Privacy Preserving Link Prediction - PPLP

Kaleb Kim, Wiam Skarki, Carson Whitehouse,
Jonah Lorenzo, Kent Manion, Aidan Bugayong

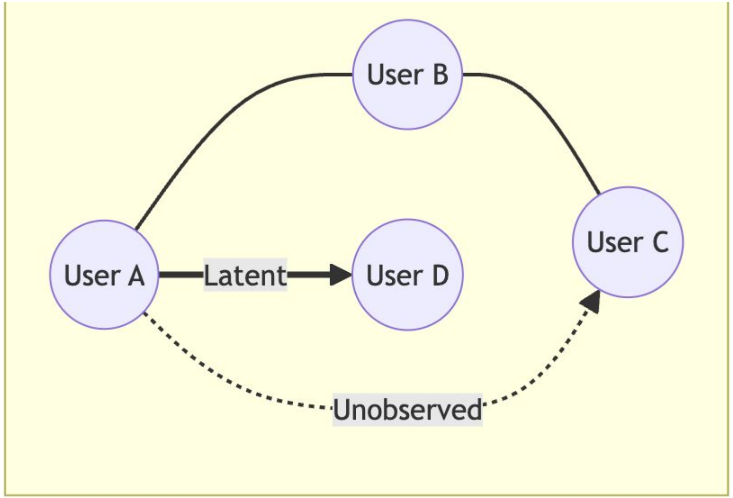
Our Goal

To build a Python library that allows anyone to apply **privacy preserving link prediction**.

Link Prediction

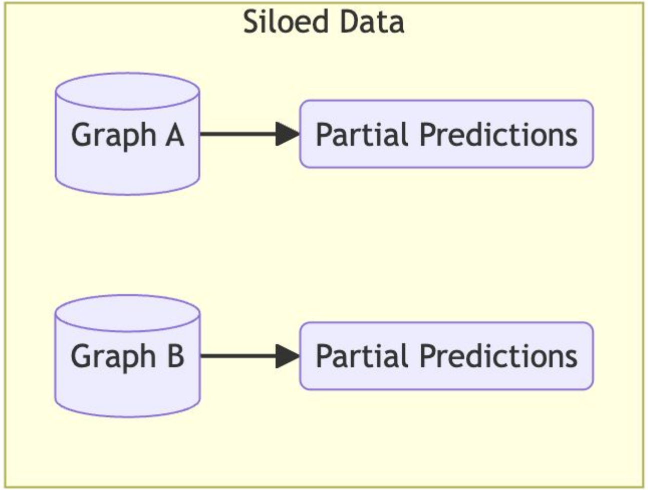
Link Prediction

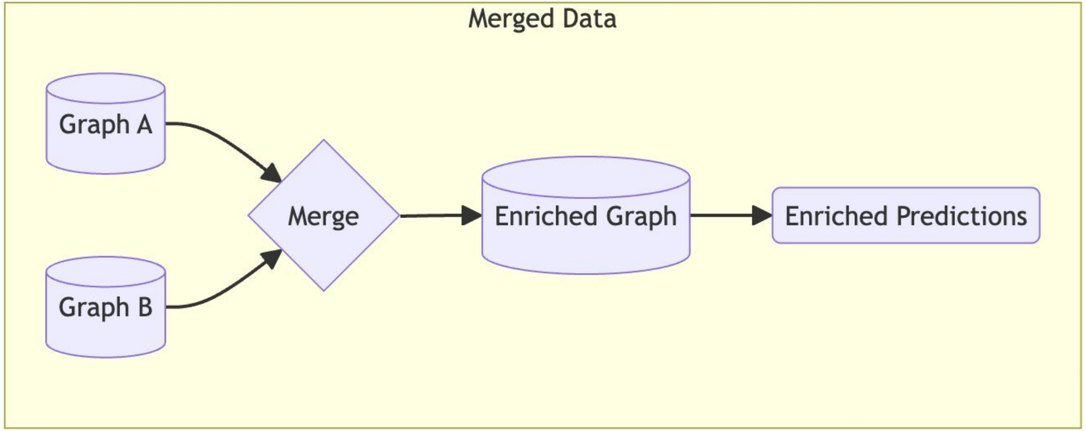
- Link prediction helps discover unobserved or latent connections between nodes in a graph
- Data holders rank the likelihood of new connections forming overtime
- Useful for social networks, e-commerce, telecomms, bioinformatics



