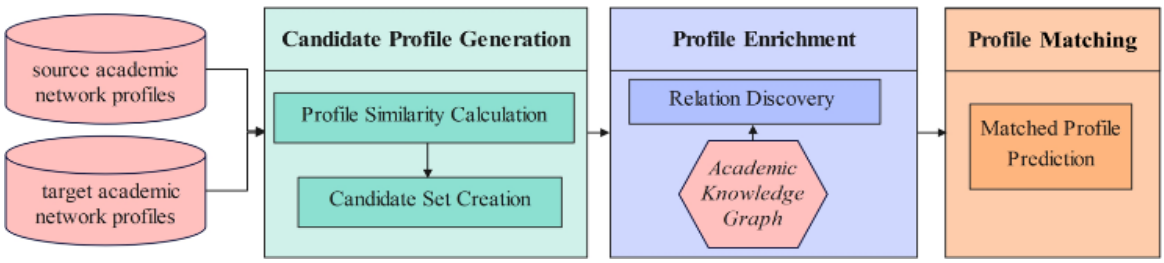

Figure 2. An academic profile matching scenario, showing the problem of matching heterogeneous profiles

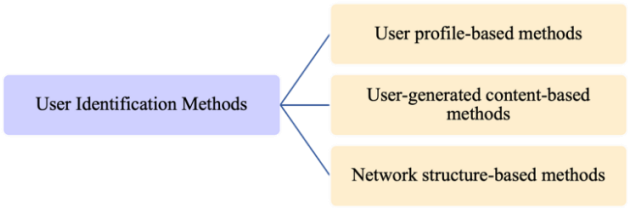

Figure 3. User identification methods

2.1. User profile-based methods

The first group focuses on exploiting user information from profile attributes [14]. Using profile information such as usernames, genders, and birthdays of users is the simplest way to facilitate user identity linkage (UIL) problems. Among all profile attributes, the username is a dominant attribute for user identification. In particular, Liu et al. concentrated on determining if cross-site user accounts with the same usernames belong to the same user [15]. Perito et al. proposed a new approach based on usernames to align multiple online profiles of the same user across different sites [16]. Li et al. examined variations in naming patterns across various social networks, developed features to leverage information redundancies, and utilized supervised machine learning to detect and classify users [17]. However, usernames are not always available or reliable sometimes they are numeric strings automatically assigned to the profiles [18]. Therefore, the username with other profile attributes can be combined to improve identification accuracy. Motoyama et al. demonstrated that accounts can be matched based on their profile attributes such as occupation, university, and gender [19]. Sharma et al. proposed a new profile-based method for aligning users by using features like username, name, bio, and profile image which are publicly available on different platforms like Instagram, Twitter, and Google+ [7]. Nahuel et al. introduced Ogmios, a new method for matching academic profiles based on features such as name, research keywords, and affiliation [20]. Nevertheless, profile inconsistency is the main challenge in this group of methods, as each network has a different profile structure. In addition, some online social networks allow users to selectively maintain some profile attributes (e.g., age or contact information) as private and others as available to the public. Also, the same attribute can be filled with different information depending on the network and user's purpose, e.g., location [5].

2.2. User-generated content-based methods

The second group collects user-generated content, such as temporal and spatial information and published content [20]. These methods analyze different behavioral patterns and construct models based on language and writing style [21]. Riederer et al. used an alignment technique to generate affinity scores to identify the most likely candidate pair using time-stamped location data [22]. Kong et al. utilized posted content by Foursquare and Twitter users to formulate the anchor links based on binary classifier scores [20]. Han et al. proposed a new framework that uses habit patterns extracted from geographic data generated by users to link up multiple profiles referring to the same user [23]. Yin et al. proposed a novel method that leverages social behaviors and contact graph to link IDs across domains, enabling the matching of adjacent and nonadjacent IDs based on a set matching algorithm and confidence scores [24]. Chen et al. proposed a method called HFUL to identify the spatiotemporal patterns of user check-in activity, which improved the effectiveness, efficiency, scalability, and robustness of cross-platform user linkage [25]. Gao et al. proposed a method to enhance text semantics by linking extracted named entities to a knowledge graph. This method incorporates temporal information via time decay functions, then extracts similarities with convolutional neural networks, and finally predicts user identity linkage with an attention mechanism [26]. The heterogeneous content information makes it very challenging to link user identities accurately since generated content can be in various types such as text, image, video, check-in, and more. In addition, users' behavior depends on the time and location of the contents they generate [5], [14].

