Intersection Attack in Non-Uniform Setting

Dongchen Zou under the instruction of Simon Langowski

Oct 15 2023

Introduction

- We all use social media to talk to people.

Introduction

- We all use social media to talk to people.
- Privacy.

Introduction

- We all use social media to talk to people.
- Privacy.
- Messages themselves? They are normally encrypted and hard to obtain.

Introduction

- We all use social media to talk to people.
- Privacy.
- Messages themselves? They are normally encrypted and hard to obtain.
- Activity pattern like logging on and off?

Introduction

- We all use social media to talk to people.
- Privacy.
- Messages themselves? They are normally encrypted and hard to obtain.
- Activity pattern like logging on and off?
- In this talk, we will explore how such information can be used to learn user connections.

User Behavior

How do users of social media behave?

User Behavior

How do users of social media behave? They behave differently and non-uniformly, so we can't treat them all the same.

User Behavior

How do users of social media behave? They behave differently and non-uniformly, so we can't treat them all the same.

- People tend to talk based on the number of common interests.

User Behavior

How do users of social media behave? They behave differently and non-uniformly, so we can't treat them all the same.

- People tend to talk based on the number of common interests.
- If talking previously, it is more likely for them to talk later.

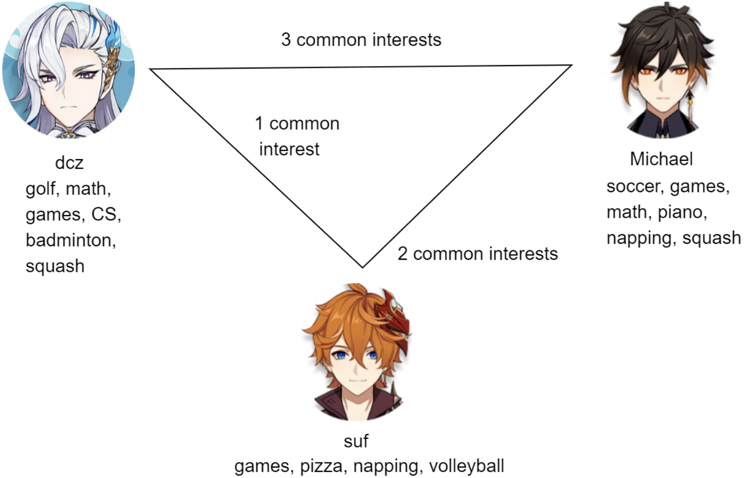
An Example of Social Media

dcz
golf, math,
games, CS,
badminton,
squash

Michael
soccer, games,
math, piano,
napping, squash

suf
games, pizza, napping, volleyball

An Example of Social Media



Eavesdropper's Observation

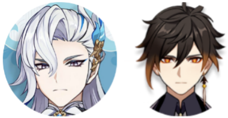
If someone is online, they are talking to someone (could be multiple)

Eavesdropper's Observation

If someone is online, they are talking to someone (could be multiple)
The eavesdropper gets to see all the people who are online in a period of time called an epoch.

Eavesdropper's Observation

Epoch 1



Epoch 2

Epoch 3

Intersection Attack in Non-Uniform Setting

Intersection attack (also known as statistical disclosure attacks) use such information to reconstruct relationships.

Intersection Attack in Non-Uniform Setting

Intersection attack (also known as statistical disclosure attacks) use such information to reconstruct relationships.

- Graph

Intersection Attack in Non-Uniform Setting

Intersection attack (also known as statistical disclosure attacks) use such information to reconstruct relationships.

- Graph
- Observation (epochs)

Intersection Attack in Non-Uniform Setting

Intersection attack (also known as statistical disclosure attacks) use such information to reconstruct relationships.

- Graph
- Observation (epochs)
- Attack

Intersection Attack in Non-Uniform Setting

Intersection attack (also known as statistical disclosure attacks) use such information to reconstruct relationships.

- Graph
- Observation (epochs)
- Attack
- Results

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.
How are we going to reconstruct the graph example?