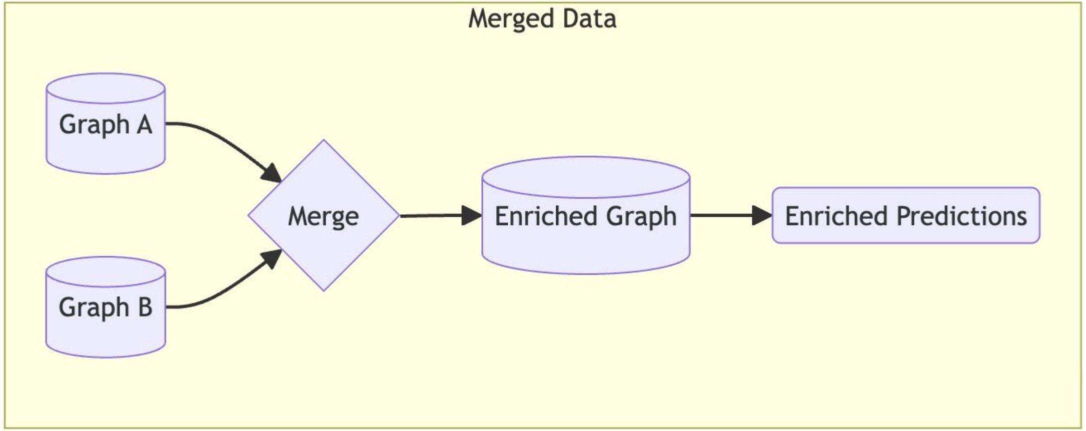
Link Prediction

- Link prediction is typically done on a single local graph
- Link prediction can be more accurate if we merge two or more graph databases that include similar information (Distributed Link Prediction)





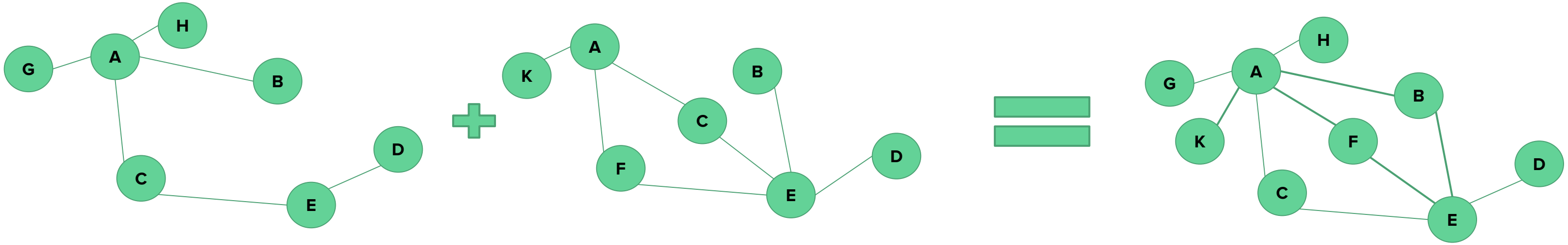
Distributed Link Prediction



- Two parties can utilize the connections in their combined graph to provide more accurate link predictions
- In other cases, distributed link prediction allows link prediction between nodes based on other attributes

Distributed Link Prediction

- In some cases, collaboration is mutually beneficial to contributing parties
- In other cases, the second party is paid to participate
- Distributed Link Prediction results in privacy concerns since it implies combining two or more different graph databases!
  - Identity Disclosure
  - Link Disclosure
  - Attribute Disclosure



Privacy Preserving Link Prediction

Privacy preserving link prediction allows multiple data holders to collaboratively forecast unobserved or latent connections between nodes **without explicitly revealing what each party knows.**

Use Cases

Social Networks:

- PPLP Used to understand the likelihood of a link between two nodes based on the similarity of these nodes in the different graphs. Here the nodes are people and the graphs represent a person's knowledge of the given social network.

