Intersection Attack in Non-Uniform Setting

Dongchen Zou under the instruction of Simon Langowski

May 19 2024

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 patterns 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 patterns like logging on and off?
- In this talk, we will explore how such information can be used to discover connections between users.

User Behavior

How do users of social media behave?

User Behavior

How do users of social media behave?

- People tend to talk based on the number of common interests. We call this clustering.

User Behavior

How do users of social media behave?

- People tend to talk based on the number of common interests. We call this clustering.
- If people have talked previously, it is more likely for them to talk again later. We call this correlation.

An Example of Social Media

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

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

suf
games, pizza, napping, volleyball

Clustering

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

3 common interests

1 common interest

2 common interests

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

suf
games, pizza, napping, volleyball

Correlation for the Example

Wanna play squash this afternoon?

How do I return your serve?

2:00pm

7:00pm

I will destroy you with my ginormous arm and leg

Haha. You need to volley them earlier.

Eavesdropper

If someone is online, they are talking to someone (possibly multiple people)

Eavesdropper

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

An Example Observation

Epoch 1

Epoch 2

Epoch 3

Intersection Attack in Non-Uniform Setting

Intersection attacks (also known as statistical disclosure attacks) use such observations to discover information about the graph.

Intersection Attack in Non-Uniform Setting

Intersection attacks (also known as statistical disclosure attacks) use such observations to discover information about the graph.
Non-uniformity comes from the two variables: clustering and correlation.

Difficulty of the Problem

Of course, both clustering and correlation come with different degrees: there can be a lot, there can be little.

Difficulty of the Problem

Of course, both clustering and correlation come with different degrees: there can be a lot, there can be little.
And they affect the difficulty of intersection attacks.

Difficulty of the Problem

Of course, both clustering and correlation come with different degrees: there can be a lot, there can be little.
And they affect the difficulty of intersection attacks. Do they make the problem easier or harder?

Hypotheses

- Clustering makes the problem easier because it allows the eavesdropper to better classify them and notice patterns.

Hypotheses

- Clustering makes the problem easier because it allows the eavesdropper to better classify them and notice patterns.
- Clustering makes the problem more difficult because it homogenizes a group of people.

Hypotheses

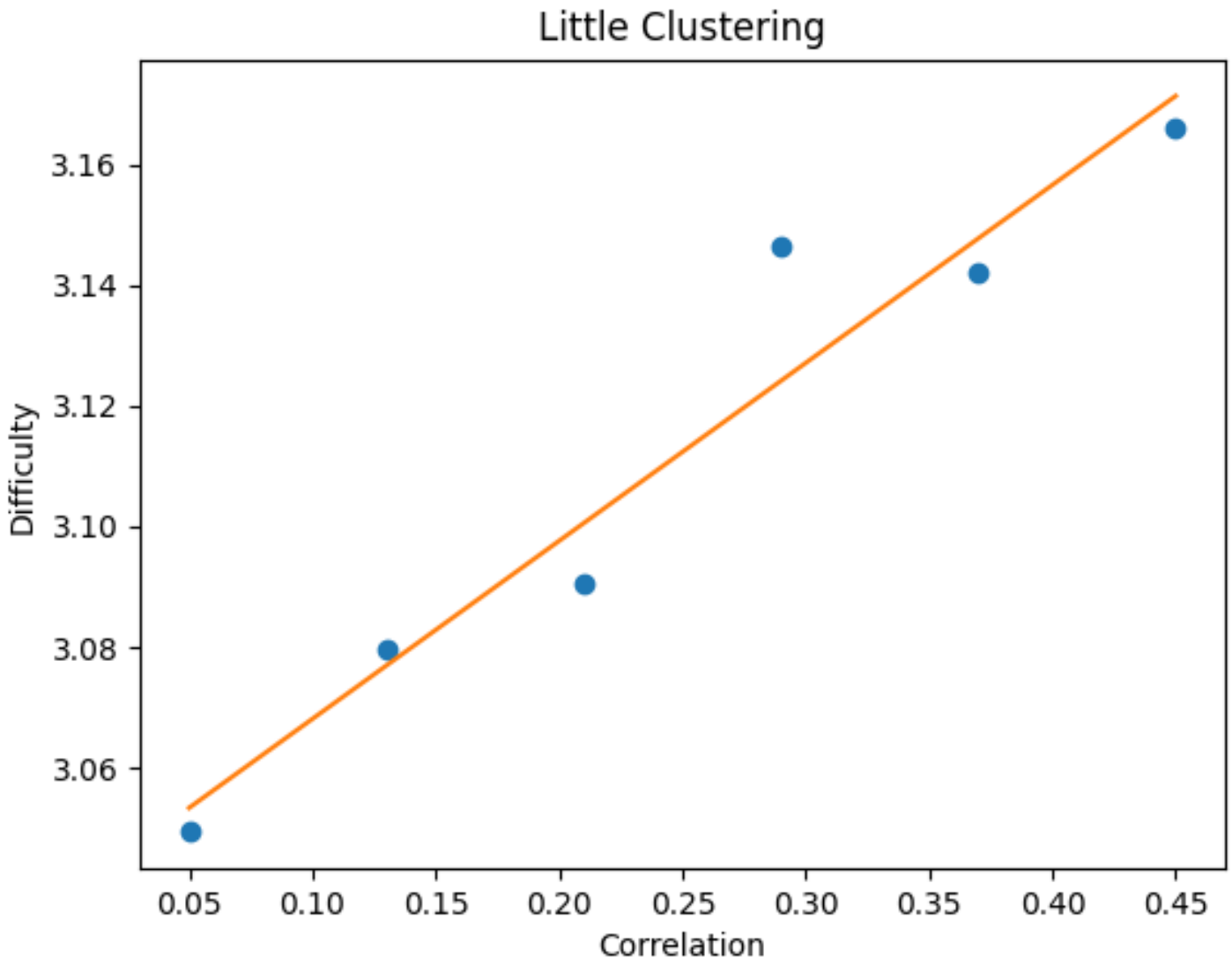
- Clustering makes the problem easier because it allows the eavesdropper to better classify them and notice patterns.
- Clustering makes the problem more difficult because it homogenizes a group of people.
- Correlation makes the problem more difficult because it gives the eavesdropper more repetitive information that confuses him.