2.3. Network structure-based methods

The third group utilizes network topology data, primarily considering user friendships, for determining the identity of an account [27]. Man et al. proposed a supervised network embedding model to capture the main structural regularities for predicting anchor links by the awareness of observed anchor links as the supervision [28]. Miao et al. proposed a supervised method, EUIA, which solves user identity linkage by learning the embedding of nodes in low-dimensional space [29]. Cheng et al. present the UASIP model, a novel approach that effectively captures both structural interaction and structural propagation to enhance the learning of a highly robust representation across diverse networks [30]. Li et al. studied the impacts of multiple friendship-based classifiers for user identification. They used features of the display names and also similarities of K-hop neighbors, to match user accounts from multiple networks [31]. These groups of methods have network diversity challenges which means that each social network is constructed for the user's specific objective. Therefore, each user has a subset of real-world connections based on the network they use. In addition, due to large-scale network structures and users' privacy issues, structure-based methods are challenging to match user entities [5].

Overall, while each group of methods has its strengths and weaknesses, matching user entities across social networks remains a challenging task. Table 1 shows different papers comparing their methods. Regarding the applications of Knowledge graphs, most papers devoting to applying KGs into specific areas have put their interests on question answering systems, recommender systems, and information retrieval systems [32]. In this paper, we propose a novel approach to incorporate KG for academic profile enrichment.

Table 1. Comparison of related works

Reference	Method	Proposed Approach	Published Year	Dataset	Evaluation Metrics
[19]	User profile methods	Counts the common words in their profile attributes, treating the attributes as bags of words	2009	Facebook, Myspace	Accuracy
[16]	User profile methods	Calculates the similarity between profile names to match accounts	2011	Google, eBay	Recall, Precision
[15]	User profile methods	Links accounts across multiple platforms utilizing username	2013	About.me	Accuracy, Precision, Recall, F1
[22]	User-generated content methods	Links users' identities, trajectory-based content features and timestamped location data.	2016	Foursquare, Twitter, Instagram	Precision, Recall
[17]	User profile methods	Checks naming patterns across social networks using profiles features	2017	Facebook, Twitter, Foursquare	Accuracy, Precision, Recall, F1, False Positive Rate
[23]	User-generated content methods	Converts user-generated geographic coordinates into semantic location words, uses Latent Dirichlet Allocation (LDA) to form user themes, and determines account identity based on trajectory co-occurrence frequency.	2017	Sina Weibo, Tencent weibo, Weixin, Renren	Accuracy, Precision, Recall, F1
[7]	User profile methods	Proposes a three-step method that involves extracting features from user profiles, optimizing feature weights using stochastic gradient descent, and performing pair-wise and multi-platform linking of user profiles.	2018	Instagram, Twitter, and Google+	Accuracy
[18]	User-generated content methods	Measures the similarity of user-generated content (UGC) in terms of space, time, and content dimensions, along with the application of supervised machine learning algorithms for user matching.	2018	Foursquare, Facebook, Twitter	Accuracy, Precision, Recall, F1, AUC
[30]	Network structure methods	Captures structural interaction and propagation for robust representation learning across diverse networks.	2018	Foursquare, Twitter	Precision, MAP
[31]	Network structure methods	Investigates the effects of friendship-based classifiers on user identification by utilizing display name features and K-hop neighbor similarities to match user accounts across multiple networks.	2020	Facebook, Foursquare, Twitter	Accuracy, Precision, Recall, F1
[33]	User profile methods	Introduces a novel method for matching academic profiles by considering features like name, research keywords, and affiliation to calculate their similarity.	2020	Málaga University, Google Scholar, ResearchGate	Precision, Recall, F1
[24]	User-generated content methods	Proposes a novel method that leverages users' social behaviors and contact graph to link IDs across domains, using a set matching algorithm to identify candidate IDs and select the best match based on confidence scores.	2022	Foursquare, Twitter	Precision, Recall, AUC

3. Proposed Approach

The objective of our work is to provide an approach for matching academic profiles across heterogeneous academic social networks. To achieve this goal, the Academic Knowledge Graph is utilized to enrich profile data to align the most appropriate candidate profile for researchers. Our problem is defined as follows:

