User Identity Linkage via Graph Neural Network-Based Stylometric Representations

Wenwen Xu

School of Computer Science

McGill University, Montreal

December 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Computer Science

Abstract

User identity linkage (UIL) is the task of aligning user identities of the same user across different social network platforms. Although existing approaches have explored various aspects such as different user profile attributes and social network structures, the writing styles from user-generated texts, which is commonly known as stylometry, remain relatively underexplored. In this thesis, we propose a novel Graph Neural Network (GNN)-based model named StyleLink, which leverages both social network structures and stylometric features derived from user-generated texts to address the UIL problem in an integrated manner. Our model utilizes GNNs to incorporate both stylometric features and the network structure for each social network, effectively embedding the network and enhancing user representation. This is the first work to incorporate stylometric features into GNNs to embed social network and then conduct UIL between two embedding spaces. Through extensive experiments conducted on real-world social network datasets, the results demonstrate that StyleLink outperforms state-of-the-art methods in both linking accuracy and identity-match ranking performance. In addition, we explore the effects of different linguistic characteristics in the identification of user profiles on social networks.

Abrégé

Le lien d'identité utilisateur (UIL) est la tâche consistant à aligner les identités d'un même utilisateur sur différentes plateformes de réseaux sociaux. Bien que les approches existantes aient exploré divers aspects tels que les différents attributs de profil utilisateur et les structures des réseaux sociaux, les styles d'écriture issus des textes générés par les utilisateurs, communément appelés stylométrie, restent relativement sous-explorés. Dans cet article, nous proposons un nouveau modèle basé sur les réseaux de neurones graphiques (GNN) nommé StyleLink, qui exploite à la fois les structures des réseaux sociaux et les caractéristiques stylométriques dérivées des textes générés par les utilisateurs pour aborder le problème de l'UIL de manière intégrée. Notre modèle utilise les GNN pour intégrer à la fois les caractéristiques stylométriques et la structure du réseau pour chaque réseau social, ce qui permet d'effectuer une meilleure représentation des utilisateurs. Il s'agit de la première étude à intégrer des caractéristiques stylométriques dans les GNNs pour encoder un réseau social et ensuite réaliser l'UIL entre deux espaces d'encodage. À travers des expériences approfondies menées sur des ensembles de données de réseaux sociaux réels, les résultats montrent que StyleLink surpasse les méthodes à la pointe de la technologie en termes de précision de liaison et de performance de classement des correspondances d'identité. De plus, nous explorons les effets de différentes caractéristiques linguistiques dans l'identification des profils d'utilisateurs sur les réseaux sociaux.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Benjamin Fung, for his unwavering support throughout my graduate studies and research. His guidance in shaping the direction of my research, along with his expertise in logical structuring, academic writing, and critical review, has been indispensable to the successful completion of this thesis.

I am also deeply thankful to my thesis examiner, Prof. Derek Ruths, for his meticulous review and insightful suggestions. His constructive feedback and thoughtful guidance have been invaluable in refining this thesis.

Furthermore, I am grateful to my colleagues in the *Data Mining and Security (DMaS)* lab. Their generous assistance in setting up computing resources, insightful feedback, and sharing of their past research experiences have been invaluable. Their support has not only enriched my academic journey but also inspired me to make the most of my graduate studies.

I am deeply grateful to my lovely friends at McGill—Yanan, Yingzi, Xiaojie, Bridget, and many others—whose unwavering support and encouragement have meant the world to me. They have stood by me during the most challenging moments, offering comfort and strength when I needed it most.

Last but certainly not least, I extend my heartfelt thanks to my parents and sister. Their unconditional support and encouragement have been a constant source of strength during my years of study at McGill.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BFS Breadth-First-Search. 6, 7

DFS Depth-First-Search. 6

GAT Graph Attention Network. 11, 42

GCNs Graph Convolutional Network. v, vii, viii, 10, 11, 26, 30, 31, 39–42

GNNs Graph Neural Networks. 8–10, 22, 23, 26, 28, 29, 39, 40, 43

LINE Large-scale Information Network Embedding. 6

OSNs Online Social Networks. 2, 3, 14, 19, 23, 43, 44

UGC User Generated Content. 16, 43, 44

UIL User Identity Linkage. vii, ix, 1, 5, 14, 16–21, 33, 35–40, 42, 44

Chapter 1

Introduction

1.1 Motivation

With the flourishing Online Social Networks (OSNs), people tend to participate in various social networks to engage in different social activities. According to reports [9, 10], roughly three-quarters of the public (73%) uses more than one OSN and the median American uses three mainstream social network sites. Each OSN serves different social networking functions in daily life. For instance, users connect with friends on Facebook and Instagram, share updates on X (formerly known as Twitter), and network with colleagues and potential employers on LinkedIn.

As a result, the same individual may have signed up multiple accounts across these diverse platforms, each account reflecting different user attributes, user-generated content (UGC), and behavior patterns, such as follows, likes, etc. User identity linkage (UIL) is a task of aligning accounts that belong to the same individual from different social networks. UIL has received increasing attention in both academia and industry, playing a crucial role in various applications, including user migration [39], recommendation systems [5, 6, 48], crime detection [82], and privacy protection [19, 42, 59, 75]. With the advancement of network embedding techniques, embedding-based methods have been widely employed to address the UIL problem. Existing approaches leverage vari-

ous dimensional attributes of user identities and can be categorized into three types: user profile-based, network structure-based, and content-based methods.

User profile-based approaches typically focus on user-provided identifiable information, including username, gender, birthday, email, education, location, etc [1,43,77]. While public profile attributes offer valuable insights for identifying users across Online Social Networks (OSNs), their effectiveness diminishes when applied to large-scale OSNs, where many attributes can be duplicated and easily impersonated.

Network-based approaches aim to link user identities with their network structures, specifically utilizing *topology consistency* [81]. Users who share similar neighborhoods in different networks could be recognized as matched identities. In social networks, social relationships, such as follower-followee, play a significant role in exploring corresponding user identities across different OSNs [44,47,81,86]. However, the assumption of topology consistency is challenged by network heterogeneity. For instance, users may prefer certain platforms, such as favoring Facebook over Twitter, leading to active engagement on one network and a subdued presence on another. Additionally, heterogeneity arises from differing semantics of relations, such as those between a career-oriented platform like LinkedIn and a co-authorship network like Google Scholar.

Content-based approaches to user identity linkage have explored various aspects of UGC and behavior patterns. These methods have analyzed tag frequencies [27], typing patterns [78], multi-modal UGC [13], and N-gram language modelling [20,66,78]. However, these approaches still have limitations. By focusing solely on UGC, they overlook the crucial network structure and user connectivity, which are the most typical characteristics of OSNs. They also face challenges with platform-specific content variations and scalability issues with large datasets. Moreover, the exclusive focus on content neglects the fundamental purpose and dynamics of social networking platforms. These limitations highlight the need for a more comprehensive approach that integrates content analysis with network structural information to achieve more robust cross-platform user identity linkage.

In OSNs, user profile attributes and network structure are closely interrelated. For instance, users with similar profile attributes are more likely to be connected as friends, and groups of users with shared characteristics often form dense communities. Due to the unique social attributes and community ecology of social networks, there exist the dual mechanisms of linguistic homophily [37]. This phenomenon manifests in two ways: firstly, individuals who share similar linguistic styles have a higher propensity to form and sustain friendships; secondly, friends tend to experience linguistic convergence over time. These mechanisms contribute to the emergence of relational echo chambers within social networks.

Drawing inspiration from the success of applying graph neural networks (GNNs) and attention mechanisms to embeddings [67, 68], we propose a novel GNN-based model for User Identity Linkage (UIL). This model leverages both social network structures and stylometric features derived from UGC to address the aforementioned limitations.

1.2 Contributions

To the best of our knowledge, this is the first work to incorporate stylometric features into GNNs to embed social network and then conduct UIL between two embedding spaces.

- We present a novel methodology that leverages both user-generated content (UGC) and network structure to establish correspondences between user accounts across OSNs. This approach reduces reliance on user-provided identifiable information, which may be inconsistent or deliberately obscured. Instead, we focus on analyzing user activities, including writing styles and social connections, which are harder to impersonate and accumulate over time with consistent social network engagement.

- We introduce StyleLink, an innovative Graph Neural Network (GNN)-based approach to tackle the UIL problem. StyleLink consists of three primary components: a) Stylometric feature extraction, where we identify distinctive linguistic patterns in UGC to capture unique writing styles; b) Network embedding: we employ GNN

models to generate user representations that incorporate both stylometric features and network structure; and c) Supervised linkage learning, where we compare the usage of cosine similarity and Triplet loss for learning and use a Multi-Layer Perceptron (MLP) as the mapping function to learn the embedding transformation between source and target networks, thus predicting aligned user identities.

- We validate the effectiveness of StyleLink on real-world datasets. Experimental results demonstrate that our method significantly outperforms existing baselines in terms of both accuracy and efficiency.

The rest of the thesis is organized as follows: in Chapter 2, we thoroughly reviewed related work for network embedding, the graph neural network models, network alignment, and writing style features in authorship identifications. Chapter 3 formally defines the UIL research problem. Chapter 4 describes the proposed StyleLink model, detailing our design architecture of three main steps. Chapter 5 illustrates the datasets, baseline models, the experiment setup, and experimental results. Finally, Chapter 6 summarizes the advantages and limitations of our approach and discusses potential directions for future work.

Chapter 2

Literature Review

Before introducing our proposed network alignment models and experiments on various social networks, we give a summary of the related work on network embedding methods, Graph Neural Networks (GNNs), the models tackling the UIL problem from different aspects, and writing styles in authorship identification.

2.1 Network Embedding

Network embedding has become a widely recognized and powerful method for generating low-dimensional representations of nodes within networks, attracting significant attention due to its effectiveness and efficiency. This section briefly introduces some classical node embedding algorithms based on random walks, including Deepwalk [53], node2vec [23], LINE [62].

2.1.1 DeepWalk

DeepWalk [53] was the first to apply random walks for representation learning in graphs, adapting concepts from the SkipGram model used in language processing. SkipGram optimizes the co-occurrence probability of words within a fixed window in a sentence. Similarly, DeepWalk leverages local information from truncated random walks to learn latent

node representations, considering these walks as equivalent to sentences. The DeepWalk algorithm comprises two main steps: generating random walks and applying the Skip-Gram algorithm to update the representations. It simulates random walks starting from each vertex, where, at each step, a neighboring vertex is uniformly selected as the next node, continuing this process iteratively until the maximum walk length is reached.

2.1.2 Node2Vec

Building upon the uniformly random walks used in DeepWalk, node2vec [23] introduces biased random walks for sampling. As depicted in Figure 2.1, it utilizes two key parameters, the return parameter p and the in-out parameter q, to effectively balance between *BFS (Breadth-First-Search)* and *DFS (Depth-First-Search)* strategies when generating these random walks. According to Grover and Leskovec [23], Breadth-First-Search (BFS) and Depth-First-Search (DFS) play a crucial role in generating representations that reflect either homophily or structural equivalence. When we use BFS to sample neighborhoods, the resulting embeddings closely align with structural equivalence. This occurs because accurately characterizing local neighborhoods is sufficient to identify structural equivalence. BFS accomplishes this by initially exploring nodes, providing a detailed view of the network at a microscopic level. The return parameter p controls the probability of immediately revisiting nodes in the generated random walk; a higher value of p reduces the likelihood of revisiting, thereby promoting exploration. The in-out parameter q determines the likelihood of remaining within the neighborhood of node v versus exploring nodes further away from node v.

