IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 October 2024, accepted 9 November 2024, date of publication 18 November 2024,
date of current version 26 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3500374

SURVEY

User Identity Linkage on Social Networks: A Review of Modern Techniques and Applications

CATERINA SENETTE[1], MARCO SIINO[2], AND MAURIZIO TESCONI[1]
[1]Institute of Informatics and Telematics, National Research Council of Italy (IIT-CNR), 56124 Pisa, Italy
[2]Dipartimento di Ingegneria Elettrica Elettronica e Informatica, University of Catania, 95124 Catania, Italy

Corresponding author: Caterina Senette (caterina.senette@iit.cnr.it)

This work was supported in part by the Project SEcurity and RIghts In the CyberSpace (SERICS) through the National Recovery and Resilience Plan-Ministero dell'Università e della Ricerca (NRRP-MUR) Program funded by the European Commission-Next GEneration EUrope (EU-NGEU) under Grant PE00000014.

ABSTRACT In an Online Social Network (OSN), users can create a unique public persona by crafting a user identity that may encompass profile details, content, and network-related information. As a result, a relevant task of interest is related to the ability to link identities across different OSNs that can have multiple implications in several contexts both at the individual level (e.g., better knowledge of users) and at the group level (e.g., predicting network dynamics, information diffusion, etc.). The purpose of this work is to provide a comprehensive review of recent studies (from 2016 to the present) on User Identity Linkage (UIL) methods across online social networks. It would offer guidance for other researchers in the field by outlining the main problem formulations, the different feature extraction strategies, algorithms, machine learning models, datasets, and evaluation metrics proposed by researchers working in this area. To this aim, the proposed overview takes a pragmatic perspective to highlight the concrete possibilities for accomplishing this task depending on the type of available data. Our analysis demonstrates significant progress in addressing the UIL task, largely due to the development of more advanced deep-learning architectures. Nevertheless, certain challenges persist, primarily stemming from the limited availability of benchmark datasets. This limitation is further compounded by current social network access policies, which prioritize privacy protection and reduce opportunities to retrieve data through APIs.

INDEX TERMS User identity linkage, social networks, network alignment, review.

I. INTRODUCTION

Everyone's social life has been changed by the recent growth of social network services of all kinds, which make it easier and more enjoyable than ever to share a variety of information (e.g., microblogs, images, videos, reviews, location check-ins). How to use this large amount of social data for improved business intelligence is undoubtedly the biggest and most fascinating topic facing all firms. People are particularly concerned with understanding each individual user more effectively, given the vast amount of social data now available. Unfortunately, a user's social scene information is fragmented, unreliable, and disruptive. Due to the wide variety of services offered by online social networks (OSNs), it seems natural for users to register for accounts (also known as user identities) on many OSNs. Having accounts (also known as user IDs) on several OSNs has grown in popularity. According to a 2023 statistic,[1] among 50 nations with internet users aged between 16 to 64, Japan had the lowest overall number of social media accounts at 3.5 per user, while India had the highest at 9, the average number around the world is 7.2 accounts per user.

User Identity Linkage (UIL) refers to the process of linking or matching user identities across different online

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

[1]https://wearesocial.com/it/blog/2023/01/digital-2023-i-dati-globali/ (2024-09-10)

TABLE 1. UIL - Formal definition.

Definition: *User Identity Linkage*
Given two online social networks P_A (first network) and P_B (second network), the task of User Identity Linkage is to predict whether a pair of user identities u_A and u_B from U_A and U_B respectively belong to the same real person. This is represented by the function: $$H : U_A \times U_B \rightarrow \{0, 1\}$$ *such that:* $$H(u_A, u_B) = \begin{cases} 1, & \text{if } u_A \text{ and } u_B \text{ are the same individual} \\ 0, & \text{otherwise} \end{cases}$$ *where H is the prediction function to be learned.*

platforms or social networks by analyzing the similarities in their profiles, behaviors, or activities. A formal definition is provided in Table 1. The problem is also known as Social Identity Linkage [1], User Identity Resolution [2], Social Network Reconciliation [3], Profile Linkage [4], Anchor Link Prediction [5]. It is worth noting that the terminology used in the literature is still consistent when considering that *Profile Linkage* is used in contexts where user profile attributes are prioritized to solve the UIL task. Conversely, the term *Social Identity Linkage* is employed to highlight the impact of social relationships and interactions in addressing it. Both concepts fall under the broader category of User Identity Linkage. UIL is used to consolidate user information from multiple sources, providing a comprehensive view of an individual across platforms. It is commonly employed in domains like personalized recommendations, cross-platform marketing, and more recently, in cyber intelligence to detect malicious actors. Linking users across social networks can have multiple implications in several contexts, both at the individual level and at the group level. At the individual level, the main interest in linking the same identity across social networks is to enable a better knowledge of each user by aggregating all the information collected from each social network platform. A more comprehensive frame of a single user simplifies strategies for cross-system personalization that in turn could be used to build user-adaptive systems able to trigger external recommendation systems and to provide personalized services in e-commerce, tourism, travel planning, and so on [6], [7]. By reasoning at the group level, linking user identities through different OSNs helps in predicting user behaviors, network dynamics, and information diffusion other than understanding migration phenomena across social media that in turn could be beneficial for social media platforms to generate revenue from suggested advertising and to grow their base with the ultimate goal to improve marketing outcomes [8], [9]. The use of UIL techniques is also crucial in the field of cyber intelligence, as it enables the identification and tracking of suspicious activities across different social platforms, facilitating the detection of malicious actors and the prevention of disinformation campaigns or coordinated attacks [10]. A set of potential real-world use cases where UIL could be successfully implemented is shown in Table 2.

For all these reasons, User Identity Linkage has become a trending topic and has attracted more and more research attention. It is not easy to take advantage of the numerous chances that people with profiles on several social media platforms present. Linking users' accounts across various online social networks provides all the aforementioned opportunities; however, the process of tying together user accounts on different OSNs is challenging. The main reasons are: (i) For the same person in the real world, user identity information on different online social networking sites can vary greatly; (ii) Online social networking data is vast, noisy, imperfect, and largely unstructured. Any single social network service can only provide a limited view of a user from a certain perspective due to limitations imposed by the features and design of each service. An otherwise disjointed user profile would be enriched by cross-platform user linking, allowing for a comprehensive grasp of a user's interests and behavioural patterns. The information posted by users on social media platforms may be incorrect, contradictory, incomplete, and deceptive for a variety of reasons. The consistency of user information can be improved by cross-referencing across several platforms.

In addition to the aforementioned challenges inherent to the task, the more critical issue lies in the scarcity of publicly available datasets, which limits the ability to train supervised systems and verify experimental results. This lack of comprehensive datasets hinders progress in effectively linking user accounts across platforms.

Although social networks rise and fall, real-world consumers continue to use them and simply switch to newer ones. Linking user identities enables the integration of important user data from platforms that have over time declined in popularity or even been abandoned.

Along with its benefits, the use of UIL also introduces considerable risks, especially regarding privacy. It enables widespread user tracking across platforms, potentially leading to unauthorized profiling, data exploitation, and breaches of user consent.

A portion of the literature on the topic has been summarized by a survey dated 2016 [11] which presents a unified framework for the UIL task consisting of two phases, feature extraction, and model construction. Moreover, the authors summarize different aspects of feature extraction and model construction techniques and discuss different datasets and evaluation metrics proposed by existing approaches.

Comparing state of the art prior to 2016 [11] with more recent works (up to 2024) permits us to provide an overall look at the UIL in terms of the following aspects: (i) The rise of new problem formulations of the User Identity Linkage problem; (ii) New methods used to extract and represent features from social networks; (iii) Up-to-date AI models built to address the UIL task; (iv) Novel algorithms

TABLE 2. Potential Real-World Use Cases of User Identity Linkage (UIL).

Use Case	Description
Cross-Platform Marketing	UIL connects user profiles across platforms like Facebook, Instagram, and Twitter. This enables companies to run targeted advertising campaigns by understanding a user's behavior across different platforms.
Fraud Detection	In the financial and e-commerce industries, UIL could help identify fraudulent activity by linking user profiles across multiple networks.
Cybersecurity	Cybersecurity professionals could use UIL to track malicious users across networks, linking hacker accounts across platforms.
Social Network Analysis for Research	Researchers use UIL to link user identities across multiple social platforms (e.g., Facebook, LinkedIn, Twitter) to study social behavior.
Law Enforcement	Law enforcement could use UIL to track individuals across social media networks, aiding criminal investigations.
Customer Support and User Service Integration	Companies could use UIL to integrate identities across platforms and provide improved services.
Content Personalization	Media platforms could use UIL to track user preferences across devices, ensuring personalized content delivery based on cross-platform behavior.

introduced; (v) Deep focus on data collection and concrete availability of datasets.

A. RESEARCH QUESTIONS AND CONTRIBUTION

It is worth noting that, in the last decade, there have been many changes in this research field which can be summarized as follows: (i) The faster growth of social networks and the higher diversity among them; (ii) The increased attention to privacy, which effectively limits the real availability of individual data thus slanting strategies towards creating detailed profiles based on user online activity rather than relying on disclosed personal data; (iii) Computational methods increasingly oriented towards deep learning, nowadays considered as a core technology of today's Fourth Industrial Revolution [12].

We believe that all these elements would require an update of the state of the art, so, the purpose of the current work is to provide a comprehensive review of very recent studies (from 2016 to date) of User Identity Linkage methods across online social networks with the intention to give guidance for other researchers working in the field in light with the current possibilities to accomplish the UIL task.

To this end, we will offer an alternative perspective that approaches the problem from a pragmatic point of view expressed by the following research questions (RQs):

- (**RQ1**)*What are the current prevailing problem formulations, methodologies, and techniques used in User Identity Linkage across social networks?*
- (**RQ2**)*What performances do they currently guarantee?*
- (**RQ3**)*What issues are still open in this field?*

By answering these research questions we provide the following contributions:

- An updated overview of the body of knowledge on UIL in order to fill the chronological gap with respect to previous literature reviews and to identify the aspects in which the major innovations have been introduced.
- A functional approach to the task with the intention to explore and give some guidance for more practical problem settings in User Identity Linkage across social networks.
- A useful collection of the datasets used for UIL task, not built ad hoc for a single experiment and shared in several research works. Each is provided with a reference.

B. ROADMAP

The paper has the following structure. After the Introduction, in Section II we describe explicit and rigorous criteria used to identify, critically evaluate, and synthesize all the available works on recent literature. In Section III the problem of User Identity Linkage is narrowed down to two possible formulations (the most common ones). Here we introduce our conceptual framework. In Section IV and V, we present all the state-of-the-art solutions to perform UIL summarizing them guided by our framework. When appropriate, collected papers are further cataloged based on the category of data used to accomplish the UIL task. Specifically, for each single data type or group of data, we explore all the research works illustrating different feature extraction strategies, different algorithms, and different machine learning models proposed. In Section VI we explore evaluation metrics used in each of the two UIL formulations. In Section VII we present a detailed catalog of all the datasets proposed in the recent literature. In Section VIII we provide a general discussion in light of the proposed conceptual framework. We also discuss some privacy and ethical concerns in UIL in Section IX.

Finally, in Section X we examine the main challenges still open, and, in Section XI we draw a conclusion for the paper.

II. METHOD

Candidate papers for inclusion in this review were gathered through four steps, as shown in Figure 1.

As a first step, we conducted a search in Google Scholar, Microsoft Academic, and IEEE Explore which are the most used academic search engines to collect the knowledge base about this topic. We carried out the search in December 2023 and again in June 2024 and October 2024 looking for any potential new entry and using search terms covering variations on ''User Identity Linkage''. Specifically, we used six different strings as key search: (a) *User Identity Linkage across social networks*; (b) *user accounts linkage across social networks*; (c) *social network alignment*; (d) *network alignment*; (e) *user profile matching across social networks*; (f) *reconciliation across social networks*. These searches yielded a total of 184 results which were downloaded. Among these, we identified only two documents representing scientific reviews that synthesize and integrate knowledge about the topic [11], [13]. The work of Shu et al. [11], dated 2016, motivates the need for a new update on the topic covering the last eight years, the second work [13] is a doctoral dissertation dated 2020 which focused on the study of conventional machine learning based approaches and more recent graph representation based approaches. Therefore the latter contribution lacks a comprehensive guide on the topic.

As a second step, we refined the search excluding research contributions dated before 2016 which were already discussed by Shu et al. [11]. We obtained a list of 99 papers.

As a third step, a careful check of titles and abstracts reduced this to a list of 89 publications selected as relevant for closer attention and consideration for inclusion in the review. Key paper citations were collected through ad-hoc software (Mendeley).

As a fourth step, for those papers that released an associated dataset, we identified and, when needed, contacted corresponding authors asking them to share their data repository. This was extremely helpful especially to verify whether the data model underlying theoretical approaches to UIL described in their work corresponds exactly to the data actually available. As shown in Table 12 we identified 16 datasets used for the UIL task in more than one research work.

A final list of 89 contributions reporting novelties on the User Identity Linkage task found via these methods is included in the current work. The associated contents are detailed in the next sections of this survey.

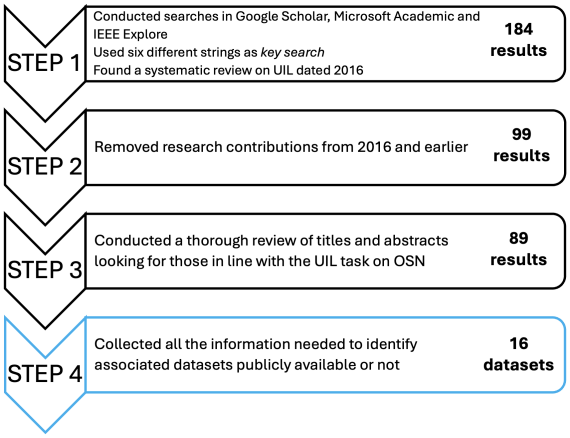


FIGURE 1. Method applied for collecting research works to include in this review.

