Spatial Bloom Filters and Privacy Preservation!

By Raj Thimmiah

Example Problem

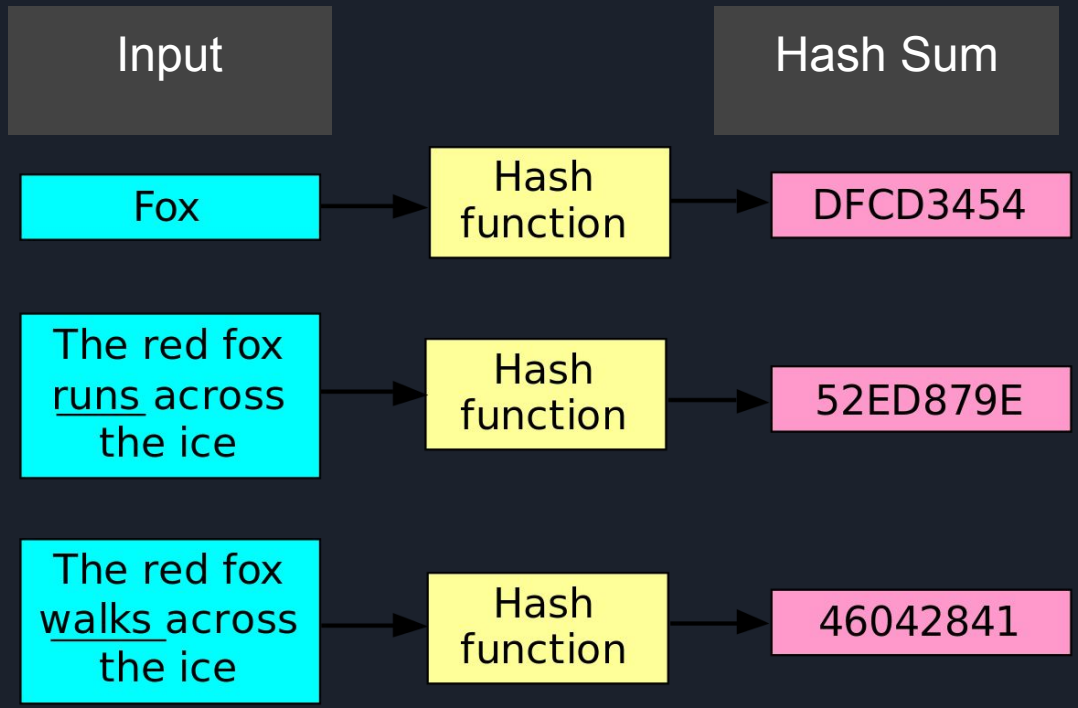
- Doing natural language processing on an online forum
- There are often spammed phrases or paragraphs that you want to omit
- As you receive new inputs, how can you efficiently filter out such unwanted text?
- Need some data structure where we can add text to a set *s* and test if an input text is in this set

Simple Item List

Size in Bits	Often Spammed Copyapasta
12312	What the fuck did you just fucking say about me, you little bitch? I'll have you know I graduated top of my class in the Navy Seals, and I've been involved in numerous secret raids on Al-Quaeda, and I have over 300 confirmed kills. I am trained in gorilla warfare and I'm the top sniper in the entire US armed forces. You are nothing to me but just another target. I will wipe you the fuck out with precision the likes of which has never been seen before on this Earth, mark my fucking words. You think you can get away with saying that shit to me over the Internet? Think again, fucker. As we speak I am contacting my secret network of spies across the USA and your IP is being traced right now so you better prepare for the storm, maggot. The storm that wipes out the pathetic little thing you call your life. You're fucking dead, kid. I can be anywhere, anytime, and I can kill you in over seven hundred ways, and that's just with my bare hands. Not only am I extensively trained in unarmed combat, but I have access to the entire arsenal of the United States Marine Corps and I will use it to its full extent to wipe your miserable ass off the face of the continent, you little shit. If only you could have known what unholy retribution your little "clever" comment was about to bring down upon you, maybe you would have held your fucking tongue. But you couldn't, you didn't, and now you're paying the price, you goddamn idiot. I will shit fury all over you and you will drown in it. You're fucking dead, kiddo.
10168	To be fair, you have to have a very high IQ to understand Rick and Morty. The humor is extremely subtle, and without a solid grasp of theoretical physics most of the jokes will go over a typical viewer's head. There's also Rick's nihilistic outlook, which is deftly woven into his characterisation - his personal philosophy draws heavily from Narodnaya Volya literature, for instance. The fans understand this stuff; they have the intellectual capacity to truly appreciate the depths of these jokes, to realize that they're not just funny- they say something deep about LIFE. As a consequence people who dislike Rick and Morty truly ARE idiots- of course they wouldn't appreciate, for instance, the humour in Rick's existencial catchphrase "Wubba Lubba Dub Dub," which itself is a cryptic reference to Turgenev's Russian epic Fathers and Sons I'm smirking right now just imagining one of those addlepated simpletons scratching their heads in confusion as Dan Harmon's genius unfolds itself on their television screens. What fools... how I pity them. And yes by the way, I DO have a Rick and Morty tattoo. And no, you cannot see it. It's for the ladies' eyes only- And even they have to demonstrate that they're within 5 IQ points of my own (preferably lower) beforehand.
5480	hey, sorry i saw your profile and i just thought you looked cute in your picture. i really wanted to tell you that)) It's really rare to see girls playing video games haha! I don't know why it's a guy thing honestly im like really against misogyny and like ill be the one in the kitchen making sandwiches. We should really play l4d2 sometime its a really cool zombie game with a lot of scary moments, but don't worry ill be there to protect you sorry that wasnt flirtring i swear im just trying to be friendly i really like your profile picture sorry was that too far? Really sorry i'm really shy i don't go out much haha add me on skype we should talk you look really nice and fun xxx
13304	I have scored between 151 and 167 on IQ tests. I have been educated to masters degree level and technically qualify as a genius so far as such things can be reliably measured I lack focus, so I am particularly interested in most things from the various technical, physical, psychological and evolutionary aspects of space colonisation to efficient forms of government and the process of making cheese taste good. I am fairly damned close to the whole know it all, doctor scientist guy being described here. Which is to say I can probably talk for about a minute on most topics before I run out of truth and start making shit up or just assuming shit. Some one with 40 IQ points less than me can trump me on knowledge by taking an interest in a topic and studying it in highschool, Don't get me wrong... Ifl find a topic interesting I can read up on it and absorb concepts far faster than most people, maybe even enough to stump an expert on a point or two (before they have time to think about or look for an answer). Intellect is an attribute, like height. Some people can reach high shelves, others can grasp esoteric theories. Neither makes you amazing at mountain climbing on their own unless you practice it and learn to understand it. That all said... memorising and understanding a wide array of stuff DOES make it easier to memorise and understand even more stuff and that means that you have a much wider basis to figure out even more shit when you have to. Which seems like magic genius superbrain stuff to most people... but I'd still rather an actual medical doctor dealt with my health rather than trying to work shit out on my own from the internet.