2.1.3 LINE: Large Information Network Embeddings

LINE (Large-scale Information Network Embedding) [62] has raised concerns about the lack of a clearly defined objective with respect to preserving specific network properties in DeepWalk. Instead of sampling random walks, Large-scale Information Network Em-

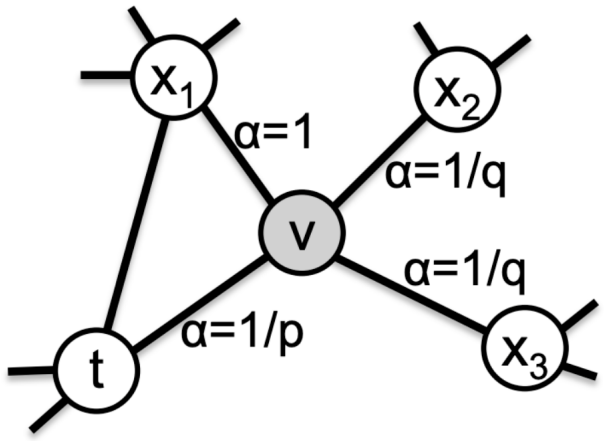


Figure 2.1: An illustration of the biased random walk process in node2vec. The parameters p and q are the two hyper-parameters that influence the walk, while α represents the probability of selecting a particular edge as the next step in the random walk. Adapted from [23]

bedding (LINE) adopts a BFS strategy, explicitly aiming to preserve both first-order and second-order proximity within the network with a carefully designed objective function. First-order proximity in a network refers to the local proximity between two vertices. Second-order proximity between a pair of vertices (u, v) refers to the similarity between their respective neighborhood structures. As implied by its name, LINE is well-suited for various types of information networks and can easily scale to millions of nodes by utilizing an edge-sampling algorithm to optimize its objective function [62].

However, the limitations of these shallow node embeddings via random walks are apparent [24]. First of all, they do not utilize node, edge, or graph features and only incorporate the topological features of the graph. Many graph datasets contain rich feature information, especially with social networks, which could be potentially informative for embedding. The second issue is that they cannot generate embeddings for nodes not in the training set and inherently transductive. Lastly, it has no sharing of parameters between nodes. To alleviate these drawbacks, shallow embedding methods can be replaced with embedding methods that depend on the structure and attributes of the graph

and nodes. Currently, the most popular and effective paradigm is *Graph Neural Networks (GNNs)*.

2.2 Graph Neural Network

As a solution to the limitations of random-walk-based embedding methods, deep representation learning and Graph Neural Networks (GNNs) have become a powerful tool to extract useful low-dimensional features in attributed networks with node features [25, 35, 65, 71].

The basic idea of applying GNNs is to encode and project the nodes in a graph into a latent space (shown as Figure 2.2), where a node is represented by a low-dimensional, continuous, and dense embedding vector. Low-dimensional embedding vectors are preferred, as the dimensionality is far smaller than the total number of nodes in the graph. Additionally, the embedding vectors must be continuous and composed of real numbers. Density is another important characteristic of embedding vectors, as we aim to avoid sparsity, a common problem of adjacency matrices. The embeddings are trained in a way that the distances or similarities in the latent embedding space correspond to the relative positions of the nodes in the original graph. The feature of a node can be aggregated and updated by its neighborhood through the message-passing mechanism.

This section includes some fundamental information about graph representation learning and fundamental GNNs to prepare readers with basic theoretical information.

2.2.1 Message Passing

Message passing is a fundamental mechanism in GNNs that allows nodes to communicate with their neighboring nodes and update their feature representations based on the information received from their neighbors [7].

In each message passing layer, each node aggregates the information from its neighboring nodes and combines it with its own features to produce a new representation. This

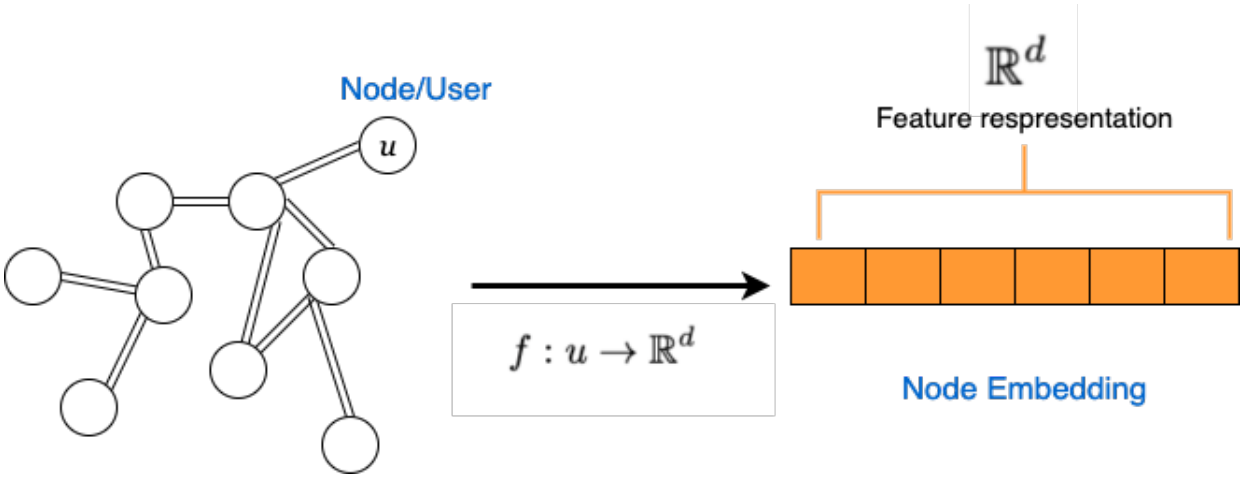


Figure 2.2: Node Embedding. Generate a low-dimensional, continuous, and dense embedding vector for each node in a graph to facilitate downstream tasks.

new representation is then passed on to its neighboring nodes for further processing. The process is repeated for multiple iterations until a stable representation is achieved.

The message-passing operation can be formally defined as follows:

$$h_v^{l+1} = AGGREGATE^l(m_u^{(l,v)} | u \in N(v)) \tag{2.1}$$

$$m_v^{l+1} = COMBINE^l(h_v^l, h_v^{l+1}) \tag{2.2}$$

where h_v^l represents the feature representation of node v at layer l, $m_u^{l,v}$ is the message sent from node u to node v at layer l, $AGGREGATE^l$ is the aggregation function that combines the messages from neighboring nodes, and $COMBINE^l$ is the combining function that produces the new feature representation for node v at layer $l+1$.

The selection of aggregation and combination functions may differ based on the specific GNNs architecture and the particular task at hand. Similar to the pooling functions used in CNN, *max*, *mean*, *sum* are typical element-wise aggregation functions in GNNs as well. However, the overall idea is to allow nodes to exchange information and update their features based on the collective information of their neighboring nodes.

Message Passing is the foundation framework for all the graph neural network models. Theoretically, the basic message passing layer representation in GNNs is defined as:

$$h_u^k = \sigma(\mathbf{W}_{self}^k h_u^{k-1} + \mathbf{W}_{neigh}^k \sum_{v \in N(u)} h_v^{k-1} + b_k) \tag{2.3}$$

2.2.2 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) [35] extend convolutional neural networks (CNNs) [40] to process graph-structured data, which lacks a fixed grid layout, enabling direct operations on the graph's structure. GCNs are proved to be effective in various downstream tasks, such as text classification [73], social network recommender systems [74], Natural Language Processing [69], etc. GCNs have also been shown to be effective in handling noisy and incomplete graph data, making them useful tools for real-world applications.

Theoretically, GCN is a multi-layer neural network that function directly on graph data., generating embedding vectors for nodes by leveraging the properties of their surrounding neighborhoods. Following our notations for the graph, Figure 2.3 is the visualization of a 2-layer version of the GCNs computation graph. By contrast with the two-layer version of a message-passing model mentioned in Sec 2.2.1, GCNs starts from the adjacency matrix $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, with diagonal elements as 1 because of adding self-loops. This $\widetilde{\mathbf{A}}$ enables the inclusion of the node's own representation when updating the node embeddings at each layer. In GCNs, the node embeddings at every layer are updated according to the following propagation rules:

$$\mathbf{H}^1 = \mathbf{X} \tag{2.4}$$

$$\mathbf{H}^{k+1} = \sigma(\mathbf{D}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \mathbf{D}^{-\frac{1}{2}} \mathbf{H}^k \mathbf{W}^k) \tag{2.5}$$

In this context, $\mathbf{H}$ represents the matrix of node embeddings $\mathbf{h}u$, $\mathbf{X}$ corresponds to the matrix of node features $\mathbf{x}_u$, $\sigma(\cdot)$ denotes the activation function (e.g., ReLU), $\tilde{\mathbf{A}}$ is the

adjacency matrix of the graph, augmented with self-loops, $\tilde{\mathbf{D}}$ is the degree matrix, also including self-loops, and Θ refers to the matrix of trainable parameters.

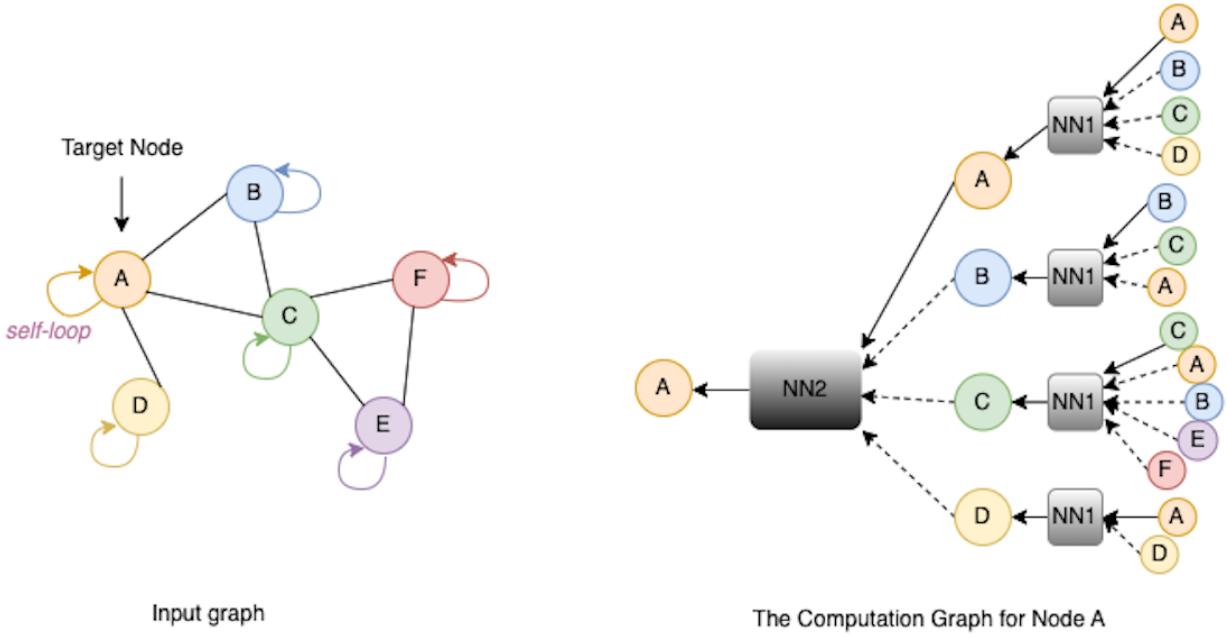


Figure 2.3: Overview of a 2-layer GCNs computation graph for a single node. NN1 and NN2 represent two different Neural Networks for layer-1 and layer-2 respectively, while the parameters are shared in each neural network.

2.2.3 Graph Attention Networks

When a graph becomes too noisy, which is normal for real social network graphs, a simple graph convolution mechanism will struggle to effectively propagate and aggregate meaningful information, as it indiscriminately combines information from neighboring nodes, including irrelevant or noisy ones [12]. However, the attention mechanism in graph structure is more robust than graph convolution and is able to alleviate the impact of noise by assigning higher weights to more relevant nodes and edges, allowing the model to focus on critical relationships while ignoring less important or noisy connections. Petar Veličković et al. [65] introduced the Graph Attention Network (GAT) in 2018. It combines graph neural networks with additional attention layers. It employs an

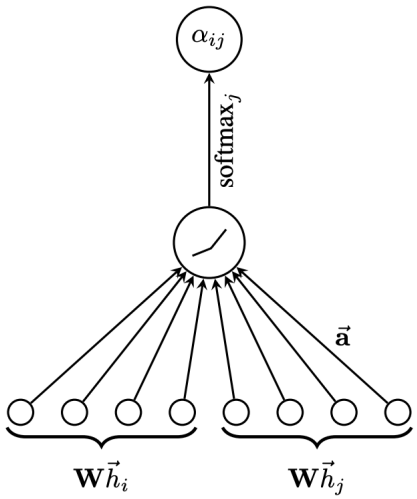


Figure 2.4: The attention mechanism $\alpha(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ are employed

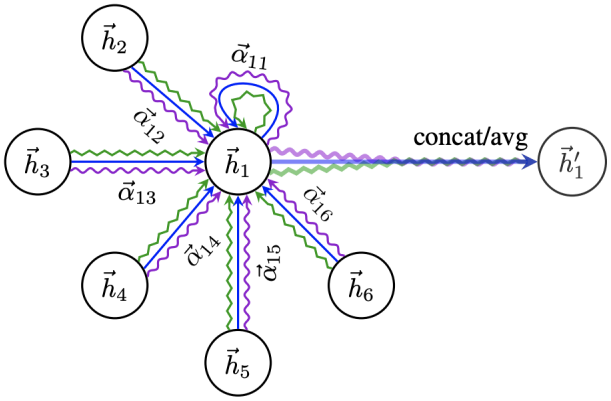


Figure 2.5: A GAT layer with multi-head attention (with $K = 3$ heads) is applied by node 1 to its neighboring nodes. The distinct arrow styles and colors illustrate separate attention computations for each head. The features gathered from each head are then either concatenated or averaged to generate the updated node representation, $\vec{h}'_1$.

attention mechanism [64] to aggregate the embeddings of neighboring nodes, where the attention mechanism assigns different weights to neighbors and edges based on their significance. Attention layers allow the model to prioritize important information from the graph rather than considering the entire graph, as shown in Figure 2.4 and 2.5.

