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

Alice and Eve

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.

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.

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.

Alice and Eve

Suppose we have two Reddit users: Alice and Eve.

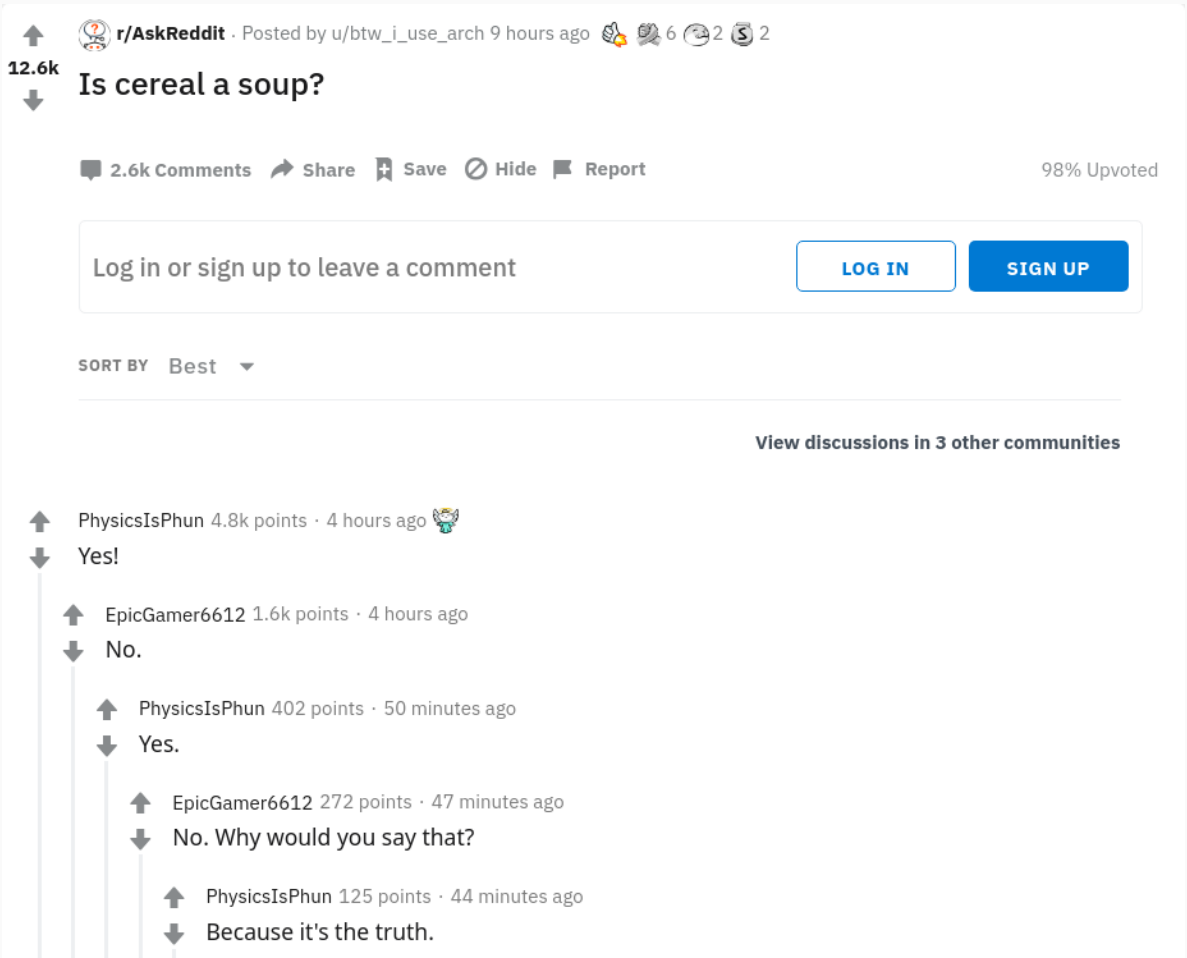
Suppose Eve wants to figure out Alice's username.

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

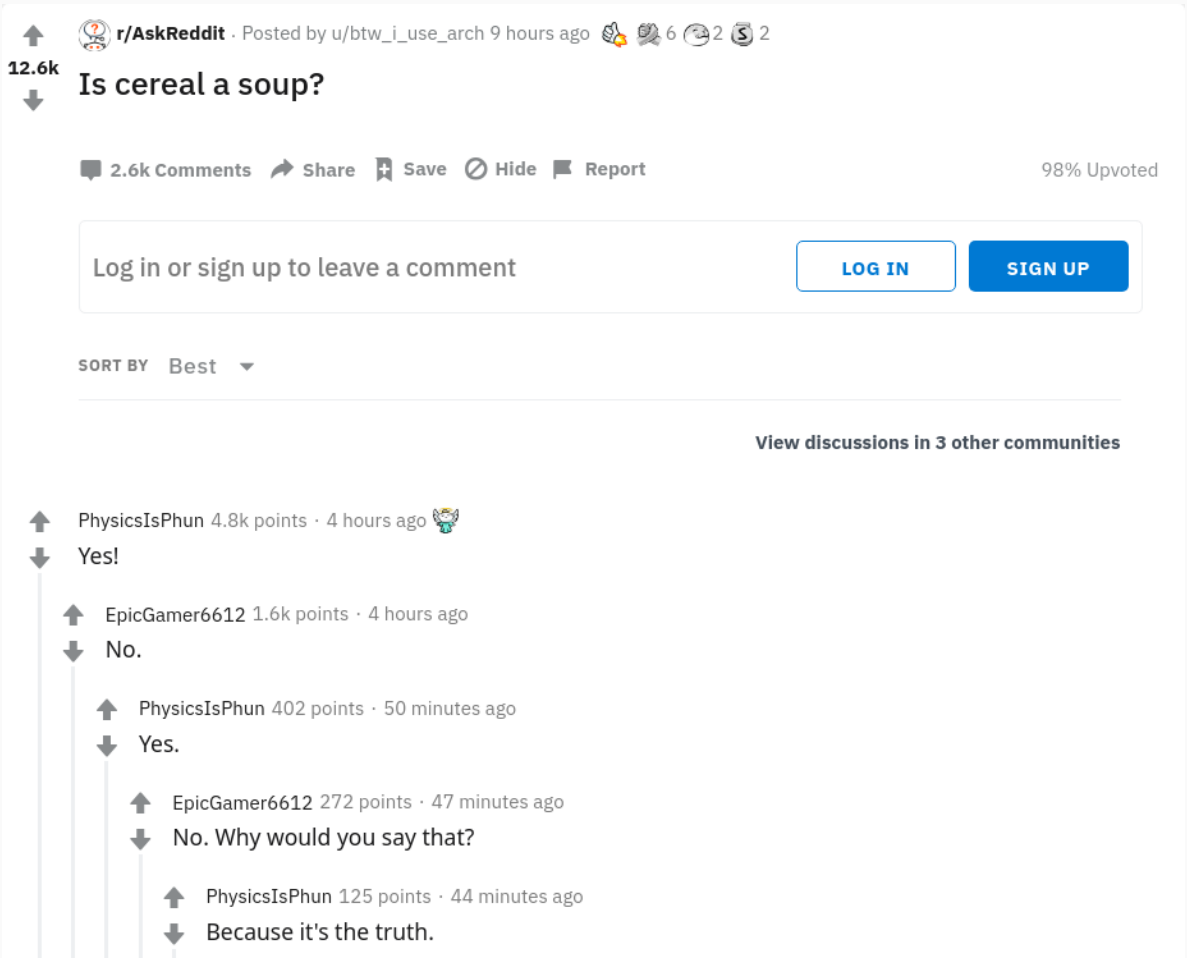
This seems okay, right? After all, Eve gets almost no info about what Alice is actually doing.