Hypotheses

- Clustering makes the problem easier because it allows the eavesdropper to better classify them and notice patterns.
- Clustering makes the problem more difficult because it homogenizes a group of people.
- Correlation makes the problem more difficult because it gives the eavesdropper more repetitive information that confuses him.
- Correlation makes the problem easier because the eavesdropper has more epochs (and thus more opportunities) to notice a connection.

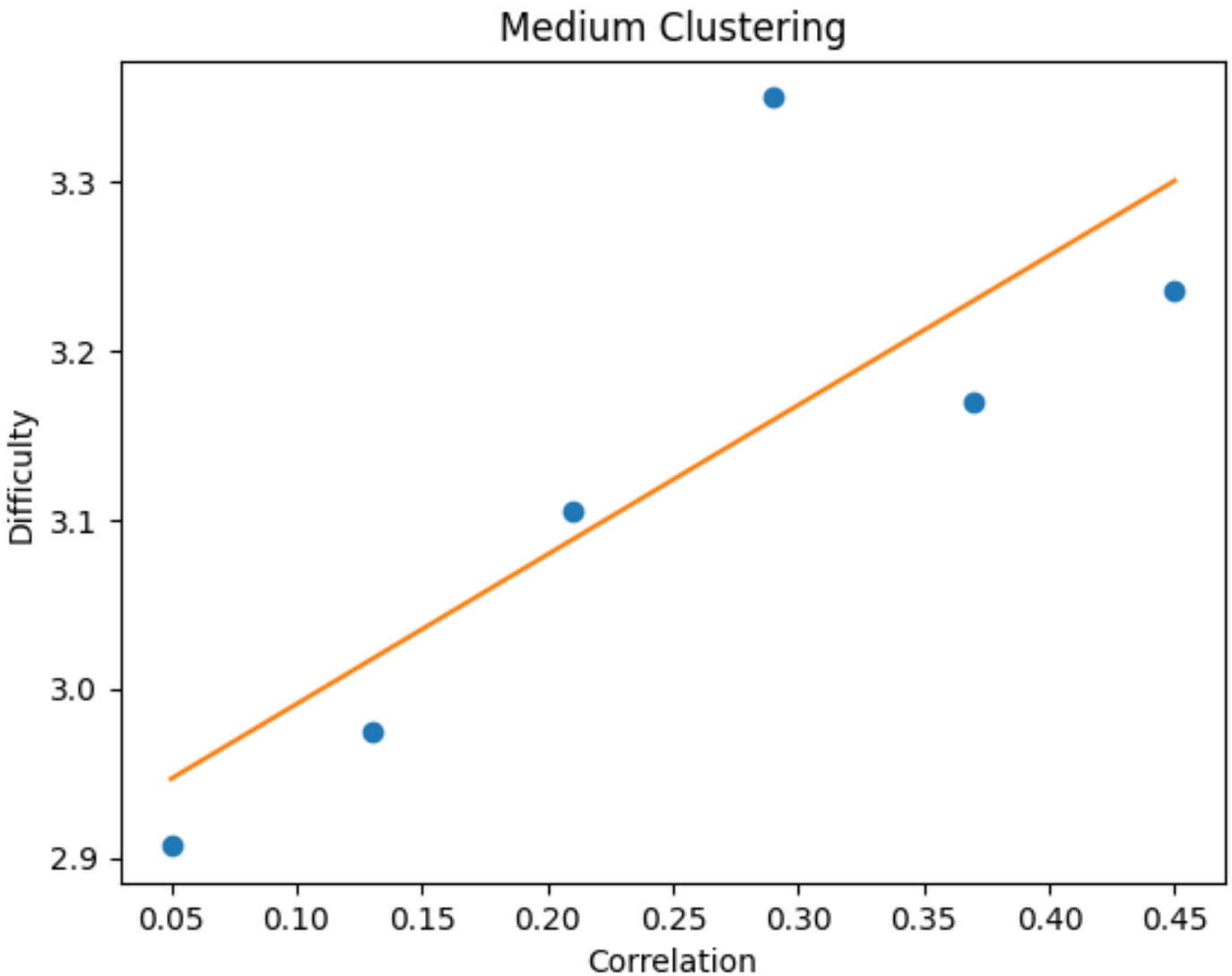
Little Clustering

$$y = 0.295x + 3.04$$



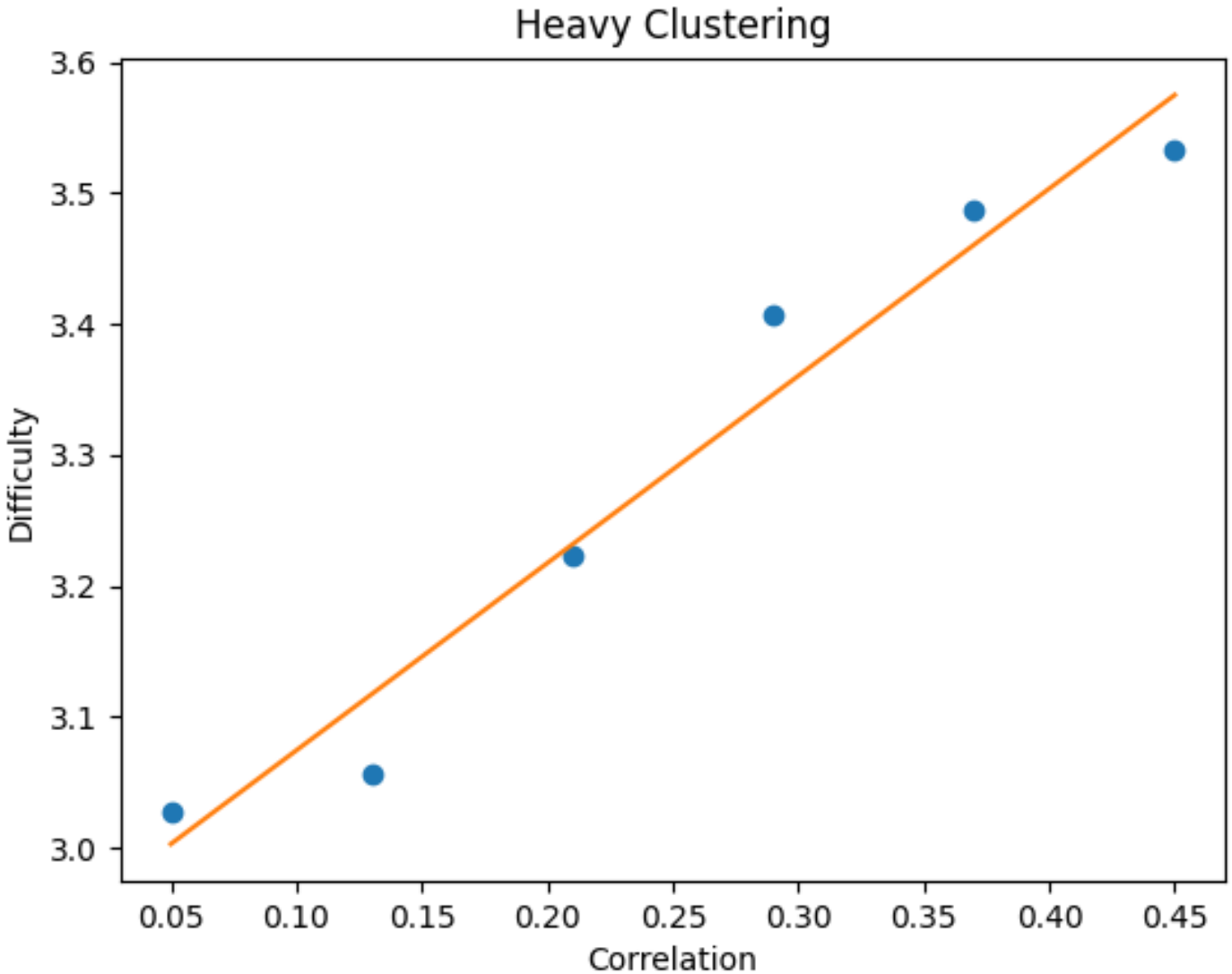
Medium Clustering

$$y = 0.885x + 2.90$$



Heavy Clustering

$$y = 1.43x + 2.93$$



Technical Part

You might have a few questions like

Technical Part

You might have a few questions like

- How exactly are "clustering" and "correlation" defined with symbols?

Technical Part

You might have a few questions like

- How exactly are "clustering" and "correlation" defined with symbols?
- How do you assess how difficult it is to extract information?

Technical Part

You might have a few questions like

- How exactly are "clustering" and "correlation" defined with symbols?
- How do you assess how difficult it is to extract information?
- What exactly is the eavesdropper trying to do?

Table of Contents

- Graph (Clustering)

Table of Contents

- Graph (Clustering)
- Epoch (Correlation)

Table of Contents

- Graph (Clustering)
- Epoch (Correlation)
- Difficulty Assessment

Table of Contents

- Graph (Clustering)
- Epoch (Correlation)
- Difficulty Assessment
- Future Work

Graph Generation

How are we going to reconstruct the graph example?

Graph Generation

How are we going to reconstruct the graph example?

- Each user k is randomly assigned $\mathbf{I}_k$.

Graph Generation

How are we going to reconstruct the graph example?

- Each user k is randomly assigned $\mathbf{I}_k$.
- The bigger the intersection $|\mathbf{I}_i \cap \mathbf{I}_j|$ between two users, the more likely they are to talk.

Graph Generation

How are we going to reconstruct the graph example?

- Each user k is randomly assigned $\mathbf{I}_k$.
- The bigger the intersection $|\mathbf{I}_i \cap \mathbf{I}_j|$ between two users, the more likely they are to talk.
- We denote the probability that i and j talk in an epoch as A[i, j].