A multi-head GAT layer can be expressed as the following equation:

$$\vec{h}'_u = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{v \in N_u} \alpha_{uv} \mathbf{W}^k \vec{h}'_v\right) \tag{2.6}$$

Specifically, K represents the number of attention heads, $\Big\Vert$ means vector concatenation, α_{uv} are the normalized attention coefficients computed by the k-th attention mechanism (a^k), and $\mathbf{W}^k$ is the weight matrix for the associated linear transformation of the input. In cases where multi-head attention is applied to the network's final (prediction) layer, *averaging* is used in place of concatenation.

$$\vec{h}'_u = \sigma\left(\frac{1}{K}\sum_{k-1}^{K}\sum_{v \in N_u} \alpha_{uv} \mathbf{W}^k \vec{h}'_v\right) \tag{2.7}$$

and the aggregation process of a multi-head graph attention layer is illustrated in Figure 2.5

2.2.4 GraphSAGE

GraphSAGE [25] is a general inductive model for learning graph representations that utilizes node attribute information, such as text features, to efficiently create node embeddings for novel data. It learns a function that generates embeddings by sampling and aggregating feature vectors from a node's neighboring connections. GraphSAGE, along with some other GNN models, is inductive in nature. This enables the model to generate embeddings for nodes that were not seen during the training phase. This is particularly useful for handling new nodes. By contrast, Node2Vec [23], being transductive, needs to be retrained to handle new nodes. GraphSAGE can generate embeddings for completely new nodes based solely on their features, without any connections.

2.3 User Identity Linkage

User Identity Linkage (UIL), also referred to as network alignment, is a concept first formalized by Zafarani and Liu. They defined UIL as the process of connecting user accounts across various Online Social Networks (OSNs) that belong to the same individual in real life. It plays a crucial role in various downstream applications, including bioinformatics such as protein-protein interaction matching [38], computer vision [15], recommendation systems [5, 6, 45, 48], drug trafficker detection [82], and privacy protection [19, 42, 59, 75], etc.

Current approaches leverage different dimensional attributes of an identity and can be classified into three categories: user profile-based, network structure-based, and content-based methods.

2.3.1 User Profile-based Approaches

User profile-based approaches typically focus on discrete user profile attributes, including username, gender, birthday, email, location, etc. These attributes are usually accessible on public platforms, prompting early studies to propose diverse approaches for leveraging these profile attributes in user identity linkage.

Usernames or screen names are a mandatory element of user profiles in nearly all social networks and have therefore been one of the first and most extensively explored methods for identifying users across different OSNs. Zafarani and Liu [77] conducted an initial investigation on the feasibility of employing usernames for mapping users across social networks, with a system relying on the public user URLs. Liu et al. [43] introduced an unsupervised method that leverages the rarity and commonality of usernames, quantified using n-gram probabilities. Similarly, Ahmad and Ali [1] focused exclusively on username as a distinctive attribute to identify matching user profiles across three distinct social networking platforms. A more in-depth study on usernames was conducted by Zafarani et al. [50], where they designed sophisticated features for identification based

on some observations about human behaviors, such as limitations in time and memory, knowledge limitations, and typing patterns, to model the behavioral patterns of users when selecting usernames.

Utilizing a combination of profile features can significantly enhance the accuracy of user identification. Motoyama and Varghese [49] matched users by calculating the similarity of users based on a number of biographical attributes, such as gender, age, occupation, hometown, etc. Carmagnola and Cena [8] proposed a method that utilizes multiple profile attributes, including username, name, location, and email address, to link user identities across platforms. More recently, Sharma and Dyreson [58] proposed *LINKSOCIAL*, which links profiles by extracting a few core attributes: *username, name, bio* and *profile image*. Goga et al. [21,22] utilized a combination of *real name, username, profile photo, location*, and *friends* to identify matching user identities between Facebook and Twitter platforms.

In order to carefully measure the reliability of user profiles across real-world social networks in user identity linkage, Goga et al. [21] introduced a framework comprising four key properties: Availability, Consistency, Impersonability, and Discriminability (ACID). Their findings support the notion that individuals generally maintain consistent personas across various social networks. However, a notable issue emerges when exclusively using profile attributes for identification, as there exists a substantial number of profiles belonging to distinct users that share similar characteristics. While public profile attributes offer valuable insights for identifying users across SNs, their effectiveness diminishes when applied to large-scale SNs, where many attributes can be duplicated and easily impersonated. A significant number of users would mask or counterfeit their personal profiles. According to a survey of 1,500 U.S. social media users conducted by USCasinos.com[1], one-third of participants create fake account profiles for various reasons.

[1]https://uscasinos.com/blog/owning-fake-social-media-accounts/

2.3.2 Content-based Approaches

Due to the potential ambiguity and unreliability of user profiles in UIL tasks, researchers have shifted their focus towards analyzing user-generated content (User Generated Content (UGC)) as an alternative approach.

Iofciu1 et al. [27] focused on tagging information and addressed the UIL problem from their user ID and tagging behavior, exploiting tag frequencies and the inverse document frequency of a tag. In terms of the textual UGCs, extracting linguistic features has proved to be effective for identifying users. Zheng et al [84] developed a framework for identifying authorship by analyzing the writing style features of online messages and employing various classification methods. Goga et al. [20] included feature extraction from textual data and built probabilistic language models for users with unigram probability distribution. The geo-location information within users' posts and the timestamp of posts are also been explored.

Srivastava and Roychoudhury [61] utilized only the publicly available content information, extracting and processing parts-of-speech, symbols, emoticons, numbers, and high-frequency words in user's posts, tweets, retweets, and URLs. *ustyle-uid* [82] leverages both writing and photography styles from the text and photo contents for drug trafficker identification. Vosoughi et al. [66] presented the effectiveness of temporal (the activity patterns of users) and linguistic (derived from TF-IDF cosine similarities and N-gram language modeling) features of users in the UIL problem. Chen et al. [13] tackled UIL between Instagram and Twitter with the multi-modal UGCs (texts and images) and temporal post correlation on different social media. Apart from handcrafted acitivity and network features, Chatzakou et al. [11] also incorporated static linguistic features to capture the writing style of tweet authors, using this information to establish connections between user accounts within the same social media platform.

However, these approaches still have limitations. By focusing solely on UGC, they overlook the crucial network structure and user connectivity, which are the most typical characteristics of OSNs. They also face challenges with platform-specific content varia-

tions and scalability issues with large datasets. Moreover, the exclusive focus on content neglects the fundamental purpose and dynamics of social networking platforms.

2.3.3 Network Structure-based Approaches

Network-based approaches seek to connect user identities with the structures of networks by leveraging the *topology consistency* [81], which means users who share similar neighborhoods in different networks could be aligned. In social networks, social relationships such as following/followee play a pivotal role in exploring corresponding user identities across different SNs [44, 47, 81, 86].

NS [51] is the first approach to demonstrate feasibility of successful re-identification based solely on the network topology. Zhou et al. [88] proposed the Friend Relationship-Based User Identification (*FRUI*) algorithm. This algorithm operates on the premise that no two users share an identical friend cycle, making it a more precise approach for cross-platform social network analysis as it leverages friendship structures. Man et al. [47] proposed a supervised model, called *PALE*, that leverages observed anchor links to identify key structural patterns. *IONE* [44] generates multiple node embeddings and conceptualizes followers and followees as input and output context vectors, respectively. This method aims to maintain the proximity of users with comparable follower/followee relationships in the embedded space. Zhou et al. presented DeepLink [86], which encodes network nodes into vector representations, capturing both local and global network structures without relying on hand-crafted features. These embeddings are then utilized to align anchor nodes through deep neural networks. Lastly, *CrossMNA* [14] tackles the multi-network alignment problem by focusing solely on network structural information, employing cross-network embedding techniques.

Treating network alignment process as Reinforcement learning (RL) has gained more attention in recent years. Li et al. [41] transformed UIL into a sequence decision problem and proposed a deep RL model, named *RLink*, for the task. They also fully utilized

UIL model	Type	Feature Involved
PALE [47], CrossMNA [14], DeepLink [86]	supervised	Network
MNA [36], DCIM [52]	supervised	Network, Content
MEgo2Vec [79]	supervised	Network, Profile
FRUI [87], IONE [44], NS [51]	semi-supervised	Network
COSNET [83]	semi-supervised	Network, Profile
HYDRA [46]	semi-supervised	Network, Profile, Content
RLINK [41]	reinforcement learning	Network

Table 2.1: Comparison of Different Network-based User Identity Linkage Models. The models with joint use of content or profile features are also listed.

both the social network structure with Node2Vec [23], and the history-matched identities, which may have long-term influences on the subsequent linkage.

However, the assumption of topology consistency can be challenged by network heterogeneity. For instance, users may exhibit personal preferences for specific social platforms, such as favoring Facebook over Twitter. Consequently, they may engage actively on one social network while maintaining a more subdued presence on another. Network heterogeneity can also arise from variations in the semantic meaning of relationships [72] across different platforms. For instance, the connections on a professional networking site such as LinkedIn differ significantly in nature from those found in academic collaboration networks like Google Scholar, where links typically represent co-authorship.

Apart from leveraging features solely from one of the categories above, the joint usage of profile information, user-generated contents, and network structures promisingly bring better results [32, 33].

MNA (Multi-Network Anchoring) [36] extracts social features, including spatial, temporal, and text content features (bag-of-words vectors weighted by TF-IDF), and neighborhood-based network features and match user identity pairs. Zhang et al. [79] proposed *MEgo2Vec*, a graph neural network model designed for alignment, where both attribute embeddings and structural embeddings are seamlessly integrated into a convolutional neural network. Zhang et al. [83] proposed the *COSNET* model, which considers both local consistency and global consistency. *FINAL* [81] leveraged not only node profile information

but also edge feature information in the graph. Liu et al. [46] proposed *HYDRA*, which incorporates and makes full use of user profile attributes, user-generated contents targeting topic and style, and all the social behavior exhibited by a user on the platforms along the timestamps to link user accounts across different OSNs. Nie et al. [52] proposed a Dynamic Core Interests Mapping (*DCIM*) algorithm, which integrates both users' social network structures and their textual content. It models each user's core interests and subsequently computes the similarity between pairs of target users based on these modeled interests.

2.4 Stylometric Features for Authorship Analysis

Writing styles can serve two contrasting application directions: revealing authorship, as in authorship attribution, where stylometric analysis is used to identify an individual to a specific text; and hiding authorship, as in authorship anonymity [4] or authorship obfuscation [3], where techniques are employed to obscure or neutralize distinctive writing features to protect the writer's identity.

Stylometric features effectively distinguish the author of texts from posts, articles, emails, and reviews [2, 16, 17]. It involves examining various linguistic and stylistic features of the text and comparing them to a known set of writing styles by the suspected author. By analyzing features from Character-, Word-, and POS-level [17, 28, 30, 31], authorship attribution aims to attribute an anonymous piece of text to its rightful author.

