The University of Texas at Austin
Center for Identity

Privacy Risk Predictions Based on Fundamental Understanding of Personal Data and an Evolving Threat Landscape

Haoran Niu
K. Suzanne Barber

UTCID Report #26-03

March 2026

Privacy Risk Predictions Based on Fundamental Understanding of Personal Data and an Evolving Threat Landscape

Haoran Niu[a] and K. Suzanne Barber[b]
The University of Texas at Austin
haoranniu@utexas.edu, sbarber@identity.utexas.edu

Keywords: Privacy Protection, Risk Prediction, Link Prediction Algorithms, Graph Neural Networks, Graph Convolutional Networks, Deep Learning, Identity Graph.

Abstract: It is difficult for individuals and organizations to protect personal information without a fundamental understanding of relative privacy risks. By analyzing over 5,000 empirical identity theft and fraud cases, this research identifies which types of personal data are exposed, how frequently such exposures occur, and what the consequences of those exposures are. We construct an Identity Ecosystem graph—a foundational, graph-based model in which nodes represent personally identifiable information (PII) attributes and edges represent empirical disclosure relationships between them (e.g., one PII attribute is exposed due to the exposure of another). Leveraging this graph structure, we develop a privacy risk prediction framework that uses graph theory and graph neural networks to estimate the likelihood of further disclosures when certain PII attributes are compromised. The results show that our approach effectively addresses the core question: Can the disclosure of a given identity attribute possibly lead to the disclosure of another attribute? The code for the privacy risk prediction framework is available at: https://github.com/niu-haoran/Privacy-Risk-Predictions-and-UTCID-Identity-Ecosystem.git.

1 INTRODUCTION

Different individuals and organizations have different sets of personally identifiable information (PII), and therefore hold distinct perspectives on which PII attributes are more vulnerable, more valuable, and in greater need of protection. An individual's PII includes personal data in four categories—What You Know (e.g., name, address, phone number, mother's maiden name), What You Have (e.g., driver's license, Social Security card, employee ID, passport), What You Are (e.g., fingerprint, voice, facial image), and What You Do (e.g., patterns of life such as websites visited, GPS location history, phone logs) (Zaeem et al., 2016).

Protecting PII data can be costly and time-consuming. Previous research has uncovered various strategies to reduce the risks of unintended data disclosure (Karr and Reiter, 2014), including statistical disclosure limitation (SDL) techniques commonly used by national statistical agencies before releasing public-use data sets. Meanwhile, research on data self-destruction focuses on protecting data privacy for users who choose cloud services (Zeng et al., 2010; Geambasu et al., 2009). Among the many existing privacy protection methodologies, this paper focuses on the first step of privacy protection: determining which set of data to protect. Protecting the most valuable and risky (i.e., likely to be exposed) set of PII promises to be a more effective and efficient method of protecting a person's personal, sensitive data. This is because individuals and institutions usually have limited time, energy, and financial resources allocated for privacy protection.

This research is based on the premise that if individuals and organizations have a more fundamental understanding and a more accurate evaluation of privacy risks resulting from disclosing PII, they will be in a better position to protect that information. Additionally, better risk analysis of personal data sharing will inform a wide range of information security and privacy applications.

Accordingly, this research determines which data require the most protection by evaluating the consequences of exposing specific PII attributes. Specifically, it analyzes the privacy risks incurred when an individual or an organization shares or loses a given

[a] https://orcid.org/0000-0003-4596-6127
[b] https://orcid.org/0000-0003-2906-6583

PII attribute. We primarily seek to answer the question: **Could the disclosure of a given PII attribute, such as date of birth, lead to the disclosure of another attribute, such as an ATM PIN**? PII attributes have associated risk values (Zaeem et al., 2016; Zaiss et al., 2019). By evaluating the risk scores of the potentially disclosed PII attributes, we provide a quantified prediction of privacy risks based on empirical data on disclosures.

To provide quantified predictions of privacy risks, we leverage graph theory. Many well-studied networks, such as transportation networks and social networks, are analyzed using graphs (Nippani et al., 2023; Ediger et al., 2010). Similar to these systems, meaningful connections and relationships exist among PII attributes. This research represents each PII attribute as a node in the graph. The directional edges capture the disclosure/exposure relationship between two PII attributes—specifically, the probability that the disclosure of one PII attribute (e.g., date of birth) could lead to the exposure of another PII attribute (e.g., address). We refer to this graph of PII nodes with directed disclosure or exposure relationships as the University of Texas Center for Identity (UTCID) Identity Ecosystem graph.

After constructing UTCID Identity Ecosystem graphs, we create and train three different link prediction models. We also develop a risk score calculation model. Together, the UTCID Identity Ecosystem graphs, the link prediction algorithms, and the risk score calculation model form a comprehensive risk prediction framework.

Suppose an individual has lost PII attributes A_1 and B_2 and wishes to determine which other PII attributes become highly risky as a result. The pipeline of the risk prediction framework is outlined below:

- The individual provides a risk score threshold (on a $[0, 100]$ scale) to indicate the level of risk of concern. The default risk score threshold is 0, meaning all potentially disclosed attributes are considered.
- The individual specifies the PII attributes A_1 and B_2 that were stolen or lost.
- The PII attribute set is defined as all nodes in the UTCID Identity Ecosystem graph associated with the individual, except for A_1 and B_2.
- The risk prediction model—composed of a link prediction module and a risk score calculation module—takes A_1, B_2, the risk score threshold, and the PII attribute set as input, and outputs the PII attributes that may be disclosed and have risk scores exceeding the specified threshold.