41,264 Bits

Hash Function



https://simple.wikipedia.org/wiki/Hash_function

Hash Table

- Store each item in set *s* as an *m* bit hash
- To add a new item, just add the hash
- To see if an input is in the item table, do h(input) and check if it matches any $h(x_s)$

http://www.michaelnielsen.org/ddi/why-bloom-filters-work-the-way-they-do/

Simple Hash Table (SHA256)

Size in Bits	Hash of Text to Filter
256	afd7f3c9b24aba22075cb07e615d47856c539a58d08479781d5cafd4a60a6d9f
256	5c5d438666f0c21db517e448f891d9175b9197bb6780fca4c65573b394de055d
256	a023061243860 92d69350075b3af13bfeba2c16b656403c065e252d82437aad1
256	8d3fcdf5810de8b67791c9b8a61465580ee1ff55c875bf98785cb27c91f90cf5

http://www.michaelnielsen.org/ddi/why-bloom-filters-work-the-way-they-do/

1,024 Bits

Around 40 times better than the original

Probability of Hash Collisions

- p = probability of false positive, m is number of bits per hash, S is size of the hash

$$p = 1 - (1 - 1/2^m)^{|S|}.$$

- 1 - (1 - ½^256)^4 = ~3.4 x 10^-77 probability of a hash collision and reporting an item as being a member of the set incorrectly
- Function to calculate number of bits needed to achieve ideal probability of false positive:

$$\# = |S| \log \frac{1}{1-(1-p)^{1/|S|}}.$$

- log(1/(1-(1-.01)^1/4)
- Based on this, we need ~9 bit hash to achieve 1% probability of false positive, meaning a total size of 112 bits (around 9 times smaller than previous hash table), 17 bits for a size of s = 1,000

(For full derivation check the link in the bottom right)

http://www.michaelnielsen.org/ddi/why-bloom-filters-work-the-way-they-do/

Bit Array Lookup Scheme for Integers

- If you want to store membership of a set of integers between 1 and 10 you could store the integers themselves or you could make a 10 bit array and flip the bits representing the numbers in the set to 1, something like this for the numbers 3, 8, 9, and 10:

0	0	1	0	0	0	0	1	1	1

http://www.michaelnielsen.org/ddi/why-bloom-filters-work-the-way-they-do/

General Bit Array Lookup Scheme: Adding Elements

http://www.michaelnielsen.org/ddi/why-bloom-filters-work-the-way-they-do/

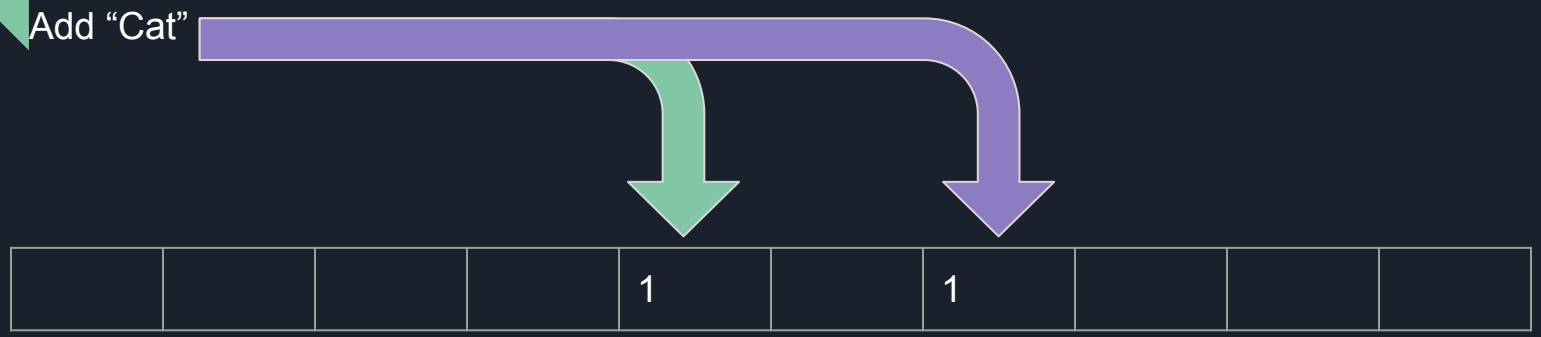
General Bit Array Lookup Scheme: Adding Elements

Add “Dog”

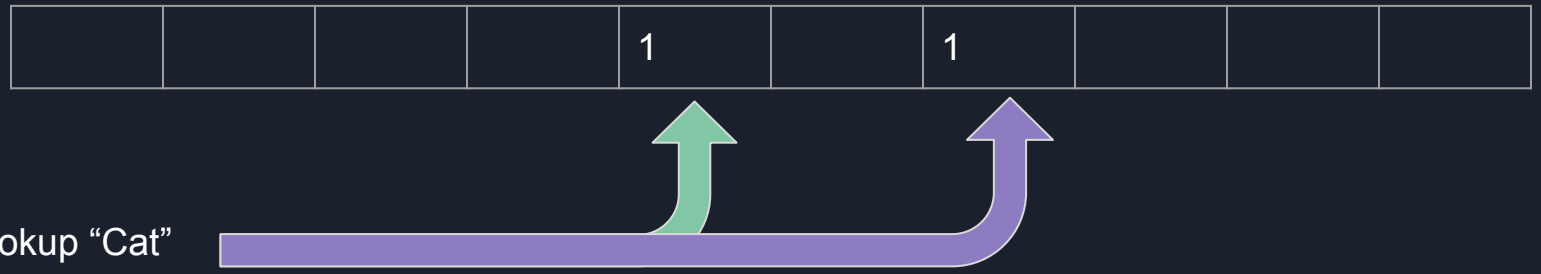
General Bit Array Lookup Scheme: Lookup

General Bit Array Lookup Scheme: Lookup

General Bit Array Lookup Scheme: Adding Elements



General Bit Array Lookup Scheme: Adding Elements



Cool Demo!

Summary

- A bloom filter can determine conclusively if an element is *not* a member of a set
- Can tell us if an element *is* a member with a small probability of a false positive

Summary

- A bloom filter can determine conclusively if an element is *not* a member of a set
- Can tell us if an element *is* a member with a small probability of a false positive

Algorithm

#Note:
b# means a Bloom Filter
Δ = the entire set of elements
δ = an individual element
H = set of hash functions used
h = single hash function from that set