Authorship attribution has a wide range of applications across various domains. These applications help especially in cyber forensics and crime investigation [29–31, 56], leveraging the ability to analyze and link texts to their authors based on writing style or other features Reasonably, it inspires content-based approaches in User Identity Linkage (UIL) approaches, with rich information extracted from the UGCs across OSNs [11,20,61,66,84].

Chapter 3

Problem Description

In this section, the main goal is to prepare the information of basic terminology and definitions for the problem of User Identity Linkage (UIL).

3.1 Social Network Graphs

Social Network Graphs (SNGs) can be represented formally as $G = \{V, E, \mathbf{X}\}$, where $V = \{v_1, v_2, ..., v_N\}$ is a set of nodes representing the users, and $E \in V \times V$ is a set of edges representing the social relationships among users, e.g., follower/followee on Instagram and Twitter. Each user v_i is associated with a d-dimensional feature vector x_i (the i-th row in $\mathbf{X}$). In addition, a set of known anchor nodes $T = \{(v_i^s, u_j^t) | v_i^s \in V^s, u_j^t \in V^t\}$ is provided, where each pair (v_i^s, u_j^t) represents accounts belonging to the same individual between the two networks. In real-life social networks, anchor links naturally exist due to users registering accounts on multiple platforms. Users may explicitly mention or link their other social network accounts in their profiles or posts, providing clear anchor links.

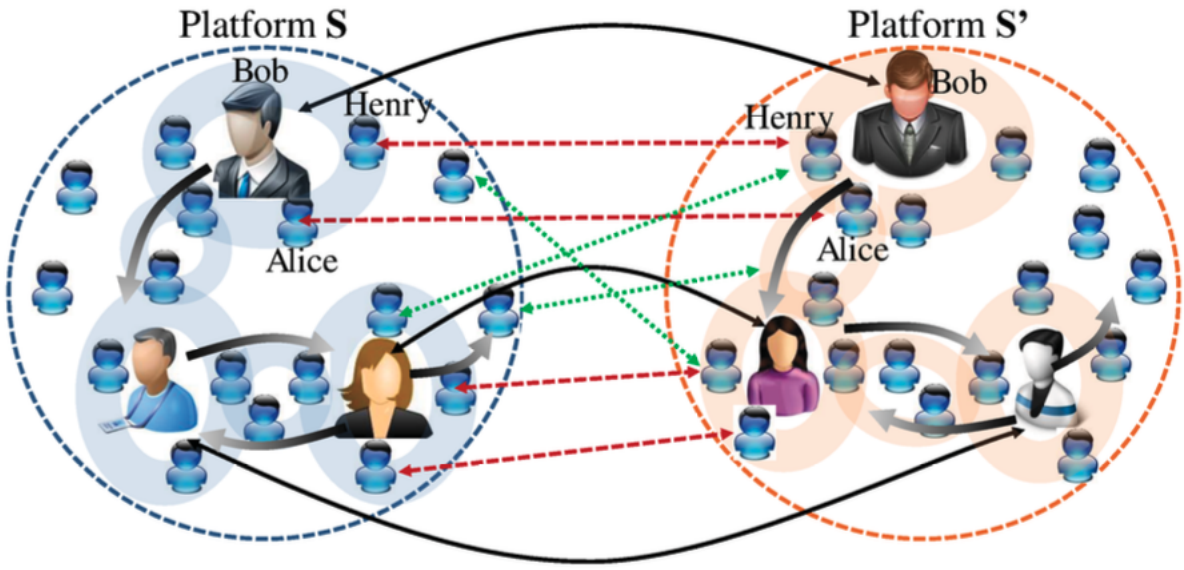


Figure 3.1: Illustrate the UIL problem

3.2 UIL Problem Definition

A social network is a graph $G = \{\mathcal{V}, \mathcal{E}, \mathcal{X}\}$, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is a set of nodes representing the users, and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is a set of edges representing the social relationships among users, e.g., follower/followee on Instagram and Twitter. Each user v_i is associated with a d-dimensional stylometric feature vector x_i (the i-th row in $\mathcal{X}$), which is extracted from the text written by the user v_i.

Let $G^s = \{\mathcal{V}^s, E^s, \mathcal{X}^s\}$ and $G^t = \{\mathcal{V}^t, \mathcal{E}^t, \mathcal{X}^t\}$ be the source and target networks, respectively, In these networks, $\mathcal{V}^s$ and $\mathcal{V}^t$ are the sets of users, $\mathcal{E}^s$ and $\mathcal{E}^t$ are the sets of edges representing connections between users, and $\mathcal{X}^s$ and $\mathcal{X}^t$ are the sets of stylometric features.

The goal of User Identity Linkage (UIL) is to predict whether a user v_i^s in the source network and a user v_j^t in the target network correspond to the same individual in the real world. Formally, the linkage can be defined as a function $f(v_i^s, v_j^t | T, G^s, G^t)$, which is

defined as:

$$f(v_i^s, v_j^t | T, G^s, G^t) = \begin{cases} 1, & \text{if } v_i^s = v_j^t \\ 0, & \text{otherwise} \end{cases}$$

The output is a binary value, where 1 indicates that v_i^s and v_j^t refer to the same person, and 0 indicates otherwise.

Table 3.1: Major Used Notations in This Thesis

Notation	**Description**
G^s and G^t	The source and target social networks.
$\mathcal{V}$ and $\mathcal{E}$	The set of users and social connections between users.
v^i and i	The target user and the index of the user.
$\mathcal{X}$	The stylometric feature matrix for a network
$f()$	The linkage function
Φ	The mapping function
Z	The embedding for network, i.e. the output from GNNs

Chapter 4

Proposed Model Architecture

To solve the User Identity Linkage (UIL) problem, we propose a Graph Neural Networks (GNNs)-based model, named StyleLink. As shown in Figure 4.1, StyleLink consists of three key components: stylometric feature engineering, network embedding via GNNs, and supervised linkage learning. We will discuss each component in detail.

4.1 Stylometric Feature Extraction

To model users' writing styles, stylometric features like word choice, frequency, punctuation, and sentence length can be easily identified [55] and assembled into sets of representative characteristics. In this thesis, we extract and characterize the writing styles of users from the following aspects, following the framework proposed and commonly used by [17,28,30,31,84]. We evaluate on 274 static features including lexical, syntactical, structural features and idiosyncratic features specially designed for UGC on OSNs, as listed in Table 4.1. Specifically, the frequency of misspellings can reflect a user's attention to detail, educational background, and language proficiency. The use of abbreviations varies greatly among users, reflecting their communication style, level of formality, and adaptation to platform norms. Users who frequently interact with various topics may exhibit a broader range of abbreviations, reflecting their engagement level. Therefore, we choose

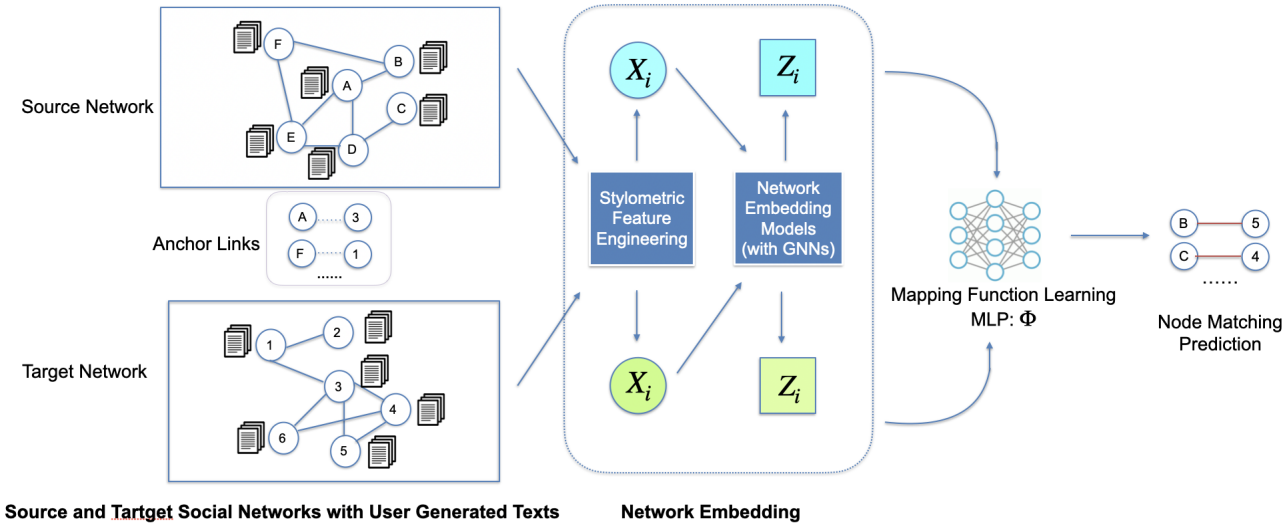


Figure 4.1: Illustration of the StyleLink model workflow. The process begins with obtaining source and target social network information, including user connections and their publicly posted texts (textual UGC). Next, stylometric features are extracted from the UGC and input into a Graph Neural Network to generate network embeddings that better represent the users. Subsequently, a mapping function is constructed to learn the relationships across the two OSNs. Finally, the user linkage results are produced.

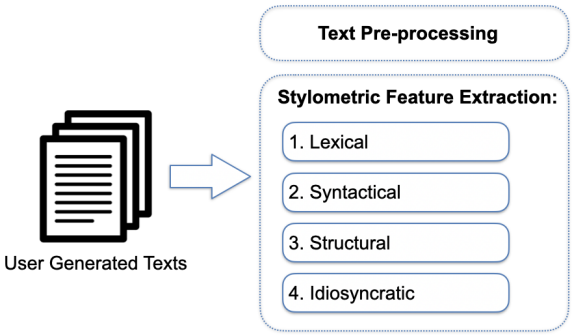


Figure 4.2: Overview of Stylometric Feature Extraction: User-generated texts undergo text pre-processing followed by extraction of four key stylometric features—Lexical, Syntactical, Structural, and Idiosyncratic—to represent writing styles.

to incorporate these features into our model to enhance the accuracy and reliability of UIL.

Table 4.1: List of Stylometric Features. The features from 1 to 242 are adapted from [84], and the corresponding abbreviations are listed in Table 4.2.

Categories	**Examples**
Lexical Features F1	*Character-based features:* 1. Total number of characters(C) 2. Ratio of alphabetic characters/C 3. Ratio of upper-case characters/C 4. Ratio of digits/C 5. Ratio of tabs/C 6–31. Frequency of letters, ignoring case (26 features: A to Z) 32–53. Frequency of special characters (22 features: ()<>%—{} []/# ˜ +-*=$ˆ&) *Word-based features:* 54. Total number of words (M) 55. Ratio of short words (less than four characters)/M 56. Total number of characters in words/C 57. Average word length (in characters) 58. Average sentence length (in characters) 59. Average sentence length (in words) 60. Total different words/M 61. Yule's K measure* (A vocabulary richness measure defined by Yule) 62–81. Word length frequency distribution / M (20 features) Frequency of words in different lengths
Syntactic Features F2	82–89. Frequency of punctuations (8 features) including " , . ? ! : ; ' 90–239. Frequency of function words (150 features) ([84])
Structural Features F3	240. Total number of sentences 241. Average sentences per post 242. Average URL per post
Idiosyncratic Features F4	243. Average Misspelled words per post 244-273. Abbreviation Frequency 274. Average Abbreviation Diversity

Table 4.2: Common Social Media Abbreviations and Their Meanings. We chose these 30 words as they are widely used and representative across various OSNs [34, 54].