Figure 1 uses an example of risk score threshold 75 to explain the pipeline. In general, the contributions of this paper include:

- We propose a method for constructing UTCID Identity Ecosystem graphs and provide mechanisms to customize the graphs for different scales and individual needs. (Section 3).
- We create and train three different link prediction models (Section 4.2, 4.3, and 4.4).
- We conduct extensive evaluations on UTCID Identity Ecosystem graphs of different sizes to demonstrate the performance of the proposed link prediction models. (Section 5).
- We construct a risk score calculation model to quantify privacy risks associated with PII disclosures (Section 6).

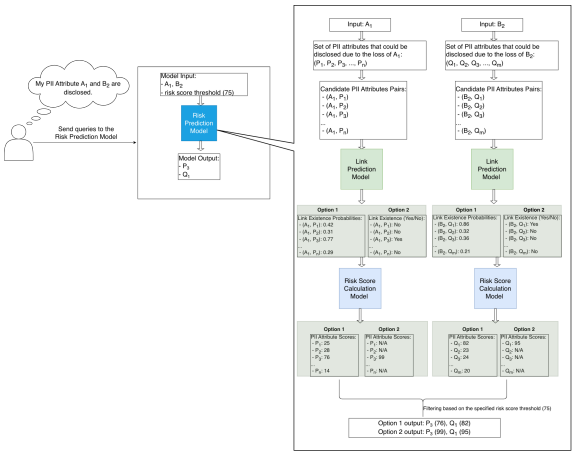

Figure 1: **Pipeline of the risk prediction framework**, illustrating the process from user query to PII attribute risk prediction results. User may choose different link prediction output formats depending on whether link prediction probabilities are heavily taken into account in the risk score calculation. Details of risk score calculation under different output options are provided in Section 6.

2 RELATED WORK

Privacy protection and risk assessment are important research topics for many market sectors that collect, store, and analyze personally identifiable information. Researchers have developed a variety of data mining techniques to find suspicious patterns in data and identity fraud transactions (Javaid, 2024). Applications of machine learning and data mining algorithms can help market sectors such as financial services, healthcare, transportation, and others mitigate financial losses and time costs.

For instance, the paper (Zhang and Guo, 2024) provides a method to assess privacy risks for medical big data. The researchers also apply the Fuzzy C-means clustering algorithm to cluster users into different groups, assign different permissions, and improve the access control accuracy.

Despite substantial research on privacy risk assessment, there is limited work on identifying the connections among different aspects of personal data. This paper not only fills this gap by uncovering interconnections between personal data attributes, but also employs graph-based models to represent and predict their interactions, thereby revealing the risks associated with data sharing and leakage.

In relation to the graph-based models utilized in this project, link prediction methods play an important role. Link prediction is commonly used to detect missing links or add future connections. There are some simple methodologies based on node similarity, such as Jaccard's coefficient (Murphy, 1996) and the Adamic/Adar measure (Adamic and Adar, 2003). These techniques are computationally efficient but do not perform well across many types of graphs. Therefore, an increasing number of researchers have turned to machine learning algorithms (Wang et al., 2007; Lichtenwalter et al., 2010; Al Hasan et al., 2006).

The framework of this paper combines both simple similarity-based properties/scores and supervised learning algorithms, enabling efficient prediction of PII risks.

3 UTCID IDENTITY ECOSYSTEM GRAPH CONSTRUCTION

We now explore the UTCID Identity Ecosystem—its content, its privacy risk analytics, and the new prediction capabilities introduced in this paper to help individuals protect their personal information from unintended and harmful disclosure. The Center for Identity at The University of Texas at Austin (UTCID) first introduced an Identity Ecosystem graph (Zaeem et al., 2016). The UTCID Identity Ecosystem graph represents PII attributes (i.e., personal data) as nodes and connects these nodes based on various types of relationships between PII attributes (Figure 2). While the work in (Zaeem et al., 2016) presents multiple types of relationships, this paper focuses specifically on the *probability of disclosure* relationship for privacy risk prediction.

Figure 2 illustrates the UTCID Identity Ecosystem graph, with nodes colored by Type (What You Know, What You Have, What You Are, What You Do) and sized by value. The value of a PII attribute node is calculated based on the degree to which that PII attribute was monetized in more than 5,000 identity theft and fraud cases analyzed in the UTCID Identity Threat Assessment and Prediction (ITAP) project (Zaeem et al., 2016; Zaiss et al., 2019). For simplicity, we refer to the collection of identity theft and fraud cases from the ITAP project as the "UTCID ITAP dataset".

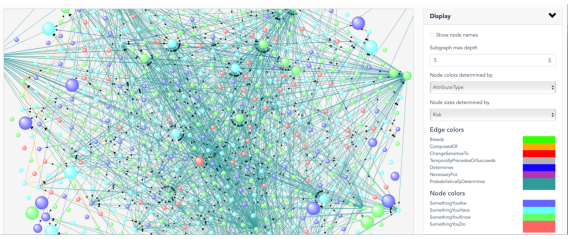

Figure 2: UTCID Identity Ecosystem Graph Representing PII Attributes and Their Relationships.

In this paper, we reconstruct the UTCID Identity Ecosystem graph and retain only a minimal set of graph features in order to predict PII disclosure risks efficiently. In real-world reports of identity theft and fraud cases, it is common to know only how attackers obtained access to certain PII attributes and which PII attributes were ultimately compromised. If a minimal set of information can be used to accurately predict PII disclosure risks, then there is no need to expend additional time and financial resources collecting other data. The construction of the new Identity Ecosystem graphs—referred to as UTCID Identity Ecosystem v2.0[1]—follows the principles outlined below.

We use nodes to represent PII attributes. A directed edge between PII attributes A and B, denoted as $A \rightarrow B$, indicates that an event disclosing attribute A leads to the disclosure of attribute B. Each edge is assigned a weight. Specifically, an edge $A \rightarrow B$ with weight w_i indicates that an event disclosing A may lead to the disclosure of B, and that such a disclosure occurred w_i times in the empirical UTCID ITAP dataset (Zaiss et al., 2019).

