

# Towards Practical Ambiguity Sets

---

Benjamin Chen, mentored by Kyle Hogan

October 2020

# Table of Contents

Introduction

Our Project

Creating Ambiguity Sets

Picking a Budget

# Introduction

---

Suppose we have two Reddit users: Alice and Eve.

## Alice and Eve

Suppose we have two Reddit users: Alice and Eve.

Suppose Eve wants to figure out Alice's username.

Suppose we have two Reddit users: Alice and Eve.

Suppose Eve wants to figure out Alice's username.

Suppose Eve can monitor Alice's traffic to Reddit's servers, but can't see the content of any transmissions.

Suppose we have two Reddit users: Alice and Eve.

Suppose Eve wants to figure out Alice's username.

Suppose Eve can monitor Alice's traffic to Reddit's servers, but can't see the content of any transmissions.

This seems okay, right? After all, Eve gets almost no info about what Alice is actually doing.

## Alice and Eve

Suppose we have two Reddit users: Alice and Eve.

Suppose Eve wants to figure out Alice's username.







Suppose Eve can monitor Alice's traffic to Reddit's servers, but can't see the content of any transmissions.

This seems okay, right? After all, Eve gets almost no info about what Alice is actually doing.

However, remember Eve uses Reddit too...



# Alice and Eve

↑  **r/AskReddit** · Posted by [u/btw\\_i\\_use\\_arch](#) 9 hours ago    6  2  2

12.6k  
↓

## Is cereal a soup?

🗨️ 2.6k Comments [Share](#) [Save](#) [Hide](#) [Report](#) 98% Upvoted


Log in or sign up to leave a comment

LOG INSIGN UP

SORT BY **Best** ▾

---

[View discussions in 3 other communities](#)

↑ [PhysicsIsPhun](#) 4.8k points · 4 hours ago 

↓ Yes!

↑ [EpicGamer6612](#) 1.6k points · 4 hours ago

↓ No.

↑ [PhysicsIsPhun](#) 402 points · 50 minutes ago

↓ Yes.






↑ [EpicGamer6612](#) 272 points · 47 minutes ago

↓ No. Why would you say that?

↑ [PhysicsIsPhun](#) 125 points · 44 minutes ago





↓ Because it's the truth.

# Alice and Eve

↑  r/AskReddit · Posted by u/btw\_i\_use\_arch 9 hours ago   6  2  2

12.6k  
↓


## Is cereal a soup?

2.6k Comments  Share  Save  Hide  Report 98% Upvoted

Log in or sign up to leave a comment [LOG IN](#) [SIGN UP](#)

SORT BY Best ▾

[View discussions in 3 other communities](#)

↑ PhysicsIsPhun 4.8k points · 4 hours ago 

↓ Yes!

↑ EpicGamer6612 1.6k points · 4 hours ago

↓ No.

↑ PhysicsIsPhun 402 points · 50 minutes ago

↓ Yes.

↑ EpicGamer6612 272 points · 47 minutes ago

↓ No. Why would you say that?

↑ PhysicsIsPhun 125 points · 44 minutes ago

↓ Because it's the truth.

Alice's traffic

251 min ago

47 min ago

# Alice and Eve

12.6k

r/AskReddit · Posted by u/btw\_i\_use\_arch 9 hours ago

## Is cereal a soup?

2.6k Comments · Share · Save · Hide · Report · 98% Upvoted

Log in or sign up to leave a comment

LOG IN SIGN UP

SORT BY Best

View discussions in 3 other communities

PhysicsIsPhun 4.8k points · 4 hours ago

Yes!

EpicGamer6612 1.6k points · 4 hours ago

No.

PhysicsIsPhun 402 points · <1 hour ago

Yes.

EpicGamer6612 272 points · <1 hour ago

No. Why would you say that?

PhysicsIsPhun 125 points · <1 hour ago

Because it's the truth.

