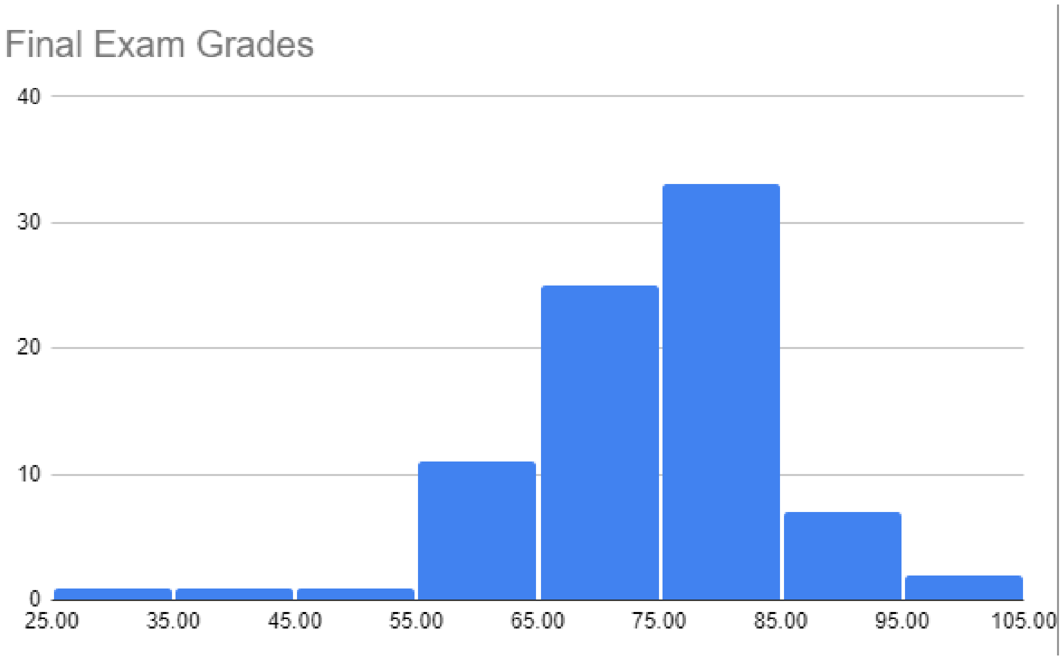
CS-523 Final Most Repeated Errors

Spring 2023



Question 1: Location privacy

- **Lack of attack.** The question clearly asks to “describe an attack the professors could launch” and lack in doing so results in point deduction. Describing the limitations of the GPS coordinates’ generalisation is not an attack: you have to explain how the professors can exploit these limitations to learn new information about the students.
- **Plausible deniability.** Spatial obfuscation as described in Q2 does not provide plausible deniability for the students. The fine-grained granularity of the data collection process (a location point every 30 minutes) and the low probability with which the random building is used instead of the correct one (¼) make the potential deniability claim *implausible* as the random noise can be easily recognized and removed.
- **Additional measures improve privacy.** The additional measures do not improve the location privacy compared to Q1: noise can be easily removed due to fine-grained granularity of the measurement and hiding is insufficient (e.g., a class lasts 2 hours). Even if these measures make the attack proposed in Q1 slightly harder (in most cases this is not the case), it is wrong to conclude that the measures improve the location privacy.
- **Partial attack.** The professors can mount an attack that maps the random unique student IDs to a physical identity by collecting some spatio-temporal points for the targeted students and link these to the anonymized locations present in the records. The effect of this attack is to reduce the anonymity set of the anonymized records to

a single person, which enables the mapping between random IDs and physical people. Identifying a set of students in a professor's class, or inferring the major of a random student from the records using the most visited buildings is a partial attack that does not enable the professor to track a precise student and learn new information about their movements on campus, since the cardinality of the anonymity set is not reduced to a single entity.