However, remember Eve uses Reddit too...

Alice and Eve

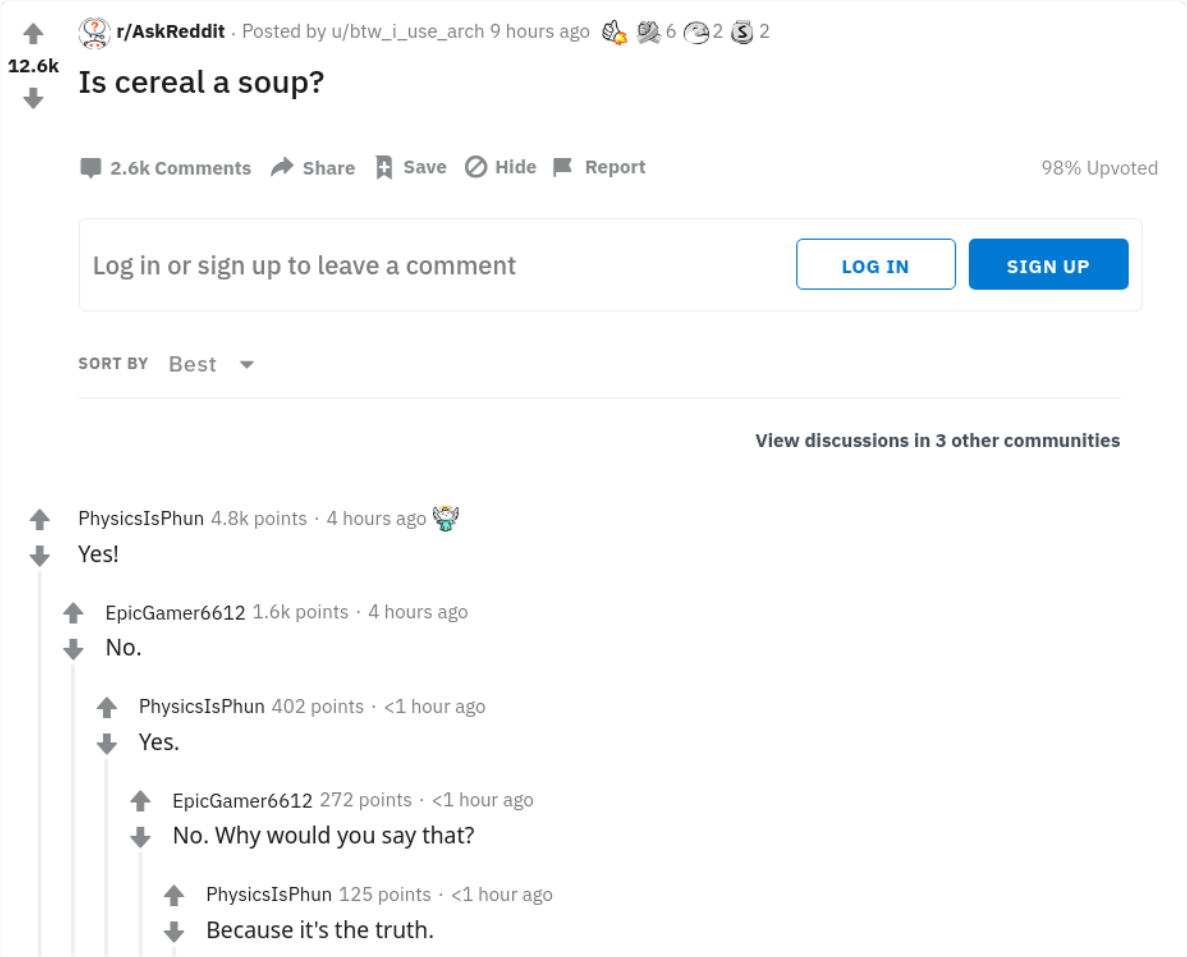


Alice and Eve



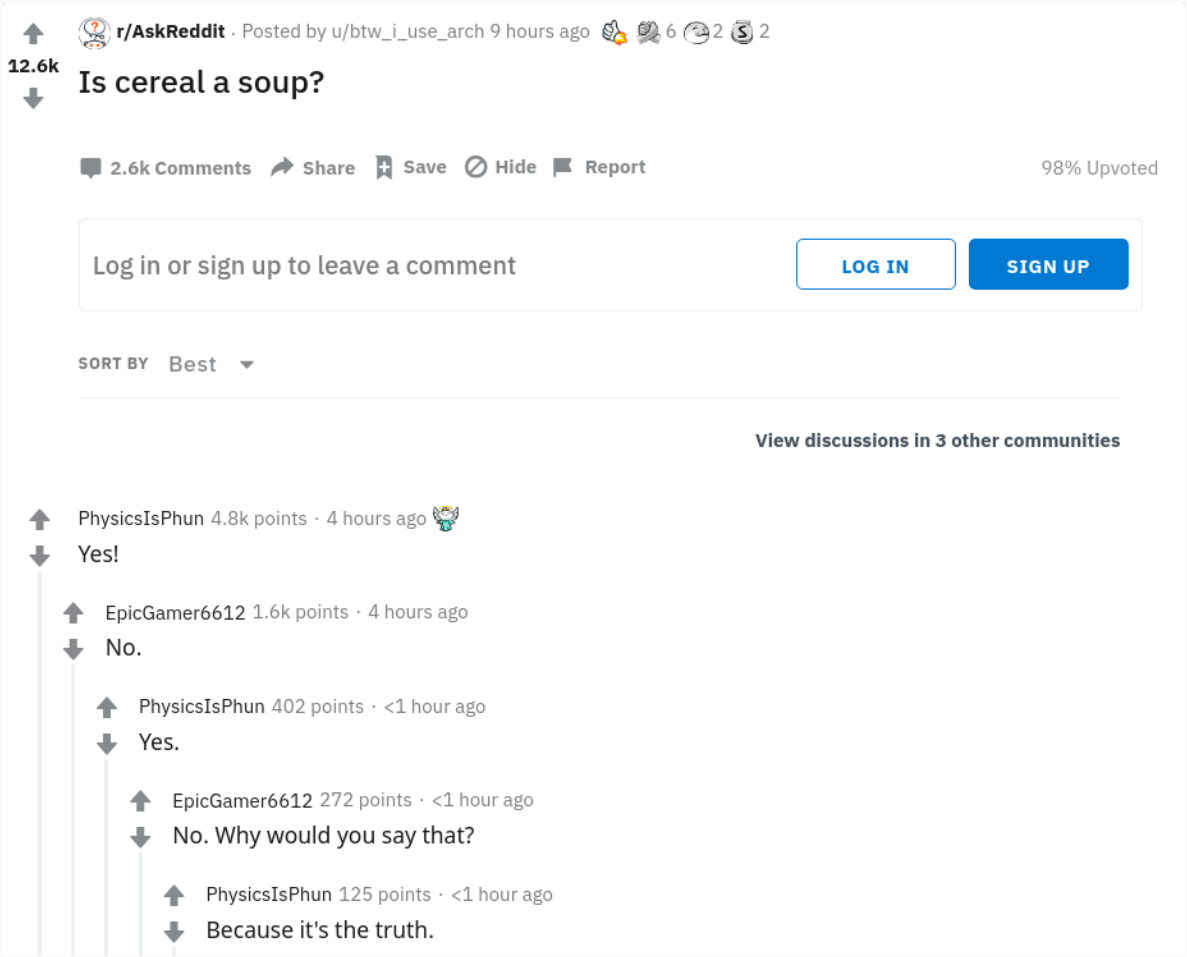
Alice's traffic
251 min ago
47 min ago

Alice and Eve



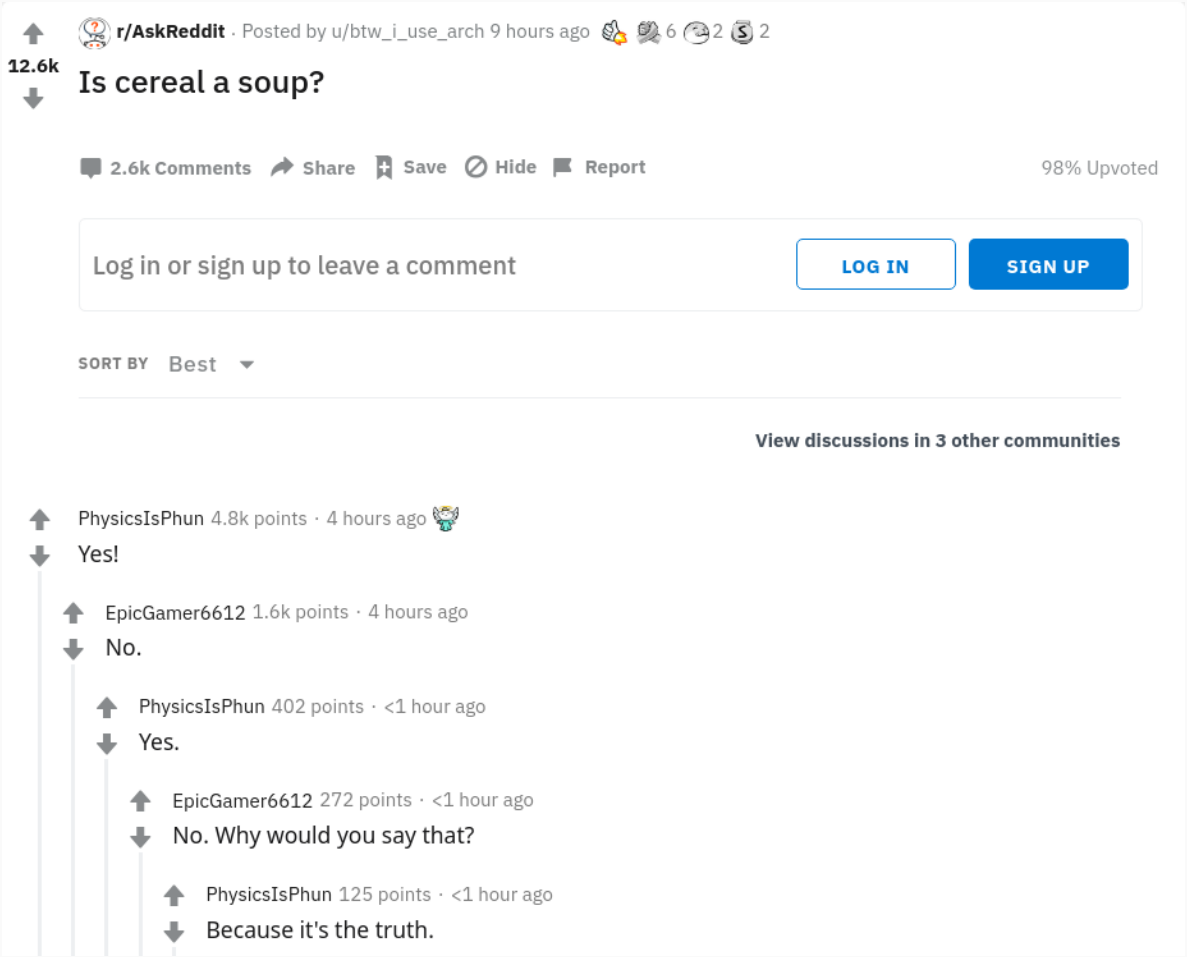
Alice's traffic
240 min ago
0 min ago

Alice and Eve



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

Alice and Eve



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?

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?

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

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?

- 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.

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?

- 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.

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).

Questions

The main questions we investigate are:

Questions

The main questions we investigate are:

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

Questions

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?