Telecommunication

- Advertising company wants to use a telecom network or phone network for an advertisement. Using PPLP and a target audience x, they can find similar audience members for their advertisement based on x's connections.

E-Commerce:

- A product would be recommended to a user if they purchase similar products or similar users buy said product

Bioinformatics:

- A graph of EHRs linking patients to a disease can be privately shared with a clustering graph to show if a disease is being spread within clusters / predict if a patient could have the disease

Common Neighbors Measure

$$CN(A, B) = |A \cap B|$$

- Computes the number of common neighbors between two parties
- More neighbors = Higher link likelihood
- Fails to account for the relative amount of neighbors

Jaccard Measure

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Computation of the relative amount of neighbors in common
- This improves on the common neighbors measure

Adamic-Adar Measure

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

- N(u) is the set of nodes adjacent to u
- the sum of the log of the intersection of the neighbors of two nodes
- Two-hop similarity helps to yield better results

Homomorphic Encryption

Homomorphic encryption allows for direct operations on encrypted data without knowing the contents of the encrypted data. This allows us to do Private Set Intersection as shown below:

$$P(X) = r \cdot (X - x_0) \cdot \ldots \cdot (X - x_9).$$

X exists in Graph 1 and Graph 2 iff P(X) = 0

Where:

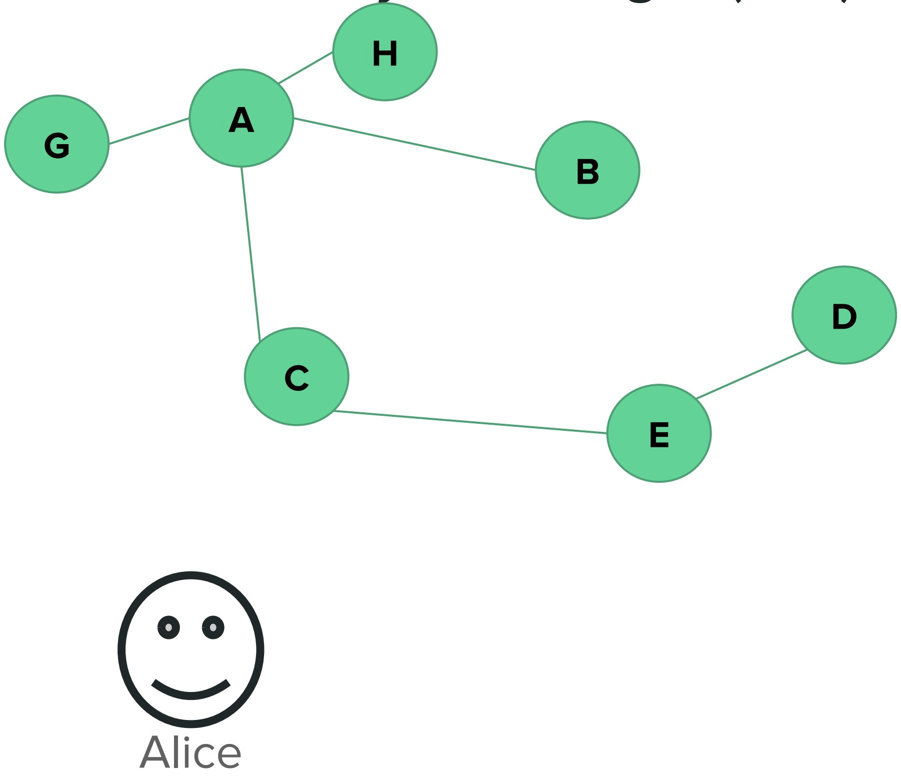
- X is a fully homomorphically encrypted node in Graph 1
- x_i is node i in Graph 2
- r is a random constant