Abbreviation	Meaning	Abbreviation	Meaning
AFAIK	As far as I know	AFK	Away from keyboard
ASAP	As soon as possible	BC, B/C	Because
BFF	Best Friend Forever	BRB	Be right back
BTW	By the way	DM	Direct message
FYI	For your information	IDK	I don't know
IMO	In my opinion	RN	Right now
JK	Just kidding	LMK	Let me know
LMAO	Laughing my ass off	LOL	Laugh out loud
NB	Not bad	NP	No problem
NVM	Never mind	OFC	Of course
OMG	Oh my God	OMW	On my way
PM	Private Message	TBH	To be honest
TMI	Too much information	HBD	Happy Birthday
TY	Thank You	WTF	What the f***
YW	You're welcome	XOXO	A term to convey affection

4.2 Graph Neural Networks for Network Embedding

In StyleLink, both source and target networks are embedded into continuous and low-dimensional spaces, represented as Z^s and Z^t respectively, and a mapping function $\Phi : Z^s \rightarrow Z^t$, which maps the latent spaces from the source to the target, is learned. Firstly, we apply Graph Convolutional Networks (GCNs) [35], one of effective GNNs that captures high-order information from neighboring nodes, to embed the source and target networks. For multi-layer GCNs, the layers can be mathematically defined as:

$$H^1 = \sigma(D^{-\frac{1}{2}}\widetilde{A}D^{-\frac{1}{2}}XW^0) \tag{4.1}$$

Table 4.3: List of English function words in our feature set, adapted from [84].

a	both	inside	off	such	what
about	but	into	on	than	whatever
above	by	is	once	that	when
after	can	it	one	the	where
all	cos	its	onto	their	whether
although	do	latter	opposite	them	which
am	down	less	or	these	while
among	each	like	our	they	who
an	either	little	outside	this	whoever
and	enough	lots	over	those	whom
another	every	many	own	though	whose
any	everybody	me	past	through	will
anybody	everyone	more	per	till	with
anyone	everything	most	plenty	to	within
anything	few	much	plus	toward	without
are	following	must	regarding	towards	worth
around	for	my	same	under	would
as	from	near	several	unless	yes
at	have	need	she	unlike	yet
be	he	neither	should	until	you
because	her	no	since	up	your
before	him	nobody	so	upon	
behind	i	none	some	us	
below	if	nor	somebody	used	
beside	in	nothing	someone	via	
between	including	of	something	we	

$$H^{k+1} = \sigma(D^{-\frac{1}{2}}\widetilde{A}D^{-\frac{1}{2}}H^kW^k) \tag{4.2}$$

where H^k is the node embedding matrix at layer k, X is the matrix of stylometric features and also the initial layer H^0, $\sigma(\cdot)$ is an activation function (e.g., we could choose ReLU $\sigma(x) = max(0, x)$), $\tilde{A}$ is the graph adjacency matrix with the addition of self-loops, ensuring that each node's own features are included in the aggregation process. $\tilde{D}$ represents the degree matrix of the graph, which includes self-loops as well. W^l is the weight matrix at layer l to learn.

Adding the attention mechanism in the process of GNNs learning, where a node involves the most relevant information from its neighborhoods and updates its own features with the learned attention weights, allows the model to focus on important nodes or edges in the graph while alleviating noise signals during message passing in the network. Here we applied Graph Attention Network (GAT) [65] as the second variant to our embedding and mathematically, the attention mechanism can be defined as follows:

Let $\mathbf{h}_i$ denote the hidden state of node i. The attention coefficient e_{ij} between node i and node j is computed as:

$$e_{ij} = \text{LeakyReLU}(\alpha^T [Wh_i \,\|\, Wh_j]) \tag{4.3}$$

where α is the attention vector, W is the weight matrix, and $\|$ denotes concatenation. The normalized attention coefficients α_{ij} are tcomputed using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{4.4}$$

Finally, the new representation of node i is computed as a weighted sum of its neighbors' representations, taking into account the attention coefficients:

$$\mathbf{h}'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W} \mathbf{h}_j \right) \tag{4.5}$$

where σ is a non-linear activation function, and $\mathcal{N}(i)$ refers to the set of neighbors of node i.

Thus, the attention mechanism enables the model to focus on important nodes or edges while aggregating the writing styles from neighbors effectively.

4.3 Supervised Linkage Learning

After obtaining the representation $Z^s \in R^{d \times n}, Z^t \in R^{d \times n}$ from GNNs for the source and target network graphs, respectively, the next step is to learn a mapping function $\Phi : Z^s \rightarrow Z^t$. This mapping function is supervised using the known anchor links T. We need to minimize the objective function during the mapping function learning.

4.3.1 Cosine Similarity

In many existing approaches, cosine similarity [13,36,70,85,86,86] has widely been adopted to compute the objective function to learn the mapping function. Cosine similarity measures the likeness between two vectors by calculating the cosine of their angle. In user identity linkage, this metric assesses the similarity of user feature representations across different social networks. Higher cosine similarity suggests greater vector alignment, indicating a higher likelihood of representing the same user. Unlike magnitude-dependent measures, cosine similarity focuses on directionality, making it particularly effective in high-dimensional spaces. This makes it a simple yet powerful tool to implement for comparing stylometric feature embeddings in user identity linkage tasks.

$$\mathcal{L}_{\text{cosine}} = \arg\min_{W_\Phi, b}(1 - \cos(\Phi(Z^s), \Phi(Z^t))) \tag{4.6}$$

4.3.2 Triplet Loss

Our proposed model adopts the Triplet Loss for the objective function in supervised linkage learning.The concept of triplet loss was initially developed for facial recognition applications [57]. It has shown an improved ability to distinguish between different items in the embedding space. Unlike cosine similarity, which only looks at pairs of items, triplet loss considers groups of three. This approach pushes the mapping function to position correct matches significantly closer together in the latent space compared to incorrect matches.

We need to minimize the objective function as follows during the mapping function learning:

$$\mathcal{L}_{\text{triplet}} = \arg\min_{W,b} \sum_{(a,p,n)\in\mathcal{T}} \big[\|\Phi(Z_a^s) - Z_p^t\|_2^2 - \|\Phi(Z_a^s) - Z_n^t\|_2^2 + \alpha\big] \tag{4.7}$$

where Anchor (a) represents a node from the source network; Positive (p) represents the corresponding node from the target network, i.e. the same user; and Negative (n) is a different node from the target network.

Based on comparisons between linear and non-linear mapping functions in [47], we also decide to employ Multi-Layer Perceptron (MLP) as our mapping function Φ, which is able to capture the non-linear mapping relationship between the source and target social networks.

4.4 Complexity Analysis

In this section, we explicitly explain the complexity of our approach with an example of applying GCNs for social network embedding and MLP for mapping function.

4.4.1 Graph Convolutional Networks)

Applying GCNs to embed the source and target networks involves several steps:

- **Graph Construction:** Building the adjacency matrix for each network. For a network with n nodes and m edges, constructing this matrix has a complexity of $O(m)$.
- **GCN Layers:** Each GCN layer performs a convolution operation on the graph, aggregating information from neighboring nodes. Given l GCN layers, the complexity per layer is $O(|E| \cdot d)$, where $|E|$ is the number of edges and d is the feature dimension. For l layers, the total complexity is $O(l \cdot |E| \cdot d)$.

- **Embedding Computation:** After l layers, the final node embeddings are computed. This process is generally linear in the number of nodes and edges, making the complexity $O(|V| \cdot d)$, where d is the dimensionality of the embeddings.

4.4.2 Multi-Layer Perceptron (MLP) for Mapping Function Φ

The MLP is used to learn the transformation between source and target network embeddings. The complexity of this step is influenced by:

- **Embedding Size:** Assuming embeddings of size d for each network, and considering an MLP with k layers and each layer having m neurons, the complexity of a forward pass through the MLP is $O(k \cdot (d^2 \cdot m))$.

- **Training:** During training, the complexity involves forward and backward passes. If the MLP has k layers with each layer having m neurons, the total training complexity per epoch is $O(k \cdot d^2 \cdot m \cdot n)$, where n is the number of training samples.

4.4.3 User Identity Linkage

Performing user identity linkage involves comparing the learned embeddings from both networks. Assuming $|V_s|$ and $|V_t|$ are the number of nodes in the source and target networks, the complexity for computing pairwise similarities is $O(|V_s| \cdot |V_t| \cdot d)$.

4.4.4 Overall Complexity

The overall complexity of the method can be summarized as:

- **GCNs:** $O(L \cdot (|E_s| + |E_t|) \cdot d)$

- **MLP Training:** $O(k \cdot d^2 \cdot m \cdot n)$

- **Linkage:** $O(|V_s| \cdot |V_t| \cdot d)$

In summary, while the method is computationally intensive due to multiple steps involving large-scale graph operations, embeddings, and neural network training, the overall complexity is manageable with respect to modern computational resources and can be optimized based on specific use cases and network sizes.

Chapter 5

Experimental Results

In this section, we evaluate the proposed StyleLink model, including both its GCN and GAT variants, through two experimental tasks. Our investigation aims to address several key aspects of the model's performance and capabilities. Firstly, we explore the effectiveness of StyleLink in predicting user identities across different OSNs. We want to see how our model performs compared to state-of-the-art (SOTA) methods in the field of UIL. In addition, we focus on how these linguistic characteristics contribute to creating more representative and informative network embeddings for OSNs. We examine various aspects of stylometric features to understand their individual and collective impact on both the quality of network embeddings and the overall performance of user identity linkage tasks.

All the experiments are carried out on a Windows Server equipped with two Xeon E5-2697 CPUs (36 cores), 384 GB of RAM, and four NVIDIA TITAN XP GPUs.

5.1 Datasets Preparation

To validate our approach, we conduct experiments using the real-world partially aligned OSN datasets: X[1] - Foursquare[2]. X (formerly Twitter) is a social media platform where

[1]https://twitter.com/home
[2]https://foursquare.com/

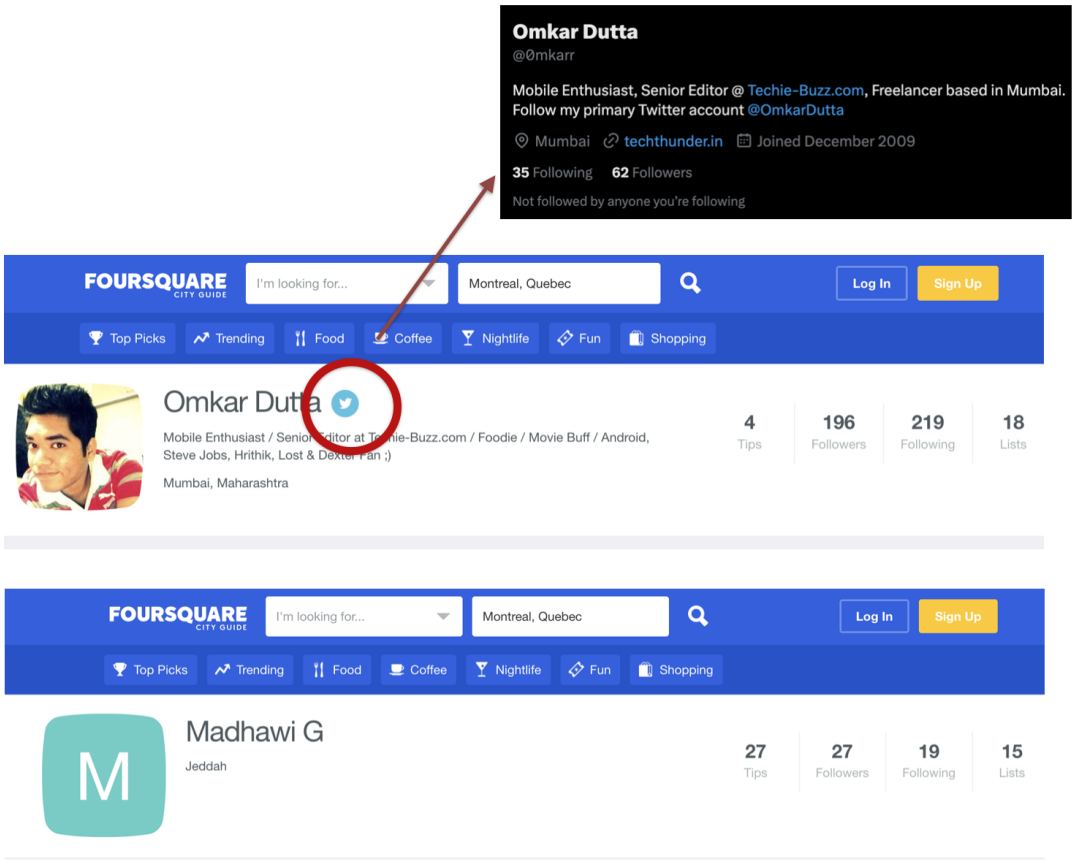


Figure 5.1: This figure illustrates the process of confirming anchor users across datasets. On a Foursquare user's profile page, a Twitter icon with hyperlink may appear to the right of the username. Clicking this icon directs you to the Twitter profile of the same individual. For users who haven't linked their Twitter accounts, it is challenging to establish them as anchor links.

users post short messages, images, and videos, and interact through likes, replies, and reposts. Foursquare is a location-based service that allows users to discover and review places, while businesses can use its analytics tools to engage customers and optimize operations. The statistics information of the datasets is shown as Table 5.1. This dataset is originally provided by Zhang et al. [80], where users of two social networks are partially aligned. The ground truth of **1,609** anchors are confirmed by users explicitly as shown in Figure 5.1.

