

Practical Anonymity Sets in Pseudonymous Forums

Benjamin Chen, mentored by Kyle Hogan

October 2020

Table of Contents

Introduction


Our Project

Creating Anonymity Sets






Picking a Budget

Introduction

What is Reddit?

↑  **r/AskReddit** · Posted by cryptonerd 6 hours ago

59.8k
↓

↑ EpicGamer6612 7.6k points · 4 hours ago     2 

↓ No :(

↑ Hadoromo 218 points · 2 hours ago

↓ Yes!

↑ EpicGamer6612 Score hidden · 14 minutes ago

↓ No.

↑ Hadoromo Score hidden · 10 minutes ago

↓ Yes.

↑ Hadoromo Score hidden · 9 minutes ago

↓ It definitely is a soup.

Alice and Eve

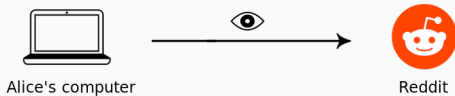
Suppose we have two Reddit users: Alice and Eve,

Alice and Eve

Suppose we have two Reddit users: Alice and Eve, and Eve wants to figure out Alice's username.

Alice and Eve

Suppose we have two Reddit users: Alice and Eve, and Eve wants to figure out Alice's username.



Alice and Eve

Suppose we have two Reddit users: Alice and Eve, and Eve wants to figure out Alice's username.



Alice's computer



Reddit

↑ **r/AskReddit** · Posted by cryptonerd 6 hours ago
59.8k
↓

↑ EpicGamer6612 7.6k points · 4 hours ago 🗨️ 😊 📧 🗑️ 2 🍌
↓ No :(

↑ Hagaromo 218 points · 2 hours ago
↓ Yes!

↑ EpicGamer6612 Score hidden · 14 minutes ago
↓ No.

↑ Hagaromo Score hidden · 10 minutes ago
↓ Yes.

↑ Hagaromo Score hidden · 9 minutes ago
↓ It definitely is a soup.

Alice and Eve

Suppose we have two Reddit users: Alice and Eve, and Eve wants to figure out Alice's username.



Alice's computer



Reddit

↑ **r/AskReddit** · Posted by cryptonerd 6 hours ago
59.8k
↓

↑ EpicGamer6612 7.6k points · 4 hours ago 2

↓ No :(

↑ Hagoromo 218 points · 2 hours ago
↓ Yes!

↑ EpicGamer6612 Score hidden · 14 minutes ago
↓ No.

↑ Hagoromo Score hidden · 10 minutes ago
↓ Yes.

↑ Hagoromo Score hidden · 9 minutes ago
↓ It definitely is a soup.

Alice's traffic

251 min ago

14 min ago

More Realistic Examples

“Eve” could be...

More Realistic Examples

“Eve” could be...

- Internet providers

More Realistic Examples

“Eve” could be...

- Internet providers
- Oppressive governments

More Realistic Examples

“Eve” could be...

- Internet providers
- Oppressive governments
- Employers

More Realistic Examples

“Eve” could be...

- Internet providers
- Oppressive governments
- Employers

Basically any adversary who can see the users' activity, but not the contents of incoming or outgoing traffic (hidden with encryption).

A Potential Solution

What if everyone's computer traffic looked the same?

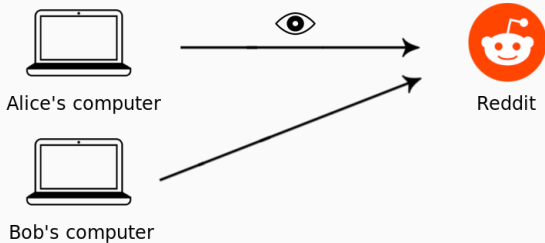
A Potential Solution

What if everyone's computer traffic looked the same?