Alice's traffic

240 min ago

0 min ago

# Alice and Eve

12.6k

r/AskReddit · Posted by u/btw\_i\_use\_arch 9 hours ago

## Is cereal a soup?

2.6k Comments · Share · Save · Hide · Report · 98% Upvoted

Log in or sign up to leave a comment

LOG IN SIGN UP

SORT BY Best

View discussions in 3 other communities

PhysicsIsPhun 4.8k points · 4 hours ago

Yes!

EpicGamer6612 1.6k points · 4 hours ago

No.

PhysicsIsPhun 402 points · <1 hour ago

Yes.

EpicGamer6612 272 points · <1 hour ago

No. Why would you say that?

PhysicsIsPhun 125 points · <1 hour ago

Because it's the truth.

Alice's traffic

240 min ago

0 min ago

0 min ago

# Alice and Eve

The screenshot shows a Reddit post in the r/AskReddit community. The post title is "Is cereal a soup?" and it has 12.6k upvotes. The post was made by user u/btw\_i\_use\_arch 9 hours ago. Below the title are options to comment, share, save, hide, and report, along with a "2.6k Comments" count and a "98% Upvoted" status. A login/sign-up prompt is visible. The comment section shows a thread of replies:

- PhysicsIsPhun (4.8k points, 4 hours ago) says "Yes!".
- EpicGamer6612 (1.6k points, 4 hours ago) replies "No.".
- PhysicsIsPhun (402 points, <1 hour ago) replies "Yes.".
- EpicGamer6612 (272 points, <1 hour ago) replies "No. Why would you say that?".
- PhysicsIsPhun (125 points, <1 hour ago) replies "Because it's the truth."

Alice's traffic
240 min ago
0 min ago
0 min ago

Alice's fake messages are called **dummy messages**.

## Alice, Bob, and Eve

What if there's another user, Bob, who also uses Reddit and posts in the same forum—but Bob posts much more frequently than Alice?

What if there's another user, Bob, who also uses Reddit and posts in the same forum|but Bob posts much more frequently than Alice?

Alice can up her dummy traffic to make her look like Bob (lots of overhead)

What if there's another user, Bob, who also uses Reddit and posts in the same forum|but Bob posts much more frequently than Alice?

Alice can up her dummy traffic to make her look like Bob (lots of overhead)

Alice can give up on looking like Bob and just post enough dummies to look like PhysicsIsPhun.



What if there's another user, Bob, who also uses Reddit and posts in the same forum|but Bob posts much more frequently than Alice?

Alice can up her dummy traffic to make her look like Bob (lots of overhead)

Alice can give up on looking like Bob and just post enough dummies to look like PhysicsIsPhun.

We would say EpicGamer6612 (Alice) and PhysicsIsPhun are in an ambiguity set, since Eve can't determine which of the two Alice is.

\Eve" could be...

"Eve" could be...

Internet providers

\Eve" could be...

Internet providers

Oppressive governments

\Eve" could be...

Internet providers

Oppressive governments

Employers

\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).

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 ambiguity sets of size at least  $k$ , for some integer  $k$ .

We make a compromise between performance and privacy:

Users are placed into ambiguity 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 make a compromise between performance and privacy:

Users are placed into ambiguity 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.









Minimizes unnecessary traffic

# The Perfect Ambiguity Sets

Minimizes unnecessary traffic

Maintains good privacy

# The Perfect Ambiguity Sets

Minimizes unnecessary traffic

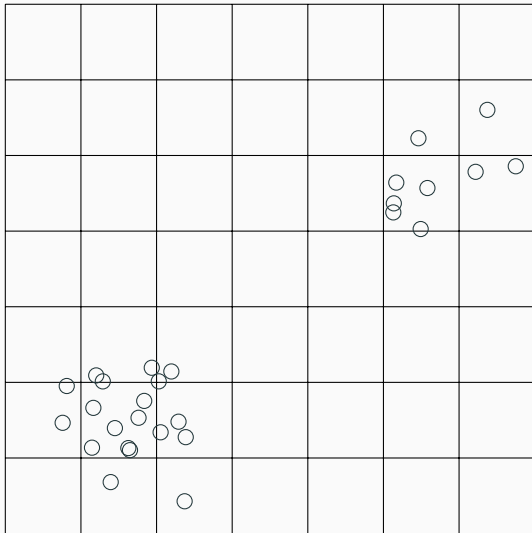
Maintains good privacy

Achieving both of these at the same time is hard.