Questions

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

Ambiguity Sets

We make a compromise between performance and privacy:

Ambiguity Sets

We make a compromise between performance and privacy:

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

Ambiguity Sets

We make a compromise between performance and privacy:

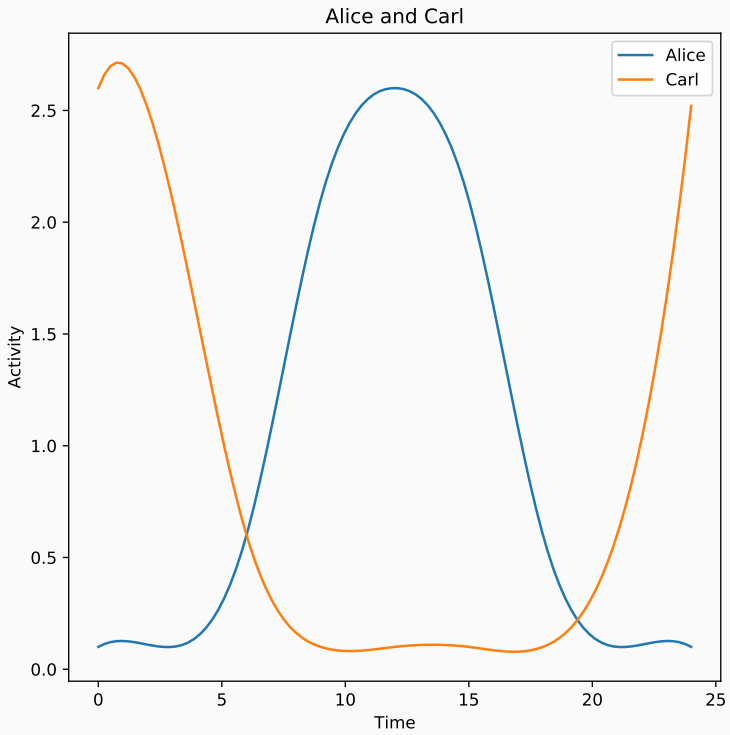
- Users are placed into ambiguity sets of size at least k, for some integer k.
- Each user in the set looks identical to every other user in the set from the adversary's point of view.

Ambiguity Sets

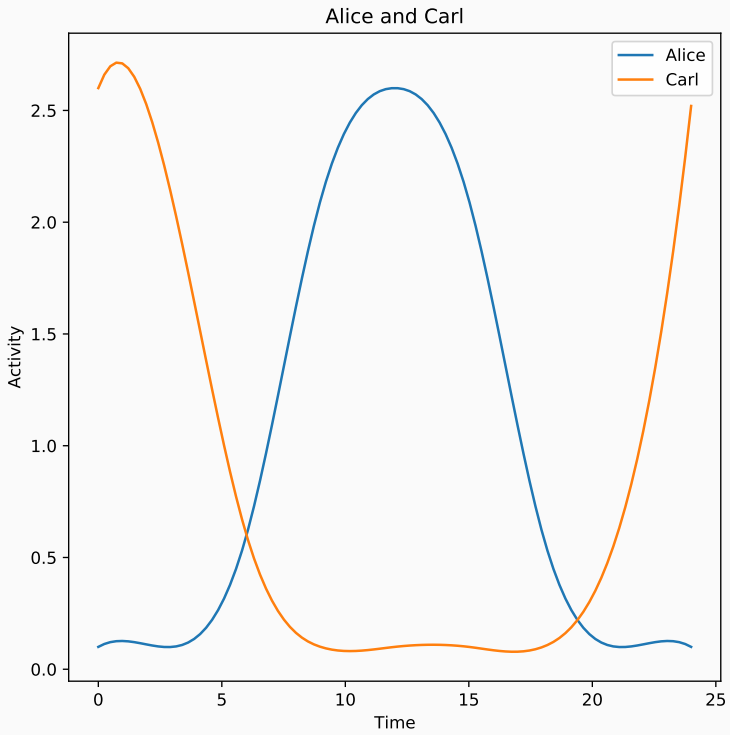
We make a compromise between performance and privacy:

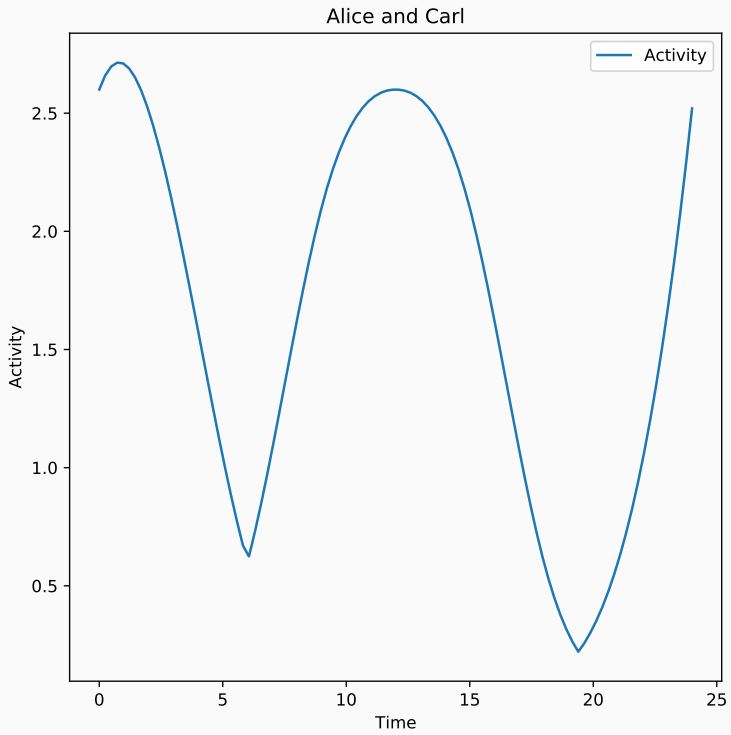
- Users are placed into ambiguity sets of size at least k, for some integer k.
- Each user in the set looks identical to every other user in the set from the adversary's point of view.
- We try to create sets to find a balance between performance and privacy.