Problem Statement Let, G^s be a source academic network andG^t be a target academic network. Denote $P^s = \{p_1, p_2, \dots, p_n\}$ the set of academic profiles in G^s and, $P^t = \{p_1, p_2, \dots, p_m\}$ the set of academic profiles in G^t. Given a $p_i \in P^s$, we aim to identify the most relevant profile $p_j \in P^t$ as a matched profile for p_i. We are particularly interested in the setting where P^s and P^t are heterogeneous

Our proposed approach uses a knowledge graph to bridge the gap between two networks. Figure 2 provides an overview of our proposed approach, which is broken down into (1) candidate profile generation, (2) profile enrichment, and (3) profile matching. In the following sections, we discuss the details of those components.

3.1. Candidate Profile Generation

The goal of the candidate profile generation component is to select a set of promising candidates from P^t, denoted by $C(i)$, as potential matches for a given profile $p_i \in P^s$. One characteristic that is expected to remain constant across the two profiles is the user's name. Since individuals on academic networks usually do not conceal their names, candidate profiles can be selected based on the similarity of their names. In our case, the set (i) includes every profile from P^t with a name similarity less than a threshold α, i.e.

$$C(i) = \{p_j \epsilon P^t | name\_similarity(p_i, p_j) > \alpha\}$$

3.2. Profile Enrichment

Academic networks often have different profile structures, making it crucial to match the corresponding parts of each profile precisely. For instance, a researcher's profile in one network (say source) may include their field of study, while their profile in another network (say target) may feature their research keywords. To match the two profiles, it is necessary to predict a user's field of study based on a set of research keywords. The profile enrichment component can use an academic knowledge graph assuming the desired fields are available in KG. There are two possible scenarios:

1. If there is a node in the KG that matches the research keyword in the target profile, then all paths in the KG starting from that node and leading to a field of study are considered, by navigating the graph paths to higher-level nodes.
2. If no node in the KG matches the research keyword in the target profile, the similarity is determined using the BERT embedding of research keywords in both the target profile and KG. Once the most similar research keywords in the KG are identified, the corresponding field of study can be obtained by navigating the path between the nodes in the KG. Each profile may have multiple research keywords, and each keyword may lead to several fields of study in the KG.

We convert knowledge graph concepts into dense vectors by leveraging the contextual understanding capabilities of BERT, allowing us to capture rich semantic information. By encoding keywords related to each other into vector representations and employing cosine similarity, we efficiently identify the most semantically similar concepts within the knowledge graph. This method enhances the retrieval and exploration of related concepts in knowledge graph analysis.

It is possible to reach one field of study from different research keywords, and the frequency of a field of study is also considered. Algorithm 1 details the process of obtaining a set of fields of study based on the research keywords in a profile. Finally, the list of fields of study is sorted based on their frequencies, and the top-K fields of study are deemed as the fields of study of the user. While our discussion here is centered around the field of study and research keywords, the approach is applicable to other fields, provided a relationship can be established in the KG.

Algorithm 1. Finding field studies corresponding to the research keywords' profile

input: a set of research keywords R in the profile, an academic knowledge graph named AKG
output: a set of field studies

```
field_studies = EmptyDictionary;
For r in R do
    r_field_studies = AKG.high_level_node(node=r)
    For f_s in r_field_studies do
        field_studies [f_s] = field_studies.get(key= f_s) + 1
    end
    end
return field_studies
```

3.3. Profile Matching

Given a source profile $p_i \in P^s$ s and a set (i) of candidate profiles, our objective is to find the most relevant $p \in C(i)$ as the matched profile. Since the number of candidate profiles is typically more than one, it is imperative that we investigate these candidates and select only one, assuming that each profile in the source network matches at most one profile in the target network. Algorithm 2 describes the details of making this section based on the field of study, assuming other fields match. If more than one candidate profile matches the field of study withp_i, the algorithm returns with no match.

4. Experimental evaluation

In this section, we focus on key aspects of our study. Firstly, we provide an overview of the dataset used in this paper. Following that, we delve into the Academic Knowledge Graph, shedding light on its structure. We then proceed to outline the experimental setup used for our research.