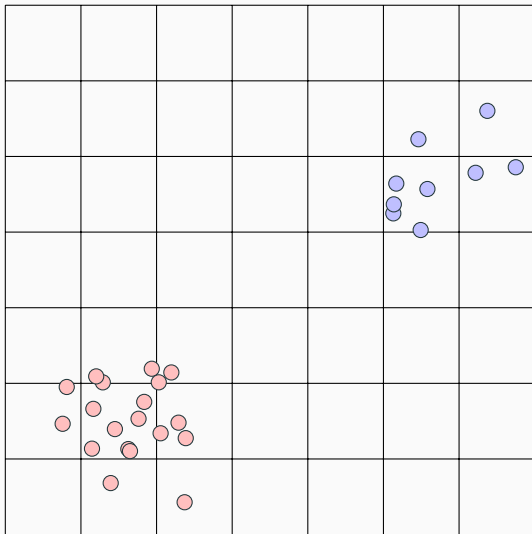
## Creating Ambiguity Sets

---

# K-Means

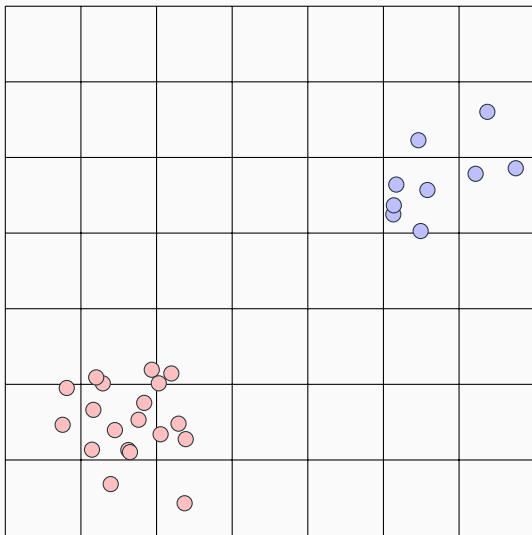


# K-Means





# 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

---

Once the ambiguity 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 ambiguity 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 ambiguity 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 ambiguity 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.

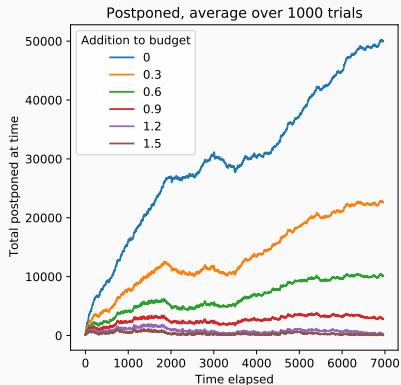
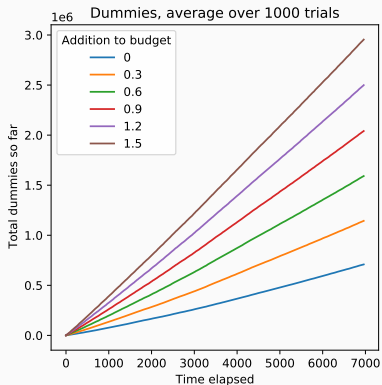
If a user sends over the budget, their messages are postponed to a later round.

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



# The Solution

budget = mean (1 + addition to budget)



Testing this on bigger datasets

Testing this on bigger datasets

Looking more closely at the people who make up the sets here

# Acknowledgements

Kyle Hogan

Dr. Gerovitch & Prof. Devadas

PRIMES Program & MIT

Thanks for listening!