CMPT 479 Assignment 3

Junchen Li

301385486

2022/8/13

1. (a) All data in below table from the https://metrics.torproject.org/bandwidth-flags.html?start=2020-02-01&end=2020-02-02

date	have_guard_flag	have_exit_flag	advbw	bwhist
2020-02-01	FALSE	FALSE	48.829315	17.433462244
2020-02-01	FALSE	TRUE	19.30846652	6.450413772
2020-02-01	TRUE	FALSE	249.635765144	118.833907772
2020-02-01	TRUE	TRUE	92.368756464	47.212829288

The advertised bandwidth of relays with "Exit" flag (but not the "Guard" flag) is 19.30846652. The advertised bandwidth of relays with "Guard" flag (but not the "Exit" flag) is 249.635765144. From above data comparison, it is clearly showing that relays with "Guard" flag (but not the "Exit" flag) is more that relays with "Exit" flag (but not the "Guard" flag) on 2020-02-01.

Reason: The "Guard" proxy has the larger advertised bandwidth than "exit" proxy because the "Exit" proxy as known as the last proxy should send the packet to the destination. It may need to make legally liable for sending packet or data out. So "Exit" is not the popular one and not a lot of people want to do.

(b)

date	filesize	source	server	q1	md	q3
2020-02-01	51200	op-hk	onion	4.715	5.315 s	6.302
2020-02-01	5242880	op-hk	onion	24.985	28.58 s	51.53

This data set shows the three measurements when downloading static files (size 50KiB and 5MiB respectively) through the op-hk onion server on February 1, 2020. The "md" means Median of time in seconds until receiving the number of bytes in file size.

Media download rate for 51200 bytes: 51200/ 5.315 = 9633.1138 bytes/s

Media download rate for 5242880bytes: 5242880/ 28.58 = 183445.766 bytes/s

Ratio1: 5242880 / 51200 = 102.4

Ratio2: 183445.766 / 9633.1138 = 19.04324

From above data, we can notice that bigger file will take more time to download. While the file size changes by a factor of 102, the media download rate only changes by a

factor of 19. The downloading time is depending on file size and speed of internet connection. More specific, time of downloading a large file is depending on bandwidth, while time of downloading a small file is depending on processing memory and latency. The large difference in the above data is caused by the latency of files in downloading files

(c)

To estimate the median download rate for a 50 KiB if only using one node. We need to know what happened if it only using one node. In order to estimate that, we need to assume the network environment should be consistent. Including the speed of internet connection, bandwidth and processing memory. One node means it is a start also end. It can know the resource and destination at the same time. We also must trust it because of knowing both the destination and the source would threaten the user's privacy security. Since one node will be having less latency than three nodes, if the file size remains the same, we can estimate the median download rate for a 50 KiB in one node scenario.

(d)

Date	dirread	dirwrite	dirauthread	dirauthwrite
2017-01-01	0.043169104	0.982572224	0.003685152	0.078933
2018-01-01	0.070707224	2.621898696	0.002545448	0.107308456
Growth Rate	63.791%	166.8403%	-30.926%	35.9487%

I calculated the growth rate by following this formula:

$$\frac{current - previous}{previous} * 100$$

From above calculation, for both authorities and mirrors the growth rate of writing is higher than growth rate of reading. We can see that over time, more and more bytes are spent responding to directory requests. Scrolling through some of Tor's events, I can see that between 2017 and 2018, not only were the next generation of Onion services released, but also ways of changing the distribution of traffic. This period also saw the release of a sanitizing process that separated TOR network processes from the rest of the computer. One reason for this increase is that it can make more replies in the same amount of time to cope with the increased service traffic.

2. (a) K-anonymity is the correct one

   K-anonymity is a feasible method to anonymize private data. It can be effectively transform data which contain the user's private information. Also for attackers, it is difficult to determine the identity of individuals in this data set. Therefore, K-anonymity can combine these data sets with similar attributes and make sure that identifying information about any individuals who contributing to the data can be obscured.

   SMPC means computing jointly on an agreed function securely on the inputs without revealing them. SMPC is not a good choice for this scenario because it doesn't have two parties with different data can jointly computer a known function on the union of data. Need 2 parities work and the computational cost involved is too large, which is not a good approach.

   (b) Differential privacy is the correct one

   Differential privacy works by adding "noise" to the data set to perform calculations. Differential Privacy is a good technology choice. Because it requires such a large data set, it is computationally intensive, and organizations the resources or personnel to deploy it (A new company making these smart devices). It can comply with data privacy regulations without compromising his ability to analyze customer behavior. And covering up information about each person's record.

   PIR is not a good solution, for this scenario the client only needs to analyze the user's exercise habits and does not need to confirm to the database owner which element is selected. And sending the entire database to the client to allow perfect privacy queries is also a difficult task to achieve.

   (c) PIR is the correct one

   K-anonymity is not a correct technique because it worried that attempting sending a DNS query for the domain and not lead to purchase it. And k-anonymity, which ensures that every set of identifiers appears times, it can't anonymize domain names reasonably and does not really help this problem.

PIR can hide the identity of the retrieved item from the database by sending a query to the database and getting an answer back. PIR is an appropriate technological solution. The DNS does not know what potential customers asking for but returns the searched answer. It can make inquiries in a privacy-protected manner.

(d) SMPC is the correct ones

Differential privacy works by adding “noise” to the data set to perform calculations. Differential privacy is not a good technique for this situation. Because adding noise into “data” as known as location information will break the nearby query. It will not guarantee the privacy security of users.

SMPC is a suitable technique. We can treat all location information holed by hospital as a database and users can input their address (other database) to check if someone living in their apartment has been infected. Both of them can jointly compute a function and do not need to reveal their own information.