We extended the original datasets with additional UGC scraping until year of 2023, such that the average number of posts are increased by 30% in Table 5.1 to avoid stylometric features being too sparse.

To improve the accuracy of our analysis, we pre-process the datasets by removing non-English UGCs first. Then for each valid tweet from X or tip from Foursquare, we remove user mentions (e.g., "@username"), hashtags (e.g., #topic), and replace URLs uniformly with a specified token. These elements are excluded because they are often generic and repetitively used by many users, making them less useful for representing the unique writing styles of users.

Table 5.1: Summary Statistics of X - Foursquare Dataset, with 1,609 anchors users.

	N	**V**	**Avg Degree**	**Avg Posts**	**Vocabulary Size**
X	5,120	130,575	60.28	1,405.5	90,661
Foursquare	5,313	54,233	26.05	270.6	480,135

5.2 Evaluation Metrics

As outlined in [60], we use several standard metrics to assess prediction and ranking performance, including *Precision@k (P@K)*, *MAP*, *AUC* and *Hit-Precision* [50].

In the setting of UIL, *Precision@k (P@k)* is the metric for evaluating the linking accuracy, which is exactly the same as *Recall@k* and $F_1@k$. It is defined as:

$$P@k = \sum_{i}^{n} \mathbb{I}_i\{success@k\}/n \tag{5.1}$$

Here, $\mathbb{I}_i\{success@k\}/n$ evaluates whether the correctly matched identity is present in the *top-k ($k \leq n$)* results, where *n* represents the total number of testing anchor nodes.

For evaluating the ranking performance of the algorithms, we use the following measurements:

$$\begin{aligned} MAP &= \frac{1}{n}(\sum^{n}\frac{1}{ra}) \\ AUC &= \frac{1}{n}(\sum^{n}\frac{m+1-ra}{m}) \\ Hit-Precision &= \frac{1}{n}(\sum^{n}\frac{k-(ra-1)}{k}) \end{aligned} \tag{5.2}$$

where ra represents the rank position of the positive matching identities, i.e., the matched target user in the return *top-k* candidate target entities, m is the number of negative user identities, and n is the number of total testing anchor nodes. Mean Average Precision (MAP) is a robust metric known for its strong discrimination and stability. Unlike precision@k, MAP places greater emphasis on the ranking of the top returned items. It is important to note that for all these metrics, a **higher** value indicates **better** model performance.

5.3 Comparing Models

We evaluate the performance of StyleLink by comparing it with the following baselines, among these baseline models, network embedding methods are employed such that the user latent space is obtained for aligning the user identities.

Table 5.2: Comparison among different baseline UIL methods. The comparison tells whether their network embedding methods involves topology and attributes or not.

UIL method	Type	Topology	Attribute
MNA [36]	supervised	×	✓
RLink [41]	reinforcement	✓	×
PALE [47]	supervised	✓	✓
DeepLink [86]	supervised	✓	×
IONE [44]	supervised	✓	×

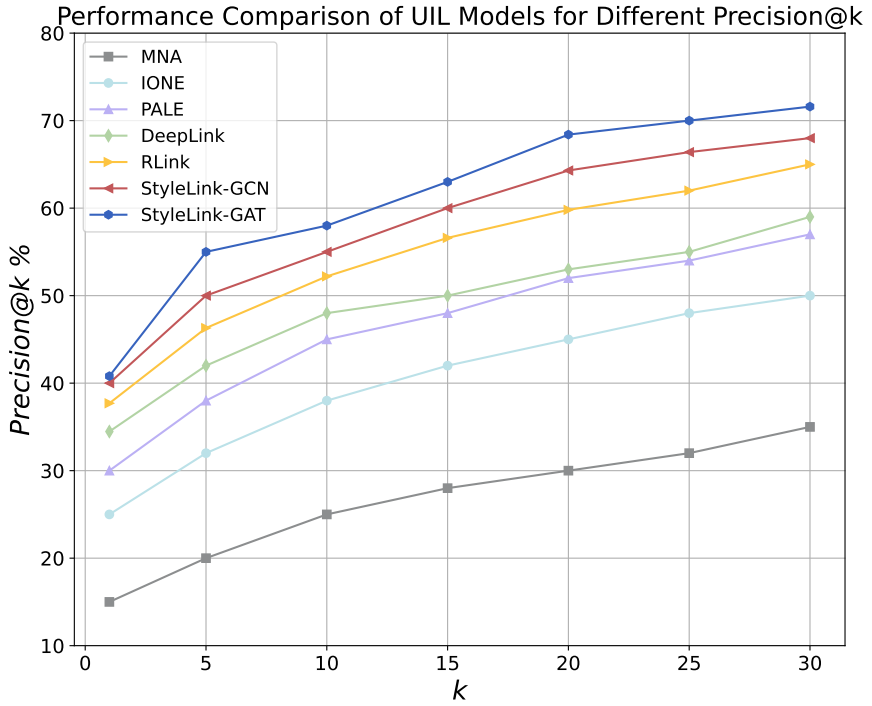
- **IONE** [44]: Input-Output Network Embedding (IONE) proposes a network embedding method to learn the follower-ship/followee-ship of each user simultaneously

and utilities input/output context vectors to preserve the proximity of anchor users. In IONE, the followers and followees of users are embedded with three vectors: node vector, input vector and output vector.

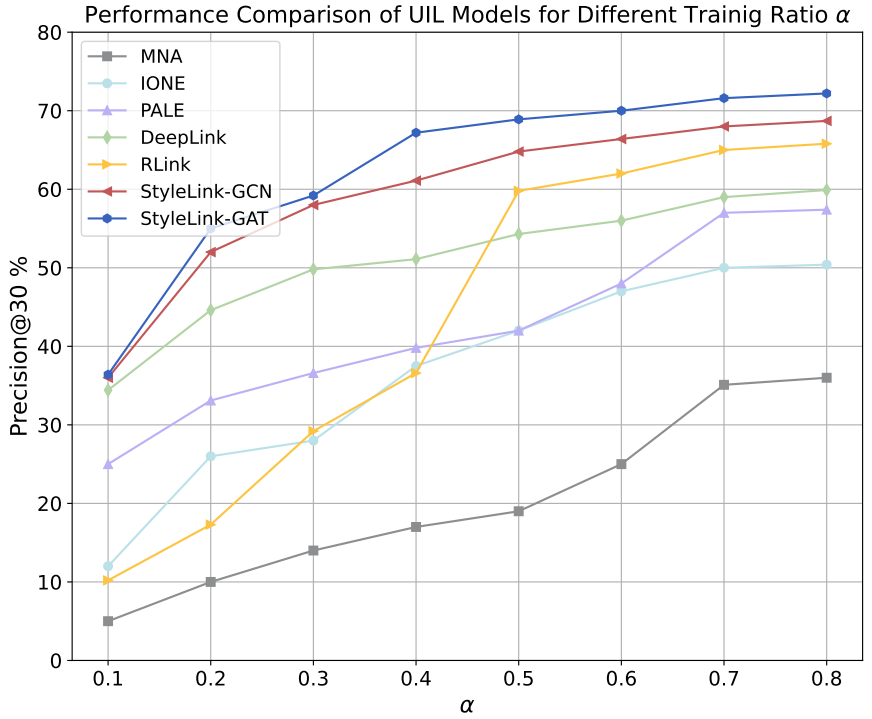
- **PALE (MLP)** [47]: Predicting Anchor Links via Embedding (PALE) conducts network embedding to capture its major structural regularity. In the matching stage, it learns a mapping function (MLP) across two low-dimensional latent spaces.
- **DeepLink** [86]: DeepLink is a deep reinforcement learning based algorithm which applies unbiased Random Walk to generate embeddings and uses MLP in a dual learning way to map users.
- **MNA** (Multi-Network Anchoring) [36]: MNA extracts social features, including spatial, temporal and text content features (bag-of-words vectors weighted by TF-IDF), and neighborhood-based network features and match user identity pairs.
- **RLink** [41]: RLink applies Node2Vec [23] to pre-train the network embedding and concatenates the embeddings of source and target networks to represent network structure information. Specifically, it is the first to consider UIL as a sequence decision problem and proposes a deep reinforcement learning model.

5.4 Experimental Performance Analysis

First of all, we compare the performances of various approaches by linking precision $P@k$, as presented in Figure 5.2a. We set the training ratio α to be 0.7 and present the results of different $P@k$. The results in Figure 5.2a show that both StyleLink-GCN and StyleLink-GAT consistently outperform the other models across all values of k, with StyleLink-GAT achieving the highest accuracy. On average, our method of both variants achieves a 9.2% improvement over the baseline model RLink and a 21.2% improvement over DeepLink on the X-Foursquare datasets. We observe that models utilizing deep learning techniques, such as DeepLink, PALE, RLink, and our proposed variants,



(a) This figure shows the performance of UIL models for different P@k, on the X-Foursquare dataset.



(b) This figure shows the performance of UIL models for different training ratios α, on the X-Foursquare dataset.

Figure 5.2: Performance Comparison on X-Foursquare Datasets: Each experiment was repeated 10 times, and the mean evaluation results were recorded.

StyleLink-GCN and StyleLink-GAT, generally achieve higher linking precision compared to models that do not employ neural networks, such as IONE and MNA. Specifically, IONE, DeepLink, PALE, RLink, and the StyleLink variants significantly outperform MNA, which achieves only 36.04% precision at $P@30$, whereas the other models achieve comparable precision at $P@5$. Compared to PALE and DeepLink, both of which use supervised mapping with deep learning methods, StyleLink demonstrates superior performance by integrating writing style features into the network structure embeddings.

Furthermore, we varied the training ratio from 0.1 to 0.8 and evaluated $P@30$ for each method. The proportion of anchor nodes T used during training significantly impacts the performance of UIL models. While RLink exhibits competitive performance, particularly at higher training ratios, it does not reach the precision levels achieved by the StyleLink models. This suggests that while considering UIL as a sequence decision problem is beneficial, the network embeddings generated via Node2Vec in RLink are not as effective as those produced by our GNNs-based embeddings.

To summarize, the above observations demonstrate that our proposed StyleLink models, both StyleLink-GCN and StyleLink-GAT, effectively address the User Identity Linkage (UIL) problem. Compared to other baseline models, StyleLink demonstrates significantly better performance with a lower proportion of training anchor nodes. It can effectively learn meaningful representations and perform well in scenarios where training data is incomplete or imbalanced, which is common for authentic social network datasets. In addition, StyleLink-GAT, which incorporates an attention mechanism, achieves superior linkage performance over other models.

5.5 Effectiveness of Social Network Embedding via GCNs

This experiment is meant to validate the effectiveness of our approach of generating the stylometric features and then applying GCNs to embed the whole network. We present

visualizations in Figure 5.3 to illustrate that applying GNNs in network embedding indeed helps generate more meaningful and distinguishable embeddings.

Firstly, the embeddings are reduced to two dimensions using t-SNE [63] for visualization. Then, we can use some density-based clustering algorithm, for example, we adopted DBSCAN [18] for large spatial databases with noise here, to group similar nodes based on the reduced embeddings. Nodes with a similarity score above a defined threshold are filtered for clarity. Positions for these nodes are determined, and colors are assigned according to their cluster labels. The filtered nodes and their connecting edges are then visualized in a network graph, with a color bar indicating cluster labels, helping to highlight similarities in writing styles.

From the visualization comparison in Figure 5.3, we observe that representing users solely with stylometric features results in several clustering communities, where users with similar embeddings tend to cluster closely together in the embedding space. However, as seen in Figure 5.3 (a), despite the same number of users being represented as in Figure 5.3 (b), most nodes overlap significantly, forming extremely dense clusters.