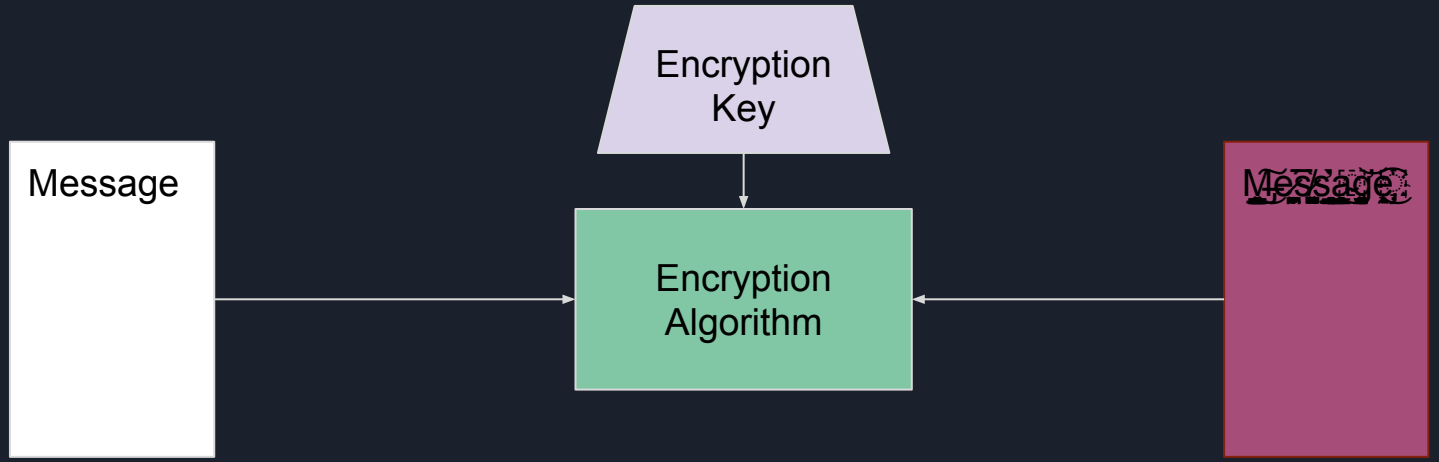
```
# Algorithm: Bloom Filter Construction
Input: Δ, H; # input all elements of the set, all hash functions to be used
Output: b#; #output the end bloom filter

foreach δ ∈ Δ do
    foreach h ∈ H do
        b#[h (δ)] ← 1;
    end
end
return b#;
```

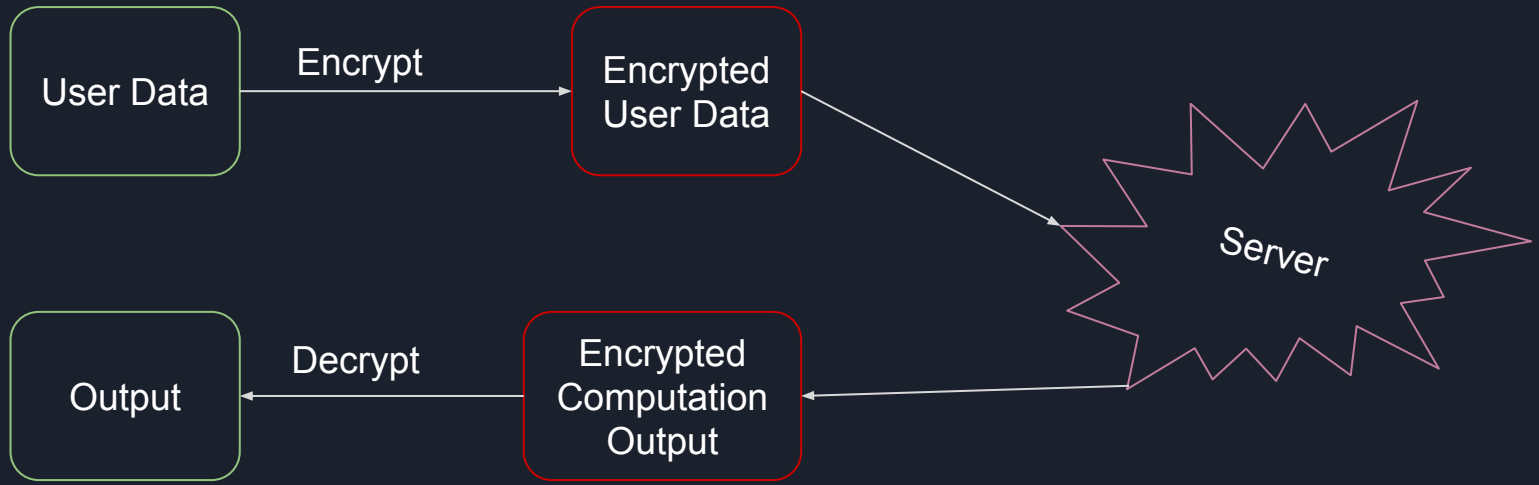
```
# Algorithm: Bloom Filter Lookup
Input: b#, H, δu, s;
Output: membership;
foreach h ∈ H do
    if b# [h (δu)] = 0 then
        return false;
end
return true
```

Homomorphic Encryption

What does encryption do?



What does homomorphic encryption do?