4.1. Dataset

In our experiments, the academic source networks were Ferdowsi University of Mashhad (FUM) (http://scimet.um.ac.ir/) and Shahid Beheshti University (SBU) (http://scimet.sbu.ac.ir/), which consisted of 811 and 895 researcher profiles respectively. The profiles included the researcher's first and last name, department, personal page link, and field of study. Our target network was obtained from Google Scholar and included the names, affiliations, and research keywords for each researcher. We collected these profiles by searching for researchers' names in Google Scholar and retrieving all profiles with similar names that were returned.

4.2. Academic Knowledge Graph

In Section 2.2, we discussed the utilization of the Academic Knowledge Graph within the profile enrichment component of our proposed approach. We employed the Microsoft Academic Knowledge Graph (MAKG) for this objective, obtaining it in the form of a large RDF dataset. The MAKG offers hierarchical associations among scientific concepts, such as research keywords and fields of study that hold relevance to our ongoing research.

Each individual concept is linked to another concept within the higher hierarchical level, ranging from 1 to 5. Figure 4 illustrates the relations between different concepts in MAKG.

For instance, Computer Science is one of the fields of study and Artificial Intelligence is a concept related to it.

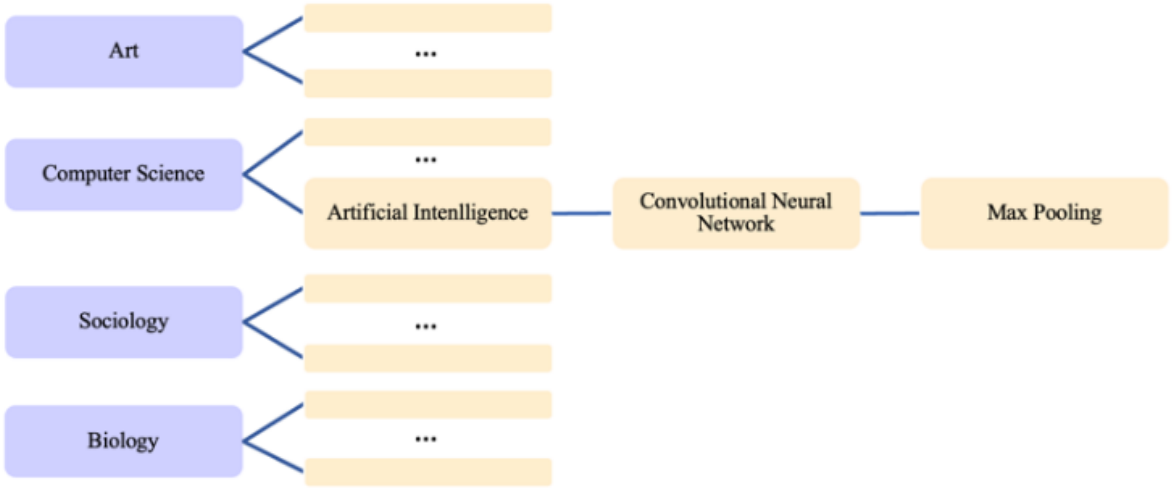

Figure 4. Hierarchical relationships between scientific concepts in the MAKG

At the time of conducting this research, the MAKG comprises a total of 714,553 distinct concepts. There are 19 concepts in the first level including Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental science, Geography, Geology, History, Materials science, Mathematics, Medicine, Philosophy, Physics, Political science, Psychology, and Sociology. The aforementioned concepts are designated as fields of study, while the remaining concepts serve as research keywords associated with these fields of study. A "hasparent" predicate links each scientific concept to its parent in the dataset. Table 2 shows the distribution of concepts utilized in our approach at different levels.

Table 2. The distribution of concepts in MAKG

Level	# Concepts
0	19
1	264
2	119707
3	124738
4	135578
5	16668

4.3. Experimental Setup

For computing the name similarity between the two profiles, as discussed in Section 2.1, we used the Levenshtein distance, to narrow down the candidate set to only those profiles whose names are similar to the source profile. Levenshtein distance, also known as edit distance, is used to measure the similarity between two strings by calculating the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other. It can be adapted for various applications, including comparing the similarity of two people's names, addresses, or other text-based attributes. We set the similarity threshold α to 0.75 in Equation 1, inspired by the choice made in the base paper, Ogmios [33]. Among various values tested, 0.75 consistently yielded optimal results in our specific context. As discussed in Section 2.2, we defined the field of study of the candidate profile as the top-K fields of study from the sorted list of extracted fields of study using Algorithm 2. After testing different values for the variable K, we set the K to 3 in our experiments. Our observation revealed that the majority of results demonstrated the highest similarity to the source network profile within the top 3 profiles. This choice aligns with the optimal performance of our approach.

