Investigating Community Detection in Arabic Scholarly Network Using Ontology-based Semantic Expansion

Sarah Al-Shareef, Rahaf Alharbi, Rawan Alharbi, Raghad Almfarriji,
Maram Alsharif, Rasha Alharthi, Lamia Althaqafi
Department of Computer Science
Umm AlQura University
Makkah, Saudi Arabia
saashareef@uqu.edu.sa, rahaf.awwad@gmail.com, rawanalrehaili1420@gmail.com, raghadsm1999@gmail.com
marammajid1999@outlook.com, alharthi.rasha99@gmail.com, lfalthaqafi@gmail.com

***Abstract*—Clustering researchers in communities is an important task to support a range of techniques for analyzing and making sense of the research environment and helps researchers find people in the same field of interest to collaborate. In computer science, ontology is commonly used to capture knowledge about a particular area using relevant concepts and relations. This study investigates the use of overlapping community detection algorithms on a multilayered Arabic scholarly network to detect communities of researchers who share their research interests. Two researchers can share an interest if they co-authored a publication or share some keywords in their publications. The set of keywords is expanded via semantic search within a cross-domain ontology, e.g. DBpedia, allowing more researchers with indirect relationships to be connected. A 2-layer scholarly network was constructed by retrieving the scholarly data of faculty members from three colleges at Umm AlQura University (UQU) with rich Arabic publications. Four versions of this network were tested: unweighted, weighted, semantically expanded, and reduced semantically expanded. It was found that weights have an insignificant role in community detection within this study. In addition, a semantically expanded network does have better clustering potentials but only if was performed selectively. Otherwise, the expanded network might suffer from generic and non-discriminative keywords, making the community detection task more challenging. To our knowledge, this is the first investigation into detecting communities within an Arabic scholarly network.**

***Index Terms*—community detection, semantic annotation, Arabic scholarly data, overlapping community detection, multilayered complex network, social network analysis, DBpedia.**

I. Introduction

In nature, many objects interact in groups to form social, technological, or biological systems. Any system in the world can be described as a complex network. A community (aka a module or a cluster) is a subset of densely connected nodes within a complex network. Identification of these communities is known as community detection, and it provides a deeper understanding of the network structure and helps extract useful information from cohesive groups [1]. In the research domain, detecting communities can help improve document retrievability, assist in constructing intelligent analytics, and support a range of techniques for analyzing and making sense of the research environment. Moreover, community detection within a research institution eases collaboration, talent recruitment, and more. [2].

Scholarly networks are groups of academic entities, e.g., scholars, researchers, papers, organizations, venues, etc., linked by one or more relationships where nodes represent researchers and documents while edges represent their relationships. Academic connections and interactions among these entities can be explored using scholarly networks, including citation networks, co-authorship networks, co-citation networks, co-word networks, and hybrid networks. For instance, researchers can be connected by co-authorship or shared interests based on the keywords from all their publications. A list of topics or keywords is not optimal for detecting research communities. For example, a researcher interested in face recognition might also be interested in activity recognition or deep learning. However, if he has no publications with these exact keywords, he will not be clustered with other researchers with similar interests. One way to achieve this is to link these terms and keywords. As a result, we have used semantic search. Semantic search is defined as "search with meaning," in opposition to lexical search, which looks for exact matches of the query words or variations of them without understanding the overall meaning of the query [3]. Some authors regard semantic search as a set of methods for extracting knowledge from ontologies and other richly structured data sources found on the Semantic Web [4]. Semantic search can help improve the accuracy of detecting communities that infer the indirect relationships that detect a community from the presence of more relationships.

The Arabic language is highly inflectional and derivational, with complex morphological, grammatical, and semantic features. Therefore, processing Arabic text requires language-specific pre-processing techniques such as tokenization, normalization, and stemming [5]. Even with challenges posed by the language, doing a semantic search is still possible from

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

Arabic text.

There is minimal work in Arabic scholarly networks. To our knowledge, this is the first work that proposed detecting communities within the Arabic scholarly network.

The rest of the paper is organized as follows. Section II provides a brief review of the related work. Section III describes the techniques used in this investigation, while the construction and processing of the Arabic scholarly network are provided in Section IV. Section V gives an overview of all implemented and evaluated experiments with their results presented and discussed in Section VI. Finally, the paper is concluded in Section VII.

II. Related Work

On scholarly networks, nodes represent researchers who can be connected via several relationships, such as co-authorship, citations, and research interests. Community detection in scholarly data can be performed using either graph characteristics (graph-based) or clustering the nodes based on their attributes (clustering-based). In both approaches, topics are represented as keywords associated with academic documents where semantic relationships between topics are not considered. Several researchers attempted to include semantic similarities to enhance the community detection process. Horta et al. [6] used ontological terms and rules to introduce new relationships between researchers. They generated their ontology by combining the domain taxonomy from several scientific repositories. Cifariello et al. [7] implemented a state-of-the-art expert finding system where they calculated the semantic relatedness between expert expertise and a query topic. They generated a graph for every expert in their system, including all the topics from the expert's publications, found in the Wikipedia knowledge graph and linked by their semantic relatedness. Similarly, Zevio et al. [8] used CSO [9] in their semantic annotation to create an attributed graph for each expert. On the other hand, Osborne et al. [10] used topic clustering techniques on the keywords and their related topics derived from CSO while considering the time of publication as another dimension to identify temporal topic-based communities.

