**实验目的**

根据所给的代码编译出简单的差分隐私拉普拉斯机制噪音产生和加噪程序和直方图加噪发布程序，使用给定的数据集，通过变换输入的隐私预算来观察不同隐私预算下的噪音规模和对数据的影响。

**实验文件树**

主文件夹下包含有本次实验各函数的代码和实验的样例数据集。

testraw.c 数据集直接加噪实验部分的主函数放置于此，提供了数据集读取，给定隐私预算的噪音生成和加噪后结果展示的功能。

testhist.c 数据集直方图发布加噪实验部分的主函数放置于此，提供了数据集读取，给定隐私预算的噪音生成和加噪后统计值输出的功能。

laplace.c 为实验代码生成拉普拉斯分布的随机数，用于加噪

csvpackage.c 实验代码的csv读取和预统计函数实现，针对样例代码提供读取为Animals结构体的功能

zoo.csv 本次实验使用的样例数据集。其描述了一个动物园喂食的场景，第一列中数据为动物名称，第二列中数据为动物每天消耗的胡萝卜数量。

zoo\_nb.csv 将zoo.csv中“Dugeng”项去除得到的相邻数据集

medicaldata.csv 本次实验直方图发布部分的样例数据集。其描述了一个简单的医疗数据场景：假设其建立于某医疗数据集之上，其第一列数据为一年龄区间，第二列数据为该年龄区间患有某种疾病的人数。要发布的直方图即是以第一列数据为直方图的桶（Bucket）发布的统计数据

md\_nb.csv 将其中“30-40”区间的统计值-1，模拟原始医疗数据集内一名患者退出数据共享。

Makefile 用于编译使用的编译规则文件

Include子文件夹下包含有本次实验的两组成员函数文件使用的头文件。

laplace.h 为laplace函数族提供函数定义。

csvpackage.h 为csvpackage函数族提供函数定义。

**实验环境**

Ubuntu 18.04虚拟机

**实验步骤**

1. 使用VMWare Workstation或VirtualBox建立一台Ubuntu18.04虚拟机。

2. 打开“终端”，安装必要的解释器软件，此处直接使用ubuntu的build-essential来安装gcc和相关依赖库：

sudo apt-get install build-essential

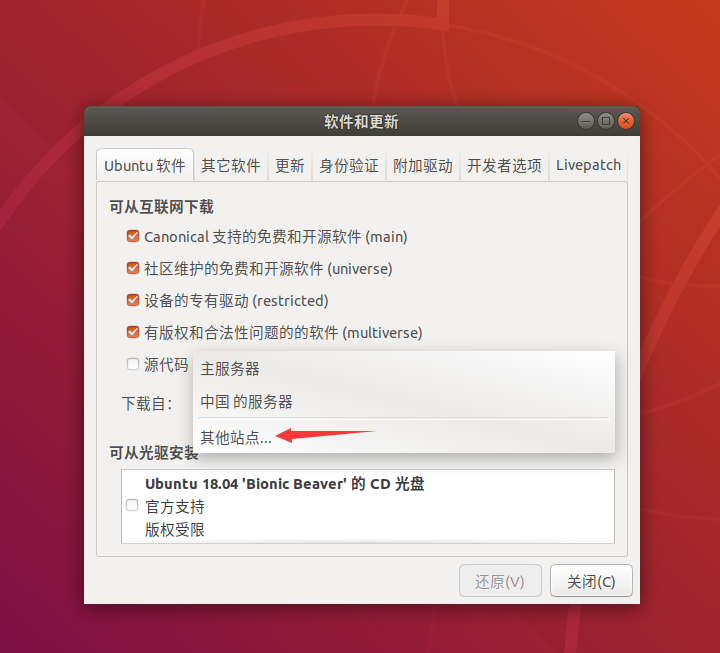


\*提示：若安装慢，可先更换到国内第三方源，具体步骤如下：

（1）在应用列表中找到“软件和更新”