III. PROBLEM FORMULATION AND GENERAL APPROACH

The number of possible problem formulations for user-identity linkage across social networks can vary based on several factors. The final goal might include identifying duplicate accounts, merging user profiles, or enhancing recommendation systems. Formulations can also vary depending on the scope, determining whether the analysis pertains to a single social network or spans multiple ones. Furthermore, the context of the linkage, the desired outcomes, and the methods employed are crucial elements that shape the problem formulation.

Focusing on the objective, the User Identity Linkage (UIL) task can be primarily formulated as a *classification* task or as a *network alignment* task.

When the objective of UIL is to determine whether two profiles from different networks belong to the same individual, the problem is framed as a classification task. This formulation is primarily addressed using supervised (and semi-supervised) methods. These methods leverage labeled data, such as Pre-aligned user Pairs or supervisory anchor pairs (SAPs), to train predictive models. The goal is to learn discriminative features that enable the prediction of whether two profiles represent the same user. This approach is more straightforward and tractable than network alignment. However, the success of supervised methods heavily depends on the availability and quality of SAPs. Despite their scarcity in real-world scenarios, these pairs are crucial for training accurate and reliable models. Results achieved by some authors [14], [15] indicate that even with a minimal amount of carefully selected SAPs, the overall performance of the models is significantly boosted.

Alternatively, when the objective of UIL is to align nodes from different social networks based on their structural attributes without relying on labeled data, the problem is conceptualized as a network alignment challenge. Unsupervised or semi-supervised approaches are employed to achieve this goal. Network alignment techniques aim to map the entire structure of one network onto another, aligning nodes based on structural similarities. This thorough alignment often results in problems that are highly complex and difficult to solve optimally. The challenge becomes particularly pronounced in networks that are either very dense or exceptionally large [16], [17].

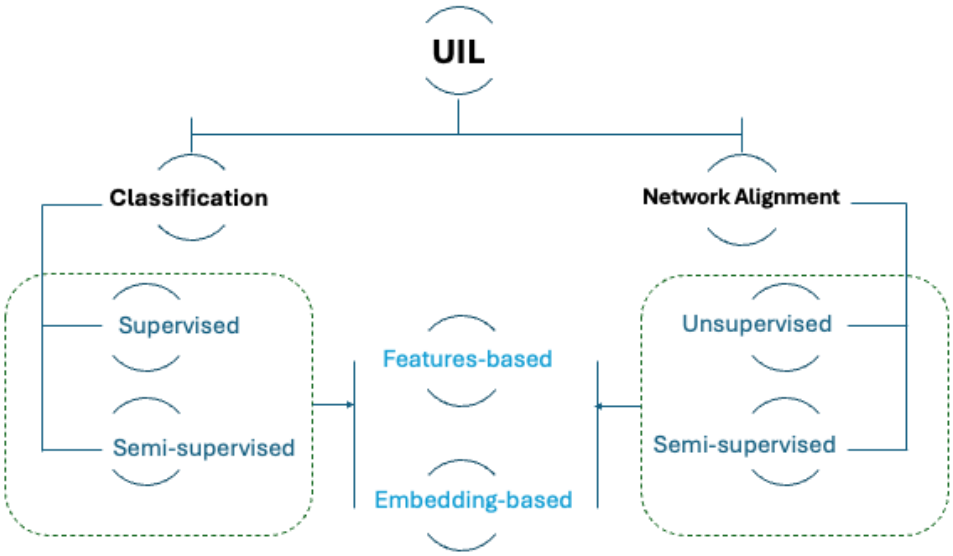


FIGURE 2. Conceptual framework that identifies two UIL formulations-Classification and Network Alignment-with sub-approaches: Feature-based and Embedding-based methods.

The narrative perspective used herein in this survey is guided by this conceptual framework as shown in Figure 2

A. FEATURES-BASED VS EMBEDDING-BASED APPROACHES

As highlighted in Figure 2 both problem formulations share two primary sub-approaches, feature-based and embedding-based strategies used to identify and match user accounts across social networks. Features are measurable attributes or properties derived from the raw data and processed/transformed to be used in machine learning models or other analytical processes. We provide a categorization of them in Table 3.

Additionally, Figure 3 shows an example of feature extraction from raw data in the scenario of UIL across X (Twitter) and Instagram.

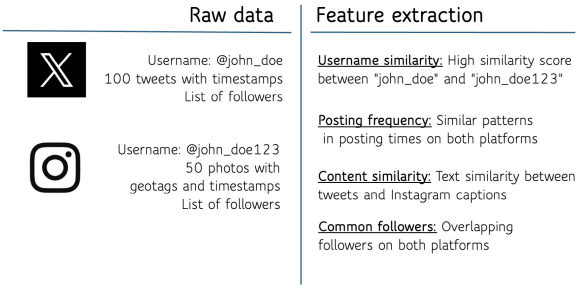


FIGURE 3. From Raw-Data to Features: an example of UIL across X (Twitter) and Instagram.

Features-based methods traditionally involve manually designing features to capture similarities and differences between user profiles across networks. These features might include profile attributes, interaction patterns, and network connections. This approach explicitly defines and extracts the features from raw data, making them interpretable but potentially limited by the quality and relevance of the chosen features.

On the other hand, embedding-based methods (particularly recent unsupervised and semi-supervised techniques) [18], [19], utilize graph embedding techniques to automatically learn low-dimensional latent vectors (embeddings) that capture the structural properties of network nodes. This latent space representation allows nodes to be represented in a continuous space where similar nodes are closer together, enhancing the methods' flexibility and ability to capture complex data patterns. Embedding-based methods generally outperform traditional feature-based approaches [11], [20], leveraging deep learning techniques to handle large-scale data and capture intricate structural relationships more effectively.

Figure 4 provides one flow diagram for each strategy. Additionally, Table 4 summarizes the aspects to consider when comparing feature-based methods with embedding-based methods in the two main User Identity Linkage problem formulations identified in this survey.

It is worth noting that while feature-based and embedding-based strategies have their strengths and weaknesses, they are not mutually exclusive and can be integrated into more robust UIL systems. For example, feature-based approaches can provide interpretable insights that help guide the design of embedding models. Conversely, embeddings can be used to enhance feature-based models by providing additional learned features [21], [22], [23].

In the next sections IV and V, we will describe the relevant literature categorizing all the strategies proposed by scholars

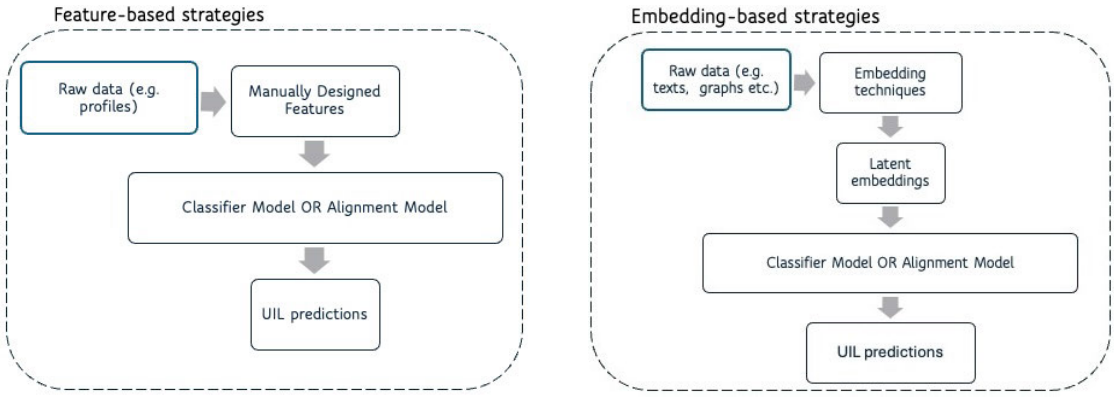


FIGURE 4. Feature-based VS Embedding-based approach. The left diagram illustrates the feature-based approach, where features are manually selected from raw data. The right diagram illustrates the embedding-based approach, which includes two initial stages: (1) raw data are transformed into embeddings and (2) subsequently projected into a latent space. All other blocks are common to both approaches.

based on the type of UIL problem formulation as introduced in our conceptual framework. This overview will answer the first research question (**RQ1**).

A synthesis of algorithms cited in this section is shown in Table 10.

IV. UIL AS A NETWORK ALIGNMENT PROBLEM

As already said, when addressing the User Identity Linkage problem as a network alignment task the process focuses on structural properties and attributes similarities between nodes without using labeled data.

A. FEATURE-BASED STRATEGIES

Table 5 provides a general scheme detailing how this type of task works given a set of input data. It begins with feature extraction, then, algorithms like graph matching or clustering are employed to align user accounts across different networks, aiming to identify the optimal alignment based on these features. Evaluation of these methods typically involves assessing how well the network structure is preserved and using indirect metrics, such as clustering quality, to gauge the effectiveness of the alignment.

An unsupervised approach called Friend Relationship-based User Identification algorithm without Prior (FRUI-P) knowledge is proposed in [24], after observing that friend relationships are trustworthy and consistent across distinct OSNs. The FRUI-P evaluates the similarities of all the possible identical users between two OSNs after extracting the friend feature of each user in an OSN into a friend feature vector. The users are then identified using a one-to-one map approach that takes into account their commonalities.

In [25] the authors proposed the CoLink framework (semi-supervised) that utilized a co-training algorithm by applying two distinct models: an attribute-based model, and a relationship-based model. Both models are designed to perform binary classification, determining whether a given pair of users is positive (linked) or negative (unlinked). The co-training algorithm iteratively enhances the performance of these two models. In each iteration, both models are retrained using the set of linked pairs, defined as S. The co-training algorithm must start with a small seed set of linked user pairs. Then the seed set is generated using specially created rules, or ''seed rules''. The attribute-based model employs sequence-to-sequence learning to handle attribute alignment, while the relationship-based model uses social connections. Despite employing sophisticated techniques, this approach fundamentally relies on manually designed features (attributes and relationships).

B. EMBEDDING-BASED STRATEGIES

Table 6 depicts an outline of how this type of task is performed using a given set of input data. The process begins with embedding the network data, transforming nodes and their connections into dense vector representations that capture both local and global structural information. Advanced algorithms, such as graph neural networks or matrix factorization techniques, are then employed to align these embeddings across different networks. The goal is to identify the optimal alignment by comparing these learned embeddings. Evaluation of these methods typically involves assessing how well the network's structural properties are preserved in the embeddings and using metrics like alignment accuracy and embedding quality to gauge the effectiveness of the alignment.

1) UNSUPERVISED METHODS

In the research described in [26], the relationship strength is measured using an improved weighted graph model. First, the authors represent the social network as a weighted graph, where the weight relates to the user interactions. Then, they suggest using the CNIL (Common Neighbors and Internal Links) index, which may also be used to represent social links, to quantify the weight. With data gleaned from second-order neighbors, the CNIL index aims to improve the CN

TABLE 3. Features for social media profile matching.

Feature Category	Details
Direct Matching Features	Identical Identifiers: Username, Email address, Phone number, Social Security number Account IDs or handles
Behavioral Features	Posting Frequency: Number of posts, Time between posts Interaction Patterns: Number of likes, Number of comments, Number of shares, Types of interactions (e.g., reactions, emojis used) Activity Timestamps: Time of day for activity, Days of the week for activity Engagement Levels: Total engagement (likes, comments, shares), Engagement per post
Attribute Similarity Features	Name Similarity: Levenshtein distance between names, Jaccard similarity of names Age or Birthdate Comparison: Age difference, Birth year comparison Education or Workplace: Comparison of educational institutions, Comparison of workplace names Gender or Pronouns: Matching gender information, Pronoun usage in profile descriptions
Graph-Based Features	Shared Connections: Mutual friends or followers, Shared groups or communities Social Network Graph Metrics: Degree centrality, Betweenness centrality, Clustering coefficients
Textual Features	Profile Description Similarity: TF-IDF similarity of profile descriptions, Word embeddings (e.g., Word2Vec) similarity Post Content Similarity: TF-IDF similarity of post content, Cosine similarity of post embeddings
Metadata Features	Interests or Hobbies: Similarity of interests, Comparison of liked pages or groups Profile Completeness: Completeness of profile information, Presence of specific attributes (e.g., "interests" section)
Image-Based Features	Profile Picture Similarity: Image hashing similarity, Deep learning feature similarity (e.g., CNN features)
Spatio-Temporal Features	Account Creation Date Difference, Last Active Timestamp Comparison, Time Since Last Activity Location Proximity: Distance between reported locations, Common locations or regions

(Common Neighbors) index. The authors then divide these social links into strong ties and weak ties based on the weights by taking into account the social theory that asserts strong ties have a tendency to draw close friends into the same social circles. Nodes that are indirectly connected by two strong linkages are given special consideration.

In [27] authors suggest GAlign as a method for unsupervised network alignment that does not require any prior understanding of the relationships between the networks (aka anchor links). Given that this paradigm is based on rich network data and multi-dimensional embeddings. Regardless of its modality, information is tied to the network structure and to node properties that can be expressed, such as age, email address, and marital status. The model proposed in [28] is made to handle diverse social connections, content types, and profile aspects of various OSNs. The fundamental tenet is that each element of a user identification describes the actual identity owner, setting that person apart from other users. The experiment results show that Factoid Embedding outperforms the state-of-the-art methods even without training data.

Another hybrid approach (namely *INFUNE*) is presented in [29]. The information fusion component and the neighborhood enhancement component make up the model. A weighted sum of node similarity and neighborhood similarity is evaluated as the unified similarity for user identification linkage. The raw properties of users, including structure, profile, and content, coupled with known anchor linkages, are initially pre-processed as distinct similarity matrices. A set of encoders and decoders are used by the information fusion component to combine heterogeneous data and produce discriminative node embeddings for preliminary matching.

TABLE 4. Comparison of features-Based vs. Embedding-based methods in user identity linkage problem.