Due to the limited coverage of research repositories for Arabic research and the challenges the Arabic language poses to the current linguistic tools, there are rarely to no papers to create Arabic scholarly networks. Hence, this work can be considered the first attempt to investigate the community detection topological approach in an Arabic scholarly network.

III. Methodology

The task of this study is to investigate the use of community detection algorithms for overlapping groups of researchers who share their interests within a multilayered Arabic scholarly network. To achieve this, a 2-layer scholarly network was constructed from co-authorship and publication keywords from online resources and repositories. Then, a semantic expansion was applied to the publication keywords layer to allow the detection of more relevant communities using semantic annotation. Finally, the impact of this expansion was investigated through four overlapping community detection algorithms. This section describes the techniques above.

A. Community Detection in Multilayered Networks

Community detection aims to uncover subsets of nodes of connected communities. The community structure can generally be classified into two main categories: Partitioning (disjoint) and overlapping community structures. In the case of a partitioning community structure, a node can belong to a maximum of one community. In contrast, a node can belong to one or more communities in the overlapping community structure. Many algorithms have been developed to identify partitioning and overlapping communities with varying strengths and weaknesses, representing various approaches.

Community detection aims to uncover subsets of nodes of connected communities. In overlapping community detection, a node can belong to one or more communities. Community detection methods in multilayered networks can be grouped into three main classes based on how multiple layers are treated [11]: flattening algorithms, layer-by-layer algorithms, and multilayered algorithms. Flattening algorithms merge edges from the different layers in the multilayered network and then detect communities using traditional single-layer techniques. The number of edges can be utilized as weights in the flattened network, generating a robust network to noise. However, the resulting communities may be biased towards edges occurring on multiple layers, and the results may be more challenging to interpret due to the weights [11]. Instead of merging layers, layer-by-layer techniques detect communities in each layer separately and then process the outcomes. The final community set is structured by including actors in the same community if they are members in at least one layer due to the layer-by-layer community detection step. Because layer-specific communities are merged, these algorithms can only detect pillar (vertical) communities in theory. The final approach works directly on the multilayered network model and locates communities by exploring the multilayered data. Multilayer approaches have been developed from techniques initially developed for single-layered networks, such as density-based methods and clique percolation.

Due to the larger technical support and the structure of the collected data, this study chose a flattening approach, and four overlapping community detection algorithms were explored. Table I summarizes the included algorithms.

- **DANMF** [12] is similar to a deep autoencoder because it uses an encoder and a decoder component to learn the hierarchical mappings between the original network and its final community assignment.
- **DPClus** [14] is a community detection algorithm that uses a common neighbor technique to project weights onto an unweighted graph where the cluster begins with a single node and gradually expands by adding nodes from its neighbors. DPClus produces non-overlapping clusters and then adds nodes from their first neighbors in the original graph.

TABLE I
COMMUNITY DETECTION ALGORITHMS WITH THEIR CHARACTERISTICS USED IN THIS STUDY.

Algorithm		Network			Communities		
Name	**Package**	**Directed**	**Undirected**	**Weighted**	**Flat**	**Multilayered**	**Overlapped**
DANMF [12]	CDLib [13]	–	✓	✓	✓	–	✓
DPClus [14]		–	✓	✓	✓	–	✓
IPCA [15]		–	✓	✓	✓	–	✓
DCS [16]		–	✓	✓	✓	–	✓

- **IPCA** [15] is a modified version of DPClus. In contrast to DPClus, IPCA calculates local vertex and edge weights by counting the number of shared neighbors between two vertices. At the start of the algorithm, IPCA calculates these values only once instead of updating them each time a discovered cluster is deleted from the graph. This allows natural overlap between clusters, as cluster nodes are not removed from the graph permanently, leading to much cluster overlap.
- **DCS** [16] is a simple technique for detecting overlapping communities. It starts by removing loosely connected edges to separate the graph into modules or sub-graphs. Then, it finds local leaders in each graph module and expands these leader sets using a scoring procedure based on conductance and internal density. Finally, the local communities are merged to form a global view of each.

B. Semantic Keyword Expansion

DBpedia [17] was built on Wikipedia infoboxes from 27 distinct language editions into a single shared ontology, making it not only cross-domain but also a multilingual ontology. It comprises 768 classes that form a hierarchy and describe by 3000 different characteristics. Moreover, it is a free resource that makes it appealing for research purposes. Many tools are available to access this resource, such as DBpedia Spotlight [18] for automatically annotating DBpedia resources. These resources are used as identifiers for most domains mentioned in the text, giving a way to connect unstructured data sources to the linked open data cloud via DBpedia; however, this tool only supports a few languages, and the Arabic language is not included. Hence, Arabic text must be translated before annotation.