Figure 3 presents an example visualization based on a simplified use case, where edge thickness corresponds to the edge weight, representing the frequency of a specific disclosure relationship (i.e., $A \rightarrow B$ with weight w_i). This figure illustrates a simple UTCID Identity Ecosystem graph with three PII attributes. The weight of the edge "name $\rightarrow$ bank account" is

[1]UTCID Identity Ecosystem v2.0 offers a new means by which Identity Ecosystem graphs are constructed to promote accurate prediction of risks resulting from identity attribute exposure and disclosures. Throughout this paper, references to the UTCID Identity Ecosystem refer to UTCID Identity Ecosystem v2.0.

3, indicating three observed events in which disclosure of a name led to the disclosure of a bank account. Similarly, the weight of the edge "name → birth date" is 7, indicating seven observed events in which disclosure of a name led to the disclosure of a birth date. Assuming that these weights are derived from distinct events, the example graph illustrates the following:

- Disclosure of the "name" attribute may lead to the disclosure of a "bank account" with probability 0.3 (calculated as $3/(3+7) = 0.3$).
- Disclosure of the "name" attribute may lead to the disclosure of a "birth date" with probability 0.7 (calculated as $7/(3+7) = 0.7$).

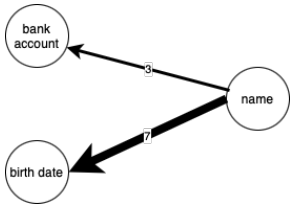

Figure 3: An Example Identity Ecosystem Graph with Three Nodes.

Throughout this paper, we use UTCID Identity Ecosystem graphs, constructed according to the rules described above, to analyze and predict PII attribute disclosures and, consequently, privacy risks.

To collect and structure the data necessary to perform the proposed risk prediction, two steps are required. First, we preprocess the UTCID ITAP dataset. For the inputs and outputs of the identity theft and fraud cases in the ITAP dataset, only identity-related information is retained. For each case, inputs are the data used by the actors to conduct the identity theft or fraud. The outputs include the consequent data that are acquired, stolen, or otherwise exposed by the conclusion of the identity theft or fraud actions.

The second step is to construct the UTCID Identity Ecosystem graphs. We previously introduced the graph construction rules using a small example graph (e.g., Figure 3, which illustrates a graph with three nodes and two edges). Here, we provide a more comprehensive explanation of these rules using multiple input and output identity attributes.

Suppose there is a reported identity theft and fraud case. From this case, three input PII attributes can be extracted: "bank account", "name", and "Social Security Number" (abbreviated as "SSN"). The corresponding output PII attributes are "credit card" and "debit card". If we construct a UTCID Identity Ecosystem graph based on this single case, five nodes and six directed edges are added to the graph. The five nodes are "bank account", "name", "SSN", "credit card", and "debit card". The six directed edges are listed below:

- From "bank account" to "credit card".
- From "bank account" to "debit card".
- From "name" to "credit card".
- From "name" to "debit card".
- From "SSN" to "credit card".
- From "SSN" to "debit card".

Because this graph is constructed from a single identity theft and fraud case, every edge in the graph has a weight 1, reflecting that each input–output attribute pair appears only once within this scope.

Now consider a more complex scenario involving three identity theft and fraud cases (Cases 1, 2 and 3). In Case 1, the input PII attributes are "bank account", "name", and "SSN", and the corresponding output attributes are "credit card" and "debit card". In Case 2, the input attributes are "bank account" and "SSN", and the output attributes are "birth date", "credit history", and "credit card". In Case 3, the input attribute is "SSN", and the output attribute is "bank account".

Across these three cases, there are seven unique identity attributes: "bank account", "name", "SSN", "credit card", "debit card", "birth date" and "credit history". Based on the three cases, the relationships among the seven attributes are listed below:

- bank account → credit card (weight: 2).
- bank account → debit card (weight: 1).
- bank account → birth date (weight: 1).
- bank account → credit history (weight: 1).
- name → credit card (weight: 1).
- name → debit card (weight: 1).
- SSN → credit card (weight: 2).
- SSN → debit card (weight: 1).
- SSN → birth date (weight: 1).
- SSN → credit history (weight: 1).
- SSN → bank account (weight: 1).

These weights reflect the occurrence frequencies of the input–output pairs among the identity theft and fraud cases. Figure 4 shows a visualization of the Identity Ecosystem graph constructed from the three-case example above.

The largest UTCID Identity Ecosystem graph we construct in this paper is based on 5,636 cases from the UTCID ITAP dataset. Following the procedures given above, the resulting graph contains 1,733 nodes and 19,483 edges. We denote this graph as G_{grand}.

Following the construction rules, UTCID Identity Ecosystem graphs can be constructed at different scales. For example, the UTCID ITAP dataset includes identity theft and fraud cases spanning a wide range of market sectors and victim demographics, involving different types of PII with varying values and losses. By filtering cases based on specific parameters (e.g., victim age, market sector, monetary loss amount), we can tailor the analysis to particular scenarios. As an illustration, filtering for cases with losses greater than $10,000 yields a smaller graph with 761 nodes and 6,413 edges. We denote this filtered graph as G_{big_loss}.