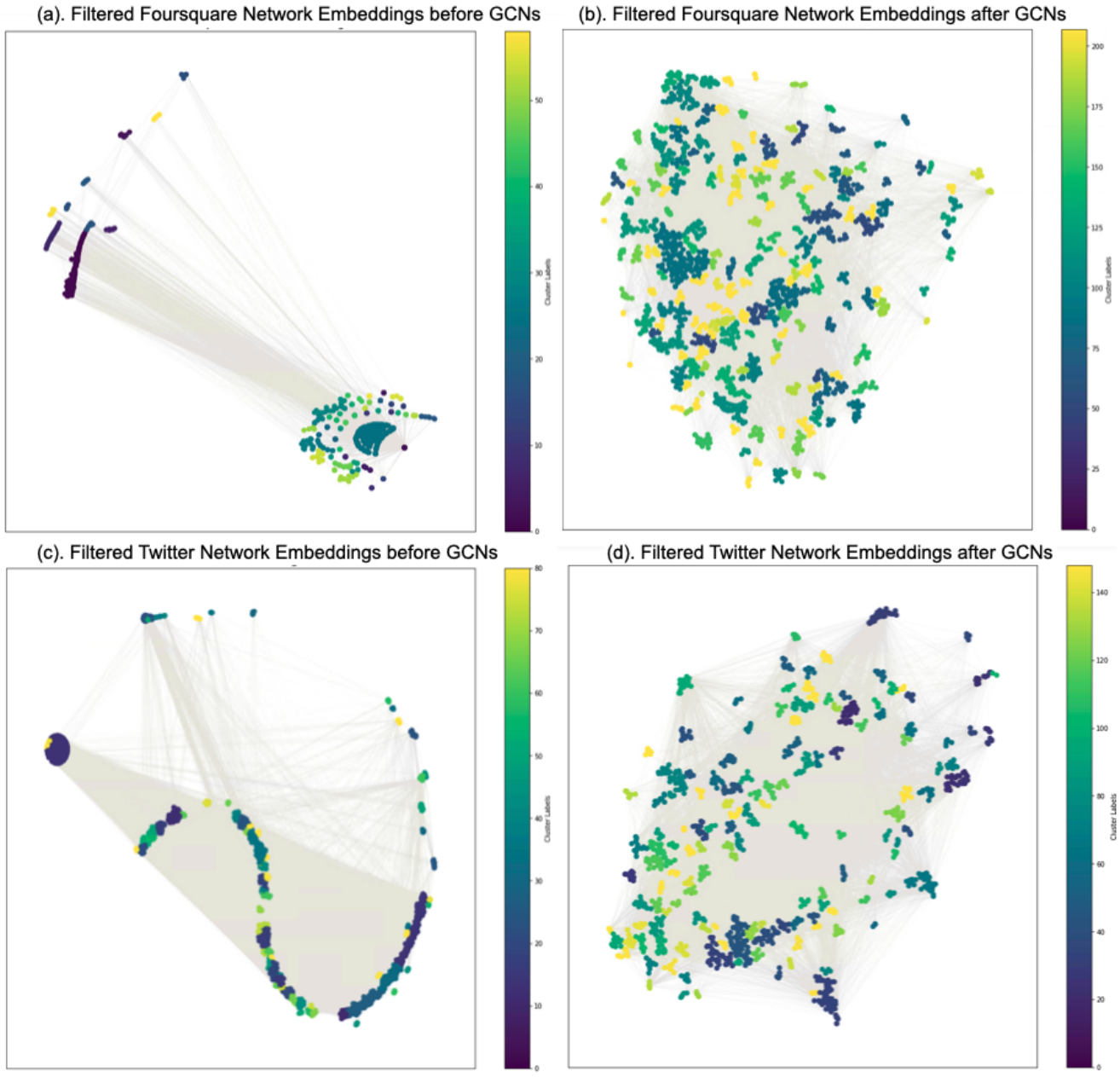
In contrast, after applying GCNs to generate network embeddings, we notice that users with similar embeddings (i.e., users with similar colors) still form clusters, but these clusters are more dispersed and better by clearer boundaries. This observation indicates that GCNs can create more distinct clusters, which helps differentiate between various users.

These observations are also evident in Figures 5.3 (c) and 5.3 (d), where similar patterns in the X dataset can be observed.

5.6 Ablation Study

An ablation study was carried out to determine the contribution of different components of stylometric features to the network embedding and UIL performance on OSNs. In Table 4.1, stylometric features are divided into 4 categories, from the perspective of different

Figure 5.3: We present embedding visualizations for the X-Foursquare datasets, comparing the representations before and after applying GCNs. High dimensional embeddings of V, X, and Z, are projected to 2D dimension and the light grey lines represent the edges E from the network graphs. To enhance clarity and improve visualization quality, we filtered out nodes with similarity scores below a certain threshold. These filtered nodes, colored in dark purple, contribute to visual clutter if not removed. After applying this filtering process, 2,004 Twitter users and 2,369 Foursquare users remain, which were used to generate the visualizations shown above.



linguistics. Therefore, we conducted experiments on X-Foursquare datasets between different variants of StyleLink, with different category of stylometric features padded zeros respectively. Negative values indicate a decrease in performance when that category of features is padded with zeros.

Table 5.3: Stylometic Feature Ablation Results.

Features	StyleLink-GCN		StyleLink-GAT	
	P@10	MAP@10	P@10	MAP@10
All features	55.3	47.1	57.5	50.6
(–) Lexical	-1.87	-2.32	-1.00	-0.89
(–) Syntactic	-1.43	-1.21	-0.60	-1.10
(–) Structural	-0.83	-0.29	-0.40	-0.50
(–) Idiosyncratic	-1.02	-1.29	-1.40	-1.00

Overall, each category of stylometric feature types contributes positively to the performance of our model, but to different extents. Lexical features remain the most critical for StyleLink-GCN according to both metrics. For StyleLink-GAT, idiosyncratic features have the largest impact in terms of P@10, while lexical and syntactic features affect MAP more. Compared to StyleLink-GAT, which demonstrates more balanced sensitivity across different feature types, StyleLink-GCN exhibits higher sensitivity to feature ablation, particularly for lexical and syntactic features. Structural features have the least impact on StyleLink-GCN but are more influential for StyleLink-GAT.

5.7 Discussion

This study aims to develop and evaluate the StyleLink model, integrating both GCNs and GAT variants, to improve user identity linkage (UIL) across various online social networks (OSNs). Our main objective was to evaluate whether incorporating stylometric features, linguistics characteristics inherent in users' writing styles, into Graph Neural Networks could enhance the linking precision and quality of network embeddings for UIL tasks. We compare StyleLink's performance against leading UIL methods, examining

how these stylometric features contribute to the model's ability to generate superior social network embeddings and accurately link user identities across different OSNs.

Our results align with previous studies highlighting the importance of linguistic features in user identification [20, 61, 84]. However, our work extends these findings by demonstrating the effectiveness of GNNs in leveraging these features for cross-platform identity linkage. While our study demonstrates the effectiveness of StyleLink, our experiments were conducted on one benchmark dataset, however with rich information on textual UGCs. The model's performance is expected to be assessed on additional OSNs with varying platform functionalities and user behaviors.

Chapter 6

Conclusion and Future Work

This research introduced and evaluated the StyleLink model, incorporating GCN and GAT variants, for user identity linkage (UIL) across different online social networks (OSNs). The results demonstrate that StyleLink, particularly the GAT variant, significantly outperforms some of state-of-the-art UIL methods in linkage precision, especially as the training data ratio increases. This performance improvement highlights the effectiveness of integrating stylometric features into graph-based models, providing a more effective and representative embedding of user identities across OSNs.

The thesis confirms that stylometric features, such as lexical, syntactical, structural, and idiosyncratic characteristics, play a crucial role in enhancing the quality of network embeddings, leading to more accurate UIL. These findings contribute to UIL across different OSNs by offering a novel approach that leverages both linguistic analysis and advanced graph neural networks. In summary, StyleLink presents a promising direction for future research and practical applications in OSNs analysis, where accurate user identity linkage is critical.

There are several directions that need to be investigated in the future. Since UGCs are always associated with timestamps, we aim to explore whether temporal writing style evolution plays a significant role in user identity linkage on social networks. This could involve developing time-aware variants of StyleLink that can capture and lever-

age temporal patterns in user behavior and writing style. A key challenge with current stylometric-based user identity linkage lies in the variability of stylometric features across users. Some users generate strong, distinct writing styles, while others produce weaker or less consistent signals, potentially affecting the model's ability to link identities effectively. Future work could also focus on developing evaluation metrics to assess the strength of a user's stylometric feature and implementing triage mechanisms, thus enhancing the model's robustness and ensuring more reliable identity linkage across diverse user populations. We also intend to explore the application of more advanced Graph Neural Network architectures, such as Graph Transformers [26,76], to potentially enhance the capability to capture noisy, complex, and large-scale social network structures. Finally, we aim to extend StyleLink to handle multi-platform scenarios beyond pairwise network alignment, enabling simultaneous user linkage across multiple social networks.

Bibliography

[1] AHMAD, W., AND ALI, R. Social account matching in online social media using cross-linked posts. *Procedia Computer Science 152* (2019), 222–229.

[2] ALTAKRORI, M. H., CHEUNG, J. C. K., AND FUNG, B. C. M. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Punta Cana, Dominican Republic, November 2021), ACL, pp. 4242–4256.

[3] ALTAKRORI, M. H., SCIALOM, T., FUNG, B. C. M., AND CHEUNG, J. C. K. A multifaceted framework to evaluate evasion, content preservation, and misattribution in authorship obfuscation techniques. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Abu Dhabi, UAE, December 2022), ACL, pp. 2391–2403.

[4] BO, H., DING, S. H. H., FUNG, B. C. M., AND IQBAL, F. ER-AE: Differentially private text generation for authorship anonymization. In *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (Online, June 2021), ACL, pp. 3997–4007.

[5] BOK, K., LIM, J., YANG, H., AND YOO, J. Social group recommendation based on dynamic profiles and collaborative filtering. *Neurocomputing 209* (2016), 3–13.

[6] BONHARD, P., AND SASSE, M. A. 'knowing me, knowing you'—using profiles and social networking to improve recommender systems. *BT Technology Journal 24*, 3 (2006), 84–98.

[7] BRONSTEIN, M. M., BRUNA, J., COHEN, T., AND VELIČKOVIĆ, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[8] CARMAGNOLA, F., AND CENA, F. User identification for cross-system personalisation. *Information Sciences 179*, 1 (2009), 16–32.

[9] CENTER, P. R. Social media use in 2018. `https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/`. Accessed: 2018.

[10] CENTER, P. R. Social media use in 2021. `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/`. Accessed: 2021.

[11] CHATZAKOU, D., SOLER-COMPANY, J., TSIKRIKA, T., WANNER, L., VROCHIDIS, S., AND KOMPATSIARIS, I. User identity linkage in social media using linguistic and social interaction features. In *Proceedings of the 12th ACM Conference on Web Science* (2020), pp. 295–304.

[12] CHEN, D., LIN, Y., LI, W., LI, P., ZHOU, J., AND SUN, X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence* (2020), vol. 34, pp. 3438–3445.

[13] CHEN, X., SONG, X., CUI, S., GAN, T., CHENG, Z., AND NIE, L. User identity linkage across social media via attentive time-aware user modeling. *IEEE Transactions on Multimedia 23* (2020), 3957–3967.

[14] CHU, X., FAN, X., YAO, D., ZHU, Z., HUANG, J., AND BI, J. Cross-network embedding for multi-network alignment. In *Proeedings of The World Wide Web Conference (WWW)* (New York, NY, USA, 2019), WWW '19, p. 273–284.

[15] CONTE, D., FOGGIA, P., SANSONE, C., AND VENTO, M. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence 18*, 03 (2004), 265–298.

[16] DING, S. H. H., FUNG, B. C. M., AND DEBBABI, M. A visualizable evidence-driven approach for authorship attribution. *ACM Transactions on Information and System Security (TISSEC) 17*, 3 (March 2015), 12:1–12:30.

[17] DING, S. H. H., FUNG, B. C. M., IQBAL, F., AND CHEUNG, W. K. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics (CYB) 49*, 1 (January 2019), 107–121.

[18] ESTER, M., KRIEGEL, H.-P., SANDER, J., AND XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (SIGNDD)* (1996), AAAI Press, p. 226–231.