For querying the DBpedia resources from DBpedia Spotlight, SPARQL[1] was used, allowing querying data from databases or any other data source that can be mapped to RDF. An English-translated publication title was sent as a query via SPARQL to request all the related resources for each word in the title. After that, all the retrieved resources were saved in a list. Finally, keywords were extracted, and all the resources were added to the publication data as new keywords. Figure 1 shows an illustration of extracting semantically related keywords from the translated publication title: ”Towards the Kufic Readers”. Newly added keywords are kept in English and preprocessed by removing all punctuation except for ”_”.

[1] *http://dbpedia.org/sparql*

Fig. 1. Extracting semantically related keywords to a publication title using DBpedia Spotlight.

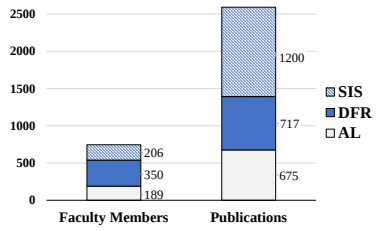

Fig. 2. The distribution of faculty members and publications among the chosen three colleges. Total faculty members = 745 and total publications = 2592.

IV. DATA: UQU NETWORK

Three colleges with rich Arabic publications were chosen from UQU: Arabic Language (AL), Da’wah and Fundamentals of Religion (DFR), and Sharia and Islamic Studies (SIS). For each faculty member, ParseHub[2] was used to extract their name, email, department and college from the university website[3] and their publication information from Dar Almandumah online database[4]. For each publication, the following data items were extracted: title, keywords, authors, affiliation, and scientific degree. ParseHub collects data from multiple pages and stores that data automatically in Excel format after removing HTML tags.

In total, 745 faculty members and 2592 publications were extracted. Figure 2 shows the distribution of faculty members and publications among the three colleges. This list was automatically checked to ensure that all authors in the publication dataset are faculty members at UQU.

A. Data Preprocessing

The order of the author’s name associated with the publication set differs from the extracted information from the UQU website. Hence, all authors’ names were rearranged

[2] *http://www.parsehub.com*
[3] *http://www.uqu.edu.sa*
[4] *http://mandumah.com*

Fig. 3. An example of data preprocessing steps applied on both author and publication sets. Substitutions are colored and underlined.

with the removal of all noisy tags. Characters with different shapes were normalized into one to reduce ambiguity, and all diacritics and punctuation were removed. Finally, each keyword is considered a separate entry. An example of processing publication information is listed in Figure 3.

Many authors share their first and last names and are sometimes affiliated with the same institution or department, causing huge ambiguity. This makes it challenging to guarantee that each researcher is linked with their publications. Consequently, ambiguity was resolved through several validation steps, and only 705 authors with resolved ambiguity were kept and linked to their publications.

B. Network Construction

Generally, multilayered networks are defined as (A, L, V, E), where A represents a set of actors, L represents a set of layers, and (V, E) represents a graph on $V \subseteq A \times L$. It is constructed by combining several layers of subnetworks. Subnetworks are all made up of the same actors, while each subnetwork is connected by a single type of edge. A 2-layer network was constructed from authorship and keywords layers for this study. Actors are fixed in both layers representing researchers. Researchers are connected with weighted and undirected edges in the authorship layer whenever they co-authored a publication. The weight equals the count of publications that the two researchers co-authored. In the keywords layer, two researchers are connected with an undirected edge when the exact keyword appears in any of their publications and is weighted by the count of the shared keyword occurrence. However, there is no interconnection across the layers. Table II lists the counts of actors and edges in each layer.

Due to the limitations of tools and packages supporting multilayers networks analysis, the two layers were constructed as individual networks. Then, each layer was converted to an adjacency matrix to sum corresponding weights and generate a flattened network that combines all two relationships in one network with the weights of all two relationships, as shown in the last row of Table II. The final weights are normalized to be within $[0, 1]$. The network construction process from the

TABLE II
UQU NETWORK. (V, E) IS THE NUMBER OF ACTORS AND EDGES, RESPECTIVELY.

