Privacy Guarantees of Aggregation in Gossip Protocols

A Project Report

Submitted for Minor Project I of 6th Semester for partial fulfilment of the requirements for the award of the degree of Bachelors of Technology in Computer Science and Engineering.

Submitted By :

Group Number : 45

Tejash Ranjan	2006007
Abhishek Kumar	2006035
RaviRanjan Kumar	2006013

Under the Supervision of

Dr. Antriksh Goswami

Asst. Professor CSE Dept, NIT Patna

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, PATNA

May 2023

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

CERTIFICATE

This is to certify that TEJASH RANJAN Roll No. 2006007, ABHISHEK KUMAR Roll No. 2006035, RAVIRANJAN KUMAR Roll No. 2006013 has carried out the Minor project entitled as “Privacy Guarantees of Aggregation in Gossip Protocols” during their 6th semester under the supervision of Dr. Antriksh Goswami, Assistant Prof., CSE Department in partial fulfilment of the requirements for the award of Bachelor of Technology degree in the Department of Computer Science and Engineering, National Institute of Technology Patna.

....

Dr. Antriksh Goswami

Assistant Professor

CSE Department

NIT Patna

……………………

Dr. M.P. Singh

Head of Department

CSE Department

NIT Patna

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION

We students of the 6th semester hereby declare that this project entitled “Attention loss driver fatigue detection system” has been carried out by us in the Department of Computer Science and Engineering of National Institute of Technology Patna under the guidance of Dr. Antriksh Goswami Department of Computer Science and Engineering, NIT Patna. No part of this project has been submitted for the award of a degree or diploma to any other Institute.

Name Signature

1. TEJASH RANJAN ……………………….

2. ABHISHEK KUMAR ……………………….

3. RAVIRANJAN KUAMR ……………………….

Place: NIT Patna Date:……………………….

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

ACKNOWLEDGEMENT

We would like to acknowledge and express my deepest gratitude to my mentor Dr. Antriksh Goswami, Assistant Professor, Computer Science and Engineering Department, National Institute of Technology Patna for the valuable guidance, sympathy, and co-operation for providing necessary facilities and sources during the entire period of this project.

We wish to convey our sincere gratitude to the Head of Department and all the faculties of the Computer Science and Engineering Department who have enlightened us during our studies. The faculties and cooperation received from the technical staff of the Department of Computer Science and Engineering is thankfully acknowledged.

1.	Tejash Ranjan	2006007
2.	Abhishek Kumar	2006035
3.	RaviRanjan Kuamr	2006013

Table of Contents

Abstract

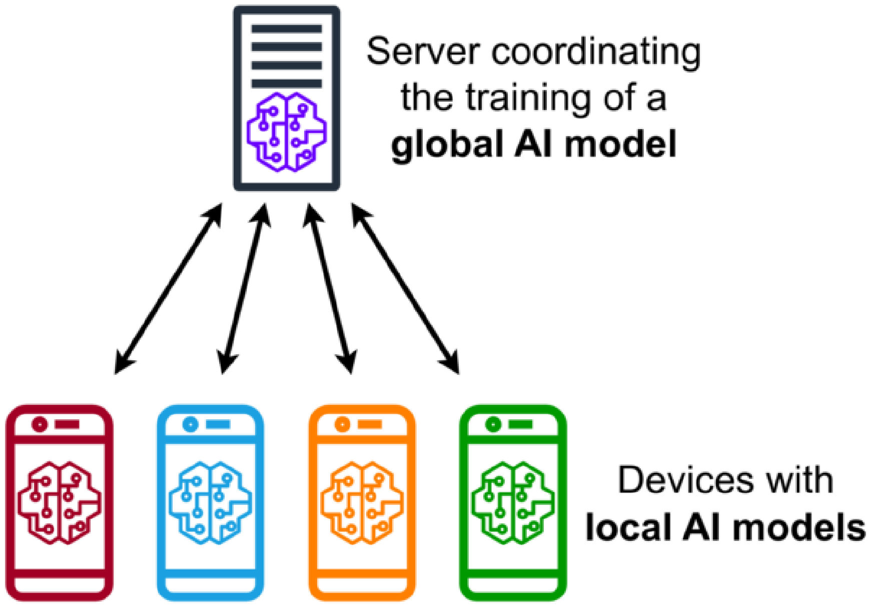
This paper explores the privacy guarantees of aggregation in gossip protocols. Gossip protocols are widely used in distributed systems to disseminate information, but the aggregation of data in such protocols can reveal sensitive information about individuals. Here the paper proves that a gossip protocol can make shared data private without adding any noise into the data.