Algorithm 2. profile matching

input: a profile from source academic network p_i,
a set of candidate profiles from the target academic network C(i)
output: a matched profile to p_i
source_study_field = get_study_field(p_i);
matched_profile_set = [];
For p_j **in** C(i) **do**
  study_field_set = get_study_field(p_j)
  For fs **in** study_field_set **do**
    if fs == source_study_field **do**
      matched_profile_set.append(p_j)
      next;
    end
  end
end
if count(matched_profile_set) == 1 **do**
  return p_j
end
else
  p_i remains unmatched;
end

4.4. Evaluation Metrics

In our study, we employ a set of three evaluation metrics, namely recall, precision, and F-Measure. These established performance benchmarks are utilized for the assessment of our system's operational effectiveness. Metrics are defined as follows:

$$Precision = \frac{tp}{(tp + fp)}$$

$$Recall = \frac{tp}{(tp + fn)}$$

$$F1 = 2 \times \frac{Precision \times Recall}{(Precision + Recall)}$$

5. Discussion

After conducting experiments we present an analysis of the results obtained from heterogeneous networks and homogeneous networks. Lastly, we conducted an error analysis to identify and examine any potential discrepancies or limitations encountered during our study.

5.1. Evaluation Result on Heterogeneous Networks

In this section, we present the result of our evaluation by comparing our approach with BERT, which stands for "Bidirectional Encoder Representations from Transformers". BERT is a natural language processing (NLP) model developed by Google AI researchers. It has had a significant impact on various NLP tasks and has become a fundamental building block for many state-of-the-art NLP models. Since the release of BERT, numerous variants and pre-trained models have been developed, each fine-tuned for specific NLP tasks, such as text classification, named entity recognition, question-answering, and more. These models have significantly advanced the state of the art in NLP and have become the foundation for various applications in NLP understanding and generation [34].

Our evaluation is conducted on heterogeneous profiles across academic networks, where the source profile schema includes the 'name' and 'field of study', and the target profile schema includes the 'name' and 'research keywords'. As depicted in Table 3, the performance of BERT, in terms of both precision and recall, in matching heterogeneous profiles was close to 30%. The reason for this low performance is that BERT cannot establish strong semantic relationships between academic concepts, such as the field of study, and keywords in our dataset, which are located at different abstraction levels. However, BERT is effective in cases where a relationship cannot be established between the concepts using the knowledge graph, as illustrated in the last two rows of Table 3.

5.2. Evaluation Result on Homogenous Networks

In this section, we demonstrate that our proposed method is also applicable to homogeneous academic profiles. For this evaluation, we compared our approach to Ogmios [33], a recently proposed method for comparing homogenous profiles. We downloaded a set of research keywords from our source networks for this comparison. These keywords were obtained by the source networks based on the tag cloud of keywords extracted from the papers of the researchers. The Ogmios framework matched profiles based on the similarity between different attributes, including name, affiliation, and research keywords. In our implementation of Ogmios, we used a customized score function based on the intersection of research keywords in source and target profiles. Our performance comparison in terms of precision, recall, and F-score is reported in Table 4. Our approach aligns significantly more profiles than the customized Ogmios method because our method explores a knowledge graph to enrich research keywords instead of using them directly. With a precision of 0.93 and 0.94 for two datasets, our approach can match profiles correctly in most cases. Additionally, in at least 54% of the cases, our approach can retrieve the matched profiles from the target network.