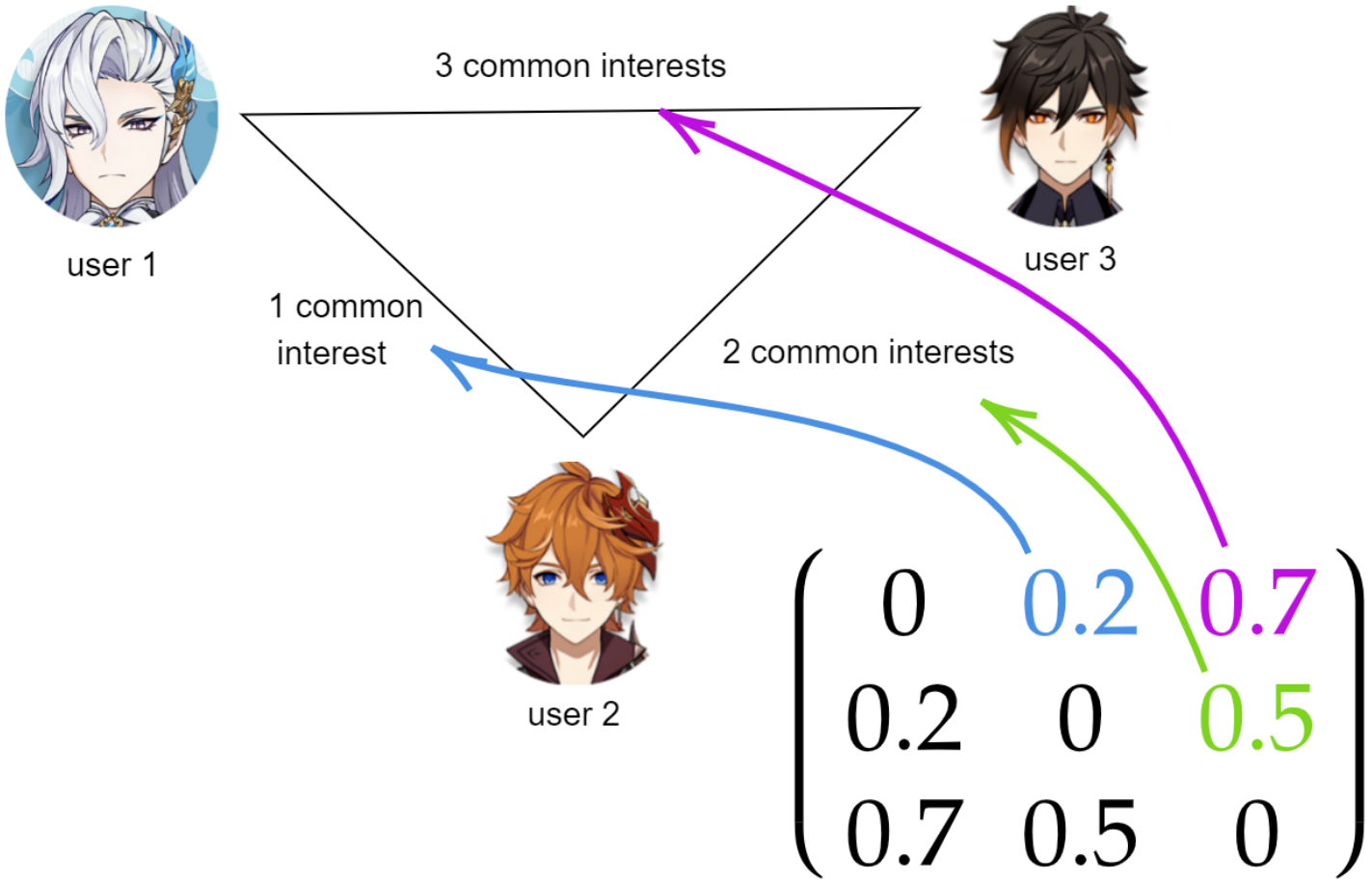
Graph Generation

How are we going to reconstruct the graph example?

- Each user k is randomly assigned $\mathbf{I}_k$.
- The bigger the intersection $|\mathbf{I}_i \cap \mathbf{I}_j|$ between two users, the more likely they are to talk.
- We denote the probability that i and j talk in an epoch as A[i, j].
- Let μ be the clustering coefficient. Then we have