###	Features-Based (Sup)	Features-Based (UNSup/Semi-Sup)	Embedding-Based (Sup)	Embedding-Based (UNSup/Semi-Sup)
Problem Type	Classification	Network Alignment	Classification	Network Alignment
Feature Extraction	Attribute and structural features	Attribute and structural features	Low-dimensional embeddings	Low-dimensional embeddings
Training Data	Pre-aligned user pairs (SAPs)	None	Pre-aligned user pairs (SAPs)	None or limited SAPs
Model	Binary classifiers (e.g., SVM, logistic regression)	Graph matching, clustering	Neural networks, advanced classifiers (BERT, MLP, GPT, etc.)	Graph neural networks, embedding matching
Evaluation Metrics	Accuracy, precision, recall, F1-score	Clustering quality, structural preservation	Accuracy, precision, recall, F1-score	Alignment accuracy, MAP, etc.
Scalability	Limited by feature complexity and classifier	Limited by network size and density	Can handle large datasets with appropriate embedding	Effective for large, complex networks
Interpretability	Generally higher due to explicit features	Higher interpretability due to explicit features	Lower due to complex models	Moderate; interpretability varies with model complexity
Performance	Good performance with high-quality features	Performance varies based on clustering quality	High performance with sufficient data	High performance, especially in dense networks
Dataset Requirements	Requires quality labeled data for training	Can work with limited data, but quality matters	Often requires large datasets for effective training	Can work with fewer labels but benefits from diverse data

2) SEMI-SUPERVISED AND SUPERVISED METHODS

Some authors proposed semi-supervised approaches. The authors in [30] focused on the trustworthiness of certain users. They examine the influence of a user's social network. The main idea is that if the majority of someone's closest friends believe certain accounts across different networks belong to them, those accounts are presumed to be theirs. An Authority-Trustworthiness Analysis Model has been developed to determine each friend's authority and the reliability of their verdict. The authors address the UIL problem only if structural information (links between users) is available.

Recently, there has been increasing interest in applying graph embedding techniques for semi-supervised learning. These methods enable the extraction and representation of the structural properties of vertices in networks through low-dimensional latent vectors [31]. Certain approaches are designed to position vectors closer together in the latent space if the corresponding vertices exhibit greater similarity in their identities during the vector representation process.

In 2022 scholars in [32] suggest a multiple consistency-based anchor link prediction approach (MC). It employs intralayer structure information through network representation learning and interlayer structure information in an iterative manner. A matrix factorization-based network representation learning technique is used to learn embedding vectors that include global structural properties of nodes

TABLE 5. UIL as a network alignment problem Features-based.

UIL as a Network Alignment problem Features-based

GIVEN:

- Two or more social network platforms denoted as $P_1, P_2, \ldots, P_n$.
- Each platform P_i contains a set of user accounts denoted as $U_i = \{u_{i1}, u_{i2}, \ldots, u_{im_i}\}$.
- Each user u_{ij} has associated attributes or features, such as:
  - -- Username, email address, or other identifiers
  - -- Profile information (e.g., name, age, location)
  - -- Behavioral patterns (e.g., posting frequency, types of interactions)
  - -- Metadata (e.g., interests, groups, friends)

TASK:

Problem Definition:

- Define the task as aligning user accounts from different platforms based on their structural and attribute similarities.

Formalization:

- Let U_i and U_j be the sets of user accounts on platforms P_i and P_j respectively.
- For a given pair of user accounts $u_{ik} \in U_i$ and $u_{jl} \in U_j$, define the alignment task as finding the best matching based on feature similarities.
- The task can be denoted as finding a mapping function $f : U_i \times U_j \rightarrow [0, 1]$, where a higher value indicates a higher likelihood of the accounts being the same user.

Feature Representation:

- Define a set of features $X_{ik,jl}$ representing the similarity or dissimilarity between the attributes of user accounts u_{ik} and u_{jl}.
- Features can include:
  - -- Direct matching features (e.g., identical identifiers)
  - -- Behavioral features (e.g., posting frequency, interactions)
  - -- Attribute similarity features (e.g., name similarity, location proximity)
  - -- Graph-based features (e.g., shared friends, groups)
  - -- Textual features (e.g., similarity of profile descriptions, posts)

Model Selection and Evaluation:

- Use algorithms such as graph matching or clustering to find the optimal alignment based on the features.
- Evaluate the performance of the alignment using metrics such as structural preservation and clustering quality.
- Implement the chosen alignment algorithm and test it on a dataset of user pairs (u_{ik}, u_{jl}) with corresponding similarity scores.
- Use the alignment algorithm to predict the likelihood of a pair of user accounts representing the same individual for new, unseen data.

when employing the intralayer structural information. The mapping function of a radial basis neural network is then trained to map embedding vectors from many spaces to a single space. Finally, by taking into account both the interlayer and intralayer structures, the anchor linkages between node pairs are predicted.

Using common profile attributes, authors in [33] employ a latent user space to solve the UIL problem starting from basic profile attributes such as gender, nationality, birthday, marital status, degree, work experience, location, and educational background. Each real person has a corresponding point into the latent user space that they relate to. If a real user maintains accounts on several social media sites, each one is just seen as a projection of the real person underneath. More specifically, anything that can be seen about a real person on a social network, such as their profile traits, is a projection of that person that is bound by the feature structures that the platform offers. It follows from this model that when the data from

TABLE 6. UIL as a network alignment problem Embedding-based.

UIL as a Network Alignment problem Embedding-based

GIVEN:

- Two or more social network platforms denoted as $P_1, P_2, \ldots, P_n$.
- Each platform P_i contains a set of user accounts denoted as $U_i = \{u_{i1}, u_{i2}, \ldots, u_{im_i}\}$.
- Each user u_{ij} has associated attributes or features, such as:
  - -- Username, email address, or other identifiers
  - -- Profile information (e.g., name, age, location)
  - -- Behavioral patterns (e.g., posting frequency, types of interactions)
  - -- Metadata (e.g., interests, groups, friends)

TASK:

Problem Definition:

- Define the task as embedding user accounts from different platforms into a common latent space and aligning them based on their embeddings.

Formalization:

- Let U_i and U_j be the sets of user accounts on platforms P_i and P_j respectively.
- For a given pair of user accounts $u_{ik} \in U_i$ and $u_{jl} \in U_j$, define the alignment task as finding the best matching based on their embeddings.
- The task can be denoted as finding a mapping function $f : (u_{ik}, u_{jl}) \rightarrow [0, 1]$ in the latent space, where a higher value indicates a higher likelihood of the accounts being the same user.

Feature Representation:

- Define a set of embeddings E_{ik} and E_{jl} for user accounts u_{ik} and u_{jl} respectively, representing their positions in the latent space.
- Embeddings can be derived from:
  - -- Structural features (e.g., network topology)
  - -- Attribute features (e.g., profile information)
  - -- Behavioral features (e.g., interaction patterns)
  - -- Combined features (e.g., joint representation of structural and attribute features)

Model Selection and Evaluation:

- Use embedding algorithms such as graph embeddings, node2vec, or GCNs to learn the embeddings for user accounts.
- Evaluate the performance of the embedding alignment using metrics such as cosine similarity, Mean Average Precision (MAP) or alignment accuracy.
- Implement the chosen embedding algorithm and test it on a dataset of user pairs (u_{ik}, u_{jl}) with corresponding embeddings.
- Use the embedding model to predict the likelihood of a pair of user accounts representing the same individual for new, unseen data.

different platforms are projected to this latent space, the data points of the same user should be close to each other (ideally, they should be projected to a single data point). In essence, the more different the two users, the greater the distance between their data points in the latent user space.

Tang et al. [34] proposed the Cross-Platform User Matcher (CPUM) framework (semi-supervised) introducing the Adaptive Graph Attention Network (AdaGAT), a GNN-based encoder that models both user attributes and network topology, capturing two alignment principles: topology consistency and attribute consistency. Derived from the spectral network alignment technique AdaGAT guarantees alignment efficacy. Additionally, the framework incorporates position encoding schemes to resolve alignment confusion typical in GNN models.

Very recently some authors proposed supervised models. Xiong et al. [35] proposed a supervised model called DSANE that optimizes network structure using a complete-filter

method. It enhances linkage performance by adding relevant edges between seed user pairs across networks and removing less useful nodes. The model integrates both local and global network information through depth and breadth traversal, enriching node features.

Zheng et al. [36] introduced the JORA model that uses an inductive GCN for low-dimensional user representations while maintaining essential features. JORA is optimized via representation learning to preserve similarities within networks and alignment learning to project and align across networks using hard and attention-based soft alignment mechanisms. This helps reduce errors from predefined similarities but struggles with unsupervised tasks and static graphs, suggesting future extensions to dynamic networks.

V. UIL AS A CLASSIFICATION PROBLEM

This section will explore and analyze modern methods that treat UIL as a binary classification problem where the goal is to predict whether two nodes (one from each network) represent the same user. Labeled data (pre-aligned user pairs - SAPs) are used to train classifiers like logistic regression, support vector machines (SVM), or deep learning models. Strategies proposed in the literature were mainly *data-oriented*. Consequently, we will summarize the main strategies associated with single data categories, such as *social connections data, profile, and content data, behavioural data, spatio-temporal data*, and *network traffic data* as well as strategies that apply to combinations of these data categories. As known, raw data provides the foundational elements extracted from social media, while features are processed, transformed, and utilized for analytical purposes that are task-specific.

A. FEATURES-BASED STRATEGIES

Table 7 illustrates an overview of how this process operates with a given set of input data. It begins with feature extraction, where relevant attributes such as user profiles, behaviors, and content are gathered from various networks. Following this, classification algorithms like decision trees or SVMs are employed to classify user accounts, aiming to identify which accounts belong to the same individual based on these features. Evaluation of these methods typically involves assessing classification accuracy and other performance metrics, such as precision, recall, and F1-score, to gauge the effectiveness of the User Identity Linkage.

1) PROFILE ATTRIBUTES AND CONTENTS DATA

One of the earliest and simplest approaches is presented in [37]. This study investigates the feasibility of connecting user profiles solely based on their usernames. It makes sense that the ''entropy'' of the username string itself has a significant impact on the likelihood that two usernames correspond to the same actual individual. This research work, which is based on crawls of actual web services, demonstrates that a sizable part of user profiles can be connected using usernames. In a more recent approach [38], authors utilize Back Propagation (BP) to change the issue into a mapping problem across several social networks, which reduces the distance between username feature vectors and, to a certain extent, eliminates the need for marked user pairs and training iterations. According to the authors, 59% of users share the same username across several social networks, and such information is typically readily available.

Employing different attributes than usernames, authors in [39] examine the user profile connection across various social platforms by putting forth an effective and efficient model named MCULK, which is different from the prior work. The model has two essential parts: 1) Producing a similarity graph using profile attributes - like username, bio, etc. - that match user profiles. Then authors utilize locality-sensitive hashing (LSH) to block user profiles and only measure the similarity for those inside the same bucket in order to accelerate the creation. The second phase is: 2) Connecting user profiles using a network of similarity.

Extending the use of common profile attributes to images, in the *Hiding Your Face Is Not Enough (HYFINE)* model, a User Identity Linking model that fully utilizes photos in profiles, is presented in [40]. The HYFINE model is divided into two sections: (1) The corpus extraction method; and (2) The classification system HYFINE-c, which fully utilizes pictures together with other features to categorize two profiles to determine if these profiles are two different identities of the same user. HYFINE-e offers the option to select several profile features for each retained profile. First name/last name, free text about the person, gender, location, profile image, and the last five posts with likes, comments, shares, and, if relevant, retweets were retrieved by the authors.

Other approaches focused on the user content. In the study by [41], authors make an effort to develop a comprehensive framework for user identity linkage across various social networks that is based solely on easily accessible textual user-generated content. Employing deep learning for NLP (Natural Language Processing), authors in [42] change the challenge to a straightforward document categorization problem utilizing text content that people have posted on a social network. Authors construct a word vector space first using the messages that all users have posted. Then, they create a document vector space. To create a user's word vector, they use Word2vec. Additionally, there are two ways to create a document vector: 1) Mean-pooling: add the word vectors in users' messages to obtain the average value, which is then used as the document vector; 2) doc2vec. Similarly, in [43] authors collect all the information of a user page including username, user profile, user content, and user behavior. In this step, they use data preprocessing such as removing inactive users, ''zombie'' users, to reduce disturbance. Then the authors turn the named entity extracted from user information into 10 categories: Location, Name, Band, Company, Facility, Product, Sport, URL, Date, and others. All useful attributes of a profile can be classified

TABLE 7. UIL as a classification problem Features-based.

UIL as a Classification problem Features-based

GIVEN:

- Two or more social network platforms denoted as $P_1, P_2, \ldots, P_n$.
- Each platform P_i contains a set of user accounts denoted as $U_i = \{u_{i1}, u_{i2}, \ldots, u_{im_i}\}$.
- Each user u_{ij} has associated attributes or features, such as:
  - -- Username, email address, or other identifiers
  - -- Profile information (e.g., name, age, location)
  - -- Behavioral patterns (e.g., posting frequency, types of interactions)
  - -- Metadata (e.g., interests, groups, friends)

TASK:

Problem Definition:

- Define the task as a binary classification problem, where the goal is to predict whether a pair of user accounts from different platforms represent the same individual or not.

Formalization:

- Let U_i and U_j be the sets of user accounts on platforms P_i and P_j respectively.
- For a given pair of user accounts $u_{ik} \in U_i$ and $u_{jl} \in U_j$, define the task as predicting the binary label $y_{ik,jl}$ where:

$$y_{ik,jl} = \begin{cases} 1 & \text{if } u_{ik} \text{ and } u_{jl} \text{ are the same individual} \\ 0 & \text{if } u_{ik} \text{ and } u_{jl} \text{ are different individuals} \end{cases}$$

- The task can be denoted as learning a function $f : (u_{ik}, u_{jl}) \rightarrow y_{ik,jl}$.