Spatial Bloom Filters

https://eprint.iacr.org/2014/531.pdf

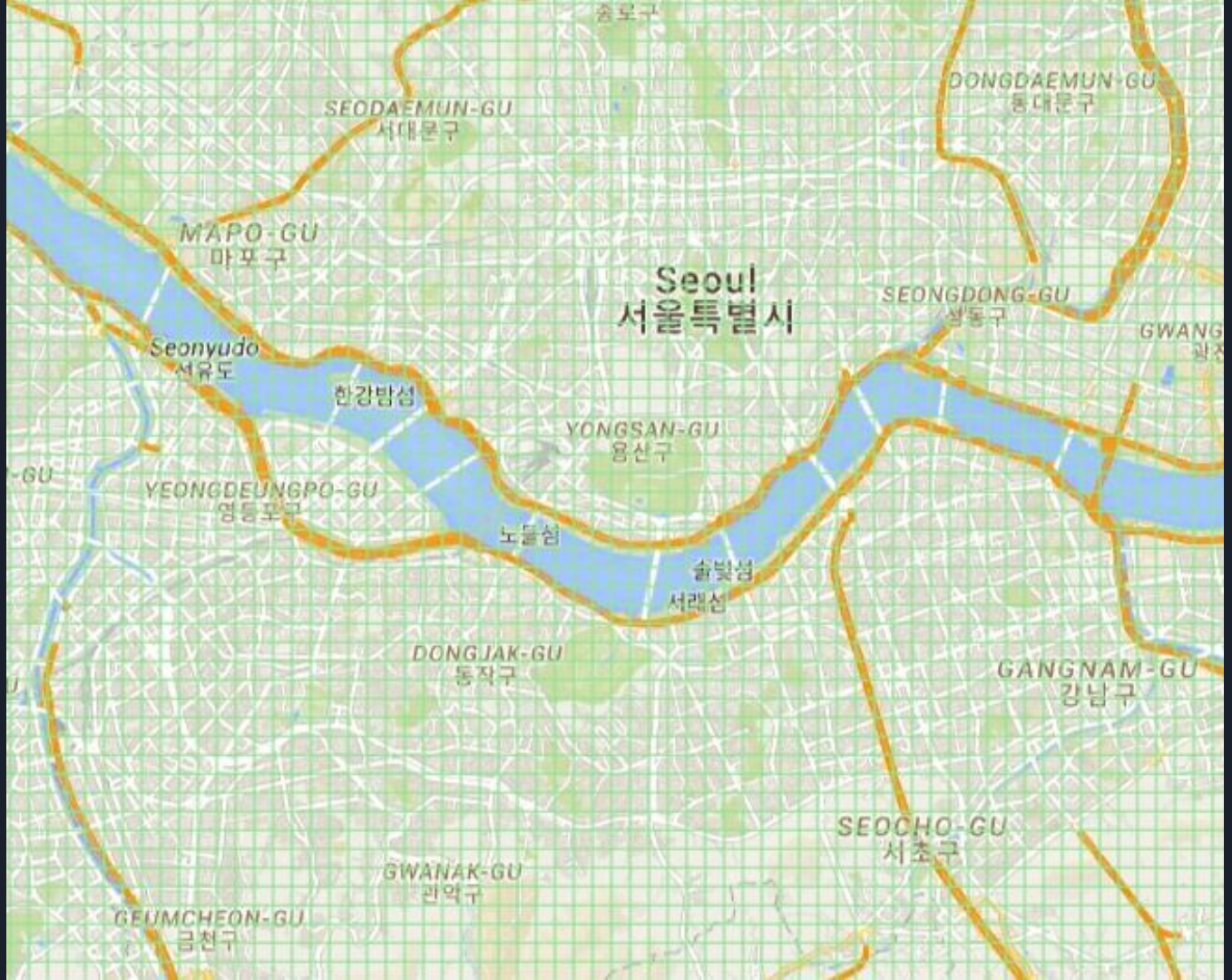
Bloom Filters vs. Spatial Bloom Filters

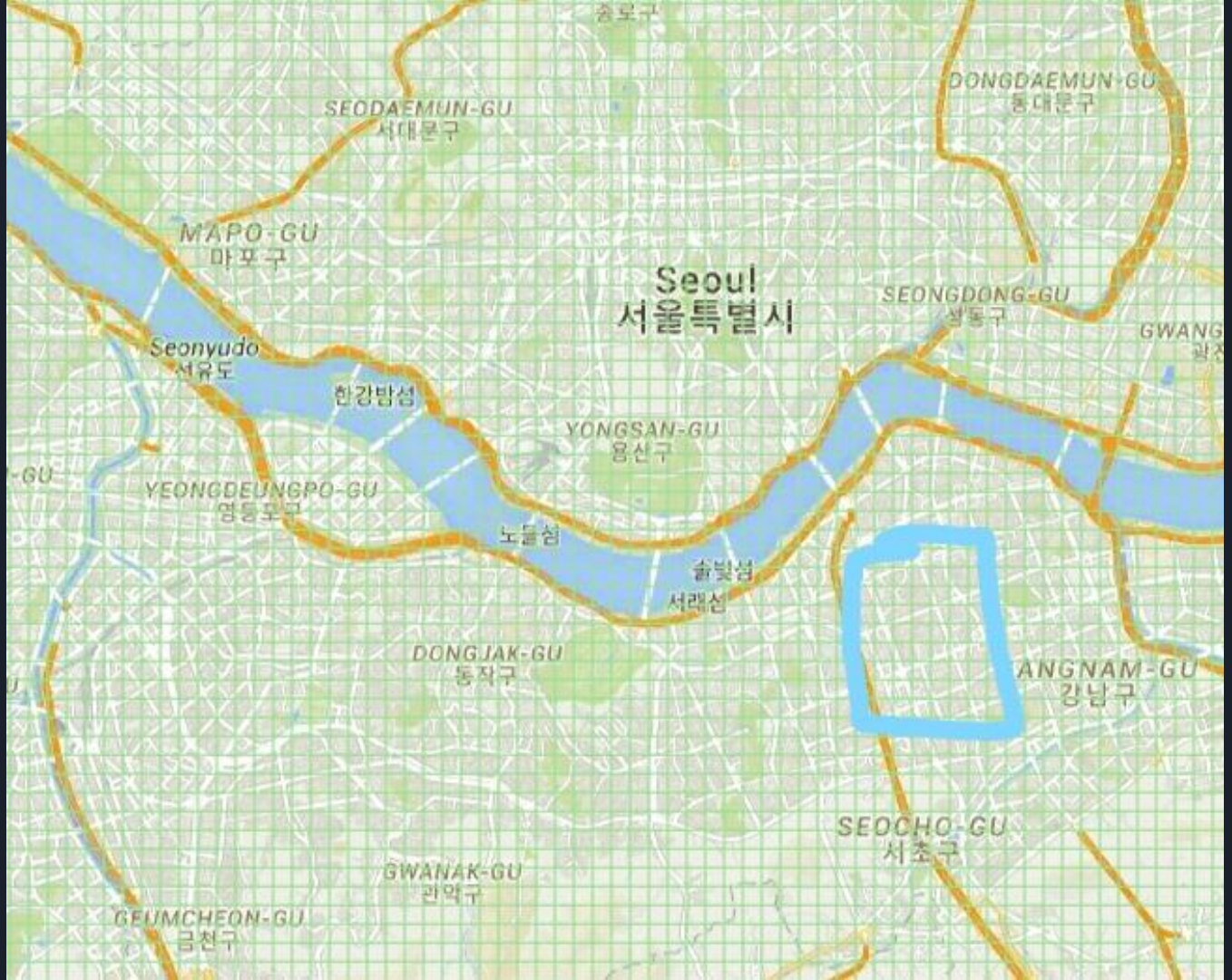
- Bloom Filters can be can be used to lookup the membership of an element in a single set
- Spatial Bloom Filters (SBF) can represent an arbitrary number of sets and can return which set an element belongs to, if any