Alice and Carl

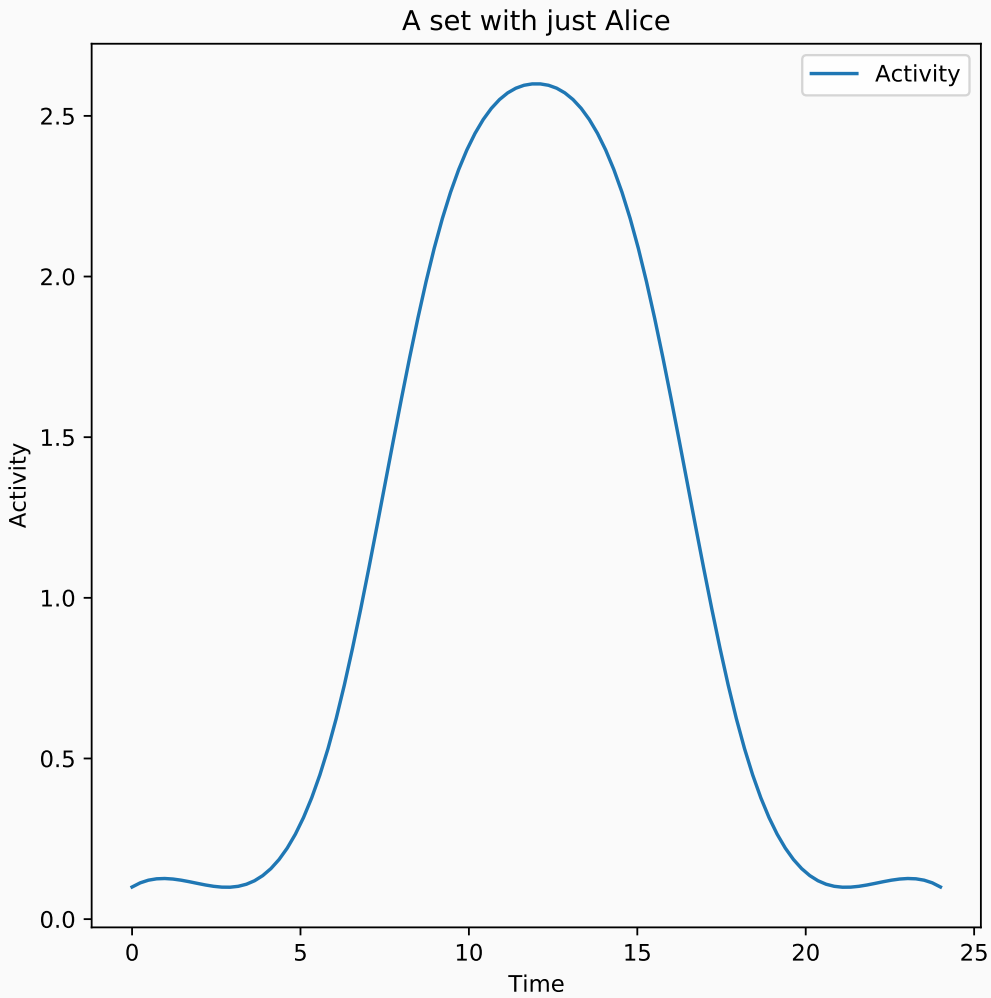


Alice and Carl





Alice and Carl



The Perfect Ambiguity Sets

- Minimizes unnecessary traffic

The Perfect Ambiguity Sets

- Minimizes unnecessary traffic
- Maintains good privacy