Feature Representation:

- Define a set of features $X_{ik,jl}$ representing the similarity or dissimilarity between the attributes of user accounts u_{ik} and u_{jl}.
- Features can include:
  - -- Direct matching features (e.g., identical identifiers)
  - -- Behavioral features (e.g., posting frequency, interactions)
  - -- Attribute similarity features (e.g., name similarity, location proximity)
  - -- Graph-based features (e.g., shared friends, groups)
  - -- Textual features (e.g., similarity of profile descriptions, posts)

Model Selection and Evaluation:

- Formulate the learning objective as optimizing a binary classification model.
- Choose an appropriate model for binary classification, such as logistic regression, SVM, or neural networks.
- Evaluate the performance of the model using metrics such as accuracy, precision, recall, F1-score, or Receiver Operating Characteristic (ROC) curve.
- Implement the chosen model and train it on a labeled dataset of user pairs (u_{ik}, u_{jl}) with corresponding labels $y_{ik,jl}$.
- Use the trained model to predict the likelihood of a pair of user accounts representing the same individual for new, unseen data.

into one category above identified. For example, user address and workplace are considered as spatio-temporal data, while keeping e-mail address and personal website in the URL category. Then the authors allocate weight to different entities for distinguishability: (1) user-generated-content: original tweets are more believable than forwarded ones; (2) Part of one site: profile is compared to user-generated-content; (3) Different social networks: information in LinkedIn are generally more serious than in Facebook. Here authors compare the similarity between profiles across different

platforms. Then the authors transform the question into a two-class classification task.

Finally, for criminal search purposes, authors in [44] created a targeted identity resolution method that uses a dataset and a single name to search for false identities of a certain target user. Data on the person's first and last names, gender, date of birth, ethnicity, and both the person's home address and the scene of the crime are necessary for the methodology.

2) SPATIO-TEMPORAL DATA

Recently, cross-device and cross-domain UIL have received a lot of attention. Getting user linkage with spatio-temporal data produced by the numerous GPS-enabled gadgets is a key area of research. The spatial-temporal localization of user actions is used in [45] to explore a more general method of linking user IDs. The essential insight is that authors can connect any online services a user uses to their physical presence, which is determined by time and location. In [46], the authors propose a novel Spatio-Temporal User Linkage (STUL) model to address this challenge. The model consists of two main components: (1) a density-based clustering method to extract users' spatial features and a Gaussian Mixture Model to capture temporal features, where distinct weights are assigned to retrieved features–downplaying similar features and emphasizing discriminative ones to enhance the accuracy of user pair linkage; (2) a method for comparing users based on the extracted attributes, which returns user pairs with similarity scores exceeding a specified threshold. Also, the authors in [47] observe that users with similar mobility patterns frequently check in at a few common sites. So their algorithm (CP-Link) entails two phases: (1) Stay Region Building: Using a DP-based clustering technique, they first create unique stay zones for every user in order to extract their movement patterns; (2) They perform UIL based on IDWT (Inverse Discrete Wavelet Transform). To compare the stay areas of cross-domain users, they offer the IDTW time series similarity matching model. After finishing UIL, the user pair with the highest similarity is chosen as the output-linked pair. However, this class of approaches usually suffers limitations due to the complexity and dimension of location and time series representations and management. Addressing these issues, the work in [48] presents a general approach that takes into account both efficacy and efficiency at the same time while performing user account linking with location data. The authors create an innovative approach based on kernel density estimation to address the data sparsity issue. They divided an area into grid cells and focused on each cell to address the issue of data missing. In addition, the author developed an entropy-based weighting mechanism for the grid cells to address the problems brought on by negative coincidence.

To summarize, all the discussed methods focus on linking user identities across various devices or domains by utilizing spatio-temporal data. They extract features or patterns from location data to establish connections between user IDs. A common step in these methods is the comparison of users to identify similarities or matches.

The main differences rely on:

- how they handle location data (grid cells, stay zones, spatial features).
- how they compare users (similarity scores, IDTW time series similarity matching, entropy-based weighting).
- the techniques they use for extracting features or patterns (density-based clustering, Gaussian Mixture Model, DP-based clustering, kernel density estimation).
- the implementation through single/multi phases, the STUL model and CP-Link algorithm involve multiple steps, while the spatial-temporal localization method and kernel density estimation approach do not have explicit phases.
- how they address data sparsity and missing data issues, the kernel density estimation approach explicitly addresses both, while the other methods do not mention these issues.

3) NETWORK TRAFFIC DATA

According to the work [20], each user action generates one or more packets from network traffic data, and these packets of cookies and other information carry a significant amount of user account correlation. In a short period, the user's actions across several network service platforms will be reflected in the network traffic data in this way, causing previously unconnected data to exhibit a particular association. As a result, the network traffic data contains much more useful hidden association information than the material conveyed by the Web. In the case of a dynamic variable IP address, this method may reliably correlate numerous accounts of a user in network traffic with more than 85% accuracy using only the IP address and the online time. Including also IP-based features, temporal features, geo-based features, device-based features, and household similarities (information of people in one household or organization), an identity graph is built [49] by discovering identity relationships using both online data traffic and offline data logs to establish links between different identities allowing for richer insights into the consumer. Then the authors of the work use a machine learning-based approach for the Identity Graph to address the UIL task.

4) MIXED DATA

As highlighted by the authors in this work [50], traits derived from the nickname have been frequently employed to identify social connections on various social media platforms. Few works, according to the authors, have relied solely on moniker traits for identification. The authors then take into account hometown similarities while noting that various social networks may have access to various forms of location data. They discuss how to calculate hometown/location similarity using various forms of location data. Finally, they

take into account user-friendliness in order to identify the same user across several social networks. Expanding this approach, the authors in [51] propose an algorithm based on network topology and just the full-name feature of the nodes. The authors expect that the user profile contains at least the full name of a user. They formulate the problem of linking user accounts from two social networks with limited profile data as an instance of maximum subgraph matching with the noisy name feature, i.e., the full name of a user.

The authors of [52] presented an algorithm that iteratively matches profiles across social networks based on people who publish the linkages to their numerous profiles using the network structure and publicly available personal information.

Building on prior research, in [53] the authors present LIAISON (reconciLIAtion of Individual profiles across Social Networks), an algorithm that iteratively reconciles profiles across n social networks based on the presence of people who disclose the links to their various profiles. LIAISON uses the network topology and publicly available personal information.

Authors in [54] formalize the association between geo-locations and texts instead of using similarity evaluation, and they suggest a brand-new User Identity Linkage framework for locating users across networks. Moreover, by using external text-location pairs, the model can solve the label scarcity issue.

A variety of features are used by authors in [55] to conduct UIL within the same social network. Two datasets are used by the authors. The first one discussed abusive behavior on Twitter, and the second one was about terrorism. The authors took into account a number of features, including a) Profile features extrapolated from a user's profile, such as demographic data, a biography, and an avatar; b) Activity features pertaining to a user's posting behavior, such as the number of posts, replies, and mentions; c) Linguistic features extrapolated from users' posted content that may be used to model users with respect to writing style or topics of interest; d) Network properties derived from interactions in social networks between users.

B. EMBEDDING-BASED STRATEGIES

Table 8 provides a general scheme detailing how this type of task works given a set of input data. It begins with embedding generation, where network data is transformed into dense vector representations that capture both local and global structural information. Following this, classification algorithms like neural networks or support vector machines are employed to classify user accounts, aiming to identify which accounts belong to the same individual based on these embeddings. The effectiveness of these methods is generally measured by examining classification accuracy alongside key performance indicators like precision, recall, and the F1-score.

1) SOCIAL CONNECTIONS DATA (GRAPH-BASED FEATURE)

Graph-based features, such as shared friends and groups, are among the most commonly used information for the UIL task. Typically, datasets for this purpose consist of a list of ID pairs from an OSN. Each pair including user u_a and user u_b represents the relationship between the two users. Depending on the OSN considered, such a relationship can be of mutual friendship as in the case of Facebook, or, for instance, the first user in the pair follows the second one. This is the case in which the social relationship between users is not mutual so the social connection is oriented (e.g., followees or followers on Twitter).

Traditional methods often rely on either interlayer structures, which refer to the connections between nodes across different layers or networks, or intralayer structures, which refer to the connections between nodes within the same layer or network. As a result, they do not fully utilize both interlayer and intralayer structures for anchor link prediction.

In [73] the authors proposed a model to capture local and global network structures. DeepLink samples the networks and learns to encode network nodes into a vector representation. This information may then be utilized to align anchor nodes using deep neural networks. The policy gradient approach is used to learn how to transmit knowledge and update the linkage utilizing a dual learning-based paradigm. The authors in [56] also include local and global properties of a network. Specifically, the first part of the proposed model encodes the social network's graph architecture into node features. Node embeddings, a common approach in network representation learning, is what this feature learning procedure entails. By projecting the network structure to the low-dimensional node space, this embedding serves to retain both the global and local graph connection patterns, resulting in rebuilt networks that are reasonably similar to the original networks and can be easily compared for UIL predictions. Recently, a neural tensor network-based approach called NUIL [14] employs the Random Walks and Skip-gram models to incorporate the network structure in a low-dimensional vector space. In NUIL, a neural tensor network model, which is better able to express the relationships between users, takes the role of a conventional neural network model. The model first creates several social sequences for each user in several rounds of random walks, encoding the social ties between users in the social networks, before embedding users into a latent space to compare latent vectors. Assessing the individual contribution of local and global properties on the same social network, authors in [57] propose the NeXLink node embedding framework, which consists of three parts. The local structure of nodes within the same social network is first preserved in order to produce local node embeddings. The global structure, which is present in the form of the common friendship displayed by nodes involved in CNLs across social networks, is preserved in order to learn the global node embeddings. Thirdly, local and global node embeddings are

TABLE 8. UIL as a classification problem Embedding-based.

UIL as a Classification problem Embedding-based

GIVEN:

- Two or more social network platforms denoted as $P_1, P_2, \ldots, P_n$.
- Each platform P_i contains a set of user accounts denoted as $U_i = \{u_{i1}, u_{i2}, \ldots, u_{im_i}\}$.
- Each user u_{ij} has associated attributes or features, such as:
  - -- Username, email address, or other identifiers
  - -- Profile information (e.g., name, age, location)
  - -- Behavioral patterns (e.g., posting frequency, types of interactions)
  - -- Metadata (e.g., interests, groups, friends)

TASK:

Problem Definition:

- Define the task as a binary classification problem, where the goal is to predict whether a pair of user accounts from different platforms represent the same individual or not.

Formalization:

- Let U_i and U_j be the sets of user accounts on platforms P_i and P_j respectively.
- For a given pair of user accounts $u_{ik} \in U_i$ and $u_{jl} \in U_j$, define the task as predicting the binary label $y_{ik,jl}$ where:

$$y_{ik,jl} = \begin{cases} 1 & \text{if } u_{ik} \text{ and } u_{jl} \text{ are the same individual} \\ 0 & \text{if } u_{ik} \text{ and } u_{jl} \text{ are different individuals} \end{cases}$$

- The task can be denoted as learning a function $f : (u_{ik}, u_{jl}) \rightarrow y_{ik,jl}$.

Embedding Representation:

- Utilize graph embedding techniques to represent user accounts u_{ik} and u_{jl} as low-dimensional vectors e_{ik} and e_{jl} respectively.
- Embeddings capture structural properties and attribute similarities of user accounts in a latent space.
- Techniques such as node2vec, DeepWalk, or graph neural networks (GNNs) can be used to generate embeddings.

Feature Representation:

- Define a set of features $X_{ik,jl}$ based on the embeddings e_{ik} and e_{jl} representing the similarity or dissimilarity between the user accounts.
- Features can include:
  - -- Cosine similarity of embeddings
  - -- Euclidean distance between embeddings
  - -- Dot product of embeddings
  - -- Concatenation of embeddings

Model Selection and Evaluation:

- Formulate the learning objective as optimizing a binary classification model.
- Choose an appropriate model for binary classification, such as logistic regression, SVM, or neural networks.
- Evaluate the performance of the model using metrics such as accuracy, precision, recall, F1-score, or Receiver Operating Characteristic (ROC) curve.
- Implement the chosen model and train it on a labeled dataset of user pairs (u_{ik}, u_{jl}) with corresponding labels $y_{ik,jl}$.
- Use the trained model to predict the likelihood of a pair of user accounts representing the same individual for new, unseen data.

TABLE 9. Data categories most used in the literature with the corresponding works listed.

Data Category	References
Social connections data	[32], [24], [56], [14], [57], [58], [30], [24], [59], [60], [61], [62], [63], [64]
Profile attributes and contents data	[37], [38], [39], [33], [40], [41], [42], [43], [44]
Behavioural data	[65], [66], [53], [54]
Spatio-temporal data	[45], [46], [47], [48]
Network traffic data	[20], [49]
Mixed data	[50], [51], [52], [55], [67], [33], [68], [69], [27], [28], [29], [26], [70], [71], [72], [25]

integrated, to keep local and global structures and make it easier to identify CNLs across social networks. Finally, in [58] authors embed graph vertices into low-dimensional vector space to investigate a multi-granular user identity alignment system. First, the higher-order structural qualities, and second, the SAP-oriented structural consistency in the topology of social networks, are preserved by a framework's two granular layers. This framework is what authors refer to as a Multi-granular Graph Embedding framework (MGGE). Furthermore, the authors extended the model—known as the ''DeepMGGE'' model to include its capacity to capture the non-linear structural characteristics of SAP-oriented structural consistency.

To provide a robust method, the authors of [59] propose a novel supervised model called PALE that uses network embedding with awareness of observed anchor links as supervised information to capture the intrinsic structural regularities of networks rather than working directly on them as most existing methods, unsupervised or supervised, do. As a drawback, the effectiveness of the method is sensitive to the high dimension and sparsity of networks. Avoiding dimensionality limitations, authors in [60] discovered that hyperbolic geometry has the advantage of describing network hierarchical structure, whereas Euclidean geometry does not, which is prompted by current developments in geometry representation learning. As a result, the authors first discuss how social networks and hyperbolic space are related in their work. After that, the authors provide a brand-new ''HUIL'' hyperbolic geometry representation learning model for user identification linking across social networks.