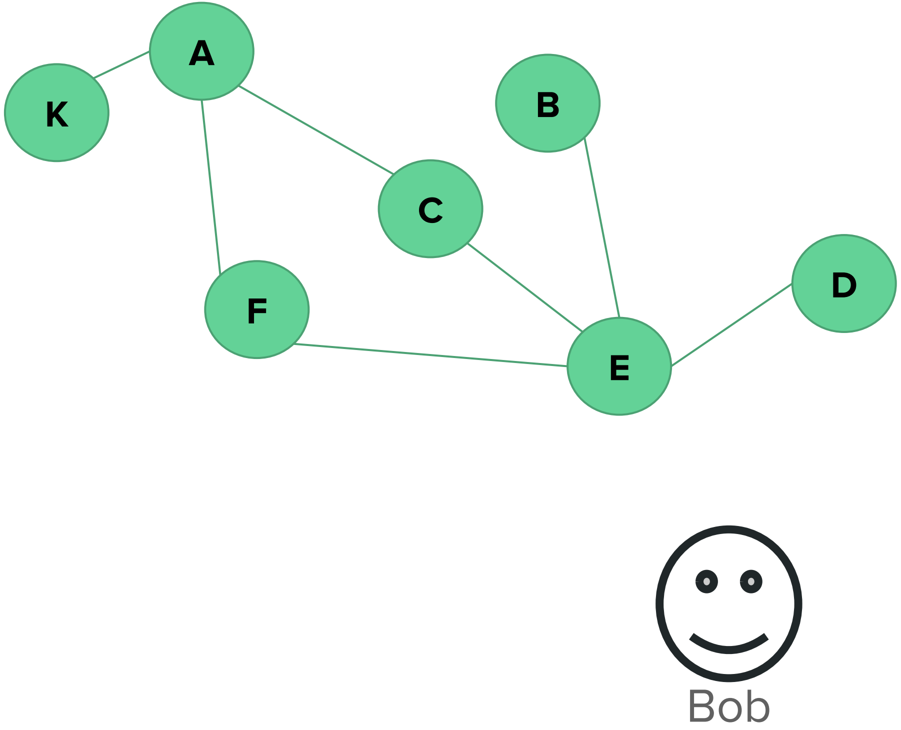
Private Set Intersection

- Private Set Intersection (PSI) protocols are the core of the privacy maintenance mechanisms used here
- There are many different algorithms that leverage both Oblivious Transfer (OT) and Fully Homomorphic Encryption (FHE), but OT is bottlenecked by communication complexity
- Lattice Based Cryptography from Craig Gentry in 2009 made FHE a practical possibility and it makes it possible to do PSI fast and with low communication overhead when optimized.

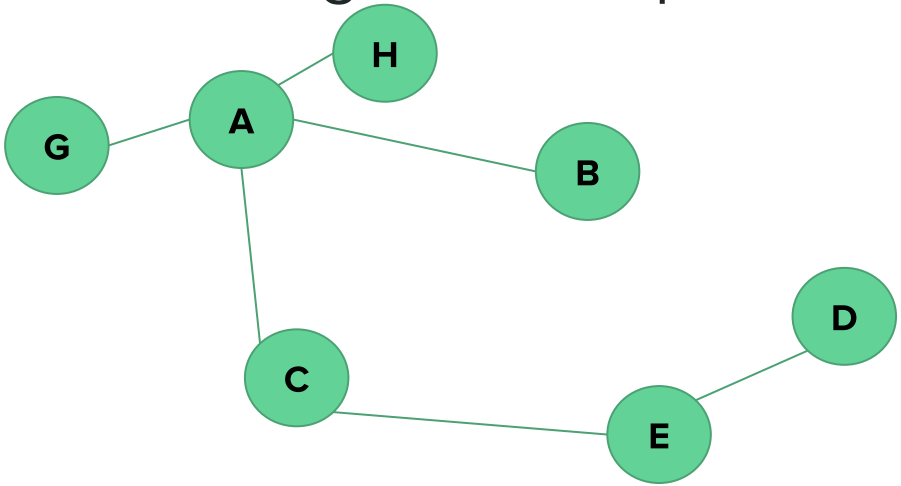
Higher Level Example

Subset only Looking N(A,E)