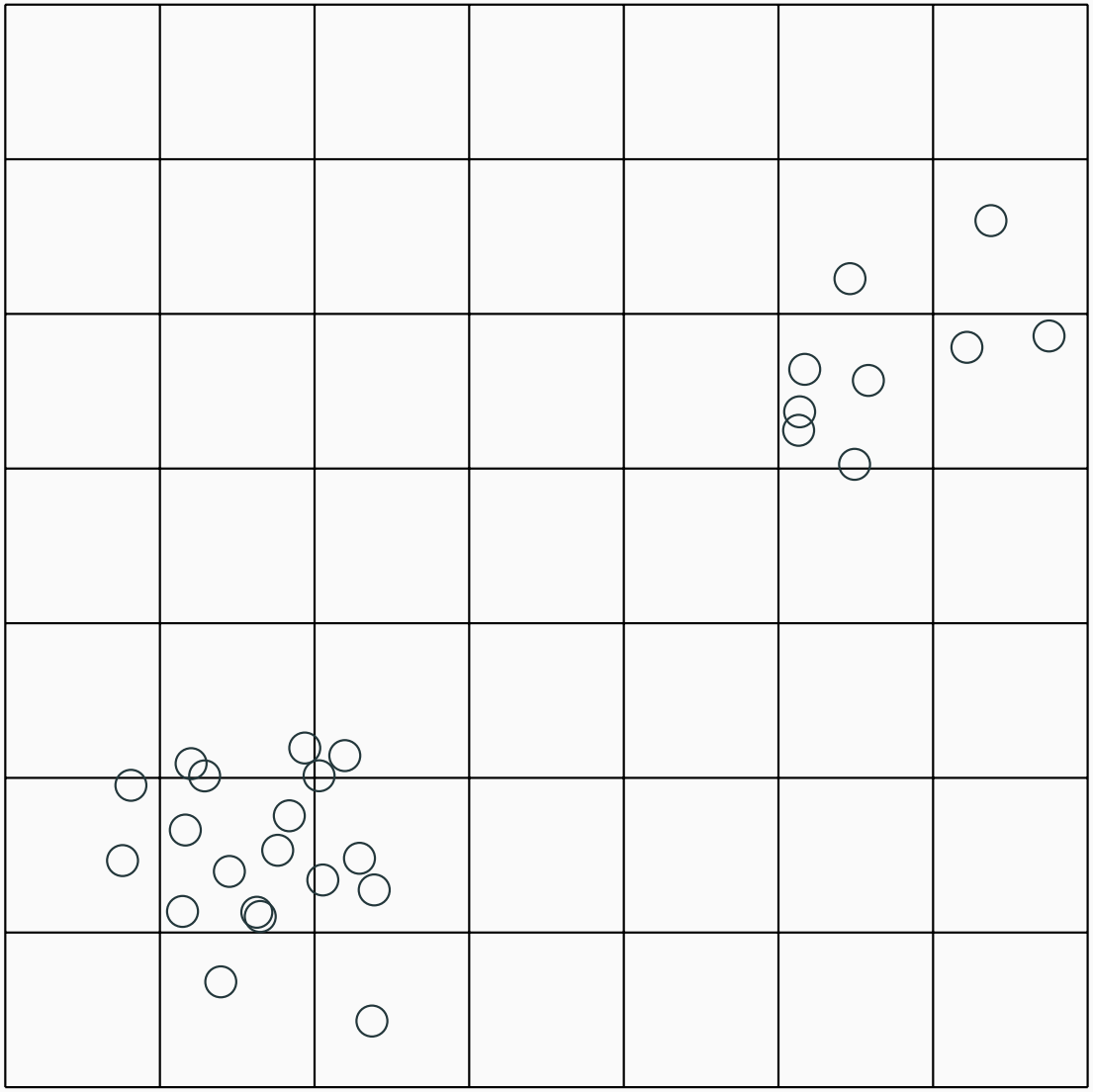
The Perfect Ambiguity Sets

- Minimizes unnecessary traffic
- Maintains good privacy

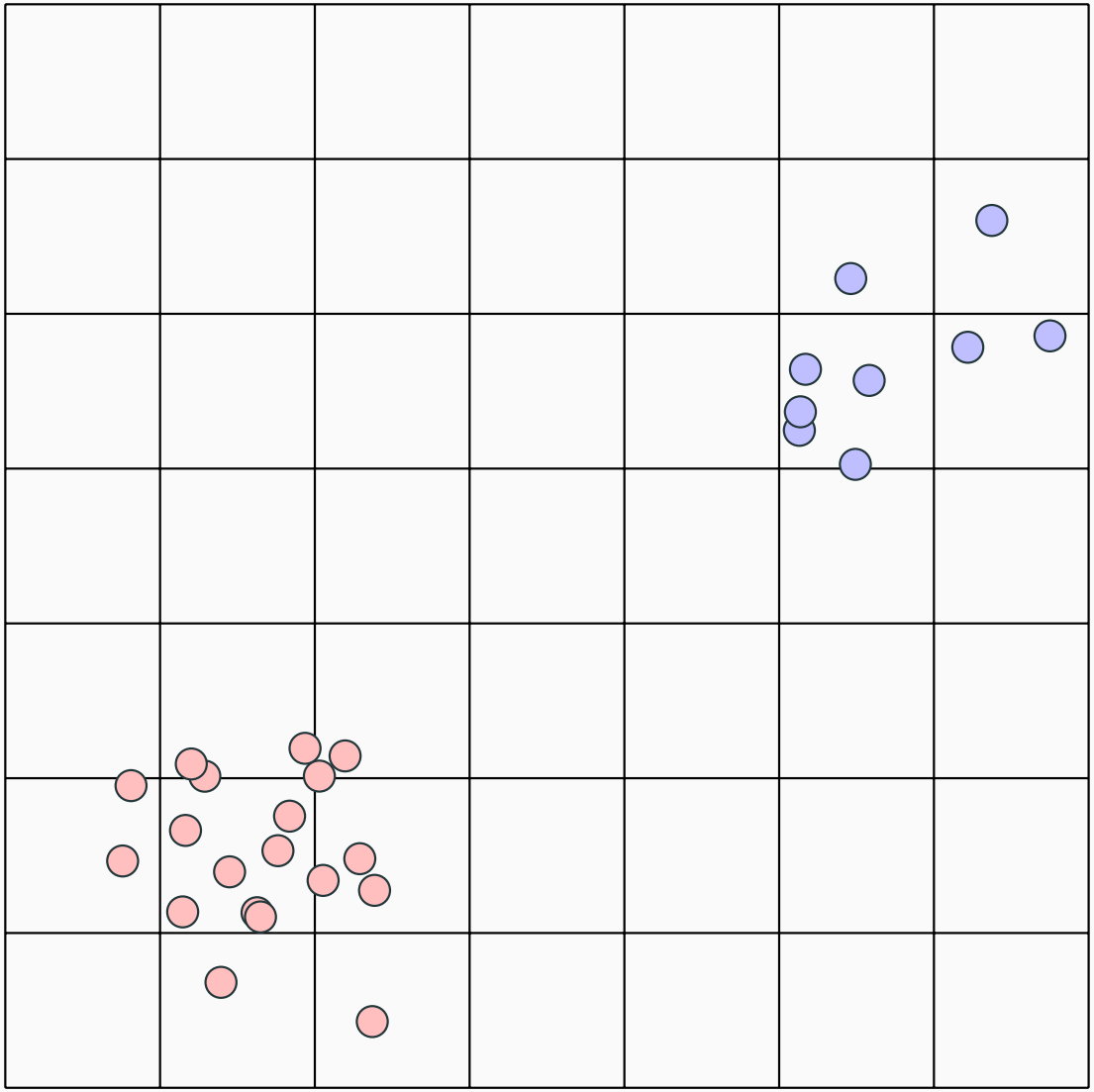
Achieving both of these at the same time is hard.