Networks	Relationship	Weights	(V, E)
Authorship	co-authored a publication	#publications	(705,702)
Keywords	share the same keyword in publications	#keywords	(705,34630)
Flattened	either share a publication or a keyword	normalized sum of #'s	(705,34645)

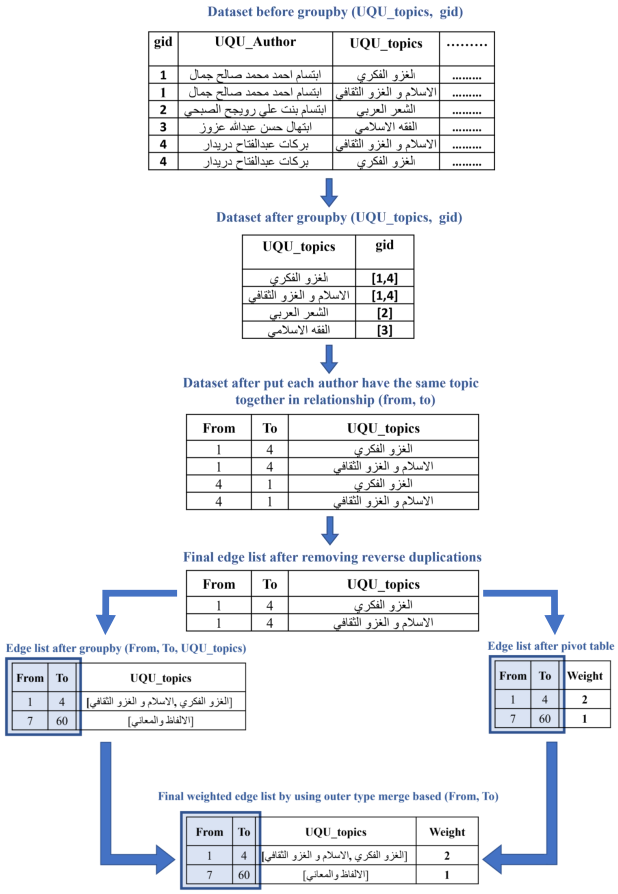

Fig. 4. A demonstration of the network construction from publication data. *gid* is a unique identifier for each researcher. The final edge list is weighted and associated with the keyword list.

publication data is demonstrated in Figure 4. The exact process was applied to construct the edge list of the authorship layer.

V. EXPERIMENTS

A. Experimental Design

All experiments were implemented and evaluated using Python and its packages. Both authorship and keywords layers were constructed using NetworkX [19] and analyzed using Multinet [20]. Community detection algorithms were applied and evaluated using NetworkX [19] and CDLib [13] packages as shown in Table I. An overview of the implementation design

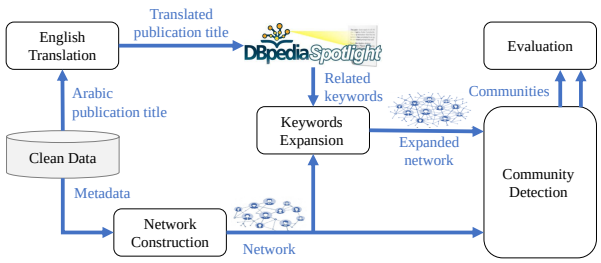

Fig. 5. The experimental design followed in this study

is depicted in Figure 5. Googletrans[5], a Python package that implements Google Translate API, was used to translate the Arabic publication title prior querying it in DBpedia Spotlight.

B. Network Construction and Semantic Expansion

For this investigation, two versions of the flattened network were implemented: weighted (F_W) and unweighted (F_U) networks. The normalized weights for F_U were ceiled to 1. Similarly, two versions of the expanded networks were implemented: weighted ($\overline{F_W}$) and unweighted ($\overline{F_U}$) networks. Table III summarizes the network characteristics of these networks in terms of average clustering coefficient, average degree, and density. There is a significant increase in the density of the semantically expanded versions over the original networks which align with the large increase in edges.

C. Evaluation

Four overlapping community detection algorithms were applied to the four constructed networks: F_U, F_W, $\overline{F_U}$, and $\overline{F_W}$. The considered measures for analyzing the outcome communities are the number of communities, the percentage of the nodes included in communities, the size of the largest community in terms of actors, and the average number of actors per community. Since the top largest communities should have comparable sizes, the ratio between the size of the second-largest community to the first-largest community is measured to indicate if the algorithm groups most of the actors together and cannot structure them into separate communities. Overlapping was measured by the average number of communities per actor, with 1 indicating no overlapping was established.

[5]*https://github.com/ssut/py-googletrans*

TABLE III
STATISTICS FOR FLATTENED UQU NETWORKS: UNWEIGHTED (F_U), WEIGHTED (F_W) AND EXPANDED NETWORKS: UNWEIGHTED ($\overline{F_U}$) AND WEIGHTED ($\overline{F_W}$). ALL NETWORKS HAVE 705 ACTORS.

Characteristics	F_W	F_U	$\overline{F_W}$	$\overline{F_U}$
Number Of Edges	34645	34645	168020	168020
Is Weighted	Yes	No	Yes	No
Average Clustering	0.07	0.76	0.02	0.83
Average Degree	8.27	98.28	16.21	476.65
Density	0.13	0.13	0.67	0.67

TABLE IV
COMMUNITY ANALYSIS FOR THE OUTCOMES OF OVERLAPPING COMMUNITY DETECTION ALGORITHMS ON THE FLATTENED UNWEIGHTED UQU NETWORK F_U.