5.3. Error Analysis

In this section, we study the situations in our observation that lead to unsuccessful matching. As explained in section 2.2, to overcome the heterogeneity problem in profile matching, we used a knowledge graph, i.e., MAKG. Specifically, the research keywords on profiles can be used to extract the field of study by MAKG. For example, a researcher named "Jim Hendler" with his research keywords given, found the correct corresponding field of study, as shown in Table 5. The research keywords are Semantic Web, Artificial Intelligence, Web Science, Ontology, and Knowledge Graph. Table 5 shows the fields of study results for each keyword using MAKG. These fields of study are varied based on the observation times, that is, 9 for Computer Science, 3 for Mathematics, and 2 for Philosophy. Therefore, the top-1 field of study for this researcher is Computer Science, which is equivalent to the correct field of study of "Jim Hendler", i.e., Computer Science.

Table 3. Performance compared to BERT on matching heterogeneous profiles

	FUM			SBU		
	Precision	Recall	F1	Precision	Recall	F1
BERT	25%	25%	25%	35%	35%	35%
Proposed Method without BERT	75%	39%	51%	80%	42%	58%
Proposed Method with BERT	76%	48%	58%	88%	49%	61%

Table 4. Performance compared to Ogmios on matching homogeneous profiles

	FUM			SBU		
	Precision	Recall	F1	Precision	Recall	F1
Ogmios	99%	20%	31%	99%	16%	27%
Proposed Method without BERT	83%	32%	46%	82%	35%	47%
Proposed Method with BERT	93%	55%	66%	94%	54%	65%

To better understand the failure cases of our approach, we took a random sample set of profiles where our approach failed to match. We identified some remaining challenges for our approach.

To begin with, unsuccessful matching often occurs due to some research keywords belonging to more than one field of study. Table 6 shows an example that the proposed approach makes a wrong prediction for the field of study of the researcher. This happens when research keywords in two field studies, e.g., Mathematics and Computer Science, overlap. In detail, the set of research keywords includes Computational Geometry and Algorithm Design, and the correct field of study is Computer Science. Table 6 shows that the top-1 field of study based on Algorithm 1 will be Mathematics. However, such cases can affect the performance of our proposed method. One solution to solve this problem can be to consider the top-K fields of study instead of the top-1.

The second reason for unsuccessful matching is that, some researchers used abbreviated research keywords, and in most cases, there was no corresponding node for them in MAKG. Third, some keywords had different meanings in various fields of study, which caused the proposed approach to extract an unrelated field of study for the keywords. Table 7 shows another example, a researcher in Aerospace Engineering has some ambiguous keywords like using the abbreviated keyword "MSDC" or a keyword with general meaning like "Turbulence" which leads to unrelated fields of study in the result.

Table 5. Finding field of study based on research keywords using MAKG

Research Keyword	Related path in MAKG
Semantic Web	Semantic Web → Ontology → Information Retrieval → Computer Science
	Semantic Web → Ontology → Epistemology → Philosophy Semantic Web→Ontology → Data Mining → Computer Science
Artificial Intelligence	Artificial Intelligence → Computer Science
Web Science	Web Science → Data Mining → Computer Science Web Science → World Wide Web → Computer Science Web Science → Data Science → Computer Science
Ontology	Ontology → Information Retrieval → Computer Science
	Ontology → Epistemology → Philosophy Ontology → Data Mining → Computer Science
Knowledge Graph	Knowledge Graph → Graph → Algorithm → Mathematics
	Knowledge Graph → Graph → Algorithm → Computer Science Knowledge Graph → Graph → Combinatorics → Mathematics Knowledge Graph → Graph → Discrete Mathematics → Mathematics

Table 6. An example of an unsuccessful matching of research keyword and field of study

Research Keyword	Related path in MAKG
Computational Geometry	Computational geometry → Geometry → Mathematics
	Computational geometry → Algorithm → Mathematics Computational geometry → Algorithm → Computer Science Computational geometry → Combinatorics → Mathematics
Algorithm Design	Algorithm Design → Algorithm → Mathematics
	Algorithm Design → Algorithm → Computer Science Algorithm Design → Machine Learning → Computer Science Algorithm Design → Mathematical Optimization → Mathematics

Table 7. An example of unrelated fields of study for different keywords

Research Keyword	Related path in MAKG
MSDC	MSDC→ Electronic engineering MSDC→Operating system→Computer science MSDC→Visual arts→Art
Direct simulation Monte Carlo	Direct simulation Monte Carlo→Dynamic Monte Carlo method→Monte Carlo method→Statistics→Mathematics
Rarefied Gas Dynamics	Metafluid dynamics→Classical mechanics→Physics Metafluid dynamics→Quantum electrodynamics→Physics Metafluid dynamics→Quantum mechanics→Physics
Cavitation	Cavitation→Composite material→Materials science Cavitation→Mechanics→Physics Cavitation→Acoustics→Physics
Turbulence	Turbulence→Flow (psychology) →Psychotherapist→Psychology Turbulence→Flow (psychology) →Social psychology→Psychology