Approaches based on clusters or community similarities have also proved to be effective on common UIL datasets. For instance, employing the proposed Foursquare-Twitter dataset ([61]) in [62] authors put forth a fresh embedding-based method that takes into account and makes use of both individual and community similarity by concurrently maximizing both in a single loss function. Authors in [63] accomplish identity alignment at the distribution level and take a holistic perspective of all the identities in a social network. The identities of the same natural person will be clustered together in the proposed model, which transforms the identity distribution in Twitter space by a set of operations (such as transposing) to minimize the distance between it and the identity distribution in Facebook. The authors' transformation of the social network alignment problem to the learning of the operation to minimize the distance between two distributions is motivated by isomorphism. In [64], in contrast to earlier efforts, the suggested model takes into account the multi-network scenario to encapsulate various anchor users' network architectures. For each social network, the authors suggest a high-dimensional base embedding and a low-dimensional social edge embedding to capture the various structural details of an anchor user from various social networks. In particular, using one of three possible aggregator functions—mean, max-pooling, or LSTM—with a self-attention mechanism, the authors develop a function that creates social edge embeddings by sampling and averaging structural data from an anchor user's neighborhood inside various social networks. As a downstream task, link prediction is utilized to assess how well the learned embeddings work.

Summarizing, the main differences among these approaches rely on:

- how they handle the network structure (interlayer, intralayer, local, global);
- the use of **supervised learning** (PALE, HUIL, NeXlink, NUIL), **semi-supervised learning** (DeepLink);
- the techniques they use for embedding (matrix factorization, deep neural networks, radial basis neural network, random walks, skip-gram, hyperbolic geometry).

Moreover, the strategy in [63] considers the distribution of identities in a social network, while the model in [64] considers a multi-network scenario, which is not the case for all the others.

2) BEHAVIOURAL DATA

Authors in [65] propose a solution framework, HYDRA, which consists of three steps: 1) modeling heterogeneous behavior by long-term topical distribution analysis and

multi-resolution temporal behavior matching against high noise and information missing. The behavior similarity is described by multi-dimensional similarity vector for each user pair; 2) building structure consistency models to maximize the structure and behavior consistency on users' core social structure across different platforms, thus the task of identity linkage can be performed on groups of users, which is beyond the individual level linkage in the previous study; and 3) proposing a normalized-margin-based linkage function formulation, and learn the linkage function by multi-objective optimization where both supervised pair-wise linkage function learning and structure consistency maximization are conducted towards a unified Pareto optimal solution.

The order of friending in actual dynamic social networks is utilized by the authors in [66]. In reality, social psychology research shows that an individual's friendship growth across social networks is predominately deterministic rather than stochastic [74].

3) MIXED DATA

With the aid of a dynamic hypergraph neural network, in [67] the feature extraction model learns node embeddings from topology space and feature space. In the WGAN training phase, the network alignment model employs a new sampling technique that places more emphasis on sample-level data. The outcomes of thorough tests conducted by the authors on the real-world dataset appear to confirm the efficiency of the suggested framework.

In [33] employing the ego networks of two users as input, authors formalize the user alignment across social networks as a classification problem. In order to align the users, the authors propose a graph neural network model called MEgo2Vec to describe the matched ego networks of the two users as a low-dimensional real-valued representation. The representation is divided into two parts: one is an embedding from the target user pairs' and their neighbor pairs' attributes, and the other is an embedding from the matching ego network's topologies. In a later work [68], the authors model the topics of user interests to represent the content information in different social networks at the same granularity and filter out the noise. Second, they capture friend-based (i.e., structure) and interest-based (i.e., content) user co-occurrence in linked heterogeneous networks using four types of sub-networks (i.e., user-user intra/inter-network and user-topic intra/inter- network). Third, they learn effective user representations by embedding the sub-networks into a unified low-dimensional space. Also in [69] - where the authors propose the MASTER framework - are integrated attribute and structure embedding for reconciliation across several social networks. In this framework, in order to define the problem as a unified optimization, authors first build a novel constrained dual embedding model by simultaneously embedding and reconciling several social networks.

The majority of approaches ignore the social network attribute data. To solve the issue, the authors in [70] suggest a brand-new semi-supervised network-embedding approach. Each node of the numerous networks is represented in the model by a vector for predicting anchor connections, which is learned with knowledge of the observed anchor links as semi-supervised information and input, as well as topology structure and attributes. The suggested model outperforms several state-of-the-art methods, as shown by experimental findings on real-world data sets.

In [71] authors study the M-NASA problem to identify the anchor links among multiple anonymized social networks. In addition to its significance, the M-NASA problem is a brand-new problem that is entirely distinct from previous efforts. The suggested procedures are as follows: (1) supervised anchor link inference across social networks, which focuses on inferring the anchor links between two social networks with a supervised learning model; (2) network matching, which investigates various heuristics to match two networks based on the known existence probabilities of potential correspondence relationships; (3) entity resolution, which aims at discovering multiple references to the same entity in one single database with a relational clustering algorithm; (4) cross-media user identification connects users from different networks based on data from multiple node attributes produced by users' social interactions.

The framework proposed in [72] uses word2vec [18] and DeepWalk [75] to first turn all textual and structural user data into low-dimensional latent spaces, then it integrates various user features and predicts empty data fields using a late fusion technique and computations based on cosine similarity. The outcomes demonstrated that by enhancing and modernizing data sources as needed, the methodology may successfully capture dynamic user data and improve the performance of identity linkage models.

Very recently, Li et al. [76] in order to address the challenge of semantic gaps across social platforms, proposed MFLink, a multimodal fusion method that combines attributes, post content, and social relationships using graph neural networks, attention mechanisms, and adversarial learning to align user representations across platforms.

VI. EVALUATION METRICS

The second research question **RQ2**, concerning current SOTA performance, is in line with the practical goal of this investigation, which aims to serve as a practical guide for potential applications. However, being able to compare the plethora of strategies presented in the literature in terms of performance, requires analyzing and comparing them under the same setting which is beyond the scope of this review.

Trung et al. [77] undertook an endeavor in this regard, presenting a comprehensive empirical examination of the effectiveness of various network alignment methods. They specifically combine a number of cutting-edge network alignment strategies in a comparable way and assess various settings to gauge the individual properties of these

TABLE 10. A synthesis of the most innovative algorithms mentioned in this survey.

Algorithm Name	Sup/UNSup Semi-Sup	Description	Data Category	Ref.
NUIL	Sup	Neural tensor network-based approach to UIL	Social Connection	[14]
NeXLink	Sup	Node Embedding Framework for Cross-Network Linkages Across Social Networks	Social connection	[57]
DeepMGGE	Sup	Deep multi-granularity graph embedding for user identity linkage across social networks	Social connection	[58]
FRUI-P	UNSup	Friend Relationship-based User Identification algorithm without Prior knowledge	Social connection	[24]
PALE	Sup	Predict Anchor Links across Social Networks via an Embedding	Social connection	[59]
STUL	Sup	Spatio-Temporal User Linkage	Spatio-Temporal	[46]
CP-Link	Sup	Check-in Patterns for User Identity Linkage	Spatio-Temporal	[47]
HYFINE	Sup	Hiding Your Face Is Not Enough. A User Identity Linking model that fully exploits images in profiles.	Profile Att. and C.	[40]
MEgo2Vec	UNSup	A graph neural network model to describe the matched ego networks of the two users as a low-dimensional real-valued representation	Mixed	[33]
MASTER	Semi-Sup	Across Multiple social networks, integrate Attribute and STructure Embedding for Reconciliation	Mixed	[69]
GAlign	UNSup	A fully unsupervised network alignment framework based on a multi-order embedding model.	Mixed	[27]
INFUNE	Semi-Sup	Information fusion component and the neighborhood enhancement component	Mixed	[29]
LIAISON	Sup	ReconciLIAtion of individuals profiles across social network	Mixed	[53]

techniques with the ultimate goal of providing a benchmark framework useful to identify the best strategy for each scenario. The benchmark findings, which were achieved using both real data and synthetic data, are then thoroughly analyzed. The datasets employed are: Douban, Flickr-lastfm, Flickr-myspace, fb-tw, fq-tw. Interestingly, for several of the models tested, the authors find that on these real datasets, accuracy is equal to 0.00 confirming that each specific scenario has its most suitable network alignment technique since no single technique consistently outperforms all others.

For these reasons, in this section, we will first give an overview of the evaluation metrics most used in the User Identity Linkage task in its two main formulations adopted in the present survey. Then, in Table 11 we will report the results achieved by different research groups (each one implementing a different UIL strategy) applying these metrics on the same dataset, the Forsquare-Twitter dataset. Reported performance values are those declared by the authors in their published research works both directly through numbers and indirectly by charts.

A. EVALUATION METRICS SPECIFIC FOR UIL AS A NETWORK ALIGNMENT TASK

1) ALIGNMENT ACCURACY

Alignment accuracy measures the proportion of correctly identified anchor links (ALs) (true matches) out of the total

number of possible anchor links. It is calculated as:

$$\text{Align. Accuracy} = \frac{\text{Number of correctly predicted ALs}}{\text{Total Number of true ALs}}$$

2) MEAN AVERAGE PRECISION (MAP)

Mean Average Precision (MAP) is a metric used to evaluate the accuracy of ranking models, considering both the precision of results at different cutoff levels and their order. The average precision (AP) for a single query or user is given by:

$$\text{AP} = \frac{1}{m} \sum_{k=1}^{m} P(k) \cdot \text{rel}(k)$$

where:

- m is the total number of true positives.
- $P(k)$ is the precision at rank k.
- $\text{rel}(k)$ is a binary indicator function that equals 1 if the item at rank k is relevant and 0 otherwise.

The MAP is then the average of these AP values across all queries or users:

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}(q)$$

where Q is the total number of queries or users.

3) NORMALIZED DISCOUNTED CUMULATIVE GAIN (NDCG)

NDCG evaluates the quality of the ranked list of results, giving higher scores to correct matches appearing higher in the ranked list. The DCG at position p is calculated as:

$$\text{DCG}_p = \sum_{i=1}^{p} \frac{2^{\text{rel}(i)} - 1}{\log_2(i+1)}$$

where $\text{rel}(i)$ is the relevance score at rank i.

NDCG is the normalized version of DCG, where DCG is divided by the ideal DCG (IDCG), which is the DCG for the ideal ordering of results:

$$\text{NDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}$$

IDCG is calculated as:

$$\text{IDCG}_p = \sum_{i=1}^{|\text{REL}_p|} \frac{2^{\text{rel}(i)} - 1}{\log_2(i+1)}$$

where $|\text{REL}_p|$ is the set of relevant items up to position p.

4) STRUCTURAL PRESERVATION

Structural preservation is an evaluation metric used to measure how well the alignment of users across different social networks preserves the structural properties of the original networks. It assesses whether the inherent relationships and connections within the networks are maintained after the linkage.

To evaluate Structural Preservation, we typically look at the consistency of the structural properties, such as the degree distribution, clustering coefficient, and shortest path length, between the original and the aligned networks.

5) DEGREE DISTRIBUTION PRESERVATION

The degree of a node in a network is the number of connections (edges) it has to other nodes. Degree distribution preservation ensures that the degree of nodes in the aligned network is similar to their degree in the original networks.

$$D(v) = \text{Degree of node } v$$

For an aligned node v across two networks G_1 and G_2, the Degree Preservation (DP) can be measured as:

$$\text{DP} = \frac{1}{|V|} \sum_{v \in V} \left| D_{G_1}(v) - D_{G_2}(v') \right|$$

where $D_{G_1}(v)$ and $D_{G_2}(v')$ are the degrees of node v and its aligned counterpart v' in networks G_1 and G_2, respectively.

6) CLUSTERING COEFFICIENT PRESERVATION - CCP

The clustering coefficient of a node measures the extent to which its neighbors form a complete graph (i.e., are interconnected). Preservation of clustering coefficients ensures that the local neighborhood structure around each node is maintained.

$$C(v) = \frac{2 \times \text{Number of closed triplets}}{\text{Number of connected triplets}}$$

For an aligned node v:

$$\text{CCP} = \frac{1}{|V|} \sum_{v \in V} \left| C_{G_1}(v) - C_{G_2}(v') \right|$$

7) SHORTEST PATH LENGTH PRESERVATION - SPLP

The shortest path length between two nodes is the minimum number of edges required to connect them. Preservation of shortest path lengths ensures that the overall connectivity and distances between nodes are maintained.

$$\text{SPL}(u, v) = \text{N. of edges in the SP between nodes } u \text{ and } v$$

For an aligned pair of nodes u and v in networks G_1 and G_2:

$$\text{SPLP} = \frac{1}{|E|} \sum_{(u,v) \in E} \left| \text{SPL}_{G_1}(u, v) - \text{SPL}_{G_2}(u', v') \right|$$

where E is the set of edges, and u' and v' are the aligned counterparts of u and v in the other network.

8) AGGREGATE STRUCTURAL PRESERVATION SCORE - AGGREGATE SPS

An aggregate score for structural preservation can be calculated by combining the individual preservation metrics, typically using a weighted sum or average:

$$\text{SPS} = w_1 \times \text{DP} + w_2 \times \text{CCP} + w_3 \times \text{SPLP}$$

where w_1, w_2, and w_3 are weights that can be adjusted based on the importance of each structural property in the specific application.

B. EVALUATION METRICS FOR BOTH UIL PROBLEM FORMULATIONS

The following are the common metrics used in literature for different formulations of UIL-related tasks. The metrics are often adapted from study to study with different meanings. Here we collect the generic definition of each metric with a specific comment on the suitability for UIL tasks.

In the field of machine learning for classification tasks, a *True Positive* (TP) is an actual positive sample correctly predicted by a model as positive. Similarly, a *True Negative* (TN) is an actual negative sample correctly predicted as negative. A *False Positive* (FP) is an actual negative sample misclassified as positive. Finally, a *False Negative* (FN) is an actual positive sample misclassified as negative. Specifically for UIL a TP usually represents a correctly predicted link between users that are actually the same real person. TN is a non-existent link correctly non-predicted, FP and FN a wrongly predicted link and a non-predicted (but actually existent) link respectively.

1) ACCURACY

Accuracy is the ratio of correct predictions on the total observations and is given by the Equation 1. *Accuracy* is one way to measure what percentage of predictions are right.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In the specific field of UIL, given a predicted list of pairs with possible links between users from different social networks, *accuracy* can measure how many links were correctly predicted by the system. However, the TNs component of the equation does not generally contribute (i.e., the interest is in linking two users that are actually the same real person, instead of predicting non-existent links) significantly for UIL-related tasks.