	F_U			
Algorithms	**IPCA**	**DPClus**	**DANMF**	**Dcs**
coverage	0.98	0.95	1	1
#communities	48	45	8	3
#singletons	0	0	0	0
avg(community/member)	6.61	1.28	1	2.39
second/first	0.65	0.51	0.80	0.71
max(community size)	354	170	136	702
avg(community size)	94.77	19.07	88.13	561.33

To compare whether two algorithms detected similar community structures, the extended normalized mutual information (NMI) [21] was adopted. The NMI measures the similarity between two community sets, with 1 as identical and 0 as totally different.

VI. RESULTS AND DISCUSSION

A. Weighted vs. Unweighted Networks

Table IV and the leftmost section of Table V show the community analysis metrics for the outcome of overlapping communities detected in F_U and F_W, respectively. All algorithms have a high coverage that includes at least 95% of the population. Both IPCA and DPClus detected significantly more communities than the rest. DPClus has a lower overlapping rate while DANMF detected only disjoint communities. In all algorithms, a slightly tailing effect can be observed from the difference between the maximum community size and the average size, in addition to the ratio of the second-largest community to the first-largest community, especially for the outcome of IPCA and DPClus.

Table VI illustrates the NMI and ONMI scores when comparing the outcome communities detected in the weighted network F_W against its unweighted counterpart F_U. Most of the studied algorithms have a high agreement; some even have identical community sets indicating that weight does not play a significant role. This was true for all overlapping community detection algorithms before and after the semantic expansion except for *DANMF*. However, with manual inspection, it seems that weight did not affect the result in terms of statistics, but communities' members differ in some of the algorithms.

B. Impact of Semantic Expansion

As shown in the middle section of Table V, the increase in the density decreased the number of detected communities in $\overline{F_W}$ compared to F_W. However, DCS failed to detect any communities. In contrast, DANMF did detect communities with no overlapping with a single large community and smaller ones, as indicated by the low ratio between the top two largest communities and average community size. While IPCA and DPClus did not assign communities to all nodes, only IPCA managed to detect overlapping communities with an average of 5 communities per actor. All algorithms suffer from generating a single large community with smaller ones.

TABLE V

COMMUNITY ANALYSIS FOR THE OUTCOMES OF OVERLAPPING COMMUNITY DETECTION ALGORITHMS ON EXPANDED FLATTENED WEIGHTED UQU NETWORK F_W, ITS SEMANTICALLY EXPANDED $\overline{F_W}$ AND REDUCED SEMANTICALLY EXPANDED $\overline{\overline{F_W}}$ VERSIONS.

	F_W				$\overline{F_W}$				$\overline{\overline{F_W}}$			
Algorithms	**IPCA**	**DPClus**	**DANMF**	**DCS**	**IPCA**	**DPClus**	**DANMF**	**DCS**	**IPCA**	**DPClus**	**DANMF**	**DCS**
coverage	0.98	0.96	1	1	1	0.99	1	1	0.99	0.99	1	1
#communities	48	46	8	3	6	16	8	1	8	18	7	1
#singletons	0	0	0	0	0	0	0	0	0	0	0	0
avg(community/member)	6.61	1.28	1	2.39	5.24	1.08	1	1	6.57	1.09	1	1
second/first	0.65	0.51	0.84	0.71	0.87	0.07	0.79	0	0.83	0.08	0.04	0
max(community size)	354	170	142	702	697	535	160	705	691	526	603	705
avg(community size)	94.77	18.70	88.13	561.33	613.0	47.19	88.13	705	577.13	42.39	100.71	705

TABLE VI

NMI SCORES FOR MEASURING THE COMMUNITIES AGREEMENT BETWEEN WEIGHTED AND UNWEIGHTED UQU NETWORK OUTCOMES. 1 IS IDENTICAL, AND 0 IS TOTALLY DIFFERENT.

Algorithm	F_W **vs.** F_U	$\overline{F_W}$ **vs.** $\overline{F_U}$
Dcs	1	1
DPClus	0.803	0.777
IPCA	1	1
DANMF	0.447	0.012

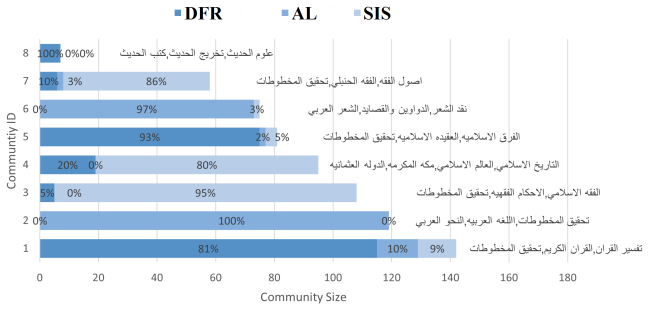

Fig. 6. Overlapping communities detected in F_W (#communities=8) using DANMF with the college affiliation distribution of each community and the top three most frequent keywords.