Creating Ambiguity Sets

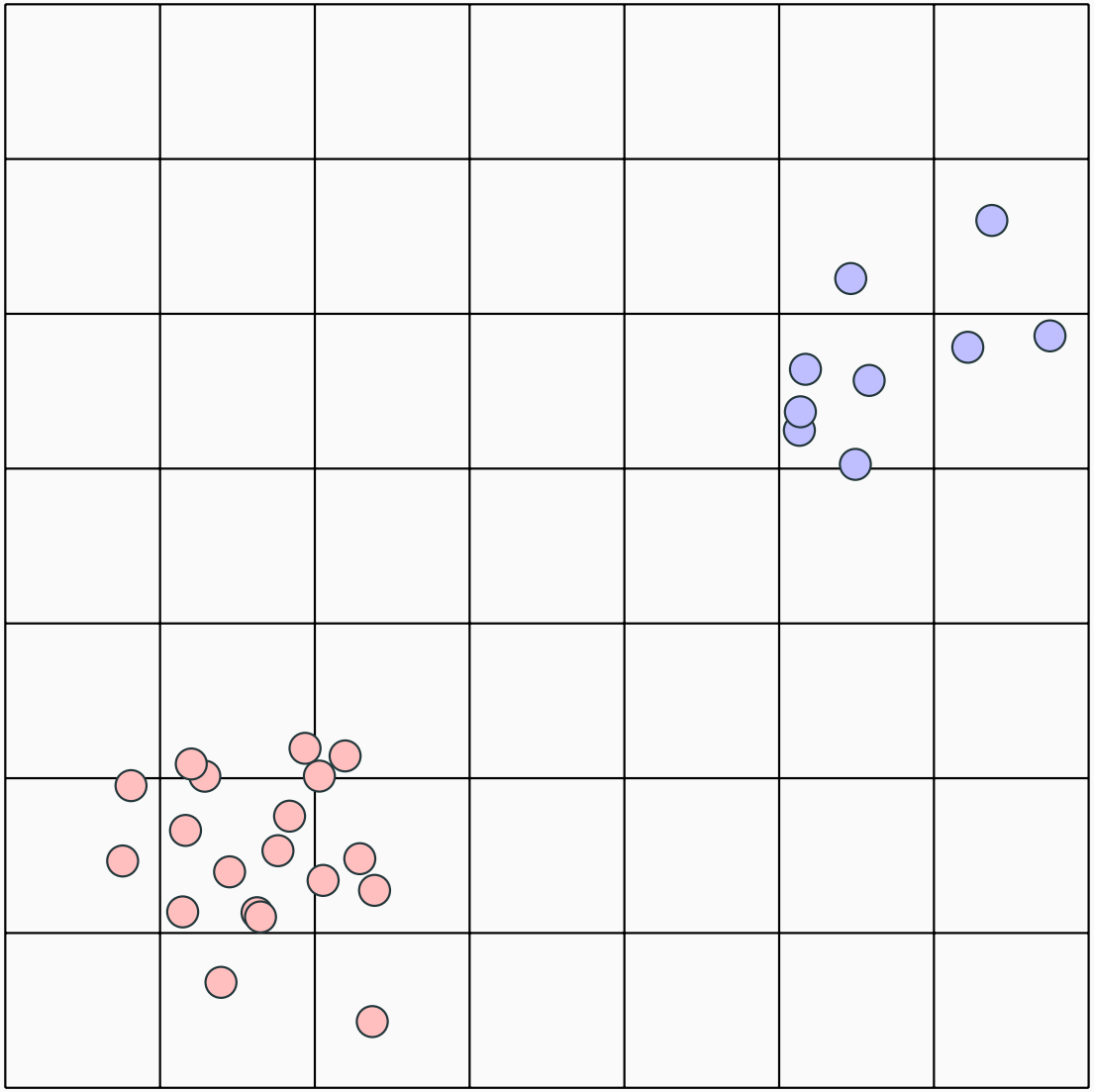
K-Means



K-Means

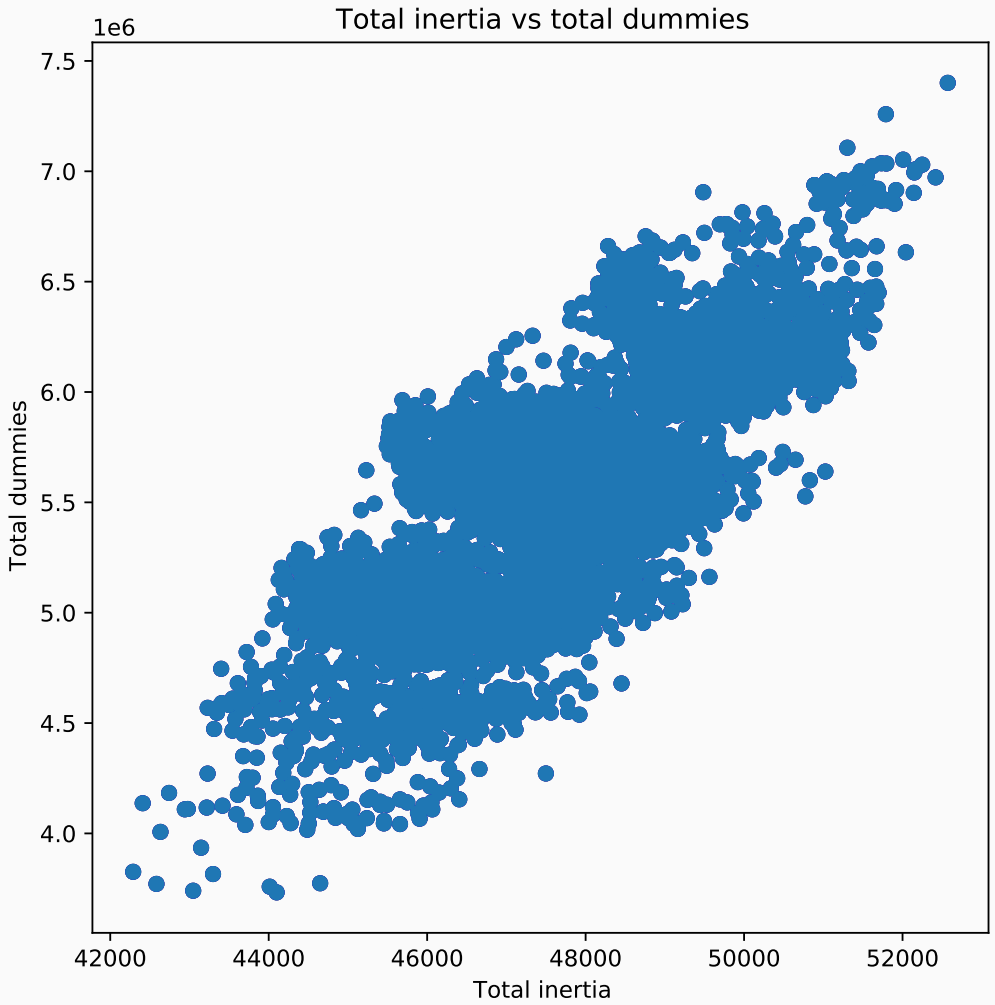


K-Means



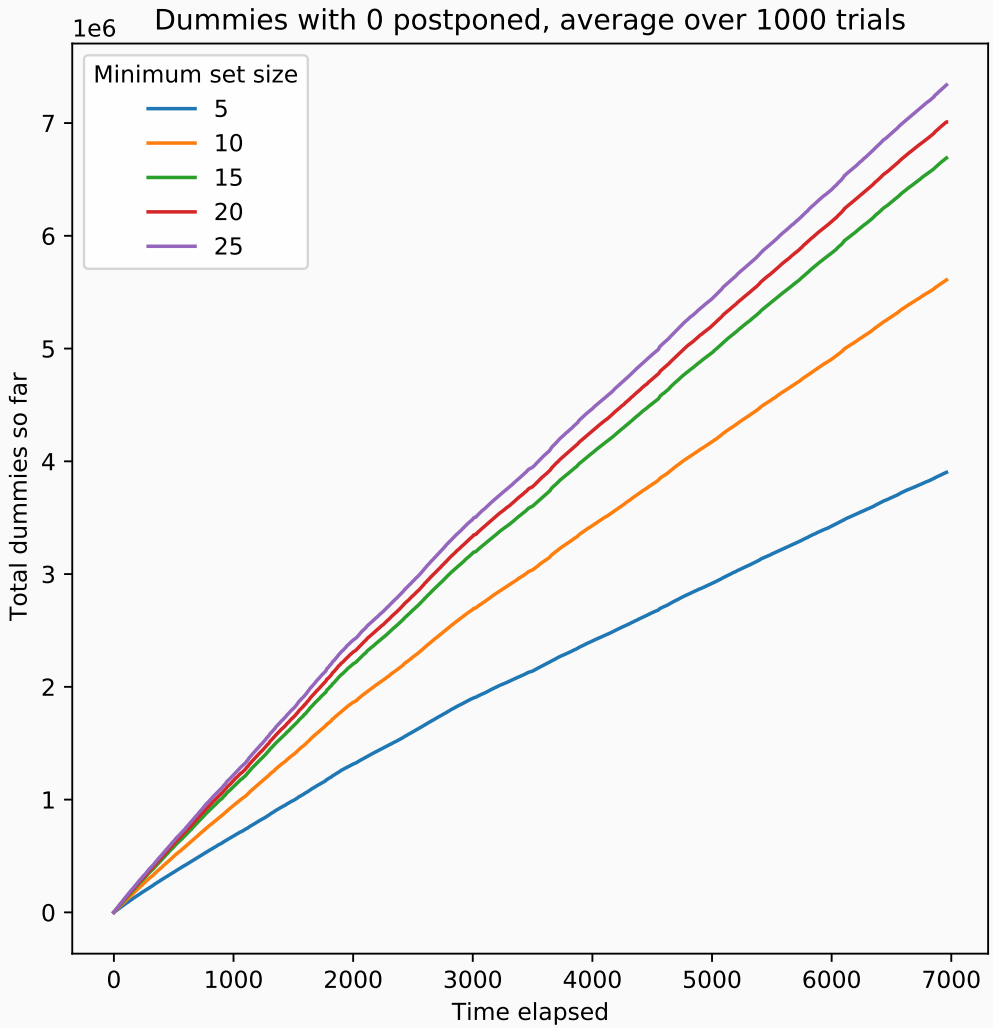
K-means attempts to minimize the **inertia** of each cluster.

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

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.

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.

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.

The Budget

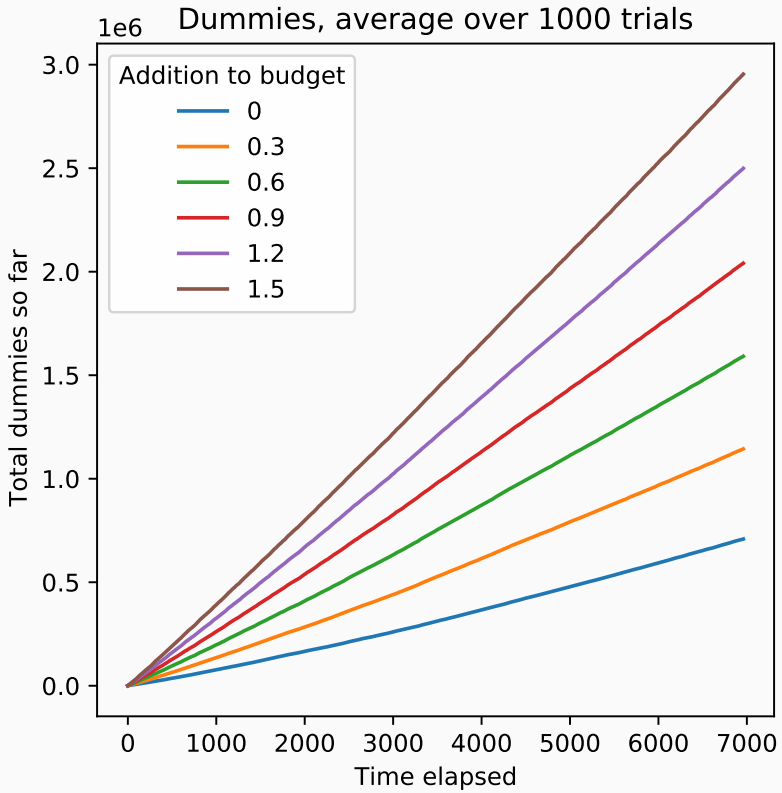