Searching Local Graph



Alice

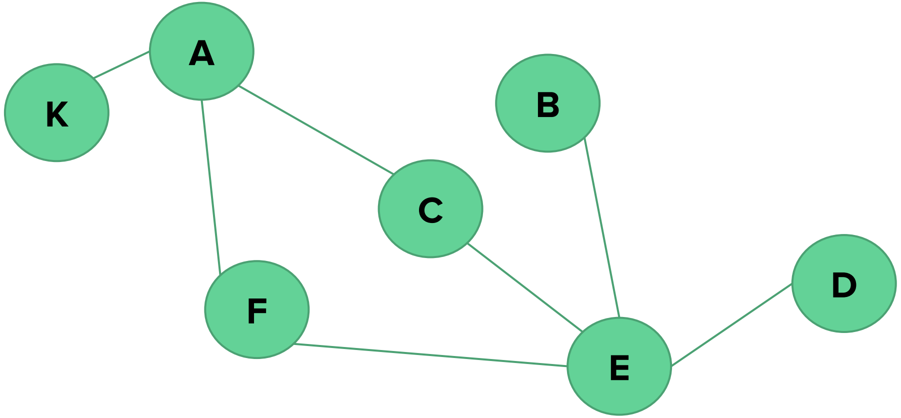
$N_A(A)=\{H,G,B,C\}$

$N_A(E)=\{D,C\}$

$N_A(A) \cap N_A(E) = I_A(A,E) = \{C\}$

$N_A(A) - I_A(A,E) =\{H,G,B\}$

$N_A(E) - I_A(A,E) =\{D\}$



$N_B(A)=\{K,F,C\}$

$N_B(E)=\{B,C,F,D\}$

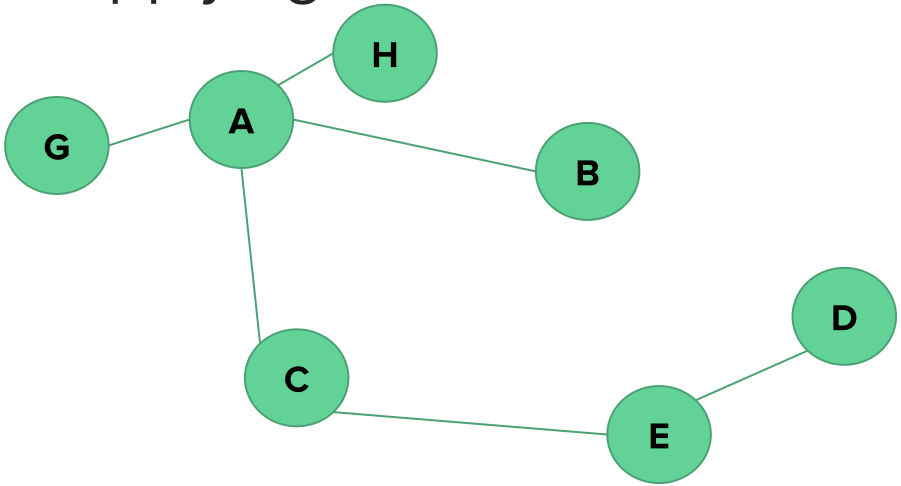
$N_B(A) \cap N_B(E) = I_B(A,E) = \{C,F\}$