2) ERROR RATE

Closely related to *accuracy* is the *error rate*. The definition is given by the Equation 2. The *error rate* expresses the percentage of predictions that are wrong.

$$ErrorRate = 1 - Accuracy = \frac{FP+FN}{TP+TN+FP+FN} \tag{2}$$

Depending on how genuine positives and negatives are defined in a multilabel scenario, the definition of this metric may differ. A prediction is deemed accurate (referred to as ''subset accuracy'') when the projected labels exactly match the actual labels. Alternately, before the *accuracy* calculation, predictions can be flattened and condensed to a single-label task. As for *accuracy*, *error rate* is not often used for UIL tasks but is more common for similar tasks such as link predictions and friend recommendation.

3) PRECISION

Equation 3 defines *precision* or *sensitivity* as the ratio of true positive (TP) observations to all-around positive predicted values (TP+FP). *Precision* is the proportion of correctly predicted events among all positively predicted events.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Precision is one of the most common metrics used in UIL-related tasks. When a model provides a list of predicted links between two users from different social networks, this metric measures how many of the predicted links are actually linking the same real person.

4) RECALL

Equation 4 gives *recall* or *specificity* as the ratio of true positive (TP) observations to all-around positive predicted values (TP+FN). *Recall* is the ratio of right predictions made overall positive predictions that should have been made.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

For scenarios involving multi-class classification, it is possible to compute the *precision* and *recall* for each class label. Also, *recall* is one of the most common metrics used in UIL-related tasks. In this case, given all the actual links between different users from different social networks, this metric allows measuring how many links were correctly predicted.

5) F1 SCORE

Equation 5 illustrates the *F1 score*, which is the harmonic mean of *recall* and *precision*. The maximum *precision* and *recall* value of an *F1 score* is 1, while the lowest value is 0.

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{5}$$

F1 score is often used in UIL-related tasks to provide, with a single scalar, the performance of the model in predicting links considering both the *precision* and the *recall* already discussed.

6) MATTHEWS CORRELATION COEFFICIENT (MCC)

The effectiveness of binary classification techniques is also measured by the *Matthews Correlation Coefficient (MCC)* [82], which collects all the data in a confusion matrix. MCC can be used to address issues with unequal class sizes and is still regarded as a balanced approach. The MCC scales from -1 to 1. (i.e., the classification is always wrong and always true, respectively). Equation 6 provides the formula for MCC.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

For the same consideration about *accuracy* (i.e., the interest in predicting non-existent links between users from different persons, the case of TNs) MCC is not frequently used in the literature for UIL.

TABLE 11. SOTA results on the Foursquare-Twitter dataset [78]. For each metric or variant of a metric (e.g. P@k) we only report the best result provided by the authors. Some results are reported using ∼ as they are extracted from graphs published in the works cited and not from explicit numerical reports.

Paper	Precision	Recall	F1	AUC
Ma et al., 2021 [79]	0.63	0.60	0.615	0.82
Wang et al., 2018 [80]	∼0.8	∼0.4	-	∼0.5
Riederer et al., 2016 [81]	∼0.85	∼0.3	-	∼0.4
Zhou et al., 2018 [73]	0.7048	0.7914	-	0.991
Zhou et al., 2019 [44]	0.7653	0.9031	-	-
Chen et al., 2018 [48]	∼0.4	∼0.4	∼0.4	-
Ding et al., 2020 [47]	∼0.7	∼0.6	∼0.65	∼0.82
Feng et al., 2020 [45]	∼0.4 [2]	-	-	-
Chen et al., 2017 [46]	∼0.8	∼0.52	∼0.62	-
Shao et al., 2021 [54]	-	-	0.8926	0.9327

7) AUC

The area under the ROC curve (AUC) evaluates the accuracy of similarity rankings by comparing the true positive rate (TPR) and false positive rate (FPR) across various thresholds. The AUC measures the probability that a randomly selected positive instance ranks higher than a randomly selected negative one, serving as an indicator of the model's effectiveness in distinguishing between the two classes. The returned user account pairs are said to be "positive" in this case if they belong to the same user.

Finally, some specific metrics related to multilabel tasks are Micro and Macro-F1 [83], Precision@k and Normalized Discounted Cumulated Gains [17].

VII. DATASETS

To evaluate the performance of AI methods, agreed benchmark datasets are needed. Until 2016 there were many datasets related to a single social network, but very few datasets were available to be used as ground truth for performing UIL tasks across social networks, thus models validation was challenging. Today the problem of obtaining a comprehensive dataset with different feature spaces still exists but a few steps forward were taken. Considering that a detailed catalog of all the datasets proposed in the literature misses, we collect in Table12 the datasets used in two or more works and built in a previous study. It is worth mentioning that not all of these studies utilized identical partitions of the same dataset. Furthermore, several different metrics are employed, making difficult an objective evaluation of models on the same dataset as clarified above. In Table12 we report when each dataset was first presented along with the year, the reference paper, and the papers reporting the studies where they were used. We also provide the dimension of each dataset (*Size* column) highlighting that *Size* refers to the number of distinct identities linked in the set of social platforms (which can include 2, 3, or more platforms) considered for each dataset. When the cross-dimension of the set is not specified by the dataset authors, we report here the greater number of identities in the set. It is worth noticing that while 16 datasets are reported in the table, several others - built-up ad-hoc for a single and specific work - are available in the literature. In some cases, also subsets of the datasets shown in the table have been used. In fact, the large majority of the works on UIL are based on novel datasets presented along with the new proposed approach discussed in the corresponding paper. From Table 12 it can be seen that the two top-referenced and used datasets in the literature are FT2 and TF2. Here we briefly introduce both of them.

For the *Foursquare-Twitter (FT2)* dataset, the scholars crawled user profiles together with their online tips. Tips and profiles are 94,187 and 5,392 respectively. The total number of places crawled from Foursquare is 38,921 and all tips can be attached to location check-ins. At Foursquare, the two unidirectional follow links that were created from the bidirectional buddy link have replaced the original follow links. Similar to this, Twitter is crawled for 5,223 people and all of their public tweets. The number of tweets crawled by authors is 9,490,707, among which 615,515 tweets contain location check-ins and they constitute around 6.5% of all the tweets. 297,182 locations in total were gleaned from the tweets.

The second top-referenced dataset is crawled from Twitter and Foursquare. The website for Foursquare, a typical location-based social network, was the first one crawled (LBSN). By performing a breadth-first search over the social graph, the authors gathered a dataset of 500 people and 7,504 tips from these users. The latitude and longitude of each tip, as well as the timestamp, are available. The Foursquare network additionally offers information on who a user is following or friending. These connections can show how socially connected the users are. Then the scholars gathered 500 people, matching the 500 Foursquare users, and 741,529 tweets from the individuals. In the Twitter network, all tweets contain a time stamp, and some tweets also contain a location stamp. In the end, the authors had 34,413 tweets in total with location information (latitude and longitude), which represents 4.6% of all the tweets we gathered.

VIII. UIL - DISCUSSION

This review reveals that User Identity Linkage on social networks has evolved significantly, with advancements in

TABLE 12. List of Datasets for UIL with evidence of year of creation, reference of the research work where it first appeared (Collected in), references to subsequent works in which the dataset was used (Used in), number of identities (Size) and publicly availability (P.A.), at the time we accessed them - January 23, 2024.

Dataset	Year	Collected in	Used in	Size	P.A.
Data Mining - Machine Learning (DM1)	2020	[58]	[32]	22,542	Yes
Douban Online - Douban offline (DD1)	2016	[84]	[27]	3,906	Yes
Douban Online - Douban offline (DD2)	2012	[85]	[70]	$\sim 50,000$	Yes
Facebook (FB1)	2009	[86]	[59], [57]	90,269	Yes
Facebook (FB2)	2014	[87]	[30]	4,039	Yes
Flickr - LastFM (FL1)	2015	[88]	[70]	215,495	Yes
Flickr - MySpace (FM1)	2016	[84]	[27]	10,733	Yes
Flickr - Twitter (FT1)	2013	[89]	[68]	1,457	Yes
Foursquare - Twitter (FT2)	2014	[78]	[79], [80], [81], [73], [44], [48], [47], [45], [16], [46], [54]	7,227	No
Instagram - Twitter (IT1)	2016	[81]	[46]	1,717	No
Instagram - Twitter - Google+ (ITG1)	2018	[90]	[39]	7,729	Yes
Lastfm-MySpace and Livejournal-MySpace (LL1)	2014	[88]	[73]	854,498 - 3,017,286	Yes
Twitter - Flickr (TF1)	2017	[91]	[72]	7,109	No
Twitter - Foursquare (TF2)	2014	[5]	[92], [62], [93], [61], [58], [69], [57], [14]	500	Yes
Twitter - LiveJournal - YouTube - Flickr	2012	[94]	[52], [53]	93,169	No

graph-based and deep-learning methods driving much of the progress. Particularly, the use of graph convolutional networks (GCNs) and (very recently) graph attention networks (GATs) has enabled more sophisticated modeling of user relationships across multiple platforms. These methods leverage the graph structure of social networks to capture latent patterns that improve linkage accuracy. More recent approaches focus on optimizing the network structure itself, improving linkages by adjusting the connections between nodes based on their relevance. Additionally, transfer learning approaches, such as domain adaptation, have shown promise in addressing the variability between platforms by adapting models trained on one platform for use on another, reducing the need for large-scale labeled datasets.

In light of the conceptual framework introduced in this survey, we can make the following considerations.

The strength of **feature-based** methods in classification lies in their ability to distinguish between users based on concrete attributes. However, these approaches are highly dependent on the quality and availability of labeled data. Additionally, the success of classification-based approaches can vary based on the types of platforms involved. For example, LinkedIn, with its professional and structured data, may provide clearer linkage opportunities compared to platforms like Facebook or Twitter, where user behavior is more casual and fragmented. Moreover, the reliance on explicit features means that feature selection becomes crucial; irrelevant or noisy features can degrade model performance. On the other hand, when UIL is framed as a network alignment problem, feature-based methods focus on structural similarities between networks. Those methods typically extract features related to user connections (e.g., friendships or interactions) and evaluate how well these relationships are preserved across networks. For instance, users who share a strong connection in one network may have a similar connection in another network, which can serve as a clue for linkage. In this context, unsupervised methods often shine, as they do not require labeled data. These methods operate by assuming that social behaviors—such as friendships or user interactions—are consistent across platforms. While these methods can be highly effective in networks where stable, predictable relationships exist, they falter in cases where the network structure is fragmented or evolving. Additionally, the absence of labeled data makes it

harder to evaluate or improve these methods, limiting their precision in dynamic environments.

Hybrid approaches in network alignment combine features such as node similarity and network structure, fusing user profile attributes with social connections to create a more robust linkage. These methods refine user matching over time by iteratively improving the alignment based on new data or connections. However, this increased complexity requires more computational resources and high-quality input data, particularly in large-scale networks with heterogeneous data sources.

Overall, feature-based methods for UIL as a classification problem emphasize user-specific attributes and rely on labeled training data, while those formulated as a network alignment task leverage structural properties and social relationships without labeled data. Both approaches face challenges, including data sparsity, complexity, and dependency on the quality of extracted features or the stability of network relationships.

Focusing on **embedding-based** methods, in the classification problem the embeddings provide a compact representation of various features, such as profile attributes, interaction patterns, or content similarity. By comparing these embeddings, models can classify profile pairs as linked or unlinked. The advantage of embedding-based methods here is that they can capture complex, non-linear relationships between users' data, which improves the classification accuracy compared to traditional feature-based methods.

In the network alignment problem formulation, embedding-based methods focus on aligning the structural properties of social networks allowing for a more flexible and scalable approach to UIL since embeddings can capture complex patterns of social connections and user behaviors. Overall, embedding-based methods in both formulations offer the advantage of reducing data dimensionality while preserving critical patterns, leading to more accurate and scalable UIL solutions. However, they can be computationally expensive, especially when handling large-scale networks, and require careful tuning to balance local and global structural information effectively.

Across all approaches, the key trade-offs involve balancing computational complexity, data quality, and accuracy, with unsupervised methods providing scalability and ease of implementation but lacking in adaptability, while supervised techniques offer greater precision at the cost of increased resource requirements. The evolution of UIL models reflects a growing integration of diverse data types and techniques aimed at improving identity linkage across increasingly complex and varied networks.

Despite all the innovations introduced, several challenges remain unresolved. Data privacy restrictions continue to limit the availability of real-world datasets, and existing models struggle to generalize across heterogeneous platforms due to the complexity of user behaviors and the diversity of features available on different social networks.

IX. UIL - PRIVACY AND ETHICAL CONCERNS

In the previous section, we highlighted one of the major challenges in solving the UIL task, which stems from the data privacy restrictions that social platforms have begun to enforce. In this section, we discuss how these restrictions represent the other side of the coin, as they respond to the urgent need to address the ethical issues raised by incorrectly implemented user identity linkage. As technology evolves to facilitate the linking and analysis of user profiles across different social networks, concerns arise regarding the use, sharing, and potential exploitation of this data. Users may be unaware of the extent to which their personal information can be aggregated, raising issues of informed consent and transparency. Importantly, some users may not want their identities linked across platforms. To address these ethical concerns, it is crucial to implement safeguards and guidelines for the responsible use of UIL technology. For example, strong data protection measures like encryption and anonymization can help mitigate privacy risks. Clear policies regarding data usage and retention are also necessary to ensure users are fully aware of how their data is being handled. Additionally, ethical governance frameworks for data practices should be established. These could include regular audits, mechanisms for user feedback, and involvement from key stakeholders to promote accountability. Lastly, promoting digital literacy education can empower users to make informed decisions about their online identities and the broader implications of UIL technologies. The present work goes beyond the scope of providing an exhaustive discussion on ethical and privacy concerns. A more in-depth analysis of these issues can be found in several works in literature [95], [96], [97], [98], [99].