（2）选择“其他站点”，在下拉列表中找到“中国”，选择任意一个服务器即可。

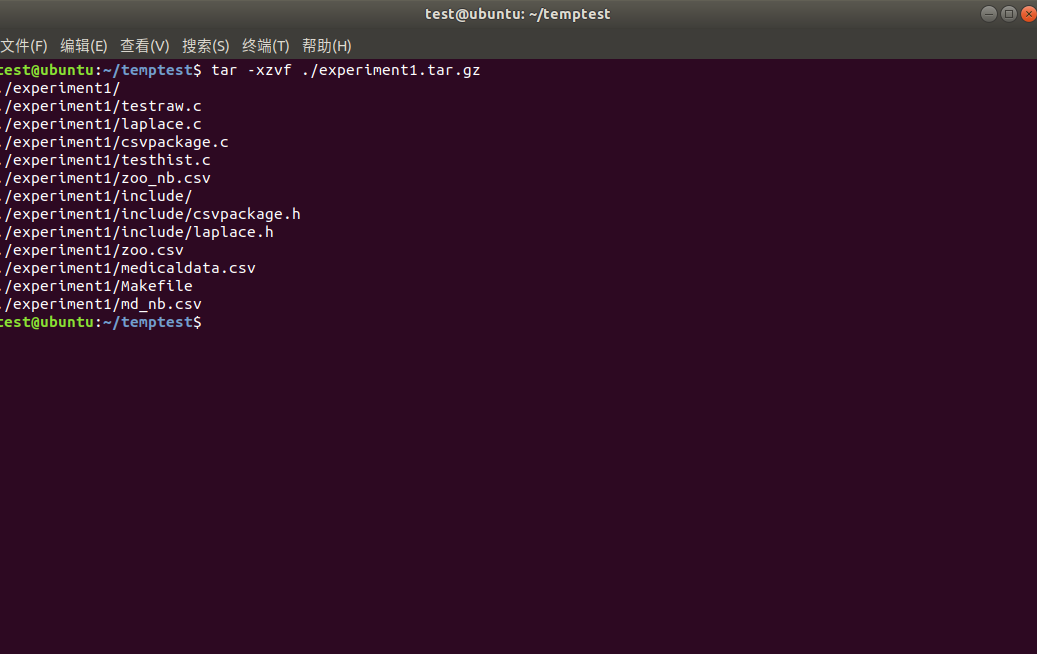


3. 执行指令gcc -v,观察到类似图中输出即为安装成功：

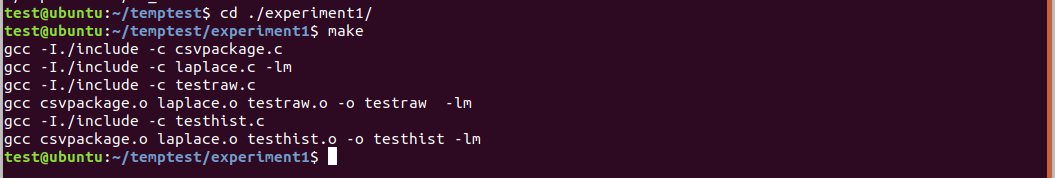


4. 在准备进行实验的文件夹内打开终端，解压提供的test.tar.gz文件。

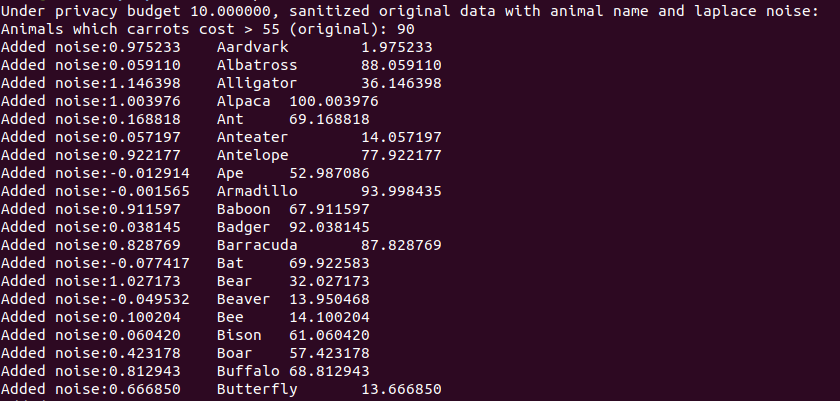
指令参考：tar -xzvf ./test.tar.gz



5. 进入实验文件夹，使用make指令完成编译。



6. 运行testraw程序，观察生成的噪音和加噪后的数据与原始数据的差别。



可以看到，在投入较大的隐私预算的情形下，添加的噪音均小于1或略大于1。对于特定查询“每日进食大于55根胡萝卜的动物个数”，在该预算下，加噪前和加噪后的响应一致，数据可用性好：