Once the ambiguity sets are created, we define a budget (how much traffic people should send) based on the mean activity over users in the set.

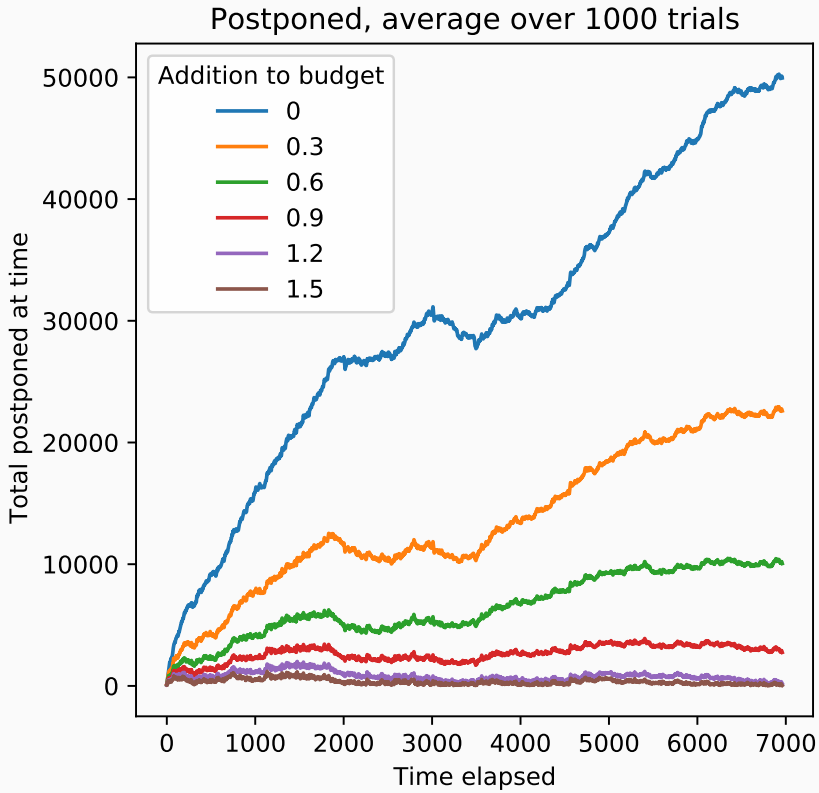
- If a user sends under the budget, they send dummy messages until the budget is reached.
- If a user sends over the budget, their messages are postponed to a later round.

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

The Solution

$$\text{budget} = \text{mean} \cdot (1 + \text{addition to budget})$$





Future Research

- Testing this on bigger datasets

Future Research

- 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!