With a closer look at the detected communities and the shared keywords between their members, it was found that the keywords with high frequency in each generated community before the expansion have some direct or indirect relation with its members' affiliated colleges. Figure 6 shows eight communities detected by DANMF, the college affiliation distribution of each community, and its three most frequent keywords. For instance, community#1 has 81% affiliated with DFR, which is a college that primarily studies Islamic religion, has *Quran interpretation*, *Quran* and *manuscript investigation* as its most frequent keywords, which align with the college's main interest. It is also worth noting that the ordered keyword list for each community differs across the detected communities. This was true for both overlapping and partitioning algorithms.

However, after semantic expansion, the list of keywords for each community does not fully align with the affiliated college of its majority members specifically. Figure 7 shows seven communities detected by DANMF and the college affiliation distribution of each community along with its top frequent

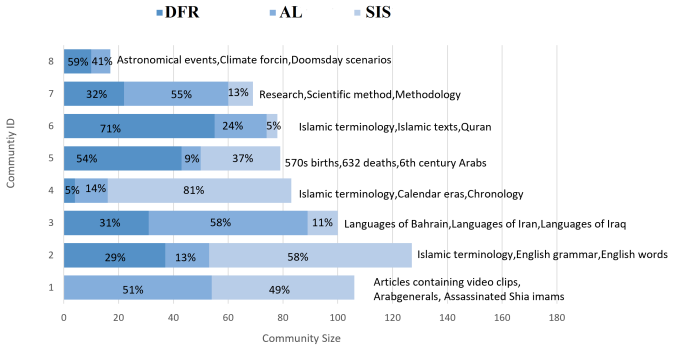

Fig. 7. Overlapping communities detected after semantic expansion in $\overline{F_W}$ (#communities=8) using DANMF with the college affiliation distribution of each community and the top three most frequent keywords.

three keywords. The list of most frequent keywords, which was the discriminative feature for a given community before the semantic expansion, almost does not include any of the original keywords. For instance, community#2 has *Islamic terminology*, *English grammar* and *English words* as the most frequent keywords, which does not fully align with the college of SIS, which studies Islamic law and history, where 58% of its member belong to it. Some generic and noisy keywords have emerged, such as *Articles containing video clips* and *Arabgenerals* for community#1 with its members should study the Arabic language or history. Again, this randomness was observed in all generated communities regardless of the used algorithm.

C. Overexpansion Issue

The analysis and evaluation show that the semantic expansion did degrade the performance. Most frequent words are generic or incompatible with the associated college and do not align with its line of studies. In the resulting semantic expansion from DBpedia using `skos:broader` and `dct:subject`, we retrieved too many relationships that are irrelevant to the original keywords, or perhaps the relationship is very far from the topic. This can be accounted to two possible causes. First, no limit was specified for this semantic expansion, allowing all related ancestors to be included. The further the ancestor from the current topic, the more generic it is, especially since the number of edges after the semantic expansion has increased

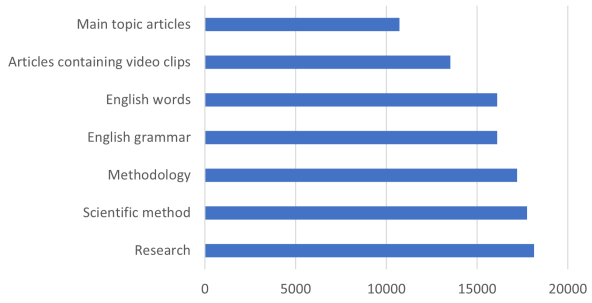

Fig. 8. Top seven most frequent keywords in the semantically expanded keywords network (Keywords) with their counts.

TABLE VII
STATISTICS OF THE KEYWORDS LAYER: ORIGINAL (KEYWORDS), SEMANTICALLY EXPANDED ($\overline{\text{KEYWORDS}}$) AND AFTER REMOVING THE EDGES REPRESENTING FOUR OF THE GENERIC KEYWORDS-*Research*, *English words*, *English grammar*, AND *Main topic articles* ($\overline{\overline{\text{KEYWORDS}}}$) ALONG WITH THE CHANGE RATIO IN THE VALUE BEFORE REMOVING GENERIC KEYWORDS AND AFTER.

Characteristics	Keywords	$\overline{\text{Keywords}}$	$\overline{\overline{\text{Keywords}}}$	$\pm\%$
Number Of Nodes	705	705	705	0%
Number Of Edges	34630	168000	160705	-4.34%
Average Clustering	0.08	0.02	0.02	0%
Average Degree	123.37	2900.48	2727.06	-5.97%
Density	0.13	0.67	0.64	-4.47%

significantly, almost triple the number of edges in the original networks F_U and F_W. The second possible cause is using an English ontology with a translated query, which might not be accurate, especially for terminologies causing the retrieval of irrelevant keywords. Figure 8 presents the count of the top seven most frequent keywords in $\overline{F_U}$ and $\overline{F_W}$. These keywords are too generic and were added to a large portion of the nodes making clustering them more challenging.