Keywords: Gossip Protocols, privacy , aggregation , differential privacy

Chapter 1

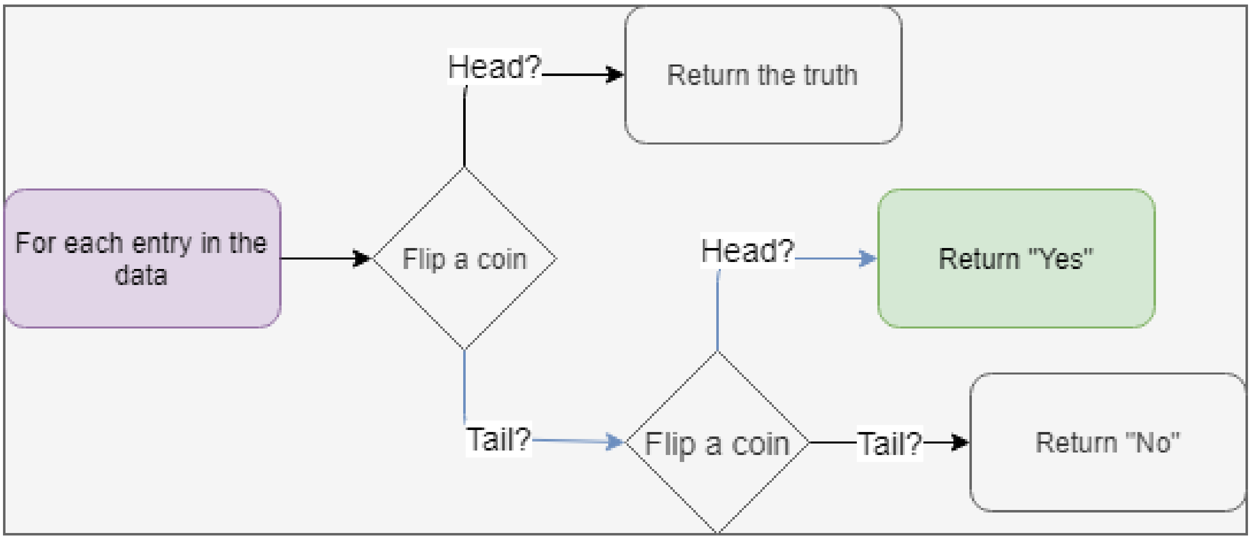
1.0 INTRODUCTION

Privacy concerns are a critical issue in distributed systems that rely on gossip protocols for disseminating information. While aggregation is an efficient way to summarize data in these systems, it can reveal sensitive information about individuals.
To understand this ,first we should know about federated learning.
Federated learning is a decentralized machine learning technique that enables multiple parties to collaboratively train a shared model while keeping their local data private. In traditional centralized machine learning, all data is collected and stored in a central server, which raises concerns about data privacy and security. Federated learning, on the other hand, allows each participating device or node to perform local model updates on its own data, and then sends only the model parameters (not the raw data) to a central server, where they are aggregated and used to update the shared model.



Let's suppose, each device is training their own parameter and after the training is over they send their parameter to the central server. The data of each will be private to each other but the central server has the access to the parameter of everyone, thus it can differentiate between the parameters and can find or extract some data.
So to make the parameters private all devices can add noise to their parameter before passing. This can make their parameter private but the accuracy of the model will be affected.