[19] FIRE, M., KAGAN, D., ELYASHAR, A., AND ELOVICI, Y. Friend or foe? fake profile identification in online social networks. *Social Network Analysis and Mining 4* (2014), 1–23.

[20] GOGA, O., LEI, H., PARTHASARATHI, S. H. K., FRIEDLAND, G., SOMMER, R., AND TEIXEIRA, R. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 447–458.

[21] GOGA, O., LOISEAU, P., SOMMER, R., TEIXEIRA, R., AND GUMMADI, K. P. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015), pp. 1799–1808.

[22] GOGA, O., PERITO, D., LEI, H., TEIXEIRA, R., AND SOMMER, R. Large-scale correlation of accounts across social networks. *University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002* (2013).

[23] GROVER, A., AND LESKOVEC, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), pp. 855–864.

[24] HAMILTON, W. L. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning 14*, 3, 1–159.

[25] HAMILTON, W. L., YING, R., AND LESKOVEC, J. Inductive representation learning on large graphs, 2018.

[26] HU, Z., DONG, Y., WANG, K., AND SUN, Y. Heterogeneous graph transformer. In *Proceedings of the web conference 2020* (2020), pp. 2704–2710.

[27] IOFCIU, T., FANKHAUSER, P., ABEL, F., AND BISCHOFF, K. Identifying users across social tagging systems. In *Proceedings of the International AAAI Conference on Web and Social Media* (2011), vol. 5, pp. 522–525.

[28] IQBAL, F., BINSALLEEH, H., FUNG, B. C. M., AND DEBBABI, M. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences: Special Issue on Data Mining for Information Security 231* (May 2013), 98–112.

[29] IQBAL, F., DEBBABI, M., AND FUNG, B. C. M. *Machine Learning for Authorship Attribution and Cyber Forensics*. Computer Entertainment and Media Technology. Springer Nature, December 2020.

[30] IQBAL, F., HADJIDJ, R., FUNG, B. C. M., AND DEBBABI, M. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation 5*, 1 (September 2008), S42–S51.

[31] IQBAL, F., KHAN, L. A., FUNG, B. C. M., AND DEBBABI, M. E-mail authorship verification for forensic investigation. In *Proc. of the 25th ACM SIGAPP Symposium*

on Applied Computing (SAC) (Sierre, Switzerland, March 2010), ACM Press, pp. 1591–1598.

[32] JAIN, P., AND KUMARAGURU, P. Finding nemo: Searching and resolving identities of users across online social networks, 2012.

[33] JAIN, P., KUMARAGURU, P., AND JOSHI, A. @ i seek'fb. me' identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 1259–1268.

[34] KHAIRUTDINOV, R. R., MUKHAMETZYANOVA, F. G., AND GAYSINA, A. R. Socio-psychological characteristics of the subject use of slang and abbreviations in english-speaking social networks. *Turk. Online J. Des. Art Commun* (2017).

[35] KIPF, T. N., AND WELLING, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations* (2017).

[36] KONG, X., ZHANG, J., AND YU, P. S. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013), pp. 179–188.

[37] KOVACS, B., AND KLEINBAUM, A. M. Language-style similarity and social networks. *Psychological science 31*, 2 (2020), 202–213.

[38] KUCHAIEV, O., AND PRŽULJ, N. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics 27*, 10 (2011), 1390–1396.

[39] KUMAR, S., ZAFARANI, R., AND LIU, H. Understanding user migration patterns in social media. *Proceedings of the AAAI Conference on Artificial Intelligence 25*, 1 (Aug. 2011), 1204–1209.

[40] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[41] LI, X., CAO, Y., LI, Q., SHANG, Y., LI, Y., LIU, Y., AND XU, G. Rlink: Deep reinforcement learning for user identity linkage. *World Wide Web 24* (2021), 85–103.

[42] LI, X., CHEN, L., AND WU, D. Adversary for social good: Leveraging adversarial attacks to protect personal attribute privacy. *ACM Trans. Knowl. Discov. Data 18*, 2 (nov 2023).

[43] LIU, J., ZHANG, F., SONG, X., SONG, Y.-I., LIN, C.-Y., AND HON, H.-W. What's in a name? an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining* (2013), pp. 495–504.

[44] LIU, L., CHEUNG, W. K., LI, X., AND LIAO, L. Aligning users across social networks using network embedding. In *IJCAI* (2016), vol. 16, pp. 1774–1780.

[45] LIU, L., QU, B., CHEN, B., HANJALIC, A., AND WANG, H. Modelling of information diffusion on social networks with applications to wechat. *Physica A: Statistical Mechanics and its Applications 496* (2018), 318–329.

[46] LIU, S., WANG, S., ZHU, F., ZHANG, J., AND KRISHNAN, R. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2014), SIGMOD '14, Association for Computing Machinery, p. 51–62.

[47] MAN, T., SHEN, H., LIU, S., JIN, X., AND CHENG, X. Predict anchor links across social networks via an embedding approach. In *Ijcai* (2016), vol. 16, pp. 1823–1829.

[48] MAZHARI, S., FAKHRAHMAD, S. M., AND SADEGHBEYGI, H. A user-profile-based friendship recommendation solution in social networks. *J. Inf. Sci. 41*, 3 (jun 2015), 284–295.

[49] MOTOYAMA, M., AND VARGHESE, G. I seek you: searching and matching individuals in social networks. In *Proceedings of the eleventh international workshop on Web information and data management* (2009), pp. 67–75.

[50] MU, X., ZHU, F., LIM, E.-P., XIAO, J., WANG, J., AND ZHOU, Z.-H. User identity linkage by latent user space modelling. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, p. 1775–1784.

[51] NARAYANAN, A., AND SHMATIKOV, V. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy* (may 2009), IEEE.

[52] NIE, Y., JIA, Y., LI, S., ZHU, X., LI, A., AND ZHOU, B. Identifying users across social networks based on dynamic core interests. *Neurocomputing 210* (2016), 107–115.

[53] PEROZZI, B., AL-RFOU, R., AND SKIENA, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), pp. 701–710.

[54] PRATIWI, I. D., AND MARLINA, L. An analysis of abbreviation in twitter status of hollywood pop singers. *English Language and Literature 9*, 1 (2020), 127–133.

[55] SARI, Y., STEVENSON, M., AND VLACHOS, A. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics* (Santa Fe, New Mexico, USA, Aug. 2018), E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Association for Computational Linguistics, pp. 343–353.

[56] SCHMID, M., IQBAL, F., AND FUNG, B. C. M. E-mail authorship attribution using customized associative classification. *Digital Investigation (DIIN) 14*, 1 (August 2015), S116–S126.

[57] SCHROFF, F., KALENICHENKO, D., AND PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015), IEEE.

[58] SHARMA, V., AND DYRESON, C. Linksocial: Linking user profiles across multiple social media platforms. In *2018 IEEE International Conference on Big Knowledge (ICBK)* (2018), IEEE, pp. 260–267.

[59] SHETTY, N. P., MUNIYAL, B., DOKANIA, A., DATTA, S., GANDLURI, M. S., MABEN, L. M., PRIYANSHU, A., AND REZAI, A. Guarding your social circle: Strategies to protect key connections and edge importance.

[60] SHU, K., WANG, S., TANG, J., ZAFARANI, R., AND LIU, H. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter 18*, 2 (2017), 5–17.

[61] SRIVASTAVA, D. K., AND ROYCHOUDHURY, B. Words are important: A textual content based identity resolution scheme across multiple online social networks. *Knowledge-Based Systems 195* (2020), 105624.

[62] TANG, J., QU, M., WANG, M., ZHANG, M., YAN, J., AND MEI, Q. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web* (2015), pp. 1067–1077.

[63] VAN DER MAATEN, L., AND HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 86 (2008), 2579–2605.

[64] VASWANI, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[65] VELIČKOVIĆ, P., CUCURULL, G., CASANOVA, A., ROMERO, A., LIÒ, P., AND BENGIO, Y. Graph attention networks, 2018.

[66] VOSOUGHI, S., ZHOU, H., AND ROY, D. Digital stylometry: Linking profiles across social networks. In *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7* (2015), Springer, pp. 164–177.

[67] WANG, Z., WU, C.-H., LI, Q.-B., YAN, B., AND ZHENG, K.-F. Encoding text information with graph convolutional networks for personality recognition. *Applied Sciences 10*, 12 (2020).

[68] WANG, Z., YE, C., AND ZHOU, H. Geolocation using gat with multiview learning. In *2020 IEEE International Conference on Smart Data Services (SMDS)* (2020), pp. 81–88.

[69] WU, L., CHEN, Y., SHEN, K., GUO, X., GAO, H., LI, S., PEI, J., AND LONG, B. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning 16*, 2 (2023), 119–328.

[70] XIE, W., MU, X., LEE, R. K.-W., ZHU, F., AND LIM, E.-P. Unsupervised user identity linkage via factoid embedding. In *2018 IEEE International Conference on Data Mining (ICDM)* (2018), IEEE, pp. 1338–1343.

[71] XU, K., HU, W., LESKOVEC, J., AND JEGELKA, S. How powerful are graph neural networks?, 2019.

[72] YAN, Y., ZHANG, S., AND TONG, H. Bright: A bridging algorithm for network alignment. In *Proceedings of the Web Conference 2021* (New York, NY, USA, 2021), WWW '21, Association for Computing Machinery, p. 3907–3917.

[73] YAO, L., MAO, C., AND LUO, Y. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 7370–7377.

[74] YING, R., HE, R., CHEN, K., EKSOMBATCHAI, P., HAMILTON, W. L., AND LESKOVEC, J. Graph convolutional neural networks for web-scale recommender

systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2018), KDD '18, Association for Computing Machinery, p. 974–983.

[75] YUAN, M., CHEN, L., AND YU, P. S. Personalized privacy protection in social networks. 141–150.

[76] YUN, S., JEONG, M., KIM, R., KANG, J., AND KIM, H. J. Graph transformer networks. *Advances in neural information processing systems 32* (2019).

[77] ZAFARANI, R., AND LIU, H. Connecting corresponding identities across communities. In *Proceedings of the International AAAI Conference on Web and Social Media* (2009), vol. 3, pp. 354–357.

[78] ZAFARANI, R., AND LIU, H. Connecting users across social media sites: A behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, Association for Computing Machinery, p. 41–49.

[79] ZHANG, J., CHEN, B., WANG, X., CHEN, H., LI, C., JIN, F., SONG, G., AND ZHANG, Y. Mego2vec: Embedding matched ego networks for user alignment across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2018), CIKM '18, Association for Computing Machinery, p. 327–336.

[80] ZHANG, J., AND YU, P. S. Pct: Partial co-alignment of social networks. In *Proceedings of the 25th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, 2016), WWW '16, International World Wide Web Conferences Steering Committee, p. 749–759.

[81] ZHANG, S., AND TONG, H. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), p. 1345–1354.

[82] ZHANG, Y., FAN, Y., SONG, W., HOU, S., YE, Y., LI, X., ZHAO, L., SHI, C., WANG, J., AND XIONG, Q. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference* (New York, NY, USA, 2019), WWW '19, Association for Computing Machinery, p. 3448–3454.

[83] ZHANG, Y., TANG, J., YANG, Z., PEI, J., AND YU, P. S. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015), pp. 1485–1494.

[84] ZHENG, R., LI, J., CHEN, H., AND HUANG, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology 57*, 3 (2006), 378–393.

[85] ZHONG, Z., CAO, Y., CAO, Y., GUO, M., GUO, M., NIE, Z., AND NIE, Z. Colink: An unsupervised framework for user identity linkage. *Proceedings of the AAAI Conference on Artificial Intelligence 32*, 1 (Apr. 2018).

[86] ZHOU, F., LIU, L., ZHANG, K., TRAJCEVSKI, G., WU, J., AND ZHONG, T. Deeplink: A deep learning approach for user identity linkage. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)* (2018), IEEE, pp. 1313–1321.

[87] ZHOU, X., LIANG, X., ZHANG, H., AND MA, Y. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE transactions on knowledge and data engineering 28*, 2 (2015), 411–424.

[88] ZHOU, X., LIANG, X., ZHANG, H., AND MA, Y. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Transactions on Knowledge and Data Engineering 28*, 2 (2016), 411–424.