6. Conclusion

Matching user profiles across multiple ASNs is a crucial task with many application areas, such as recommendation systems and link prediction. This paper proposes an academic profile-matching approach that effectively addresses the challenges of profile structure diversity by leveraging an Academic Knowledge Graph (in our case, MAKG).

The proposed approach consists of three key components: candidate profile generation, profile enrichment, and profile matching. Our experimental results on two real-world datasets from different sources demonstrate that the proposed approach achieves very strong performance in matching academic profiles compared to our baselines.

Moving forward, our research is focused on two main directions. Firstly, we aim to provide an explanation for our profile-matching approach by demonstrating the paths between a source node and a target node in the knowledge graph. Secondly, we plan to enhance MAKG by adding missing concepts. These future directions will help further improve our approach's effectiveness and explainability and support its broader applicability.

7. References

[1] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *Journal of Network and Computer Applications*, vol. 132. Academic Press, pp. 86–103, Apr. 15, 2019. doi: 10.1016/j.jnca.2019.01.029.

[2] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data: A Survey," *IEEE Trans Big Data*, vol. 3, no. 1, pp. 18–35, Jan. 2017, doi: 10.1109/tbdata.2016.2641460.

[3] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, 2014, pp. 51–62. doi: 10.1145/2588555.2588559.

[4] L. Wang, K. Hu, Y. Zhang, and S. Cao, "Factor Graph Model Based User Profile Matching across Social Networks," *IEEE Access*, vol. 7, pp. 152429–152442, 2019, doi: 10.1109/ACCESS.2019.2948073.

[5] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User Identity Linkage across Online Social Networks: A Review." [Online]. Available: http://www.pewinternet.org/2015/01/09/

[6] Y. Huang, P. Zhao, Q. Zhang, L. Xing, H. Wu, and H. Ma, "A Semantic-Enhancement-Based Social Network User-Alignment Algorithm," *Entropy*, vol. 25, no. 1, Jan. 2023, doi: 10.3390/e25010172.

[7] V. Sharma and C. Dyreson, "Linksocial: Linking user profiles across multiple social media platforms," in *Proceedings - 9th IEEE International Conference on Big Knowledge, ICBK 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 260–267. doi: 10.1109/ICBK.2018.00042.

[8] Y. Li, W. Ji, X. Gao, Y. Deng, W. Dong, and D. Li, "Matching user accounts with spatio-temporal awareness across social networks," *Inf Sci (N Y)*, vol. 570, pp. 1–15, Sep. 2021, doi: 10.1016/j.ins.2021.04.030.

[9] M. Färber, "The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data", doi: 10.5281/zenodo.2159723.

[10] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A User Identification Algorithm Based on User Behavior Analysis in Social Networks," *IEEE Access*, vol. 7, pp. 47114–47123, 2019, doi: 10.1109/ACCESS.2019.2909089.

[11] A. T. Hadgu, J. Kumar, and R. Gundam, "Learn2Link: Linking the Social and Academic Profiles of Researchers," 2020. [Online]. Available: www.aaai.org

[12] Y. Qu, L. Xing, H. Ma, H. Wu, K. Zhang, and K. Deng, "Exploiting User Friendship Networks for User Identification across Social Networks," *Decis Support Syst*, vol. 14, no. 1, Jan. 2022, doi: 10.3390/sym14010110.

[13] R. Wang, H. Zhu, L. Wang, Z. Chen, M. Gao, and Y. Xin, "User identity linkage across social networks by heterogeneous graph attention network modeling," *Applied Sciences (Switzerland)*, vol. 10, no. 16, Aug. 2020, doi: 10.3390/app10165478.

[14] X. Chen, X. Song, G. Peng, S. Feng, and L. Nie, "Adversarial-Enhanced Hybrid Graph Network for User Identity Linkage," in *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2021, pp. 1084–1093. doi: 10.1145/3404835.3462946.