HOW ADDING NOISE CAN HELP IN MAKING OUR DATA PRIVATE ?
In the figure given below , let's suppose there exist an algorithm that runs that way.



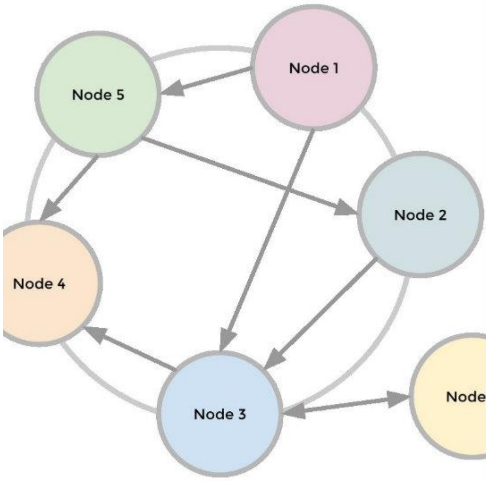
Assume if there is a survey that asks for their relationship status. The main key is that, before answering , one must toss a coin and if it is heads then he must answer correctly but if he got tails , then he will toss again and if he gets heads then answer correctly otherwise gives the wrong answer.
This will collect data which will be approx. 75% correct. This will give an overall overview of the population but it can't helps anyone to get someone's personal information.
This way the way by adding extra noise in the data.

Another way of doing is just using the gossip protocols to transfer the data.

WHAT GOSSIP PROTOCOLS DO ?
Lets' understand through an example, say there exist 3 devices and instead of sharing their parameter to central server they just pass it to next device. The device that gets the parameter passes the average of its parameter and the received one to its next.
After some round , no one will be able to know what each other parameter is.
That's how gossip protocols make our data private without any noise.

Chapter 2

LITERATURE



There exist some nodes that perform gossip within them. One of them perform as attacker that wants to know the source of the data that it received.

2.1 THREAT MODEL

The threat model described assumes that an attacker is attempting to identify the source of data in a distributed system. The attacker is assumed to have the ability to monitor the ongoing communication within the system, meaning they can potentially observe the messages being exchanged between nodes.

In this model, it is further assumed that the attacker has a certain probability, represented by $0 < \alpha \leq 1$, of correctly observing the source of the data. This means that there is a chance that the attacker can correctly identify the node that generated the data.

The observed event is represented by $S = ((x, i))$, where $i \in V$ represents the node that the attacker believes the data originated from, and x represents the data received by the attacker. It is important to note that the attacker may not have access to the entire message, but only a portion of it.

This threat model highlights the need for privacy-preserving techniques in distributed systems to protect against attacks that seek to identify the source of data. Techniques such as secure multi-party computation and onion routing can be used to ensure that sensitive information remains private and protected from attackers who may be monitoring the communication within the system.

2.2 PRIVACY MODEL

The privacy model adopted for the gossip protocols is differential privacy. Differential privacy is a framework that provides strong privacy guarantees for statistical data analysis by ensuring that the inclusion or exclusion of an individual's data does not significantly affect the output of the analysis.

In this model, the privacy loss of the gossip protocol is measured by the probability of a node i being identified as the source of the data x when it is observed, represented by $P_G^i(S_{x,i})$. The probability of node i as source is then compared to the probability of another node j, represented by $P_G^j\left(S_{x,i}\right)$, with a tolerance level of "δ". The tolerance level "δ" indicates the acceptable level of error in the privacy loss estimation.

The privacy budget ε, where ε >= 0, determines the maximum amount of privacy loss that can be tolerated in the system. The privacy loss of the gossip protocol is considered bounded by ε with a probability of at least 1 − δ. This means that the probability of a node being identified as the source of data is constrained within an acceptable range of privacy loss, ensuring that the privacy of individual data is protected.