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.
How are we going to reconstruct the graph example?

- Every user is assigned a probability p_i.

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.
How are we going to reconstruct the graph example?

- Every user is assigned a probability p_i.
- For each possible interest (integer from 0 to 99), we do a coin flip with probability p_i which decides whether the user will have that interest or not.

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.
How are we going to reconstruct the graph example?

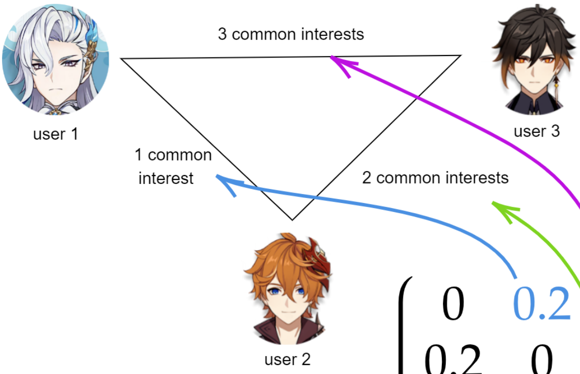
- Every user is assigned a probability p_i.
- For each possible interest (integer from 0 to 99), we do a coin flip with probability p_i which decides whether the user will have that interest or not.
- For each pair $[i, j]$, the larger the intersection, the more probable it is that they talk.

Graph Generation (Clustered?)

Previous papers all worked on a uniform graph.
How are we going to reconstruct the graph example?

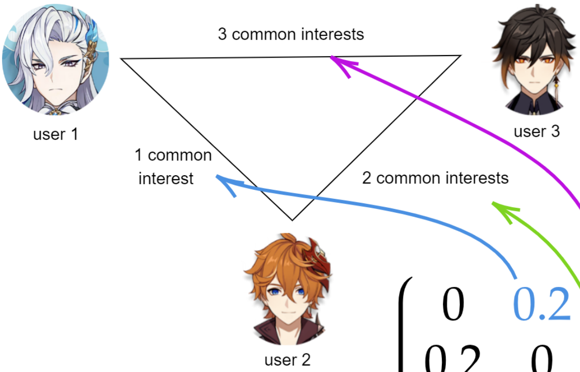
- Every user is assigned a probability p_i.
- For each possible interest (integer from 0 to 99), we do a coin flip with probability p_i which decides whether the user will have that interest or not.
- For each pair $[\mathrm{i}, \mathrm{j}]$, the larger the intersection, the more probable it is that they talk.
- We will denote probability that i and j talk in an epoch with $\mathrm{A}[\mathrm{i}, \mathrm{j}]$

Probability Matrix A for the Example Graph



$$\begin{pmatrix} 0 & 0.2 & 0.7 \\ 0.2 & 0 & 0.5 \\ 0.7 & 0.5 & 0 \end{pmatrix}$$

Probability Matrix A for the Example Graph



$$\begin{pmatrix} 0 & 0.2 & 0.7 \\ 0.2 & 0 & 0.5 \\ 0.7 & 0.5 & 0 \end{pmatrix}$$

The eavesdropper is trying to figure out the probability matrix.

Epoch Generation (Correlated?)

Previous papers simply did a coin flip on $A[i, j]$ for all $[i, j]$.

Epoch Generation (Correlated?)

Previous papers simply did a coin flip on $A[i, j]$ for all $[i, j]$.

- If *i* and *j* talked in the previous epoch, it is more likely for them to keep talking in this epoch.

Epoch Generation (Correlated?)

Previous papers simply did a coin flip on $\mathrm{A}[\mathrm{i},\mathrm{j}]$ for all $[\mathrm{i},\mathrm{j}]$.