Original Use Case

- Want to see when a user is in the vicinity of the area of our meetup

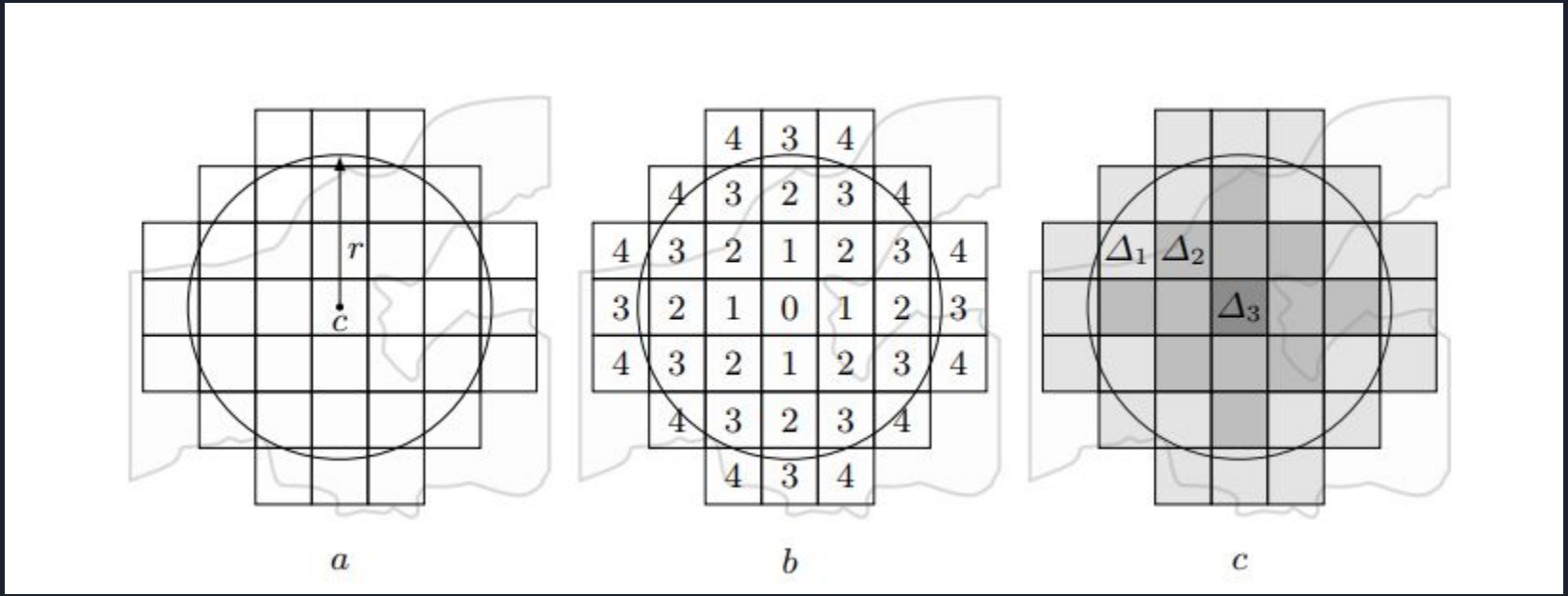

종로구
SEODAEMUN-GU
서대문구
DONGDAEMUN-GU
동대문구
MAPO-GU
마포구
Seoul
서울특별시
SEONGDONG-GU
성동구
Seonyudo
선유도
한강밤섬
YONGSAN-GU
용산구
YEONGDEUNGPO-GU
영등포구
노들섬
솔빛섬
서래섬
DONGJAK-GU
동작구
GANGNAM-GU
강남구
SEOCHO-GU
서초구
GWANAK-GU
관악구
GEUMCHEON-GU
금천구


종로구
SEODAEMUN-GU
서대문구
DONGDAEMUN-GU
동대문구
MAPO-GU
마포구
Seoul
서울특별시
SEONGDONG-GU
성동구
GWANG
Seonyudo
선유도
한강합성
YONGSAN-GU
용산구
YEONGDEUNGPO-GU
영등포구
노들섬
솔빛섬
서래섬
DONGJAK-GU
동작구
GANGNAM-GU
강남구
SEOCHO-GU
서초구
GWANAK-GU
관악구
GEUMCHEON-GU
금천구


종로구
SEODAEMUN-GU
서대문구
DONGDAEMUN-GU
동대문구
MAPO-GU
마포구
Seoul
서울특별시
SEONGDONG-GU
성동구
GWANG
Seonyudo
선유도
한강탐섬
YONGSAN-GU
용산구
YEONGDEUNGPO-GU
영등포구
노들섬
솔빛섬
서래섬
DONGJAK-GU
동작구
ANGNAM-GU
강남구
SEOCHO-GU
서초구
GWANAK-GU
관악구
GEUMCHEON-GU
금천구

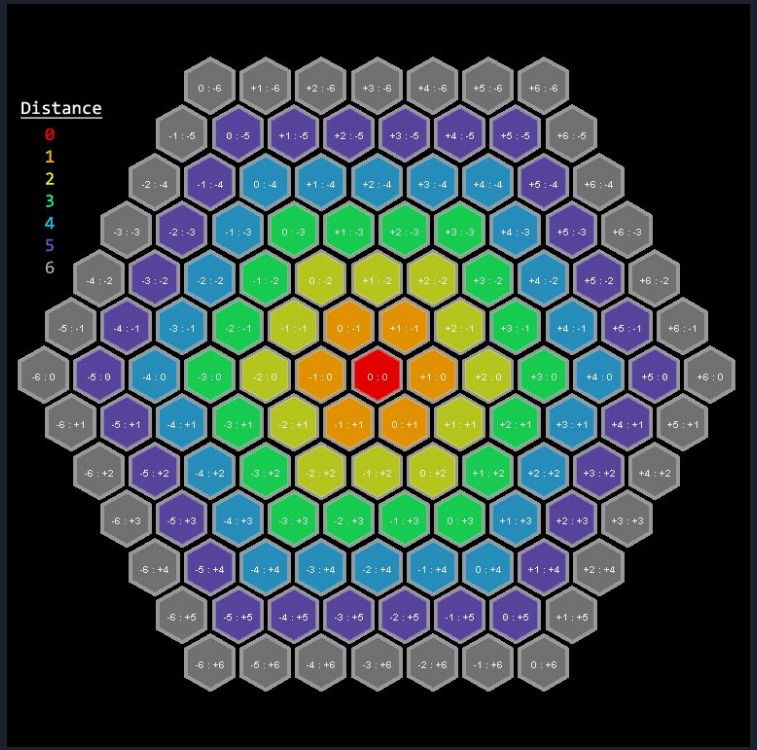
Can represent only one set of areas!

Point of Interest



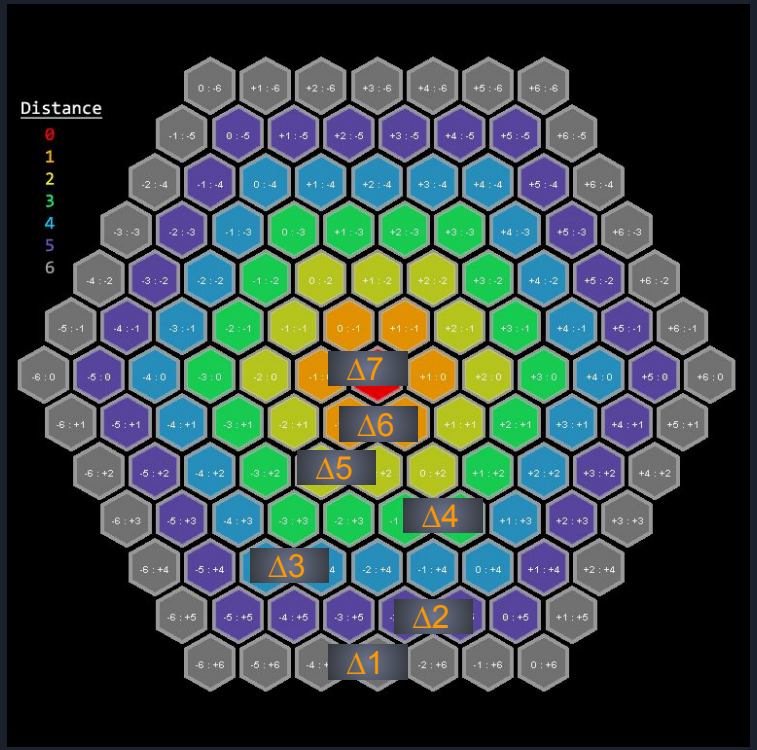
https://eprint.iacr.org/2014/531.pdf

Spatial Bloom Filter Construction



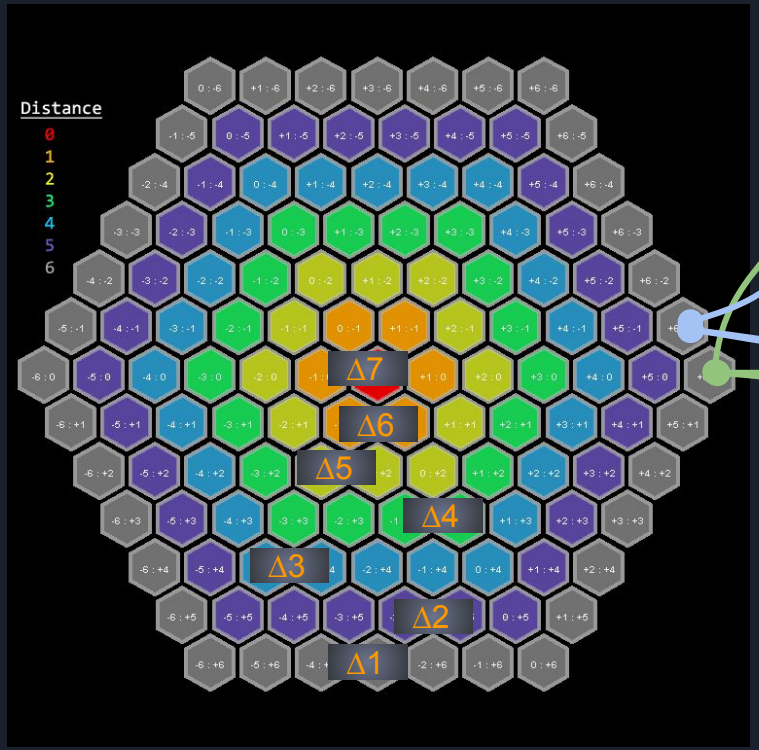
https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