Overall, the adoption of differential privacy in the gossip protocol ensures that the privacy of individual data is maintained even when the data is aggregated and disseminated across the network. This approach can provide valuable privacy guarantees for distributed systems that rely on gossip protocols for information dissemination.

2.3 Loss of Privacy(L)

- L = ln ($\frac{P_G^i(S_{x,l})}{P_G^j(S_{x,l})}$)
- If probability of i being the source of x is high, then loss will be positive.
- If probability of j being the source of x is high, then loss will be negative.
- More similar the values of probability , lesser will be the loss.
- When $P_G^i(S_{x,l})$ = $P_G^j\left(S_{x,l}\right)$, L = 0 .

Chapter 3

3.0 Proof

In the similar way as of privacy book has defined (ε, δ) differential privacy, we have

- $P_G^i(S_{x,i})$ ≤$e^{\epsilon}.P_G^j(S_{x,i}) + \delta$

  Here we will try to find the bounds of ***ϵ* and** δ

  3.1 Finding the bound of delta

- $\delta \geq P_G^i(S_{x,i}) - e^{\epsilon}.P_G^j(S_{x,i})$
- $P_G^i(S_{x,i}) = \alpha$

  $O_{d(i,j)}$ = communication is detected after d(i,j) gossip action have been executed.

  d(i,j) = length of shortest path between i and j.

- $P_G^j(S_{x,i}) = P_G^j(S_{x,i} \cap \bar{O}_{d(i,j)}) + P_G^j(S_{x,i} \cap O_{d(i,j)})$
- $P_G^j(S_{x,i}) = P_G^j(S_{x,i} \cap O_{d(i,j)})$ <= $P_G^j(\, O_{d(i,j)}) = (1 - \alpha)^{d(i,j)}$
- So, $\delta \geq max(\alpha - e^{\epsilon}.(1 - \alpha)^{d(i,j)})$
- $\delta \geq \alpha - e^{\epsilon}.(1 - \alpha)^{D_G}$, D_G is the diameter of the nodes.
- Since, $\sum_{j \in V} P_G^i(S_{x,j}) = 1$

  So , there always exist a node l ∈ V such that,

- $P_G^i(S_{x,l})$ <= $\frac{1}{n-1}.\sum_{j \in V, j \neq i} P_G^i(S_{x,j}) = \frac{1 - P_G^i(S_{x,i})}{n-1}$
- $P_G^i(S_{x,l})$ <= $\frac{1 - P_G^i(S_{x,i})}{n-1}$
- $P_G^i(S_{x,l})$ <= $\frac{1-\alpha}{n-1}$
- $\delta \geq \alpha - e^{\epsilon}.\frac{1-\alpha}{n-1}$
- So, δ ≥max($\alpha - e^{\epsilon}.\frac{1-\alpha}{n-1}$, $\alpha - e^{\epsilon}.(1 - \alpha)^{D_G}$) = lower_bound
- δ ∈ [lower_bound , 1]

This is a mathematical derivation of the lower bound on the tolerance level "δ" for differential privacy in the context of a gossip protocol. The equations presented show how "δ" can be bounded based on the probability of a node i being the source of data x, the communication distance between nodes i and j, and the diameter of the network. The final result is that "δ" must be greater than or equal to the maximum of two terms, which represents the lower bound on "δ". This result is useful for ensuring that the privacy budget is appropriately set to achieve the desired level of privacy protection in the gossip protocol.

As delta is bounded and is less than 1. It depicts that gossip protocols has less than 1 tolerance which means some of the data has been differentially private throughout the gossiping.