To cite a few of them: the work by Chandok et al. [97] introduces a framework to help users control their *linkability*, which is the likelihood that an adversary can identify them across multiple networks. This framework uses a *linkability* calculator and provides notifications or visual alerts when users may be at risk, helping 75% of study participants make more informed decisions, thus minimizing unintended exposure. The study by Backes et al. [96] explores whether anonymity within a single platform can prevent cross-site identity linking. The researchers find that anonymity alone is insufficient to assess risks and propose an absolute *linkability* measure that accurately predicts the likelihood of identity linkage across platforms in 75% of cases. A broader discussion of the problem can be found in the book by Christen et al. [95], which explores effective methods for preserving linked data securely. The book is written from the perspective that data linkage is a valuable tool for scientific research, acknowledging that, like any tool, it could be misused if not carefully regulated. Consequently, a societal consensus on the responsible use of such techniques is essential. The methods presented are crafted to minimize the risks associated with improper use of linked data, emphasizing ethical standards and protective measures.

In summary, it is clear, that while an important part of the academic community focuses on finding optimal strategies to address the User Identity Linkage task, another equally important segment is dedicated to the study of systems designed to hinder the reckless application of UIL, which poses a serious threat to users' data.

X. UIL - OPEN ISSUES AND RESEARCH AGENDA

Concerning the **RQ3**, *What open issues still remain in UIL*, as detailed in the previous sections, researchers are proficiently working to provide more and more solutions to address feature extraction and model construction steps in the UIL framework in order to adequately face noisy, imperfect and largely unstructured data coming from different social networks. However, some open issues in this field still remain in this field, primarily concerning: (i) data and datasets; (ii) evaluation; (iii) dynamic UIL; (iv) unsupervised models.

Regarding the latter, it is useful to cite the new potential offered by the advent of Large Language Models (LLMs) [100], [101], [102], [103] X-0e as briefly described in the following paragraphs.

a: OPEN ISSUE: DATA AND DATASETS

There is no established benchmark dataset for assessing and going beyond existing approaches. The number of publicly accessible datasets that have complete profiles, content, and network information is limited. However, existing datasets with partial features (like user names and network architecture) are available. A relevant issue related to the task concerns the ground truth. Finding user identity pairs that match across social media websites has become even more complex than before, especially when users make content private for the purpose. Also, getting a large dataset for research purposes is an open and challenging task. After the 2016 European GDPR, several concerns and limitations about user privacy were issued. Accessing user identity attributes and using them without violating the user's privacy is more challenging than before. From a practical point of view, OSN restrictions also limit the ability to crawl data via APIs. While some social network websites offer APIs for adequate data access, they frequently impose rate limits and place restrictions on permission, which makes it challenging to collect data on a big scale.

b: OPEN ISSUE: EVALUATION

The effectiveness of a UIL model could be assessed using a wide range of metrics. The review of the literature reveals that choosing a metric frequently depends on the kind of model used, making it difficult to consistently compare different models. The model, in turn, depends not only on the data sources but also on the specific application domains involved. It is still true that there is no definitive method for User Identity Linkage that is universally applicable. For instance, the models suitable vary based on whether you want a top-k matching or a perfect matching between pairs, and as a result, the metrics to utilize differ. Furthermore, the substantial imbalance between matching and non-matching user identity pairs, which is a structural component of every dataset being worked on, has a significant impact on performance evaluation.

c: RESEARCH AGENDA: DYNAMIC UIL

Furthermore, OSNs are always evolving in a dynamic manner. As time passes, profile, content, and network features for user identities continue to evolve and new links between friend users or links between the same person across different OSNs can be created. For this reason, the poor performance of a model for the UIL task could be motivated by a not-yet available online link but actually existing in the real world between the same person. At the same time, several advancements have been accomplished in addressing the UIL task thanks to the advent of more effective deep learning architectures. Given the modern large pre-trained models [100] and the embedding-based models available [18], [19], it is easier to involve user content on an OSN in the UIL task and compare embeddings of different users for tasks like link prediction and recommendation [104], [105], [106].

d: RESEARCH AGENDA: UNSUPERVISED MODELS

In terms of unsupervised techniques, only three works had been examined up until 2016, and this field had been deemed to be understudied. The current review demonstrates that efforts to propose unsupervised methods have increased. We gathered five studies from after 2016 and three of them suggest novel unsupervised methods based on multi-dimensional embeddings and rich network data [27], factoid embedding [28], and co-training algorithms [25] that manipulate two independent models (attribute-based model and the relationship-based model), and makes them reinforce each other iteratively. However, although they exceed the state of the art of the previous unsupervised ones in performance, they do not outperform the state-of-the-art supervised, demonstrating that this area remains open for future research.

e: RESEARCH AGENDA: POTENTIAL ROLE OF LLMS IN ACCOMPLISHING UIL TASK

Large Language Models (LLMs) such as GPT-4, PaLM 2, Claude, and LLaMA to cite the most powerful (according to ChatBot Arena's leaderboard[3]): offer advanced natural language processing capabilities that could significantly enhance User Identity Linkage across social networks. These models could be used to efficiently extract and represent features from user-generated content, profiles, and behaviors, enabling the detection of similarities indicative of the same user across different platforms. By converting textual content into dense vector representations, LLMs could create a unified embedding space that captures semantic nuances and

[3]https://chat.lmsys.org/ (2024-07-15)

contextual information, improving the accuracy of matching user profiles.

Additionally, LLMs could effectively handle entity resolution and disambiguation, identifying and differentiating between users with similar or identical names based on contextual cues. Their integration with graph-based methods, such as Graph Neural Networks (GNNs), should further enhance UIL by leveraging both textual and structural data, thus providing a more comprehensive analysis. Moreover, LLMs could address issues of data sparsity and heterogeneity through transfer learning and the imputation of missing data, ensuring better generalization across different platforms with varying data formats.

Cutting-edge strategies such as *In-context learning* [107], [108], [109] should be explored in conjunction with generative models. In-context learning refers to the ability of machine learning models to understand and utilize the surrounding context of an input prompt to enhance the accuracy and relevance of their predictions. This approach can be particularly beneficial in User Identity Linkage, where context may include information about user behaviors, preferences, and interactions, enabling the creation of more accurate links between different digital identities. On the other hand, crafting specific inputs (prompt engineering [110]) that guide large language models (LLMs) to generate precise and contextually relevant responses can help models to effectively detect and associate user identities across platforms, harnessing the full scope of reasoning and knowledge capabilities in generative models like GPT. When used strategically, both approaches can greatly enhance the accuracy and efficiency of UIL systems, expanding potential applications in digital identity management and analysis.

As already said, privacy considerations are also crucial, and LLMs can be deployed in privacy-preserving frameworks like federated learning, allowing for decentralized data learning without compromising sensitive user information. They could also facilitate anonymization and pseudonymization, enabling identity linkage without directly exposing personal identifiers. Overall, LLMs have the potential to enhance the accuracy, robustness, and scalability of UIL solutions, making them invaluable in the ongoing efforts to link user identities across diverse and continually evolving social networks.

To the best of our knowledge, there are currently no studies in the literature that utilize these language models to address the UIL task.

XI. CONCLUSION

The process of tying together user accounts on different OSNs is challenging and attracted more and more research attention in the last two decades. The current work provides a comprehensive review of recent studies (from 2016 to the present) on User Identity Linkage (UIL) methods across online social networks by outlining various feature extraction strategies, algorithms, machine learning models, datasets, and evaluation metrics proposed by researchers working in this area. The proposed overview takes a pragmatic perspective to highlight the concrete possibilities for accomplishing this task depending on the type of available data. To this purpose, we offer a practical guide for other researchers in the field enriched with useful points of reference regarding algorithms, models, datasets, and evaluation metrics. Our work provides the following contributions: (i) An updated overview of the body of knowledge on UIL to fill the chronological gap with respect to previous literature reviews and identify the aspects in which the major innovations have been introduced; (ii) A functional approach to the task that identifies two main problem formulations: as either a classification task or a network alignment task, to explore and provide guidance for more practical problem settings in UIL across social networks; (iii) A useful collection of the datasets used for UIL task, not built ad hoc for a single experiment and shared in several research works. Each entry includes a reference and indicates whether it is publicly available.

The proposed excursus shows that several advances have been accomplished in addressing the UIL task thanks to the advent of more effective deep learning architectures. However, some issues remain open and they mainly rely on the limited availability of benchmark datasets whose construction is even more complicated by the current social network access policies that reinforce privacy protection and reduce the possibility of accessing the data through API (see recent updates on Twitter APIs[4]).

REFERENCES

[1] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 51–62.

[2] S. Bartunov, A. Korshunov, S.-T. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *Proc. 6th SNA-KDD Workshop*, 2012, pp. 1–9.

[3] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proc. VLDB Endowment*, vol. 7, no. 5, pp. 377–388, Jan. 2014.

[4] H. Zhang, M.-Y. Kan, Y. Liu, and S. Ma, "Online social network profile linkage," in *Proc. Asia Inf. Retr. Symp.*, Kuching, Malaysia. Cham, Switzerland: Springer, 2014, pp. 197–208.

[5] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2013, pp. 179–188.

[6] F. Carmagnola and F. Cena, "User identification for cross-system personalisation," *Inf. Sci.*, vol. 179, nos. 1–2, pp. 16–32, Jan. 2009.

[7] Z. Deng, J. Sang, and C. Xu, "Personalized video recommendation based on cross-platform user modeling," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[8] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migration patterns in social media," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 1204–1209.

[9] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *Proc. 3rd Int. ICWSM Conf.*, 2009, pp. 354–357.

[10] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, and P. Zola, "Coordinated inauthentic behavior and information spreading on Twitter," *Decis. Support Syst.*, vol. 160, Sep. 2022, Art. no. 113819. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923622000902

[11] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explor. Newslett.*, vol. 18, no. 2, pp. 5–17, Mar. 2017.

[4]https://developer.twitter.com/en/docs/twitter-api/migrate/whats-new (2024-09-10)

[12] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, p. 420, Nov. 2021.
[13] R. Kaushal and P. Kumaraguru, "A systematic review on user identity linkage across online social networks," Doctoral dissertation, Indraprastha Inst. Inf. Technol. Delhi, New Delhi, India, 2020.
[14] X. Guo, Y. Liu, X. Meng, and L. Liu, "User identity linkage across social networks based on neural tensor network," in *Proc. Int. Conf. Secur. Privacy New Comput. Environ.* Cham, Switzerland: Springer, 2020, pp. 162–171.
[15] M. Qiao, H. Chen, X. Xu, W. Zhang, J. Wang, and M. Xie, "Dual-neighborhood attention network for user identity linkage," in *Proc. 14th Int. AAAI Conf. Web Social Media (ICWSM)*, 2020, pp. 474–483.
[16] J. Zhang, J. Chen, S. Zhi, Y. Chang, P. S. Yu, and J. Han, "Link prediction across aligned networks with sparse and low rank matrix estimation," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 971–982.
[17] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.
[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[20] Y. Liu, "User identity linkage method based on user online habit," *ICST Trans. Secur. Saf.*, vol. 7, no. 26, Oct. 2020, Art. no. 170240.
[21] Y. Xu, B. Zong, H. Xu, C.-T. Hsieh, T. Chakraborty, and W.-S. Ku, "Deep neural networks for user identity linkage," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2020, pp. 684–692.
[22] L. Zhang, Z. Liu, C. Yang, and M. Sun, "Combining structured and unstructured data for social network-based user identity linkage," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, 2018, pp. 1938–1947.
[23] L. Zhang, L. Wu, J. Zhang, and X. Du, "Cross-network user identification via graph embedding," *Expert Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112869.
[24] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.
[25] Z. Zhong, Y. Cao, M. Guo, and Z. Nie, "CoLink: An unsupervised framework for user identity linkage," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1. [Online]. Available: https://aaai.org/proceeding/01-thirty-second-aaai-conference-on-artificial-intelligence-2018/
[26] T. Qin, Z. Liu, S. Li, and X. Guan, "A two-stagse approach for social identity linkage based on an enhanced weighted graph model," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1364–1375, Aug. 2020.
[27] H. T. Trung, T. Van Vinh, N. T. Tam, H. Yin, M. Weidlich, and N. Q. V. Hung, "Adaptive network alignment with unsupervised and multi-order convolutional networks," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 85–96.
[28] W. Xie, X. Mu, R. K. Lee, F. Zhu, and E.-P. Lim, "Unsupervised user identity linkage via factoid embedding," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 1338–1343.
[29] S. Chen, J. Wang, X. Du, and Y. Hu, "A novel framework with information fusion and neighborhood enhancement for user identity linkage," in *Proc. ECAI*, 2020, pp. 1754–1761.
[30] X. Li, Y. Su, W. Tang, N. Gao, and J. Xiang, "User identity linkage with accumulated information from neighbouring anchor links," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Cham, Switzerland: Springer, 2018, pp. 335–344.
[31] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowl.-Based Syst.*, vol. 151, pp. 78–94, Jul. 2018.
[32] Y. Yang, L. Wang, and D. Liu, "Anchor link prediction across social networks based on multiple consistency," *Knowl.-Based Syst.*, vol. 257, Dec. 2022, Art. no. 109939.
[33] X. Mu, F. Zhu, E.-P. Lim, J. Xiao, J. Wang, and Z.-H. Zhou, "User identity linkage by latent user space modelling," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1775–1784.
[34] W. Tang, H. Sun, J. Wang, C. Liu, Q. Qi, J. Wang, and J. Liao, "Identifying users across social media networks for interpretable fine-grained neighborhood matching by adaptive GAT," *IEEE Trans. Services Comput.*, vol. 16, no. 5, pp. 3453–3466, Feb. 2023.
[35] Z. Xiong, X. Xie, X. Wu, Y. Peng, and Y. Lu, "DSANE: A dual structure-aware network embedding approach for user identity linkage," in *Proc. IEEE 8th Int. Conf. Big Data Analytics (ICBDA)*, Mar. 2023, pp. 193–198.
[36] C. Zheng, L. Pan, and P. Wu, "JORA: Weakly supervised user identity linkage via jointly learning to represent and align," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3900–3911, Apr. 2024.
[37] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Proc. 11th Int. Symp. Privacy Enhancing Technol.*, Waterloo, ON, Canada. Cham, Switzerland: Springer, 2011, pp. 1–17.
[38] Z. Yuan, L. Yan, G. Xiaoyu, S. Xian, and W. Sen, "User naming conventions mapping learning for social network alignment," in *Proc. 13th Int. Conf. Comput. Autom. Eng. (ICCAE)*, Mar. 2021, pp. 36–42.
[39] M. Wang, W. Chen, J. Xu, P. Zhao, and L. Zhao, "User profile linkage across multiple social platforms," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Cham, Switzerland: Springer, 2020, pp. 125–140.
[40] L. Ranaldi and F. M. Zanzotto, "Hiding your face is not enough: User identity linkage with image recognition," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–9, Dec. 2020.
[41] Y. Benkhedda, F. Azouaou, and S. Abbar, "Identity linkage across diverse social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Dec. 2020, pp. 468–472.
[42] Y. Sha, Q. Liang, and K. Zheng, "Matching user accounts across social networks based on users message," *Proc. Comput. Sci.*, vol. 80, pp. 2423–2427, Jan. 2016.
[43] A. Li, S. Mu, L. Zhang, and Y. Jia, "A practical approach to construct profile linkage framework," in *Proc. Int. Conf. Intell. Comput., Autom. Syst. (ICICAS)*, Dec. 2019, pp. 81–84.
[44] J. Zhou and J. Fan, "TransLink: User identity linkage across heterogeneous social networks via translating embeddings," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 2116–2124.
[45] J. Feng, M. Zhang, H. Wang, Z. Yang, C. Zhang, Y. Li, and D. Jin, "DPLink: User identity linkage via deep neural network from heterogeneous mobility data," in *Proc. World Wide Web Conf.*, May 2019, pp. 459–469.
[46] W. Chen, H. Yin, W. Wang, L. Zhao, W. Hua, and X. Zhou, "Exploiting spatio-temporal user behaviors for user linkage," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 517–526.
[47] F. Ding, X. Ma, Y. Yang, and C. Wang, "User identity linkage across location-based social networks with spatio-temporal check-in patterns," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Dec. 2020, pp. 1278–1285.
[48] W. Chen, H. Yin, W. Wang, L. Zhao, and X. Zhou, "Effective and efficient user account linkage across location based social networks," in *Proc. IEEE 34th Int. Conf. Data Eng. (ICDE)*, Apr. 2018, pp. 1085–1096.
[49] L. Jalali, M. Khan, and R. Biswas, "Learning and multi-objective optimization for automatic identity linkage," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4926–4931.
[50] Y. Zhang, L. Wang, X. Li, and C. Xiao, "Social identity link across incomplete social information sources using anchor link expansion," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2016, pp. 395–408.
[51] I. Nurgaliev, Q. Qu, S. M. H. Bamakan, and M. Muzammal, "Matching user identities across social networks with limited profile data," *Frontiers Comput. Sci.*, vol. 14, no. 6, pp. 1–14, Dec. 2020.
[52] N. Bennacer, C. Nana Jipmo, A. Penta, and G. Quercini, "Matching user profiles across social networks," in *Proc. 26th Int. Conf. Adv. Inf. Syst. Eng.*, Thessaloniki, Greece. Cham, Switzerland: Springer, 2014, pp. 424–438.
[53] G. Quercini, N. Bennacer, M. Ghufran, and C. Nana Jipmo, "LIAISON: Reconciliation of individuals profiles across social networks," in *Advances in Knowledge Discovery and Management*. Cham, Switzerland: Springer, 2017, pp. 229–253.
[54] J. Shao, Y. Wang, H. Gao, H. Shen, Y. Li, and X. Cheng, "Locate who you are: Matching geo-location to text for user identity linkage," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 3413–3417.
[55] D. Chatzakou, J. Soler-Company, T. Tsikrika, L. Wanner, S. Vrochidis, and I. Kompatsiaris, "User identity linkage in social media using linguistic and social interaction features," in *Proc. 12th ACM Conf. Web Sci.*, Jul. 2020, pp. 295–304.
[56] W. Zhang, K. Shu, H. Liu, and Y. Wang, "Graph neural networks for user identity linkage," 2019, *arXiv:1903.02174*.