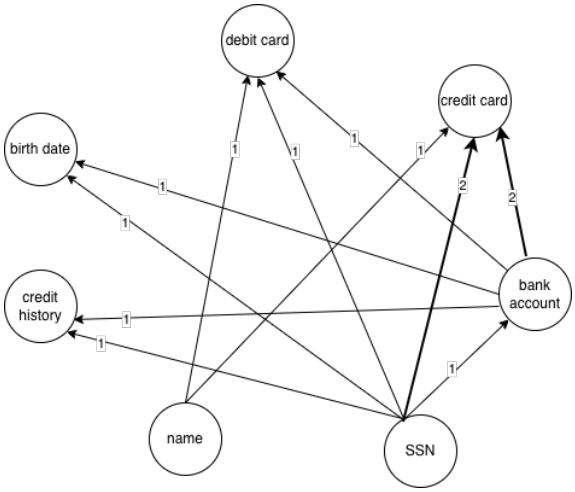

Figure 4: An Example of UTCID Identity Ecosystem Graph Construction Using Three Cases.

4 LINK PREDICTION ALGORITHMS

To answer a question such as **Could the disclosure of a given identity attribute possibly lead to the disclosure of another identity attribute of concern?**, we convert the question into a link prediction task (Kumar et al., 2020) and a risk score calculation process, as shown in Figure 1. For the link prediction task, we aim to determine whether a directed edge exists between an initial PII attribute (input) and a target PII attribute (output).

Suppose a UTCID Identity Ecosystem graph G_{TX} is constructed based on identity theft and fraud cases that occurred in Texas. If an individual living in Texas is involved in a breach and is informed that their driver's license number was disclosed, then, given knowledge of G_{TX}, the link prediction algorithms can help determine whether a possible link (i.e., a potential privacy risk) exists between the disclosed "driver's license number" and other PII attributes, such as a bank account or a credit card.

In general, link prediction algorithms can be used to examine whether possible directed links exist between initially exposed PII attributes and the remaining PII attributes in a UTCID Identity Ecosystem graph. An overview of link prediction algorithms is shown in Figure 10 (Kumar et al., 2020; Arrar et al., 2024) in Appendix 8.

Based on prior experimental studies conducted on ten benchmark homogeneous graphs (e.g., Ecoli (Salgado et al., 2001), FB15K (Newman, 2006)), GNN-based link prediction algorithms have demonstrated superior performance compared with similarity-based methods (Islam et al., 2020). Additional research shows that graph convolutional networks (GCNs) and Word2Vec combined with a multilayer perceptron (MLP) offer significant computational efficiency advantages over exponential random graph–based approaches while maintaining strong prediction performance (Sosa et al., 2024).

Moreover, because the nodes of a UTCID Identity Ecosystem graph represent PII attributes and carry background information related to individuals' identities, a major limitation of non-learning-based approaches is their difficulty in incorporating contextual information. Given the potentially large scale of UTCID Identity Ecosystem graphs and the importance of accurate privacy risk prediction, it is necessary to develop link prediction algorithms that are both efficient and accurate while accounting for graph structural complexity. Therefore, we narrow the link prediction algorithm choices to GNN-based and general deep-learning-based models, which are highlighted with red boxes in Figure 10. In the following subsections, we discuss the features of the UTCID Identity Ecosystem graphs that are used as inputs to the deep models. We also explain the three link prediction models that we propose.

Note that the link prediction models we propose have advantages and disadvantages. However, all three models demonstrate robust performance under different conditions.

4.1 Semantic Processing for PII Attribute Nodes

The nodes of UTCID Identity Ecosystem graphs correspond to PII attributes. As we discussed earlier, PII attributes are represented in natural language to describe identity- and privacy-related information. In other words, each PII attribute consists of one or more English words. In this paper, we focus exclusively on

English-language PII attributes.

Each English word has a corresponding definition that can be obtained from standard dictionaries. Accordingly, we define the *contextual information* of PII attributes as the collection of word-level definitions associated with the words composing that attribute. We refer to the process of transforming a PII attribute from its original surface form into its word-level semantic representation as *semantic processing*.

Our hypothesis is that, similar to inherent node properties of the UTCID Identity Ecosystem graphs (e.g., node degrees and node centrality), semantic features derived from PII attribute descriptions—such as token IDs obtained from word-level representations—provide meaningful node features. These semantic features can assist in predicting whether a link exists between two PII attribute nodes. We evaluate this hypothesis using the proposed model described in Section 4.4.

The semantic processing toolkit we use is the Natural Language Processing Toolkit (NLTK) (Bird et al., 2009). For each node in a UTCID Identity Ecosystem graph, the semantic processing procedure consists of the following steps:

- Extract the words composing the PII attribute.
- Process each word iteratively using the NLTK *synsets* and *definition* functions to obtain a textual explanation.
- Concatenate the explanations of all constituent words into a single paragraph representing the context information of the PII attribute.

Furthermore, we use the BERT uncased tokenizer (Devlin et al., 2019) to obtain token IDs from the context information associated with each node. We treat the resulting token IDs as weak contextual signals—deterministic numerical representations that reflect semantic features and capture contextual patterns to a limited degree. Since our focus is not on comparing tokenization techniques or optimizing semantic embeddings, we do not evaluate alternative tokenizers in this paper. Our objective is to demonstrate that incorporating weak semantic information can improve link prediction performance. We refer to the token IDs generated from context information using the BERT tokenizer as *semantic encoded signals* or *semantic embeddings*, noting that this usage differs from the conventional definition of semantic embeddings.

To illustrate the process of converting PII attributes written in English into semantic embeddings, we consider a concrete example: the PII attribute "employee credential" (Figure 5). This attribute consists of two words: "employee" and "credential". As described above, our semantic processing algorithm applies the NLTK *synsets* and *definition* functions to each word. Using this procedure, the context information for "employee" is "a worker who is hired to perform a job", and the context information for "credential" is "a document attesting to the truth of certain stated facts". Concatenating these two strings yields the context information for "employee credential": "a worker who is hired to perform a job a document attesting to the truth of certain stated facts". This text is then passed to the pretrained BERT uncased tokenizer (Devlin et al., 2019) to generate the corresponding token ID sequence. The token ID sequences associated with different PII attribute nodes may vary in length. In our experiments, the median semantic embedding length across all nodes in G_{grand} is 162. Accordingly, we truncate sequences longer than 162 and apply zero padding to sequences shorter than 162.