Question 2: Machine learning

Q1:

- **Changing the goal of the attack from learning whether the classmate has your notes to whether your classmate has *any* cs-523 related-content.** While related, this other attack does not fulfil the objective of the question statement. To repeat the statement: your objective is to know whether the classmate kept your notes. Inferring whether your classmate has cs-523 material on their laptop does not help you in knowing whether this classmate kept their word and deleted your notes or not, as it can be some slides, other notes, or project source code.
- **Explaining the knowledge and capabilities of your classmate and not you.** The threat model needs to be about the adversary, in this case, you. Answers talking about the classmate only (i.e., *"classmate is honest-but-curious and cannot deviate from the protocol"*) received no points. Moreover, in many answers, the threat model was incomplete with regards to capabilities. For example, if your attack requires sending crafted gradients, it is important to highlight that (and why) you can deviate from the protocol and send modified gradients to your classmate. Please make sure to use all the hints we give you. :)

Q2:

- **Proposing FL as a system-based defense.** The task for this question is to reduce the attack surface that you exploited in question 1, i.e., reduce the privacy leakage of the DL protocol. Unfortunately, when proposing to use a server to aggregate the gradients/do the training as a defense, you are introducing another adversary (the server) in your system, but not removing the privacy leakage. Furthermore, it fails to provide sufficient protection because of the same reason as explained below.
- **Using MPC (Secure Aggregation/HE) to aggregate the gradients each round as a defense.** In this 2-party scenario, this technique does not hide the individual gradients. Since the output is simply *your_grad* + *classmate_grad*, the adversary (your classmate) can retrieve the gradient by looking at the output and removing its own contribution *classmate_grad*. Note: even if the aggregated gradient is applied to the model as a second step of the MPC protocol, the adversary can still recover the aggregated gradient by looking at their local model across iterations, and compute the aggregated gradient as the difference between two steps, and then extract your gradient as described above.

Question 3: Anonymous communications and censorship

- **Q1: Limiting the adversary to launch passive attacks.** In some answers only passive traffic analysis was considered. Relying only on passive attacks leads either to unrealistic assumptions (e.g. assuming that there is only one message in

SandCave network that day) or stating that it is impossible to prove that Alex is a sender. However, an active attack in which the government isolates Alex' message in a batch would actually provide evidence that Alex is the sender.

- **Q1: Block other Sandcave users.** Some attacks assume that the government can just block some Members of Sancave from sending their messages to Blend, without any explanations how the government can do it. Partial or full points were reduced depending on other steps in the proposed attack.
- **Q2: DNS or BGP hijacking.** Some measures proposed in this part contained censorship mechanisms listed in the course. However mechanisms like BGP hijacking required control over ISP infrastructure (or essentially building your own ISP infrastructure) which, unless such assumption was explicitly stated, were not in the capabilities of the adversary of this question..

Question 4: Online tracking

- **Q1: Not providing details about how the tracking is done:** Not describing how the adversary knows the request is from Alice (ex. Source IP address, email/username gathered from packet inspection). Partial points were reduced for the subquestion.
- **Q1: DNS-request based tracking:** This attack indeed works even with HTTPS, but it doesn't leverage all the available information to the adversary to perform the tracking especially when HTTP is used by Alice. Full points were awarded for discussing the attack's lower effectiveness when not using the plaintext request content or combining using this information with the DNS-request information in the attack, while partial points were reduced for not doing so.
- **Q2.1: Complaining to the GDPR:** The GDPR is a set of regulations/laws, not an entity to complain to (such as a data protection regulator or the national court for instance).
- **Q2.2: Ambiguity about what information is encrypted with HTTPS:** The information related to client software, device, and plugins from the HTTP headers is encrypted with HTTPS. Stating that the attack works because this information is still available resulted in full point reduction. Stating that the attack works with only the information available after switching to HTTPS without discussing that the attack's effectiveness gets reduced resulted in partial point reduction.
- **Q2.2: CNAME Cloaking works with HTTPS:** This is incorrect, unless the adversary can forge valid SSL certificates.