[57] R. Kaushal, S. Singh, and P. Kumaraguru, "NeXLink: Node embedding framework for cross-network linkages across social networks," in *Proc. Int. Conf. Netw. Sci.* Cham, Switzerland: Springer, 2020, pp. 61–75.

[58] S. Fu, G. Wang, S. Xia, and L. Liu, "Deep multi-granularity graph embedding for user identity linkage across social networks," *Knowl.-Based Syst.*, vol. 193, May 2020, Art. no. 105301.

[59] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach," in *Proc. IJCAI*, vol. 16, 2016, pp. 1823–1829.

[60] F. Wang, L. Sun, and Z. Zhang, "Hyperbolic user identity linkage across social networks," in *Proc. GLOBECOM - IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[61] J. Zhang, P. S. Yu, and Z.-H. Zhou, "Meta-path based multi-network collective link prediction," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1286–1295.

[62] Z. Wang, T. Hayashi, and Y. Ohsawa, "A community sensing approach for user identity linkage," in *Proc. Annu. Conf. Jpn. Soc. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 191–202.

[63] C. Li, S. Wang, Y. Wang, P. Yu, Y. Liang, Y. Liu, and Z. Li, "Adversarial learning for weakly-supervised social network alignment," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 996–1003.

[64] A. Amara, M. A. Hadj Taieb, and M. Ben Aouicha, "Cross-network representation learning for anchor users on multiplex heterogeneous social network," *Appl. Soft Comput.*, vol. 118, Mar. 2022, Art. no. 108461.

[65] S. Liu, S. Wang, and F. Zhu, "Structured learning from heterogeneous behavior for social identity linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 7, pp. 2005–2019, Jul. 2015.

[66] L. Sun, Z. Zhang, P. Ji, J. Wen, S. Su, and P. S. Yu, "DNA: Dynamic social network alignment," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 1224–1231.

[67] Z. Hu, J. Wang, S. Chen, and X. Du, "A semi-supervised framework with efficient feature extraction and network alignment for user identity linkage," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2021, pp. 675–691.

[68] Y. Wang, C. Feng, L. Chen, H. Yin, C. Guo, and Y. Chu, "User identity linkage across social networks via linked heterogeneous network embedding," *World Wide Web*, vol. 22, no. 6, pp. 2611–2632, Nov. 2019.

[69] S. Su, L. Sun, Z. Zhang, G. Li, and J. Qu, "MASTER: Across multiple social networks, integrate attribute and STructure embedding for reconciliation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3863–3869.

[70] S. Wang, X. Li, Y. Ye, S. Feng, R. Y. K. Lau, X. Huang, and X. Du, "Anchor link prediction across attributed networks via network embedding," *Entropy*, vol. 21, no. 3, p. 254, Mar. 2019.

[71] J. Zhang and P. S. Yu, "Multiple anonymized social networks alignment," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 599–608.

[72] F. Lei, Q. Li, S. Sun, L. Wang, and D. D. Zeng, "Catching dynamic heterogeneous user data for identity linkage learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[73] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "DeepLink: A deep learning approach for user identity linkage," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2018, pp. 1313–1321.

[74] N. E. Friedkin, "Bounded structure and social influence: A theoretical model of social network change," *Amer. J. Sociol.*, vol. 104, no. 6, pp. 1402–1424, 1998.

[75] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 701–710.

[76] S. Li, D. Lu, Q. Li, X. Wu, S. Li, and Z. Wang, "MFLink: User identity linkage across online social networks via multimodal fusion and adversarial learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 5, pp. 3716–3725, Oct. 2024.

[77] H. T. Trung, N. T. Toan, T. V. Vinh, H. T. Dat, D. C. Thang, N. Q. V. Hung, and A. Sattar, "A comparative study on network alignment techniques," *Expert Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112883.

[78] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, Feb. 2014, pp. 303–312.

[79] X. Ma, F. Ding, K. Peng, Y. Yang, and C. Wang, "CP-link: Exploiting continuous spatio-temporal check-in patterns for user identity linkage," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4594–4606, Aug. 2023.

[80] H. Wang, Y. Li, G. Wang, and D. Jin, "You are how you move: Linking multiple user identities from massive mobility traces," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 189–197.

[81] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 707–719.

[82] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[83] C. D. Manning, *Introduction to Information Retrieval*. Oxford, U.K.: Syngress Publishing, 2008.

[84] S. Zhang and H. Tong, "FINAL: Fast attributed network alignment," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1345–1354.

[85] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li, "ComSoc: Adaptive transfer of user behaviors over composite social network," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2012, pp. 696–704.

[86] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. 2nd ACM Workshop Online Social Netw.*, Aug. 2009, pp. 37–42.

[87] J. Leskovec and A. Krevl, "Snap datasets: Stanford large network dataset collection," Stanford Univ., Tech. Rep., 2014. [Online]. Available: http://snap.stanford.edu/data/index.html

[88] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "COSNET: Connecting heterogeneous social networks with local and global consistency," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1485–1494.

[89] M. Yan, J. Sang, T. Mei, and C. Xu, "Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[90] V. Sharma and C. Dyreson, "LINKSOCIAL: Linking user profiles across multiple social media platforms," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Nov. 2018, pp. 260–267.

[91] S. Sun, Q. Li, P. Yan, and D. D. Zeng, "Mapping users across social media platforms by integrating text and structure information," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 113–118.

[92] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1774–1780.

[93] J. Zhang and S. Y. Philip, "Integrated anchor and social link predictions across social networks," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2125–2132.

[94] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Discovering links among social networks," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Bristol, U.K. Cham, Switzerland: Springer, 2012, pp. 467–482.

[95] P. Christen, T. Ranbaduge, and R. Schnell, *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Cham, Switzerland: Springer, 2020.

[96] M. Backes, P. Berrang, O. Goga, K. P. Gummadi, and P. Manoharan, "On profile linkability despite anonymity in social media systems," in *Proc. ACM Workshop Privacy Electron. Soc.*, Oct. 2016, pp. 25–35.

[97] S. Chandok and P. Kumaraguru, "User identities across social networks: Quantifying linkability and nudging users to control linkability," Ph.D. thesis, Indraprastha Inst. Inf. Technol., Delhi, 2017. [Online]. Available: https://repository.iiitd.edu.in/xmlui/handle/123456789/831

[98] A. Andreou, O. Goga, and P. Loiseau, "Identity vs. attribute disclosure risks for users with multiple social profiles," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 163–170.

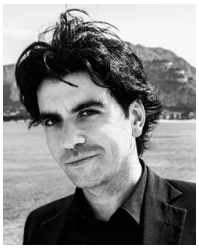
[99] Y. Shen, F. Wang, and H. Jin, "Defending against user identity linkage attack across multiple online social networks," in *Proc. 23rd Int. Conf. World Wide Web*, New York, NY, USA, Apr. 2014, pp. 375–376, doi: 10.1145/2567948.2577208.

[100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[101] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[102] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. (2019). *Language Models Are Unsupervised Multitask Learners*. openAI. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[103] T. Brown, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.
[104] C.-H. Ting, H.-Y. Lo, and S.-D. Lin, "Transfer-learning based model for reciprocal recommendation," in *Proc. 20th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, Auckland, New Zealand. Cham, Switzerland: Springer, 2016, pp. 491–502.
[105] X. Zheng, G. Zhao, L. Zhu, J. Zhu, and X. Qian, "What you like, what I am: Online dating recommendation via matching individual preferences with features," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 5400–5412, May 2023.
[106] Z. Li, X. Fang, and O. R. L. Sheng, "A survey of link recommendation for social networks: Methods, theoretical foundations, and future research directions," *ACM Trans. Manage. Inf. Syst.*, vol. 9, no. 1, pp. 1–26, 2017.
[107] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in *Proc. Deep Learn. Inside Out (DeeLIO), 3rd Workshop Knowl. Extraction Integr. Deep Learn. Architectures*, 2022, pp. 100–114.
[108] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, and S. Kurohashi, "GPT-RE: In-context learning for relation extraction using large language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 3534–3547.
[109] T. Nguyen-Mau, A.-C. Le, D.-H. Pham, and V.-N. Huynh, "An information fusion based approach to context-based fine-tuning of GPT models," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102202.
[110] P. Sahoo, A. Kumar Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," 2024, *arXiv:2402.07927*.

CATERINA SENETTE received the master's degree in biomedical engineering and the Ph.D. degree in information engineering from the University of Pisa. Since 2008, she has been a Researcher with the Institute of Informatics and Telematics–Italian National Research Council (IIT-CNR). Over the past 12 years, her research activity has focused mainly on human–computer interaction (HCI). Currently, she is part of the Cyber Intelligence Research Unit. She has ten years of experience as a Coordinator/the Project Manager in several technological and research projects. She is the co-author of more than 48 papers in international journals and conferences. Her research interests include analyzing large-scale interactions among multiple users, leveraging artificial intelligence, web mining, and social network analysis techniques. She has been a member of the Association for Computing Machinery (ACM), since 2014.



MARCO SIINO received the bachelor's and master's degrees (cum laude) in computer engineering from the University of Palermo, and the Ph.D. degree in information and communication technologies from the University of Palermo, in 2023. He is currently a Research Fellow with the University of Catania, and a freelance full-stack Developer. His work involves social-network-related tasks (e.g., sentiment analysis, hate speech detection, and fake news detection). His main interests include machine learning, deep learning, natural language processing, and recommender systems. He is an IEEE IMS Member.

MAURIZIO TESCONI received the Ph.D. degree. He is a Senior Researcher in computer science and leads the Cyber Intelligence Laboratory, Institute of Informatics and Telematics of Italian National Research Council (IIT-CNR). He teaches cyber intelligence within the master's in cyber security. He is responsible for projects funded by major public administrations on topics related to AI for secure societies. His main research interests include artificial intelligence, big data, web mining, and social network analysis within the context of open source intelligence. He is a member of the permanent team of the European Laboratory on Big Data Analytics and Social Mining.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY' within the CRUI CARE Agreement