4.2 MLP-Based Model for Link Prediction – FeatureMLP

For each node in a directed graph, basic node properties such as in-degree and out-degree can be computed. These properties can be used to characterize node similarity. Since links are more likely to exist between similar nodes (Al Musawi et al., 2022), and because powerful yet simple classifiers such as multilayer perceptrons (MLPs) (Rumelhart et al., 1986) and support vector machines (SVMs) (Cortes and Vapnik, 1995) are effective in modeling nonlinear relationships, we construct an MLP-based link prediction model that uses basic node properties as features (Al Hasan et al., 2006). An additional motivation for using an MLP is to maintain architectural consistency among the deep learning models developed in this paper. We refer to this model as *FeatureMLP*.

The overall model structure is shown in Figure 6. For FeatureMLP, the node features (illustrated by the orange block in Figure 6) include:

- **in-degree**: the number of incoming links for each node;
- **out-degree**: the number of outgoing links for each node;
- **betweenness centrality** (Freeman, 1977): a measure of how often a node lies on shortest paths between other nodes;
- **closeness centrality** (Sabidussi, 1966): a measure of how close a node is to all other nodes in the graph.

FeatureMLP is computationally efficient and straightforward to implement, relying on low-level node statistics to perform link prediction. However, while these features capture global structural

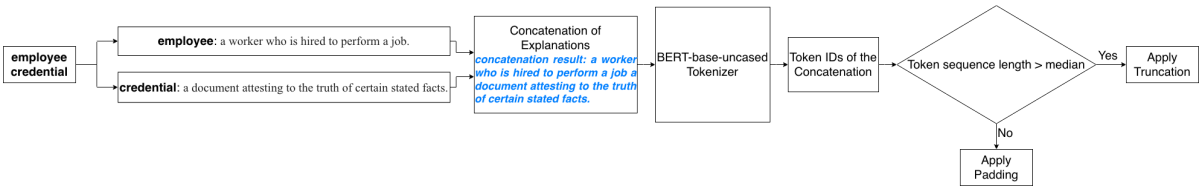

Figure 5: Process of Converting PII Attributes Expressed in English into Semantic Embeddings.

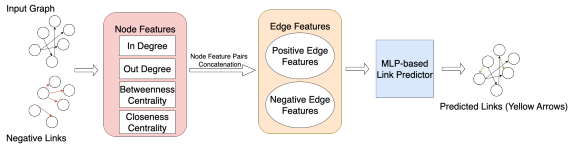

Figure 6: Overview of the FeatureMLP Model Architecture.

summaries, FeatureMLP does not explicitly model higher-order relational patterns or neighborhood-level interactions in the graph. To address scenarios that require capturing more complex structural dependencies and to maintain robust performance on large-scale graphs, we further develop additional link prediction models in Section 4.3 and 4.4.

4.3 GCN-Based Model for Link Prediction – FeatureGCN

Building on the first model (FeatureMLP), we next seek to leverage graph structural information. Therefore, we replace the MLP with a two-layer graph convolutional network (GCN). Note that the term GCN is used here in a broad sense, encompassing successors of the original GCN architecture, including GraphSAGE (Hamilton et al., 2017).

The model structure is shown in Figure 7. We use two SAGEConv layers (based on GraphSAGE), implemented with the PyTorch Geometric package (Fey and Lenssen, 2019), to generate node embeddings that capture local graph structural information. We then combine the node embeddings to form edge embeddings. For each edge embedding, the corresponding node embeddings are denoted as (ne_1, ne_2). We apply element-wise multiplication to aggregate the two node embeddings, resulting in an edge representation $ne_1 \odot ne_2$. This aggregation result is denoted as $A1$ in Figure 7.

FeatureGCN simultaneously leverages low-level node properties and graph structural information. However, as discussed earlier, the semantic information associated with PII attributes also contains valuable signals. Therefore, a more advanced GCN-based model is required to incorporate semantic information, which we introduce in the next section.

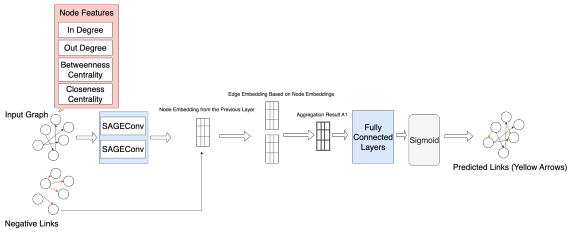

Figure 7: Overview of the FeatureGCN Model Architecture.

4.4 GCN with Semantic Embeddings for Link Prediction - SeeGCN

As discussed in Section 4.1, we obtain weak semantic embeddings—namely, the token ID sequences derived from the explanations of PII attributes—for each node in the input graphs. Building on the FeatureGCN model, we incorporate and leverage these semantic embeddings within the framework and demonstrate that they can improve the link prediction performance. The proposed model structure is shown in Figure 8, and we refer to this new model as SeeGCN.

For each pair of nodes (n_1, n_2) corresponding to an edge in the input graphs, we obtain a pair of node semantic embeddings (se_1, se_2). We refer to this pair of node semantic embeddings as an edge semantic embedding of the graph. We then aggregate se_1 and se_2 through concatenation followed by fully connected layers. Note that normalizing the concatenated semantic token ID sequence pair (se_1, se_2) leads to more robust performance across different graphs. The resulting semantic aggregation is denoted as $A2$ in Figure 8. The aggregation $A1$ in Figure 8 is identical to that defined in Figure 7. We concatenate $A1$ and $A2$ to generate the final aggregation result $A3$, as shown in Figure 8.