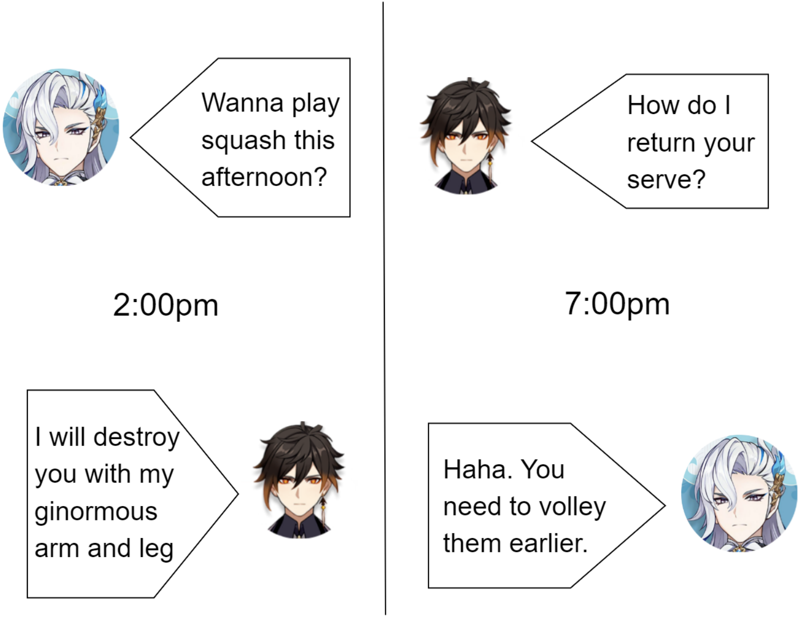
- If *i* and *j* talked in the previous epoch, it is more likely for them to keep talking in this epoch.
- How about $\mathrm{A}[\mathrm{i},\mathrm{j}] + \delta$ if *i* and *j* talked in last epoch?

Epoch Generation (Correlated?)

Previous papers simply did a coin flip on $A[i, j]$ for all $[i, j]$.

- If i and j talked in the previous epoch, it is more likely for them to keep talking in this epoch.
- How about $A[i, j] + \delta$ if i and j talked in last epoch?
- This change is apparently temporary and will vanish once $A[i, j] + \delta$ flips to tail

Correlation for the Example Graph



Some Math to Consider

- Given A, what is probability that user i and j appear online at the same time?

Some Math to Consider

- Given A, what is probability that user *i* and *j* appear online at the same time?
- $$\begin{cases} \text{Edge } [\mathrm{i}, \mathrm{j}] \text{ is active} \\ \text{They each have an edge active other than } [\mathrm{i}, \mathrm{j}] \end{cases}$$

Some Math to Consider

- Given A, what is probability that user *i* and *j* appear online at the same time?
-

$$\begin{cases} \text{Edge } [\mathrm{i}, \mathrm{j}] \text{ is active} \\ \text{They each have an edge active other than } [\mathrm{i}, \mathrm{j}] \end{cases}$$

- User *i* having at least one conversation with someone other than *j*

Some Math to Consider

- Given A, what is probability that user i and j appear online at the same time?
-

$$\begin{cases} \text{Edge } [\mathrm{i}, \mathrm{j}] \text{ is active} \\ \text{They each have an edge active other than } [\mathrm{i}, \mathrm{j}] \end{cases}$$

- User i having at least one conversation with someone other than j
- $g(i, j) = 1 - \prod_{k \notin j}(1 - \mathrm{A}[i, k])$

Some Math to Consider

- Given A, what is probability that user i and j appear online at the same time?
- $$\begin{cases} \text{Edge [i, j] is active} \\ \text{They each have an edge active other than [i, j]} \end{cases}$$
- User i having at least one conversation with someone other than j
- $g(i,j) = 1 - \prod_{k \notin j}(1 - \mathrm{A}[i,k])$
- $$\mathrm{F}_{[\mathrm{i,j}]}(A) = \mathrm{A}[\mathrm{i,j}] + (1 - \mathrm{A}[\mathrm{i,j}]) \cdot g(i,j) \cdot g(j,i)$$

What Does the Eavesdropper See

- We just calculated the theoretical probability of i and j appearing online together using A.