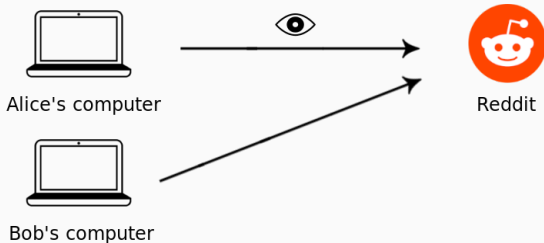
A Potential Solution

What if everyone's computer traffic looked the same?



A Potential Solution

What if everyone's computer traffic looked the same?



- Users send **dummy traffic** until their total traffic reaches a certain value (called the **budget**)

How do we pick a budget?

How do we pick a budget?

- Bandwidth

How do we pick a budget?

- Bandwidth
- Latency

How do we pick a budget?

- Bandwidth
- Latency

What if we try to divide up the people into “good” sets?

Bob also uses Reddit in a similar way to Alice.

Bob also uses Reddit in a similar way to Alice.

- Few dummies (low overhead)

Bob also uses Reddit in a similar way to Alice.

- Few dummies (low overhead)
- Few postponed messages (low latency)

Bob also uses Reddit in a similar way to Alice.

- Few dummies (low overhead)
- Few postponed messages (low latency)

We would say Alice and Bob are in an **anonymity set**, since they have the same traffic patterns and are indistinguishable.

The main questions we investigate are:

The main questions we investigate are:

- How do we group people to look the same in a good way (and what does “good” entail)?

The main questions we investigate are:

- How do we group people to look the same in a good way (and what does “good” entail)?
- How do we pick the budget for a group of people?

The main questions we investigate are:

- How do we group people to look the same in a good way (and what does “good” entail)?
- How do we pick the budget for a group of people?
- Is such a system practical in real life?

Our Project

We make a compromise between performance and privacy:

We make a compromise between performance and privacy:

- Users are placed into anonymity sets of size at least k , for some integer k .

Anonymity Sets

We make a compromise between performance and privacy:

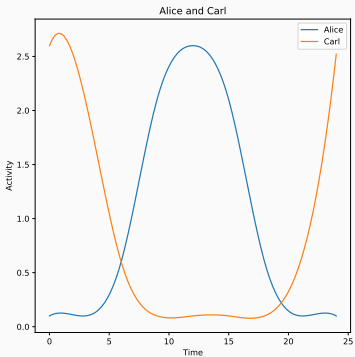
- Users are placed into anonymity sets of size at least k , for some integer k .
- Each user in the set looks identical to every other user in the set from the adversary's point of view.

Anonymity Sets

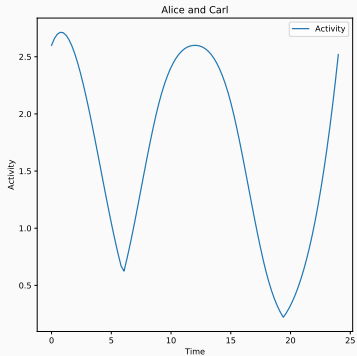
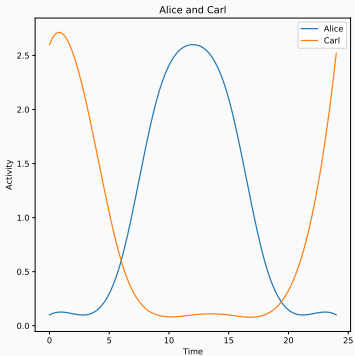
We make a compromise between performance and privacy:

- Users are placed into anonymity sets of size at least k , for some integer k .
- Each user in the set looks identical to every other user in the set from the adversary's point of view.
- We try to create sets to find a balance between performance and privacy.