This link prediction model effectively integrates low-level node properties, graph structural information, and semantic information. Moreover, built from FeatureGCN, this model remains straightforward to implement.

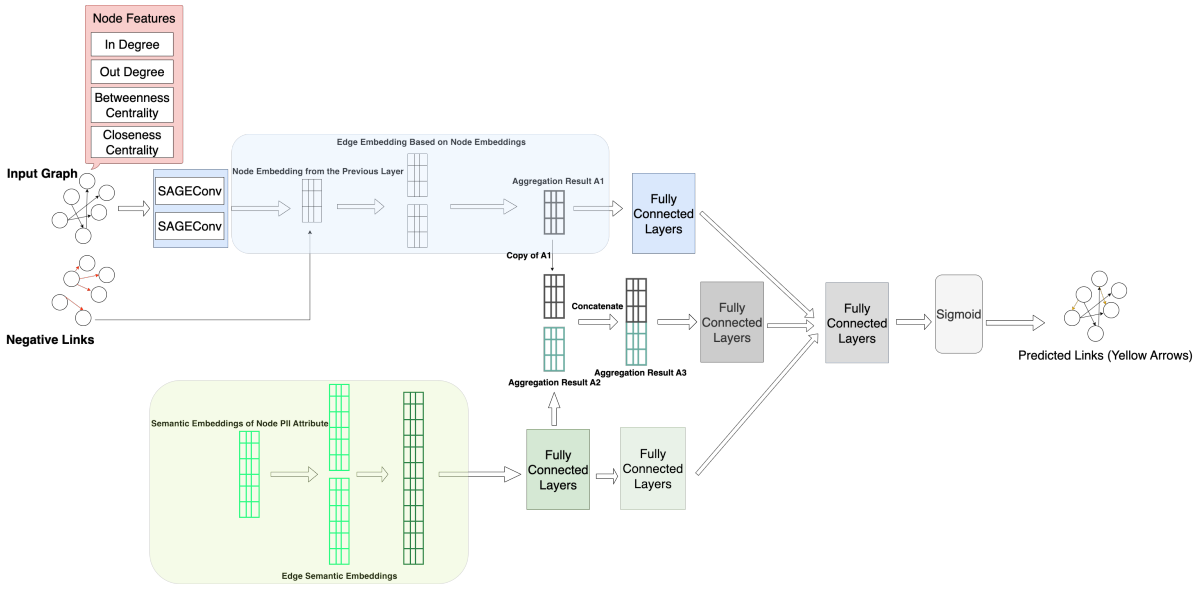

Figure 8: **Overview of the SeeGCN Model Architecture.** Normalizing the edge semantic embeddings before subsequent processing improves robustness across different graphs.

5 RESULT ANALYSIS

To compare the performance of the three proposed models, we evaluate them on the G_{grand} and G_{big_loss} graphs, as described in Section 3. We also randomly sample identity theft and fraud cases from the UTCID ITAP dataset and construct graphs based on these samples. We denote these graphs as $G_{(\#nodes,\#edges)}$, where *#nodes* represents the number of nodes and *#edges* represents the total number of edges in the graph. To examine model robustness across graphs of different sizes, the number of randomly sampled cases ranges from 500 to 5,000, with evenly spaced intervals. Graphs are constructed using the NetworkX (Hagberg et al., 2008) Python library and converted to PyTorch Geometric (PyG) data using the *from_networkX* function provided by PyG (Fey and Lenssen, 2019). The number of training epochs is set to 100 and the learning rate is set to 0.001, which is also the default value for the Adam optimizer. For each graph, edges are randomly split into training and validation sets using the *RandomLinkSplit* function from the *transforms* module of PyG, with a training-to-validation ratio of 9:1. The model performance results are reported in Table 1.

The table summarizes the performance of different models across the evaluated graphs. For each experiment, we report the best validation performance achieved over 100 training epochs in terms of ROC AUC (Hanley and McNeil, 1982) and accuracy of link existence prediction. For each graph, the highest AUC and accuracy values are highlighted in bold. Values below 0.7 are highlighted in red to indicate relatively poor results.

Across all experiments, at least two of the three proposed models achieve strong predictive performance, with AUC scores above 0.8. Although we designate scores below 0.7 as relatively poor, all reported performances remain above the random-guessing baseline of 0.5. These results indicate that all three proposed models demonstrate robust predictive capability. Notably, the SeeGCN model does not exhibit any results below the 0.7 threshold, suggesting strong robustness across both AUC and accuracy metrics. Overall, the findings show that link relationships in UTCID Identity Ecosystem graphs can be effectively predicted with the proposed models.

In this paper, UTCID Identity Ecosystem graphs are constructed from empirical cases in the UTCID ITAP dataset. The original dataset contains manual annotation errors and inconsistencies. While some of the errors are removed during preprocessing, the dataset is intentionally not heavily cleaned, resulting in graphs that contain noise. Despite this, the models achieve AUC scores as high as 0.95, and at least two of the three proposed models attain AUC values above 0.80 in every experiment. These results suggest that the proposed link prediction framework is robust to data noise. Moreover, they indicate that prediction performance could further improve when using cleaner data or when users construct customized UTCID Identity Ecosystem graphs from higher-quality empirical cases.