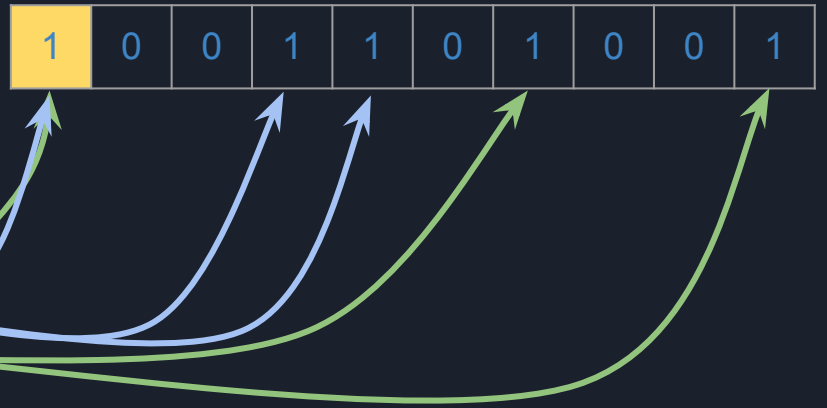
Spatial Bloom Filter Construction



https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

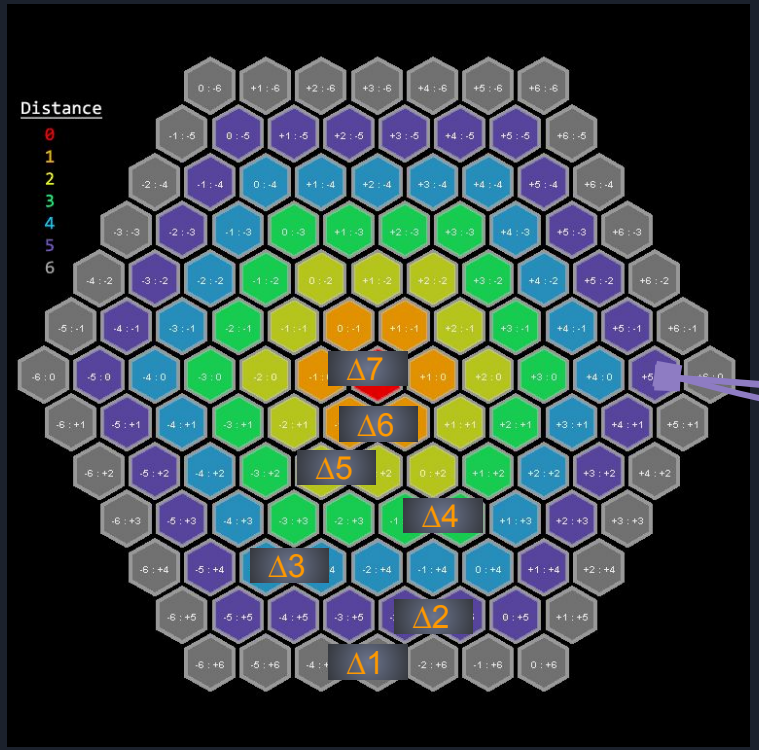
Spatial Bloom Filter Construction





https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

Spatial Bloom Filter Construction



1	0	2	1	1	2	1	0	0	2

https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

Spatial Bloom Filter Lookup Errors

https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

Spatial Bloom Filter Lookup Errors

https://stackoverflow.com/questions/38015645/a-pathfinding-in-a-hexagonal-grid

Summary of Flaws

- Same possibility for saying an element from outside the set is from inside the set
- Also probability of one set being mistaken for another
  - Dependent on order of set encoding
- These probabilities can again be tuned depending on the accuracy you want

Algorithm

Δ_i = a specific set
b# = Bloom Filter
s = number of sets
δ = a set's element
H = set of hash functions
h = individual hash function
Δ_u = input element

Algorithm 2: Spatial Bloom Filter construction.

Input: $\Delta_1, \Delta_2, \ldots, \Delta_s, H;$
Output: $b^{\#};$

```
1 for i ← 1 to s do
2 |   foreach δ ∈ Δ_i do
3 |   |   foreach h ∈ H do
4 |   |   |   b#[h(δ)] ← i;
  |   |   end
  |   end
  end
5 return b#;
```

Algorithm 3: Spatial Bloom Filter verification.

Input: $b^{\#}, H, \delta_u, s;$
Output: $\Delta_i;$

```
1 i = s;
2 foreach h ∈ H do
3 |   if b#[h(δ_u)] = 0 then
4 |   |   return false;
  |   else
5 |   |   if b#[h(δ_u)] < i then
6 |   |   |   i ← b#[h(δ_u)];
  |   |   end
  |   end
  end
7 return Δ_i;
```