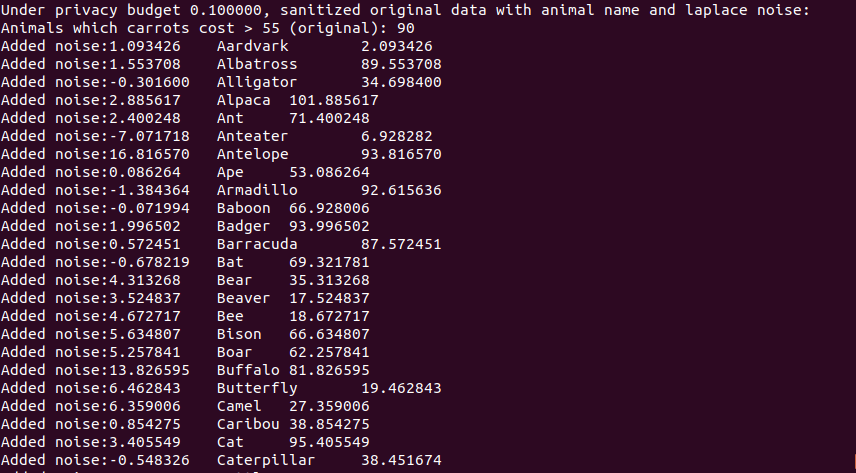


但是，观察标记后对相邻数据集的处理情况，我们可以发现，加噪后数据集对该查询的响应仍与数据集的变化一致，均为89，体现出了Dugeng离开数据集造成的差异，不能有效抵御对该查询的差分攻击。





我们此时再输入0.1作为隐私预算。可以看到，在该预算下，产生的拉普拉斯噪音也增大了，这使得加噪后的查询结果也受到了影响：



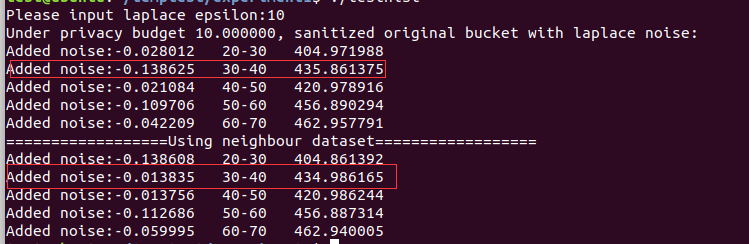


但是同时，观察对相邻数据集进行加噪的结果，可以发现，虽然相邻数据集的直接查询结果受到了“Dugeng”项移除的影响，但加噪后的相邻数据集查询结果与加噪前相比变化巨大，不再能反映出“Dugeng”项移除的影响：

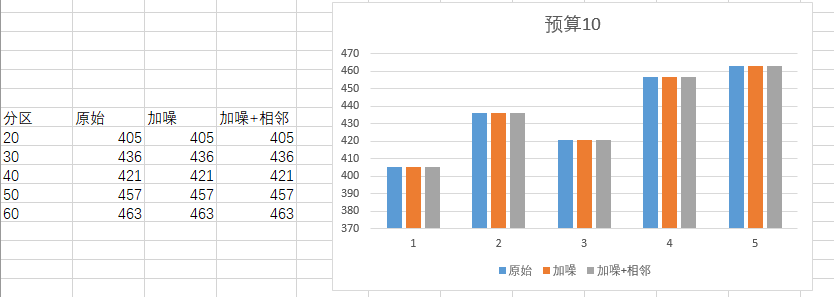


即，投入较少的隐私预算时，虽然数据的可用性降低了，但是能够更好地抵御差分攻击的影响。这里需要说明的是，由于算法存在一定随机性且比较简单，所以在某些时候可能无法取得上述结果，可以多试几次。

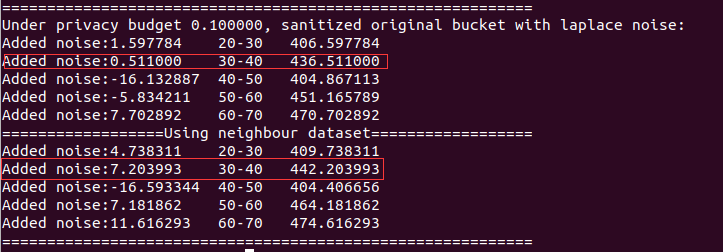
7. testhist程序提供了另一种差分隐私发布方法的演示，即差分隐私的直方图发布。在该发布方式下，加噪的对象不再是数据本身，而是对数据进行分桶统计后的计数值进行加噪。运行testhist程序，仍先以10作为隐私预算：



可以看到，当隐私预算为10时，由于加入噪音量级较小，相邻数据集的变化仍能被体现。使用预算为10的差分隐私算法生成的数据和直方图如下：



此时，使用0.1为预算，再次进行实验：



可以看到，由于噪音规模的提高，在相邻数据集的变化影响下，查询结果不减反增。即，虽然数据可用性变差，但能保护实际数据的变化不被攻击者获取，可抵御差分攻击。使用预算0.1的差分隐私算法生成的数据和直方图如下：