What Does the Eavesdropper See

- We just calculated the theoretical probability of i and j appearing online together using A.
- What is this probability, call it $C_{[i,j]}$, as observed by the eavesdropper?

What Does the Eavesdropper See

- We just calculated the theoretical probability of i and j appearing online together using A.
- What is this probability, call it $C_{[\mathrm{i,j}]}$, as observed by the eavesdropper?
- Number of times divided by total number of epochs!

Attack

Now we want to consider the entire graph. Let $\mathrm{F}(\mathrm{A}) = \sum \mathrm{F}_{[\mathrm{i},\mathrm{j}]}(\mathrm{A})$ and $C = \sum C_{[\mathrm{i},\mathrm{j}]}$

Attack

Now we want to consider the entire graph. Let $\mathsf{F}(\mathsf{A}) = \sum \mathsf{F}_{[i,j]}(\mathsf{A})$ and $C = \sum C_{[i,j]}$

- The eavesdropper tries to find a A' for which $\mathsf{F}(\mathsf{A}')$ is the closet to C. In a sense, $\mathsf{F}(\mathsf{A}') = C$.

Attack

Now we want to consider the entire graph. Let
$\mathsf{F}(\mathsf{A}) = \sum \mathsf{F}_{[i,j]}(\mathsf{A})$ and $C = \sum C_{[i,j]}$

- The eavesdropper tries to find a A' for which $\mathsf{F}(\mathsf{A}')$ is the closet to C. In a sense, $\mathsf{F}(\mathsf{A}') = C$.
- In this way, the guess A' matches the observation the most.

Attack

Now we want to consider the entire graph. Let $\mathsf{F}(\mathsf{A}) = \sum \mathsf{F}_{[i,j]}(\mathsf{A})$ and $C = \sum C_{[i,j]}$

- The eavesdropper tries to find a A' for which $\mathsf{F}(\mathsf{A}')$ is the closet to C. In a sense, $\mathsf{F}(\mathsf{A}') = C$.
- In this way, the guess A' matches the observation the most.
- Why does it work?

Attack

Now we want to consider the entire graph. Let
$\mathrm{F}(\mathrm{A}) = \sum \mathrm{F}_{[\mathrm{i,j}]}(\mathrm{A})$ and $C = \sum C_{[\mathrm{i,j}]}$

- The eavesdropper tries to find a A' for which $\mathrm{F}(\mathrm{A}')$ is the closet to C. In a sense, $\mathrm{F}(\mathrm{A}') = C$.
- In this way, the guess A' matches the observation the most.
- Why does it work?
- $$\lim_{t \to \infty} C_t = \mathrm{F}(\mathrm{A})$$

How Fast?

- Let $\mathbf{C}$ be a random variable representing the observations with $\mathbf{C}_i$ the sample for epoch i.

How Fast?

- Let $\mathbf{C}$ be a random variable representing the observations with $\mathbf{C}_i$ the sample for epoch i.
- Then $\{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_t\}$ are independent observations of an unknown distribution $\mathsf{Dist}(\mu = \mathsf{F}(\mathsf{A}), \sigma^2)$

How Fast?

- Let $\mathbf{C}$ be a random variable representing the observations with $\mathbf{C}_i$ the sample for epoch i.
- Then $\{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_t\}$ are independent observations of an unknown distribution $\text{Dist}(\mu = \text{F(A)}, \sigma^2)$
- We want to look at the sample average which is

$$\left(\bar{\mathbf{C}} = \frac{\mathbf{C}_1 + \mathbf{C}_2 + \ldots + \mathbf{C}_t}{t}\right) \rightarrow \text{F(A) when } t \rightarrow \infty$$

How Fast?

- Let $\mathbf{C}$ be a random variable representing the observations with $\mathbf{C}_i$ the sample for epoch i.
- Then $\{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_t\}$ are independent observations of an unknown distribution $\text{Dist}(\mu = \text{F(A)}, \sigma^2)$
- We want to look at the sample average which is

$$\left(\bar{\mathbf{C}} = \frac{\mathbf{C}_1 + \mathbf{C}_2 + \ldots + \mathbf{C}_t}{t}\right) \to \text{F(A) when } t \to \infty$$

- We are interested in the rate of convergence.