Spatial Bloom Filters

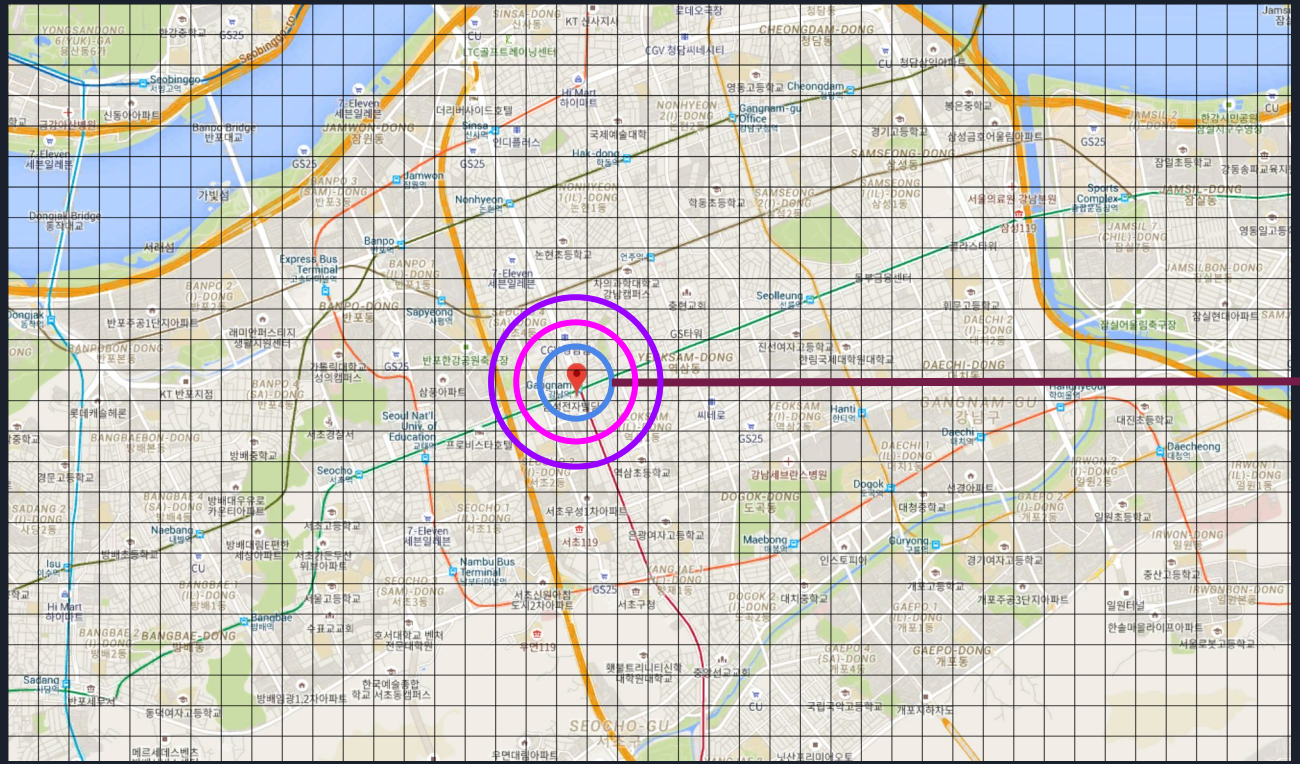
What we want to do to judge distance is create multiple concentric circles around a point of interest, each concentric circle's area being considered a separate set.

We want to

Alerting Martin to Nearby Friends

- Martin wants to be notified if any of his friends are near him
- He doesn't want them to know his exact location
- They don’t want him to know their exact location
- They still both want to know if they are near each other

Spatial Bloom Filter Construction



SBF Encryption

enc(Spatial Bloom Filter) = 

Sharing and Processing Encrypted Bloom Filter

Spatial Bloom Filter + my_location = enc(result)

Result

dec(result) = result

Location Privacy Applications

- Tell 2 users when they are in the vicinity of each other but otherwise give no information
- Allow for location based services (ads, restaurant recommendations, etc.)without violating user privacy
- Track congestion by seeing if users are in an area with high traffic

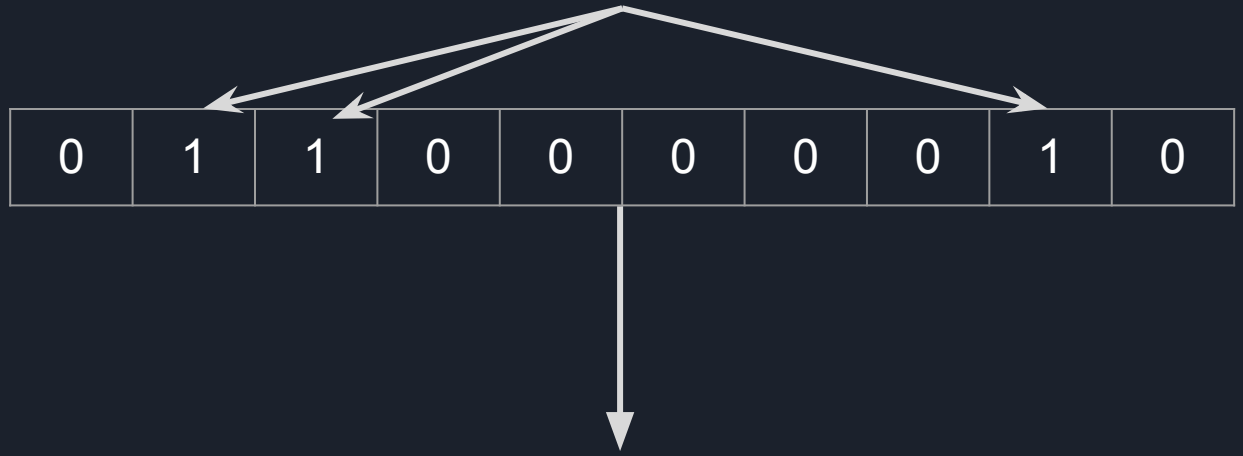
Other Applications

Alternate Password Scheme

- Server side password verification/authentication is risky
- With Bloom Filters, can offload verification/authentication to users

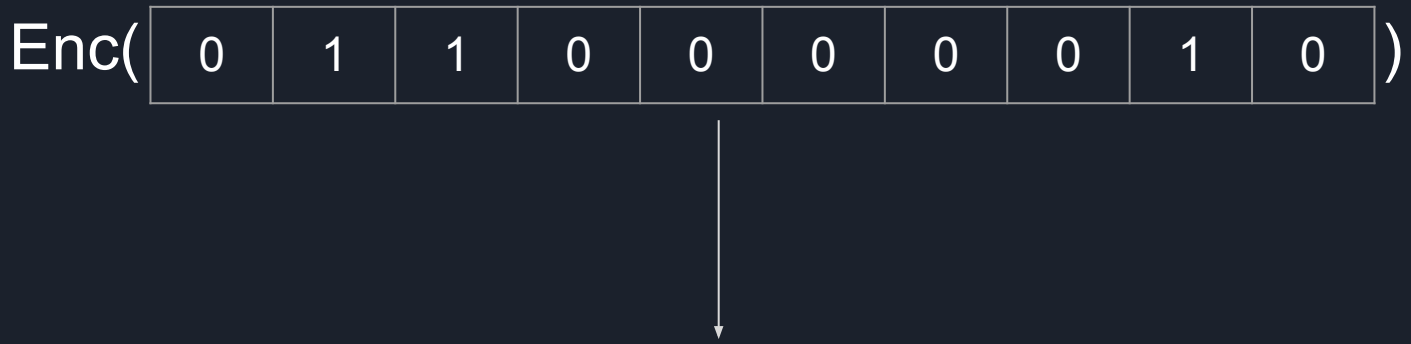
User Generates Bloom Filter of their Password

User Computes: “Password”