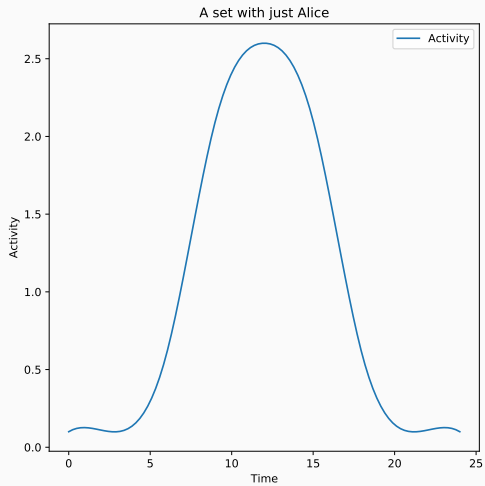
Alice and Carl



Alice and Carl



Alice and Carl



The Perfect Anonymity Sets

- Minimizes performance losses

The Perfect Anonymity Sets

- Minimizes performance losses
- Maintains good privacy

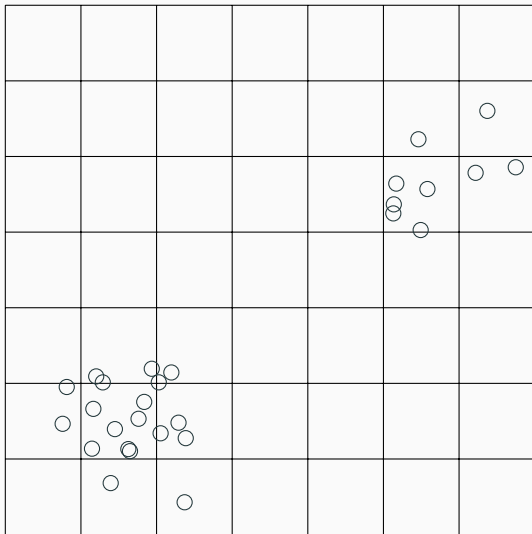
The Perfect Anonymity Sets

- Minimizes performance losses
- Maintains good privacy

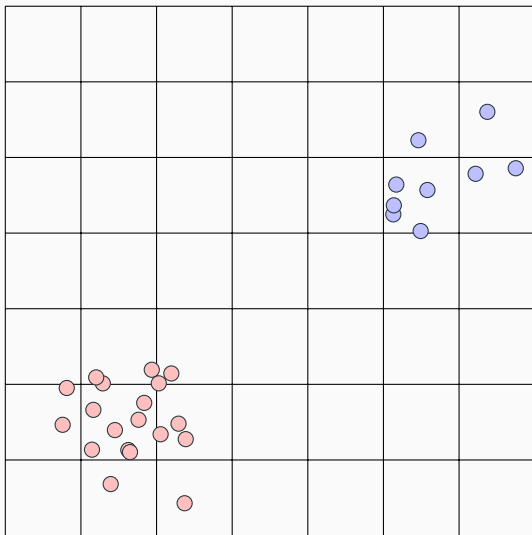
Achieving both of these at the same time is hard.