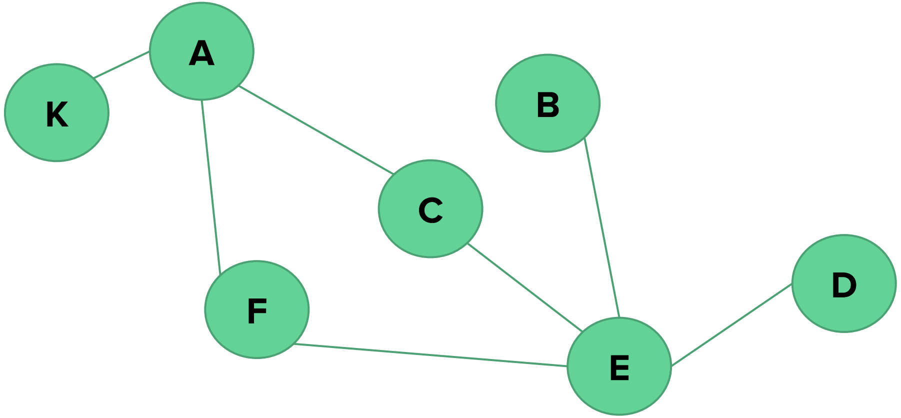
$N_B(E) - I_B(A,E) =\{B,D\}$

$N_B(A) - I_B(A,E) =\{K\}$

Bob

Applying PSI





$N_A(A) \cap N_A(E) = I_A(A,E) = \{C\}$

$N_A(A\text{-}E)=\{H,G,B\}$

$N_A(E\text{-}A)=\{D\}$

Alice

$N_B(A) \cap N_B(E) = I_B(A,E) = \{C,F\}$

$N_B(E\text{-}A)=\{B,D\}$

$N_B(A\text{-}E) =\{K\}$

Bob

PSI

$I_A \cap I_B = I_{A+B} = \{C\}$

$N_A(A\text{-}E) \cap N_B(E\text{-}A)=\{B\}$

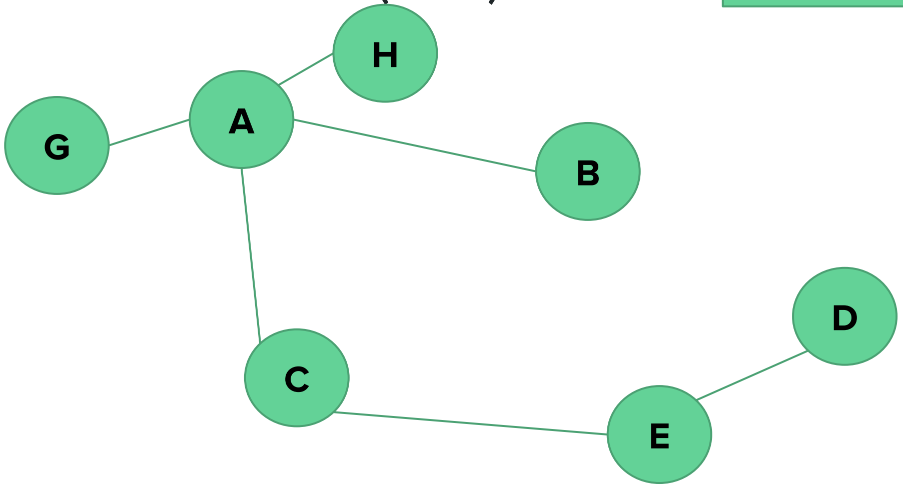
$N_A(E\text{-}A) \cap N_B(A\text{-}E)=\{\phi\}$

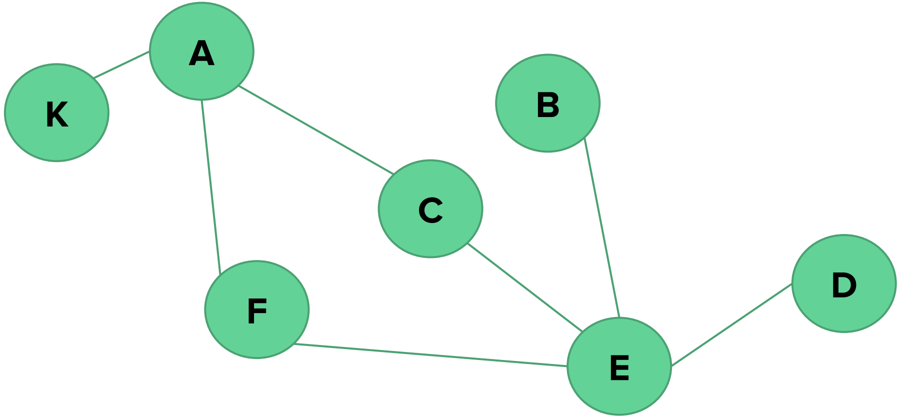
Size of CN(A,E)

|CN(A,E)| =

$$|I_A| + |I_B| - |I_{A+B}|+|\{B\}|+|\{\phi\}|$$

= 3





$N_A(A)=\{H,G,B,C\}$
$N_A(E)=\{D,C\}$
$I_A(A,E) = \{C\}$

$N_B(A)=\{K,F,C\}$
$N_B(E)=\{B,C,F,D\}$
$I_B(A,E) = \{C,F\}$

Alice

Bob

PSI
$I_A \cap I_B = I_{A+B} = \{C\}$
$N_A(A\text{-}E) \cap N_B(E\text{-}A) = \{B\}$
$N_A(E\text{-}A) \cap N_B(A\text{-}E) = \{\phi\}$

Previous Work Done

Liben-Nowell & Kleinberg (2004): The Link Prediction Problem

- Defines link prediction as a formal problem:
- Given a graph snapshot at time t, predict which edges will appear by time t' using only network topology and no node attributes.