Table 1: Model Evaluation Results on Different Graphs; AUC—ROC AUC score, ACC—Accuracy.

graph	number of cases involved	FeatureMLP		FeatureGCN		SeeGCN	
		AUC	ACC	AUC	ACC	AUC	ACC
G_{grand}	5,636	**0.94**	0.84	0.88	0.83	0.93	**0.85**
G_{big_loss}	897	**0.92**	0.80	0.86	0.82	0.90	**0.83**
$G_{(611,3094)}$	500	**0.95**	**0.87**	0.89	0.84	0.91	0.83
$G_{(823,5135)}$	1,000	**0.94**	**0.87**	0.91	0.84	0.92	0.85
$G_{(1003,7221)}$	1,500	**0.93**	**0.85**	0.90	0.77	0.92	0.83
$G_{(1128,9095)}$	2,000	**0.93**	0.65	0.91	**0.85**	**0.93**	**0.85**
$G_{(1213,10647)}$	2,500	**0.94**	**0.83**	0.91	**0.83**	0.90	**0.83**
$G_{(1318,12279)}$	3,000	**0.94**	0.61	0.89	0.82	0.92	**0.84**
$G_{(1398,13689)}$	3,500	**0.94**	0.53	0.89	0.84	0.91	**0.85**
$G_{(1471,15018)}$	4,000	**0.94**	0.77	0.86	0.82	0.93	**0.86**
$G_{(1554,16413)}$	4,500	**0.93**	0.61	0.90	**0.85**	**0.93**	**0.85**
$G_{(1625,17694)}$	5,000	**0.94**	**0.86**	0.68	0.64	0.92	**0.86**

6 RISK SCORE CALCULATION

For the same PII attribute, people may assign different levels of importance. Some people may spend most of their time on social media, while others may only check social media occasionally. It is therefore reasonable that individuals in the former group would assign higher risk scores to usernames and passwords associated with their online accounts. After identifying potentially disclosed nodes based on the link prediction results from Section 4, risk scores can be assigned either manually according to individual preferences or automatically using quantitative evaluation metrics.

In this section, we propose a risk score calculation method. Suppose the query is to assess the risk scores of nodes related to the disclosure of a PII attribute α. Let the PII attribute set be defined as all nodes in a UTCID Identity Ecosystem graph except for α, denoted as $\{n_1, n_2, n_3, \ldots, n_m\}$, where m is the total number of attributes in the set.

We apply the PageRank algorithm (Page et al., 1999) to the graph of interest to obtain PageRank coefficients, denoted as pr_i for each node n_i. We also calculate the reverse PageRank coefficients by applying the reverse PageRank algorithm on the graph (Bar-Yossef and Mashiach, 2008), denoted as rpr_i for each node n_i. The score S_i for each node is defined as the sum of its forward and reverse PageRank coefficients:

$$S_i = pr_i + rpr_i. \tag{1}$$

Let $maxS$ denote the maximum value of S_i across all nodes. We normalize and scale S_i to the range $[0, 100]$ as follows:

$$S_i = \frac{S_i}{maxS} \times 100. \tag{2}$$

Next, for each node pair (α, n_i), the link prediction algorithm outputs either a link existence probability p_i or a binary indicator of link existence. If a user chooses to output the link existence probability, the risk score of node n_i under the disclosure of α is defined as:

$$RS_i = p_i \times S_i. \tag{3}$$

Let $maxRS$ denote the maximum value of RS_i across all nodes. To map the risk scores to the range $[0, 100]$ scale, we normalize RS_i as:

$$RS_i = \frac{RS_i}{maxRS} \times 100. \tag{4}$$

Alternatively, if the user chooses to consider only whether a link exists, the risk score of node n_i is set equal to S_i when the link prediction result indicates a positive link from α to n_i. Figure 9 illustrates the risk score calculation process under the two link prediction output options.

We do not evaluate which link prediction output option is superior, as the two options are intended to provide flexibility to accommodate different user preferences. The difference between the two approaches is illustrated as follows. Suppose the link prediction probability for (α, n_{10}) is 0.42, and the scaled PageRank-based score for n_{10} is $S_{10} = 97$. Assume the user wishes to protect the PII attributes with risk scores above 40.

When the link prediction probability is incorporated, the unscaled risk score is $97 \times 0.42 = 40.74$. Applying Equation 4 yields a final risk score of 81.48, assuming $maxRS = 50$ in this example. In this case,

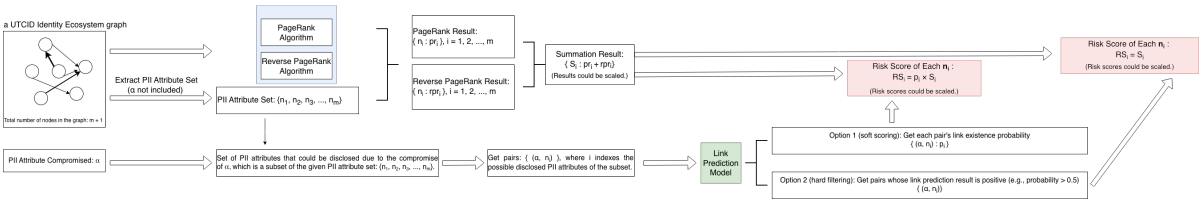

Figure 9: Process of Risk Score Calculation.

the user should allocate resources (e.g., money, time, or effort) to protect the PII attribute corresponding to n_{10}. In contrast, if the user considers only positive link prediction outcomes—using a probability threshold of 0.5 as an example in this paper—then n_{10} would not be selected for protection, since its link prediction probability is 0.42.

7 CONCLUSION

The goal of this paper is to analyze and predict privacy risks incurred when individuals share personal data. We introduce MLP-based and GCN-based algorithms—FeatureMLP, FeatureGCN, and SeeGCN—to answer the following question: **Can the disclosure of a given identity attribute possibly lead to the disclosure of another attribute**?

This work is motivated by the premise that individuals who better understand the risks associated with sharing specific personal data (identity attributes) are better equipped to make informed decisions and protect their privacy. Experimental evaluations on multiple UTCID Identity Ecosystem graphs demonstrate that the proposed framework can answer this question with strong predictive performance, robustness, and consistency. Furthermore, the UTCID Identity Ecosystem graphs, together with the proposed privacy risk prediction framework, offer flexibility and customization to support a wide range of practical use cases.