Creating Anonymity Sets

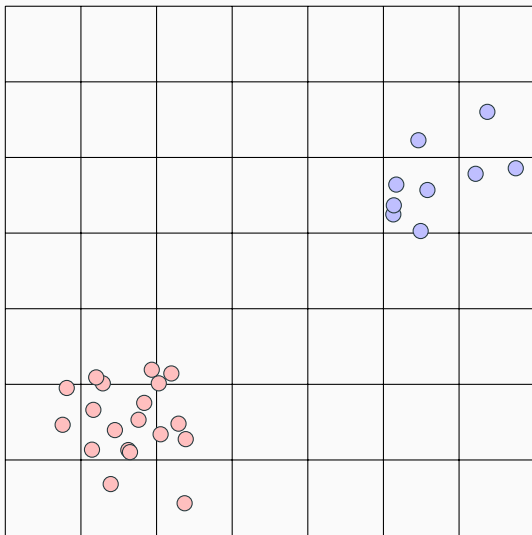
K-Means



K-Means

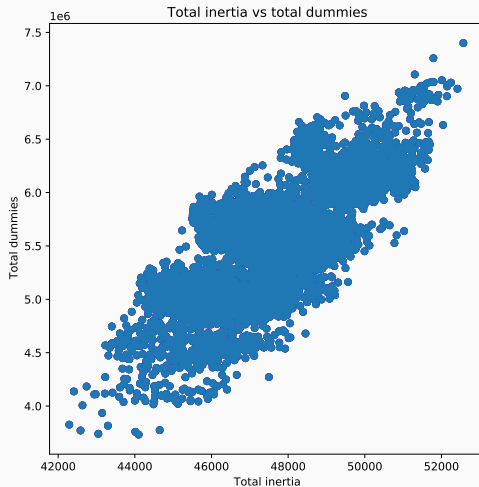


K-Means



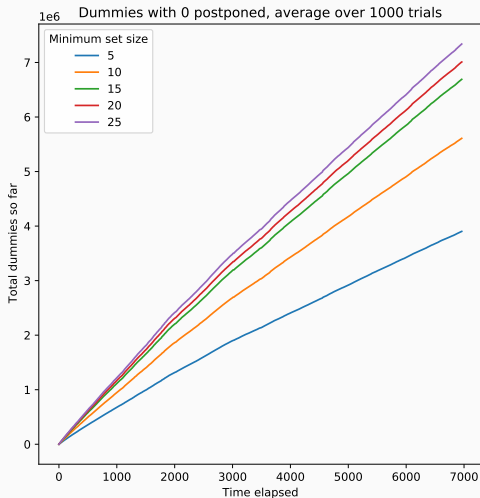
K-means attempts to minimize the **inertia** of each cluster.

Inertia as an Indicator of Performance



Each dot here represents 1 clustering setup (with a different random seed)

Cluster Sizes



(On a dataset of 100 users)

Picking a Budget

The Budget

Once the anonymity sets are created, we define a budget (how much traffic people should send) based on the mean activity over users in the set.

Once the anonymity sets are created, we define a budget (how much traffic people should send) based on the mean activity over users in the set.

- If a user sends under the budget, they send dummy messages until the budget is reached.

Once the anonymity sets are created, we define a budget (how much traffic people should send) based on the mean activity over users in the set.

- If a user sends under the budget, they send dummy messages until the budget is reached.
- If a user sends over the budget, their messages are postponed to a later round.

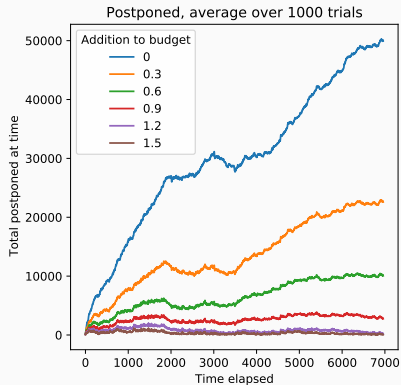
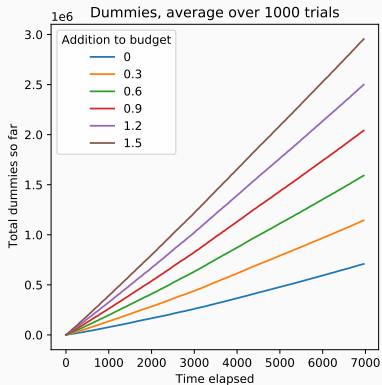
The Budget

Once the anonymity sets are created, we define a budget (how much traffic people should send) based on the mean activity over users in the set.

- If a user sends under the budget, they send dummy messages until the budget is reached.
- If a user sends over the budget, their messages are postponed to a later round.

In general, we care more about reducing postponed messages over reducing dummy messages.

The Solution



- Using machine learning to create anonymity sets

- Using machine learning to create anonymity sets
- Looking more closely at the people who make up the sets here

- Using machine learning to create anonymity sets
- Looking more closely at the people who make up the sets here
- Applying this to other cases (websites, bidirectional communication)

Acknowledgements

- Kyle Hogan
- Dr. Gerovitch & Prof. Devadas
- PRIMES Program & MIT
- My family

Thanks for listening!