[15] Stefano. Leonardi, ACM Digital Library., Association for Computing Machinery. Special Interest Group on Information Retrieval., H. and Web. Association for Computing Machinery. Special Interest Group on Hypertext, Association for Computing Machinery. Special Interest Group on Knowledge Discovery & Data Mining., and Association for Computing Machinery. Special Interest Group on Management of Data., "Hon, What's in a name?: An unsupervised approach to link users across communities," p. 798, 2013.

[16] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How Unique and Traceable Are Usernames?," *Proceedings of 11th International Conference on Privacy Enhancing Technologies, 2011, pp. 1–17.*.

[17] Y. Li, Y. Peng, W. Ji, Z. Zhang, and Q. Xu, "User Identification Based on Display Names Across Online Social Networks," *IEEE Access*, vol. 5, pp. 17342–17353, Aug. 2017, doi: 10.1109/ACCESS.2017.2744646.

[18] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Generation Computer Systems*, vol. 83, pp. 104–115, Jun. 2018, doi: 10.1016/j.future.2018.01.041.

[19] C. Yong. Chan, ACM Digital Library, H. and Web. Association for Computing Machinery. Special Interest Group on Hypertext, and Association for Computing Machinery. Special Interest Group on Information Retrieval, *I Seek You: Searching and Matching Individuals In Social Networks*. ACM, 2009.

[20] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in

International Conference on Information and Knowledge Management, Proceedings, 2013, pp. 179–188. doi: 10.1145/2505515.2505531.

[21] C. Shi and R. Duan, "Multiresolution Mutual Information Method for Social Network Entity Resolution," in *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 240–247. doi: 10.1109/ICDMW.2015.94.

[22] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *25th International World Wide Web Conference, WWW 2016*, International World Wide Web Conferences Steering Committee, 2016, pp. 707–719. doi: 10.1145/2872427.2883002.

[23] L. W. S. X. G. L. and D. Z. Xiaohui Han, "Linking social network accounts by modeling user spatiotemporal habits," in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2017.

[24] Z. Yin, Y. Yang, and Y. Fang, "Link User Identities Across Social Networks Based on Contact Graph and User Social Behavior," *IEEE Access*, vol. 10, pp. 42432–42440, 2022, doi: 10.1109/ACCESS.2022.3165568.

[25] W. Chen, W. Wang, H. Yin, L. Zhao, and X. Zhou, "HFUL: A Hybrid Framework for User Account Linkage across Location-Aware Social Networks," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.02830

[26] H. Gao, Y. Wang, J. Shao, H. Shen, and X. Cheng, "User Identity Linkage across Social Networks with the Enhancement of Knowledge Graph and Time Decay Function," *Entropy*, vol. 24, no. 11, Nov. 2022, doi: 10.3390/e24111603.

[27] Y. Qu, H. Ma, H. Wu, K. Zhang, and K. Deng, "A Multiple Salient Features-Based User Identification across Social Media," *Entropy*, vol. 24, no. 4, Apr. 2022, doi: 10.3390/e24040495.

[28] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict Anchor Links across Social Networks via an Embedding Approach."

[29] Q. Miao, L. Wang, D. Duan, X. Guo, and X. Li, "Embedding Based Cross-network User Identity Association Technology," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, ICST, 2019, pp. 138–143. doi: 10.1145/3316551.3316571.

[30] A. Cheng, C. Y. Liu, C. Zhou, J. Tan, and L. Guo, "User Alignment via Structural Interaction and Propagation," in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., Oct. 2018. doi: 10.1109/IJCNN.2018.8489228.

[31] Y. Li, Z. Su, J. Yang, and C. Gao, "Exploiting similarities of user friendship networks across social networks for user identification," *Inf Sci (N Y)*, vol. 506, pp. 78–98, Jan. 2020, doi: 10.1016/j.ins.2019.08.022.

[32] X. Zou, "A Survey on Application of Knowledge Graph," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Apr. 2020. doi: 10.1088/1742-6596/1487/1/012016.

[33] N. Verdugo, E. Guzmán, and C. Urdiales, "Integrating researchers' scientific production information through Ogmios," *Knowl Inf Syst*, vol. 62, no. 11, pp. 4199–4222, Nov. 2020, doi: 10.1007/s10115-020-01479-8.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805