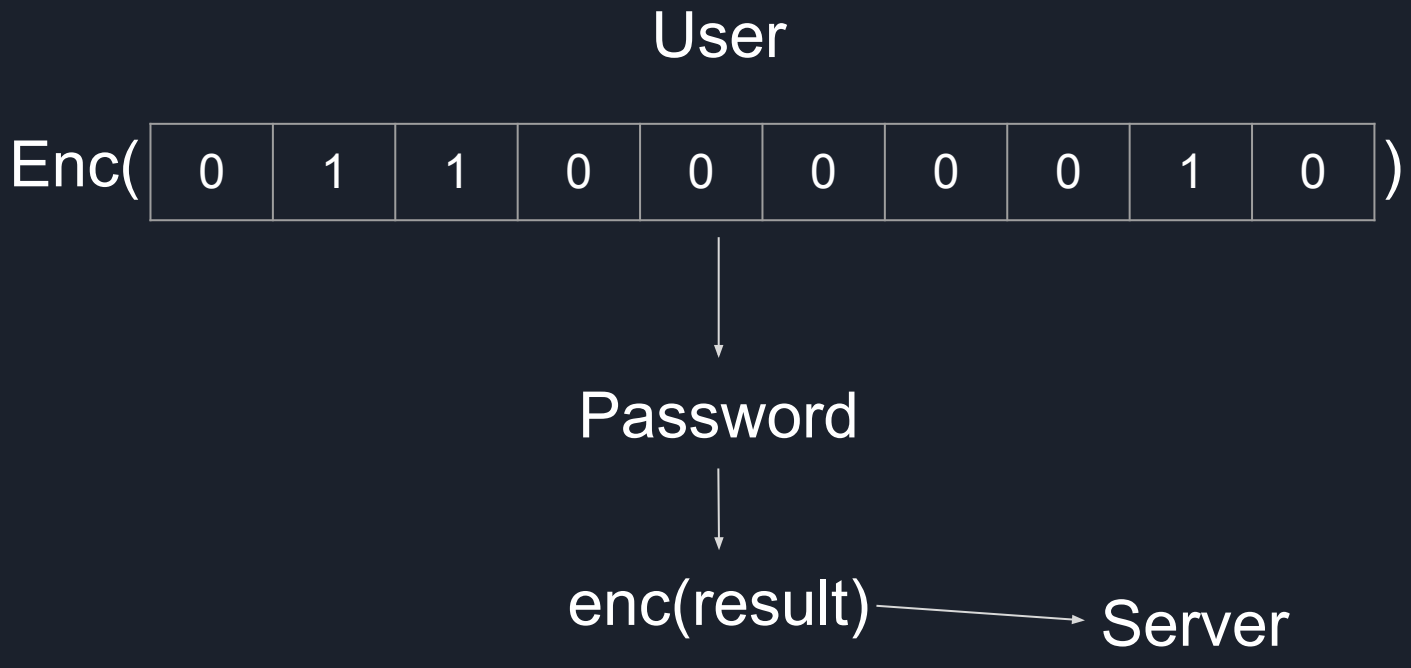


Future Password Checking by Server

Server creates:



Future Password Checking by Server



Future Password Checking by Server

enc(result)

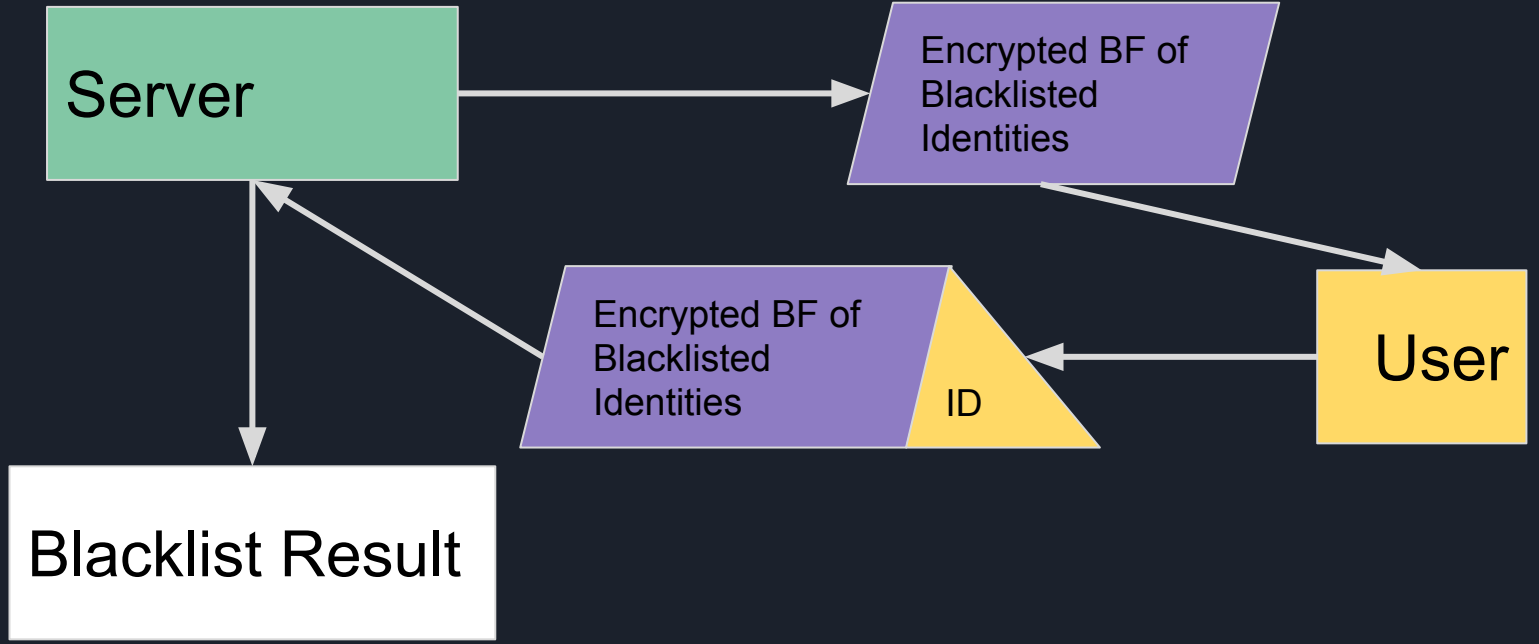
result

Similar Applications

- Could have sets of users with different privileges
- Could hide which user logged in while still providing appropriate access using a SBF

Identify Blacklisted Users

- Document validation done by human, blacklist verification done by database
- Insecure database can leak blacklist
- Insecure to upload user’s identities to check against blacklist


Server
Encrypted BF of Blacklisted Identities
User
Encrypted BF of Blacklisted Identities
ID
Blacklist Result

Smart Contract Use Case

- A smart contract should send the first 500 people who send it a token a different token in exchange

160 Bits	0x281055afc982d96fab65b3a49cac8b878184cb16
160 Bits	0x6f46cf5569aefa1acc1009290c8e043747172d89
160 Bits	0xf4b51b14b9ee30dc37ec970b50a486f37686e2a8
160 Bits	0xf27daff52c38b2c373ad2b9392652ddf433303c4

Smart Contract Use Case

160 Bits	0x281055afc982d96fab65b3a49cac8b878184cb16
160 Bits	0x6f46cf5569aefa1acc1009290c8e043747172d89
160 Bits	0xf4b51b14b9ee30dc37ec970b50a486f37686e2a8
160 Bits	0xf27daff52c38b2c373ad2b9392652ddf433303c4

Demo

Further Applications

- An Anonymous Inter-Network Routing Protocol for the Internet of Things

Recommended

- Why Bloom filters work the way they do
- A Gentle Introduction to Bloom Filter
- Spatial Bloom Filters: Enabling Privacy in Location-aware Applications
- Probabilistic Properties of the Spatial Bloom Filters and Their Relevance to Cryptographic Protocols
- Author’s Website on SBF

Interesting Alternatives

- The Bloomier Filter: An Efficient Data Structure for Static Support Lookup Tables *
- Cuckoo Filter
- Scalable Bloom Filter
- BF with deletions