Central Limit Theorem

Central Limit theorem states that

$$(\bar{\mathbf{C}} - \mathsf{F}(\mathsf{A})) \sim \frac{\mathcal{N}(0, \sigma^2)}{\sqrt{t}}$$

Central Limit Theorem

Central Limit theorem states that

$$(\bar{\mathbf{C}} - \mathsf{F}(\mathsf{A})) \sim \frac{\mathcal{N}(0, \sigma^2)}{\sqrt{t}}$$

- It essentially says that $\bar{\mathbf{C}} - \mathsf{F}(\mathsf{A})$ converges to 0 at the speed of $\frac{1}{\sqrt{t}}$.

Central Limit Theorem

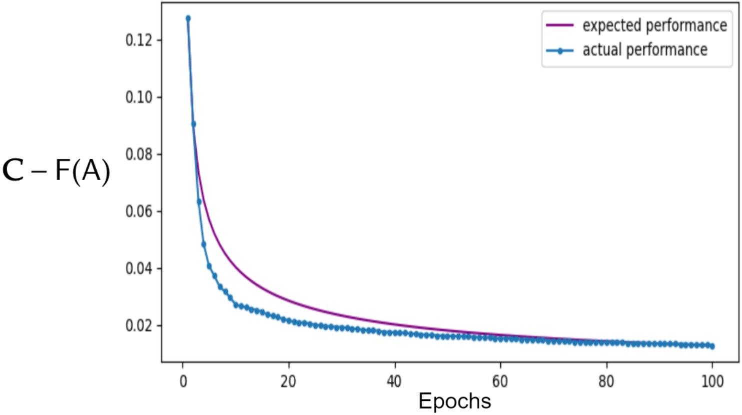
Central Limit theorem states that

$$(\bar{\mathbf{C}} - \mathsf{F}(\mathsf{A})) \sim \frac{\mathcal{N}(0, \sigma^2)}{\sqrt{t}}$$

- It essentially says that $\bar{\mathbf{C}} - \mathsf{F}(\mathsf{A})$ converges to 0 at the speed of $\frac{1}{\sqrt{t}}$.
- If I double the number of epochs given, the new difference should be $\frac{1}{\sqrt{2}} \approx 70.7\%$ of the previous difference.

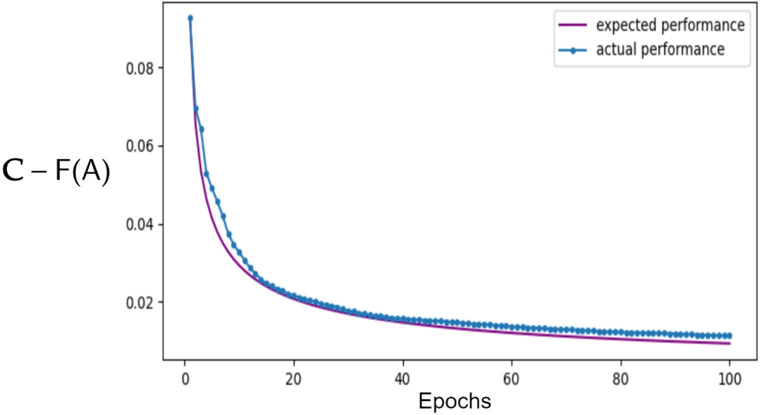
Results

Only clustered



Results

Clustered and Correlation



Future Work

- We can apply attacks in other papers on our setting and compare the results.

Future Work

- We can apply attacks in other papers on our setting and compare the results.
- We can modify F by adding in extra terms to better accommodate the graph.

Acknowledgement

- My dearest mentor Simon Langowski

Acknowledgement

- My dearest mentor Simon Langowski
- MIT PRIMES

Acknowledgement

- My dearest mentor Simon Langowski
- MIT PRIMES
- You guys for coming to my talk!