In an attempt to validate the negative impact of adding such generic keywords to the network, the following four keywords were removed from the network: *Research*, *English words*, *English grammar*, and *Main topic articles*. Consequently, a new keywords layer was constructed ($\overline{\overline{\text{Keywords}}}$) with more than 4% relative fewer edges. Moreover, the average degree and density are less than those in the semantically expanded keywords layer ($\overline{\text{Keywords}}$), as listed in Table VII.

$\overline{\overline{\text{Keywords}}}$ was used to construct the new weighted flattened network ($\overline{\overline{F_W}}$). Then, community detection algorithms were applied, and the outcome communities' analysis was listed in the rightmost section of Table V. A slight increase in the number of detected communities compared to $\overline{F_W}$ can be observed for IPCA and DPClus, while DCS still failing in detecting any communities. Unlike in $\overline{F_W}$ where DANMF detected communities of similar sizes, it detected disjoint communities with a very large community along with smaller ones. The rest has no or insignificant changes compared to the outcomes of $\overline{F_W}$.

Looking at one of the detected communities, shown in Figure 9, the keyword list becomes more specific and aligned with the line of studies of the college its majority are affiliated with.

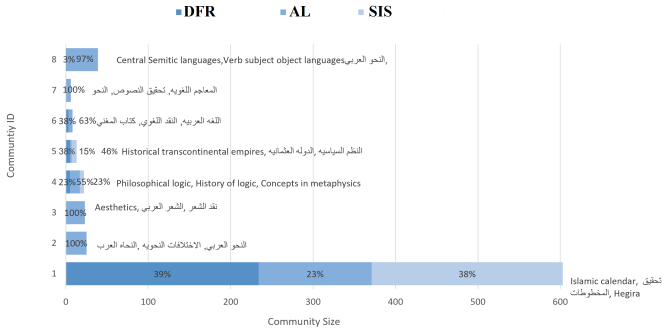

Fig. 9. Overlapping communities detected after semantic expansion in $\overline{\overline{F_W}}$ (#communities=8) using DANMF with the college affiliation distribution of each community and the top three most frequent keywords.

For instance, community#2 with the majority of members affiliated with AL who study the Arabic language, has *Arabic grammar*, *Grammatical differences* and *Arab grammarians* as their frequent keywords are aligned with the line of the college studies.

D. Colleges vs. Communities

Before applying any community detection algorithms, the only communities known for the faculty members in the UQU network are those affiliated with their colleges. Hence, members' college information was used to compare how far are the detected communities from the communities formed by the colleges. Table VIII lists how similar the outcome communities are to the college affiliation information measured using NMI scores. As aforementioned, the outcomes from the original networks, F_U and F_W, when considering the weights and without them, are almost identical for all algorithms. Both are not similar to the college affiliation, even when the number of detected communities is equal to the number of selected colleges, as in DCS case. This can be accounted to the overlapping effect, and one can conclude that being in the same college does not mean sharing the same research interest. The community assignment was driven even further when semantic expansion was applied. This impact was caused by the addition of generic and irrelevant keywords discussed in Section VI-C. A such diversion was reduced when removing four of the most frequent generic keywords in the semantically expanded network.

As a result, one can conclude that semantically keyword expansion could help if the expansion was made selectively.

VII. CONCLUSION

This study attempts to detect communities within the Arabic scholarly multiplex network with two layers: co-authorship and publication keywords. Hence, in this work, two faculty members might be clustered in a single community if they shared their research interests which can be indicated by their co-authorship of a publication or a shared keyword, which is retrieved from their publications. However, the used keywords on a publication might be limited or too specific to the publication. Hence, using an ontology, one can provide

TABLE VIII
NMI SCORES FOR COMPARING OF THE OUTCOMES OF OVERLAPPING COMMUNITY DETECTION ALGORITHMS ON EXPANDED FLATTENED UQU NETWORK F_U, F_W, ITS SEMANTICALLY EXPANDED $\overline{F_W}$ AND REDUCED SEMANTICALLY EXPANDED $\overline{\overline{F_W}}$ VERSIONS WITH THE COLLEGE AFFILIATION.

Algorithms		F_U	F_W	$\overline{F_W}$	$\overline{\overline{F_W}}$
Colleges vs.	IPCA	0.16	0.16	0.03	0.03
	DPClus	0.09	0.09	0.04	0.05
	Dcs	0.05	0.05	0.0	0.0
	DANMF	0.19	0.22	0.08	0.07
IPCA vs.	DPClus	0.47	0.46	0.34	0.42
	DCS	0.09	0.09	0.0	0.0
	DANMF	0.39	0.40	0.04	0.29
DPClus vs.	DCS	0.11	0.11	0.0	0.0
	DANMF	0.26	0.23	0.02	0.24
Dcs vs.	DANMF	0.06	0.08	0.0	0.0

related keywords which represent the field properly. Due to the unavailability of an Arabic domain ontology, English DBpedia was used to extract semantically related keywords to a translated publication title. For this study, the faculty members of three colleges from UQU with primarily Arabic publications were chosen to build the scholarly network regardless of their affiliated colleges. After constructing the UQU network, four overlapping community detection algorithms were used after flattening the multiplex network: DANMF, IPCA, DPClus and DCS. Mainly, three settings were investigated: unweighted and weighted UQU networks and a semantically expanded version of the latter.