Future work will focus on identifying optimal GCN architectures for graphs of varying sizes, exploring integration between graph neural networks and reinforcement learning, and developing more advanced methods for incorporating semantic information from PII attributes into GCN-based link prediction models.

ACKNOWLEDGMENTS

The authors used ChatGPT and Google Search AI tools solely for grammar check, language polishing, and manuscript review. All research ideas, methods, analyses, and conclusions are entirely those of the authors.

REFERENCES

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*, volume 30, pages 798–805.

Al Musawi, A. F., Roy, S., and Ghosh, P. (2022). Identifying accurate link predictors based on assortativity of complex networks. *Scientific Reports*, 12(1):18107.

Arrar, D., Kamel, N., and Lakhfif, A. (2024). A comprehensive survey of link prediction methods. *The journal of supercomputing*, 80(3):3902–3942.

Bar-Yossef, Z. and Mashiach, L.-T. (2008). Local approximation of pagerank and reverse pagerank. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 279–288.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ediger, D., Jiang, K., Riedy, J., Bader, D. A., Corley, C., Farber, R., and Reynolds, W. N. (2010). Massive social network analysis: Mining twitter for social good. In *2010 39th international conference on parallel processing*, pages 583–593. IEEE.

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Geambasu, R., Kohno, T., Levy, A. A., and Levy, H. M. (2009). Vanish: Increasing data privacy with self-destructing data. In *USENIX security symposium*, volume 316.

Hagberg, A., Swart, P. J., and Schult, D. A. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos Na-

tional Laboratory (LANL), Los Alamos, NM (United States).

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Islam, M. K., Aridhi, S., and Smail-Tabbone, M. (2020). A comparative study of similarity-based and gnn-based link prediction approaches. *arXiv preprint arXiv:2008.08879*.

Javaid, H. A. (2024). Improving fraud detection and risk assessment in financial service using predictive analytics and data mining. *Integrated Journal of Science and Technology*, 1(8):1–11.

Karr, A. F. and Reiter, J. P. (2014). Using statistics to protect privacy. *Privacy Big Data, and the Public Good: Frameworks for Engagement*.

Kumar, A., Singh, S. S., Singh, K., and Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289.

Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252.

Murphy, A. H. (1996). The finley affair: A signal event in the history of forecast verification. *Weather and forecasting*, 11(1):3–20.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(3):036104.

Nippani, A., Li, D., Ju, H., Koutsopoulos, H., and Zhang, H. (2023). Graph neural networks for road safety modeling: Datasets and evaluations for accident analysis. *Advances in neural information processing systems*, 36:52009–52032.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millán-Zárate, D., Díaz-Peredo, E., Sánchez-Solano, F., Pérez-Rueda, E., Bonavides-Martínez, C., and Collado-Vides, J. (2001). Regulondb (version 3.2): transcriptional regulation and operon organization in escherichia coli k-12. *Nucleic acids research*, 29(1):72–74.

Sosa, J., Martínez, D., and Guerrero, N. (2024). An unified approach to link prediction in collaboration networks. *arXiv preprint arXiv:2411.01066*.

Wang, C., Satuluri, V., and Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 322–331. IEEE.

Zaeem, R. N., Budalakoti, S., Barber, K. S., Rasheed, M., and Bajaj, C. (2016). Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–8. IEEE.

Zaiss, J., Nokhbeh Zaeem, R., and Barber, K. S. (2019). Identity threat assessment and prediction. *Journal of Consumer Affairs*, 53(1):58–70.

Zeng, L., Shi, Z., Xu, S., and Feng, D. (2010). Safevanish: An improved data self-destruction for protecting data privacy. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pages 521–528. IEEE.

Zhang, X. and Guo, T. (2024). Privacy risk assessment of medical big data based on information entropy and fcm algorithm. *IEEE Access*.

8 APPENDIX - LINK PREDICTION METHODS OVERVIEW

Figure 10 shows an overview of link prediction algorithms, as discussed in Section 4. The red boxes highlight the focus of this paper.

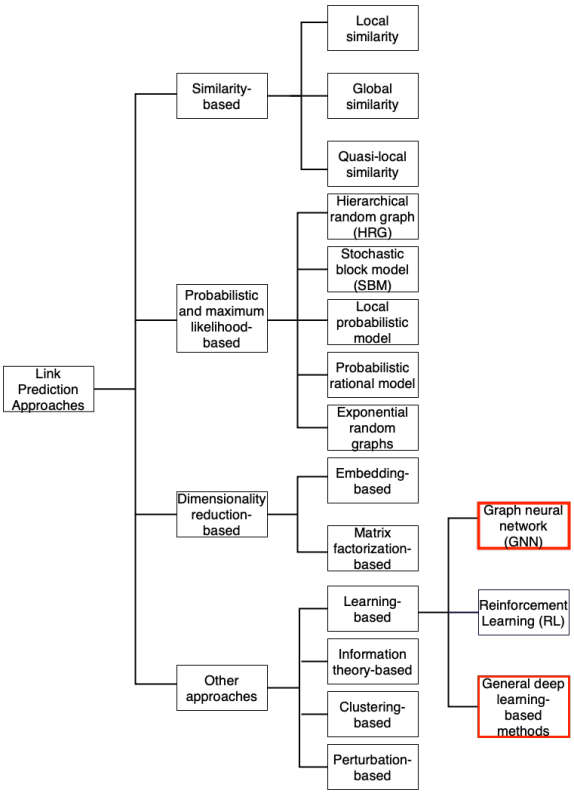

Figure 10: **Overview of Link Prediction Algorithms.** Reinforcement learning is shown as a representative learning-based algorithm in this discussion.

WWW.IDENTITY.UTEXAS.EDU