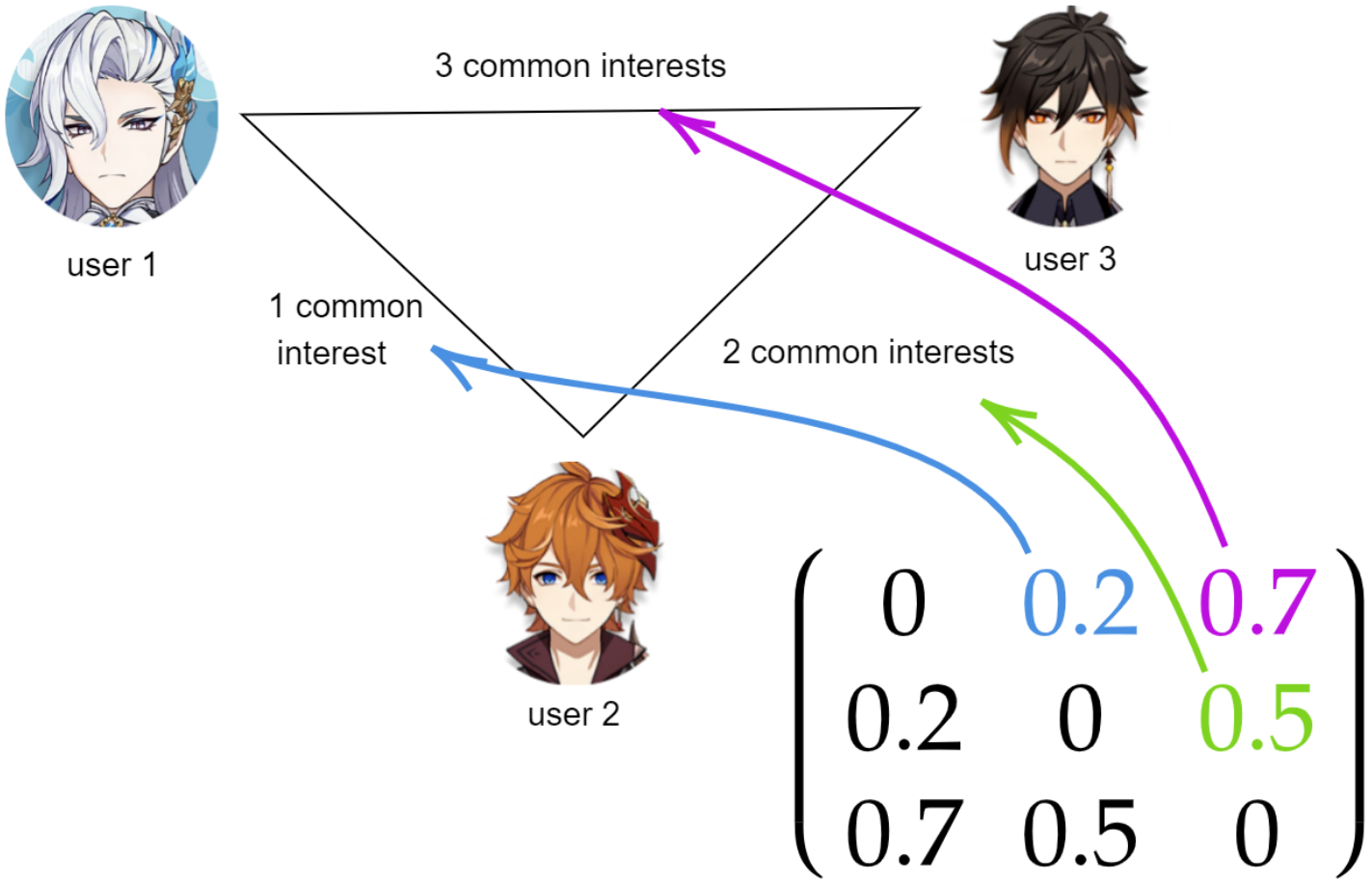
$$\mathrm{A[i,j]} = f_\mu(|\mathbf{I}_i \cap \mathbf{I}_j|)$$

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



The eavesdropper is trying to find the probability matrix.

Epoch Generation

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

Epoch Generation

- If i and j talked in the previous epoch, it is more likely for them to keep talking in this epoch.
- How about $\mathrm{A[i,j]} \rightarrow (1+\delta)\mathrm{A[i,j]}$ if i and j talked in last epoch?

Epoch Generation

- If i and j talked in the previous epoch, it is more likely for them to keep talking in this epoch.
- How about $\mathrm{A[i,j]} \rightarrow (1+\delta)\mathrm{A[i,j]}$ if i and j talked in last epoch?
- We call δ the correlation coefficient.

Difficulty Assessment

We need a way to know: "How difficult would it be for an eavesdropper to extract information from this graph, from this configuration?"

Intersection Attack

What if you ask the eavesdropper this question: "What is the probability that users i and j appear online at the same time?"

Intersection Attack

What if you ask the eavesdropper this question: "What is the probability that users i and j appear online at the same time?" He can give an answer with his observations (aka epochs)! Just the number of times it happened over the total number of epochs!

Intersection Attack

What if you ask the eavesdropper this question: "What is the probability that users i and j appear online at the same time?"
He can give an answer with his observations (aka epochs)!
Just the number of times it happened over the total number of epochs!
We can answer the same question with the matrix A, using some probability formula.

Intersection Attack

What if you ask the eavesdropper this question: "What is the probability that users i and j appear online at the same time?"
He can give an answer with his observations (aka epochs)!
Just the number of times it happened over the total number of epochs!
We can answer the same question with the matrix A, using some probability formula.
When the number of epochs goes to infinity, the two probabilities should be equal.

So what?

We can characterize the difficulty of a configuration as follows:

So what?

We can characterize the difficulty of a configuration as follows: **How many epochs does the eavesdropper need to observe until the two probabilities are within a certain range?**

Future Work

- Come up with a better attack that takes clustering and correlation into account.

Future Work

- Come up with a better attack that takes clustering and correlation into account.
- Come up with a better explanation for the difficulty of the problem.

Acknowledgements

- My dearest mentor Simon Langowski

Acknowledgements

- My dearest mentor Simon Langowski
- MIT PRIMES

Acknowledgements

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