The outcomes from these algorithms in the three settings were analyzed using several community analysis metrics along with an extended NMI to compare the outcome of the two algorithms against each other. First, it was observed that considering weights for the same multi-edge network generated similar community sets with minor differences. However, with manual inspection, it was found that the member assignments of overlapping community detection algorithms significantly differed from each others.

After the semantic expansion, the networks become harder to cluster as many of the added keyword relationships were either too generic and were added to most of the members or noisy and irrelevant to the original publication title. This can be rooted in two leading causes: (1) the addition of the semantically related keywords was without any constraints or cut-off; hence, root topics were added, and (2) the translation of terminology might not be accurate, causing unrelated topics to be added. When only four of the top most frequent newly added keywords were removed, the network was reduced by 7%, slightly improving the clustering qualities. Consequently, one can learn that semantic expansion is promising. Still, it should be used with caution not to increase the network's connectivity and decrease its potential to be clustered into smaller communities.

These findings are significant as they will direct future work. For instance, one might try to improve the technique of semantic expansion by either improving the topical or domain ontology and deriving one from researchers' publications or using a pre-trained Arabic topic model to add more related keywords to the given publications. To our knowledge, this is the first investigation into detecting communities within an Arabic scholarly network.

REFERENCES

[1] Z. Zhao, S. Zheng, C. Li, J. Sun, L. Chang, and F. Chiclana, "A comparative study on community detection methods in complex networks," *Journal of Intelligent & Fuzzy Systems*, vol. 35, no. 1, pp. 1077–1086, 2018.

[2] A. A. Salatino, F. Osborne, T. Thanapalasingam, and E. Motta, "The cso classifier: Ontology-driven detection of research topics in scholarly articles," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2019, pp. 296–311.

[3] H. Bast, B. Buchhold, and E. Haussmann, "Semantic search on text and knowledge bases," 2016.

[4] E. C. Hai Dong, Farookh Khadeer Hussain, "A survey in semantic search technologies," 2008.

[5] A.-K. Al-Tamimi, E. Bani-Isaa, and A. Al-Alami, "Active learning for arabic text classification," in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. IEEE, 2021, pp. 123–126.

[6] V. Horta, V. Ströele, R. Braga, J. M. N. David, and F. Campos, "Analyzing scientific context of researchers and communities by using complex network and semantic technologies," *Future Generation Computer Systems*, vol. 89, pp. 584–605, 2018.

[7] P. Cifariello, P. Ferragina, and M. Ponza, "Wiser: A semantic approach for expert finding in academia based on entity linking," *Information Systems*, vol. 82, pp. 1–16, 2019.

[8] S. Zevio, G. Santini, H. Soldano, H. Zargayouna, and T. Charnois, "A combination of semantic annotation and graph mining for expert finding in scholarly data," in *Proceedings of the Graph Embedding and Mining (GEM) Workshop at ECML PKDD*, 2020.

[9] A. A. Salatino, F. Osborne, T. Thanapalasingam, and E. Motta, "The cso classifier: Ontology-driven detection of research topics in scholarly articles," 2019.

[10] F. Osborne, G. Scavo, and E. Motta, "Identifying diachronic topic-based research communities by clustering shared research trajectories," in *European Semantic Web Conference*. Springer, 2014, pp. 114–129.

[11] M. Magnani, O. Hanteer, R. Interdonato, L. Rossi, and A. Tagarelli, "Community detection in multiplex networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–35, 2021.

[12] F. Ye, C. Chen, and Z. Zheng, "Deep autoencoder-like nonnegative matrix factorization for community detection," pp. 1393–1402, 2018.

[13] G. Rossetti, L. Milli, and R. Cazabet, "Cdlib: a python library to extract, compare and evaluate communities from complex networks. applied network science." Mar. 2022. [Online]. Available: https://github.com/GiulioRossetti/CDlib

[14] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[15] M. Li, J.-e. Chen, J.-x. Wang, B. Hu, and G. Chen, "Modifying the dpclus algorithm for identifying protein complexes based on new topological structures," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–16, 2008.

[16] S. A. Muhammad and K. V. Laerhoven, "Dcs: Divide and conquer strategy for detecting overlapping communities in social graphs."

[17] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, vol. 6, no. 2, pp. 167–195, 2015.

[18] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th international conference on semantic systems*, 2011, pp. 1–8.

[19] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," 1 2008. [Online]. Available: https://www.osti.gov/biblio/960616

[20] M. Magnani, L. Rossi, and D. Vega, "Analysis of multiplex social networks with R," *Journal of Statistical Software*, vol. 98, no. 8, pp. 1–30, 2021.

[21] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New journal of physics*, vol. 11, no. 3, p. 033015, 2009.