3.2 Finding bound of epsilon

- Transition probability matrix = A
- Probability of j contacting i node as it becomes active =A[j , i]
- $\hat{A}_i$[j , k] = { A[j , k] , j≠i

  0 , i=j , j≠k

  1 , i=j=k }
- $\hat{A}_i^m$[j , i] = probability of node i being active at time m.
- Probability of node i being active for first time at time m = $\hat{A}_i^m$[j , i] - $\hat{A}_i^{m-1}$[j , i]
- P_m(j -> i) = probability that i becomes active for first time and source node is j.
- P(j -> i) = $\sum_{m=1}^{\infty} P_m$(j -> i)
- P(j -> i) = $\sum_{m=1}^{\infty}$ (1- $\alpha)^m$. [$\hat{A}_i^m$[j , i]- $\hat{A}_i^{m-1}$[j , i]] , with the decay centrality logic
- P(j -> i) = $\sum_{m=1}^{\infty}$ (1- $\alpha)^m$. $\hat{A}_i^m$[j , i] - $\sum_{m=1}^{\infty}$ (1- $\alpha)^m$ $\hat{A}_i^{m-1}$[j , i]
- P(j -> i) = $\sum_{m=1}^{\infty}$ (1- $\alpha)^m$.$\hat{A}_i^m$[j , i] - $\sum_{m=1}^{\infty}$ (1- $\alpha)^m$.$\hat{A}_i^{m+1}$[j , i] - (1- $\alpha)^m$. $\hat{A}_i^0$[j , i]
- P(j -> i) = $\sum_{m=1}^{\infty} \alpha$(1- $\alpha)^m$.$\hat{A}_i^m$[j , i] - (1- α)I[j, i]
- P(j -> i) = α $\sum_{m=0}^{\infty}$ (1- $\alpha)^m$.$\hat{A}_i^m$[j , i]
- $P_G^j(S_{x,i})$ will happen if these two points will be followed :

  1. Node i becomes active for first time before attacker observes .
  2. Attacker observes node i as its first observation after node i is active for first time.

  These points results in :

- $P_G^j(S_{x,i})$ = P(j -> i) . $P_G^i(S_{x,i})$
- $P_G^i(S_{x,i})$ = $\frac{P_G^j(S_{x,i})}{\alpha.\sum_{m=0}^{\infty}(1-\alpha)^m.\,\hat{A}_i^m[\text{ j , i }]}$

  As per,

- $P_G^i(S_{x,i}) \leq P_G^j(S_{x,i})$. e^{ε} + $(\delta = 0)$

- $e^{\varepsilon} \geq max_{j \neq i} \left(\frac{P_G^i(S_{x,i})}{P_G^j(S_{x,i})} \right)$
- $e^{\varepsilon} \geq max \left(\frac{1}{\alpha . \sum_{m=0}^{\infty} (1-\alpha)^m . \widehat{A}_i^m [\mathrm{j}, \mathrm{i}]}, 1 \right)$
- $\varepsilon \geq \ln \left(max \left(\frac{1}{\alpha . \sum_{m=0}^{\infty} (1-\alpha)^m . \widehat{A}_i^m [\mathrm{j}, \mathrm{i}]}, 1 \right) \right)$

  This epsilon range is when the tolerance is zero. By inserting the value of delta in the main equation we can get the corresponding range of epsilon.

Chapter 4

4.0 Future Works

- Integrating the algorithm with the browsers.
- Finding more optimized values of privacy budget.

4.1 Conclusion

In conclusion, privacy guarantees in aggregation through gossip protocols play a crucial role in ensuring the security and confidentiality of data. Our report has highlighted the various techniques and mechanisms that are currently being employed to guarantee privacy. From the analysis, it is evident that these techniques have their strengths and weaknesses and that they need to be implemented judiciously to ensure that they are effective.

Overall, the report emphasizes the importance of having a clear understanding of the privacy requirements of the data being shared through the gossip protocol. It also emphasizes the need for collaboration among various stakeholders to ensure that privacy guarantees are implemented effectively. Finally, the report points out that the evolution of gossip protocols and the growing sophistication of data analysis techniques underline the need for constant innovation and vigilance in ensuring the privacy and security of data through gossip protocols.

References :-

1. C. Dwork, A. Roth et al., "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, no. 3–4, pp. 211–407, 2014