Ayday et al. (2022): Privacy-Preserving Link Prediction

- Two parties (Graph 1, Graph 2) compute Common Neighbours across their joint graphs without revealing their graph to each other (link prediction)
- CN = local1 + local2 + crossover1 + crossover2 - overlap computed using 3 PSI calls
- Leaks some intermediate values and proposes a heavier homomorphic implementation for zero leakage

Chen, Laine & Rindal (2017): Fast PSI from Homomorphic Encryption

- A high-performance PSI protocol using Fully Homomorphic Encryption (FHE)
- Communication complexity is O(N_small · log N_large)
- While FHE makes it asymptotically efficient, it is not efficient in practice and still leaks minor information

Ling et. al (2025): Ultra-Fast Private Set Intersection From Efficient Oblivious Key-Value Stores

- An ultra-fast PSI protocol built upon a novel, bucket-based Oblivious Key-Value Store (OKVS) and Vector Oblivious Linear Evaluation (VOLE)
- Communication complexity is O(n)
- Minimizes network communication overhead compared to previous leading protocols: 30% → 1% redundancy
- Used for our implementation

Project Details

We’ll build a Python library that allows anyone to apply **privacy preserving link prediction**.

Implementation

This slide should highlight the "hybrid" nature of your library—prioritizing performance through C++ while maintaining accessibility through Python.

- **Core Backend (High Performance):**
    - **C++ Engine:** Leverages the **Ultra-Fast PSI protocol** from Ling et al. (2025) which utilizes **Vector Oblivious Linear Evaluation (VOLE)** and **Oblivious Key-Value Stores (OKVS)** to achieve $O(n)$ communication complexity.
    - **Cryptographic Primitives:** Optimized implementations of **Fully Homomorphic Encryption (FHE)** and **Lattice-Based Cryptography** to minimize communication overhead.
- **The Bridge (Python Bindings):**
    - **Pybind11 / Cython:** Bridges the C++ backend to a user-friendly Python interface, allowing data scientists to run complex PPLP tasks without

Use Cases I

- Social Networks
  - **Privacy-First Recommendations**
  - Discovering latent connections (friend suggestions) between users based on structural similarity across different social graphs without exposing the full contact list of either party.
- Telecommunications
  - **Targeted Advertising**
  - An advertiser finds target audience members similar to an existing set (x) by performing link prediction on phone networks while keeping the actual call/link records private.

Use Cases II

- **E-Commerce**
  - Collaborative Filtering
  - Recommending products to a user based on similarities found between their purchasing graph and those of other "similar" users, without a central server seeing individual transaction histories.
- **Bioinformatics**
  - **Secure EHR Research**
  - Merging Electronic Health Records (EHR) with clustering graphs to predict disease spread or patient diagnoses without revealing sensitive protected health information (PHI).

ref.

- Privacy Preserving Link Prediction
- Fast Private Set Intersection from Homomorphic Encryption (PDF)
- Feather: Lightweight Multi-party Updatable Delegated Private Set Intersection
- The Link Prediction Problem for Social Networks
- Lucidchart

https://github.com/ShallMate/fastpsi – Ultra-Fast Private Set Intersection (PDF)

What slides do we need?

- Introduction / Overview
- Project specification
    - Previous Work Done
    - PPLP Building Blocks
        - HE, PSI, LP Algos/Statistics
        - Adamic-Adar and Jaccard
    - Topic and Research
    - Auxiliary / Supporting Resources
    - End Goal (Python Library)
    - Practical applications / use cases (What we will implement 3 applications(?))
        - What are they
        - How do they implement this topic
        - Why is it useful for this specifically / what benefit does this provide

Pro Max Ultra Lightning Fast Privacy Preserving Link Prediction for Iphone 2000 pro max