Profiling internet users

APRIL 22

Information Security Privacy Dist Sys

CIS6930.009S19

Spring 2019

Objective

Profiling internet users project main goal is to check if the usage of each user which we were given 54 users are statistically distinguishable (i.e., p value <=0.5) or indistinguishable (i.e., p value > 0.5) when comparing to each other. This analysis has been done by selecting 3-time windows and comparing each to come up with how it plays a big role in profiling.

Technical details

I have developed my project using dev c++, in c++ language. Below include how I developed this project

1) Excel data: converted xlsx to csv

2) Database: convert data tables into objects; also known as dataset

Analysis

Hence, we had to decide if the internet usage of each subject is indistinct contrast with their internet usage over a period of time. Hence, we had to find P value of the 3-time window [10 seconds, 227 seconds, and minutes] for week 1 and week 2, where one

user is compared with the other 54 users that we were given for this project and are attached in the zip folder as an excel sheets.

Analyzing: Each excel file it was for a one-time window;

10 seconds: **Distinguishable** 1271 **Indistinguishable** are: 1645

Time taken to execute code: 4824 seconds = 1.34 hours

227 seconds: **Distinguishable** are: 980 **Indistinguishable** are: 1936

Time taken to execute code: 169.50 seconds = 2.83 minutes

300 seconds: **Distinguishable** are: 935 **Indistinguishable** are: 1981

Time to execute the code: 138.9 seconds = 0.017 minutes

Therefore, we can see that the time window for 10 seconds is the best in terms of authentication as the number of users who are indistinguishable are lesser in comparison to the rest of the time window times. Although, 10 second produce lesser number of indistinguishable it is execution time was higher than the rest.

Hence, we can see that the p values are presented in the diagonal of the table, that is because r2a2a (week 2 user 2 vs week 2 user 2) is always 1 because having the difference between their ranks would be 0, due to this, the value of z will be divided by 0, hence I used to get a ‘nan’ value when we try and calculate p value, but then I changed the 1 value to 0.99 and everything was fixed.

Executing code

I used c++ language to complete this project. Each 54 user profile peruse in one sitting to analyze against each user. First, download all the files attached from the zip which has a total of 57 files; 54 of for user excel files which I converted it to csv for easy read using c++ using getline and the rest 3 related to code and object files. Second, open the file main.cpp and in line 41 put the directory in which the csv file and the codes are all in the same file in order to run the codes and be able to open the files; otherwise the

files wont open. In line 77 you can change the times window (10sec, 227sec, and 300sec).

Note: to run the file of code, I have changed 2 values in files: *mmg9n6* row 138900 to 0 and *tjdpzd* row 106255 to 0 because it was crashing my program when I was trying to run.

In line 41 I used a vector size, so it can dynamically be changed, and we can have as many files in the directory. To read each file, I opened a for lop and had some operations on the data. Since, the file take is in .csv it is separated by a comma. This makes it easy to collect data of columns: Doctets, first packets and duration into vectors, while the rest are temp's meaning they are skipped.

Data in first packet are converted to human readable form using epoch where we can get data in the form of "Mon Feb 04 20:01:52 2013":

```
const char *d = FirstPacket[i].c_str();
std::stringstream is(d);
is >> epoch64;
epoch64 /= 1000;
time_t t = epoch64;
char *epo;
char delims[] = " ,:";
char * tok;
epo = ctime(&t);
tok = strtok(epo, delims);
```

After that, we tokenized each data to store in its vector of Day, month, date, hours, min, seconds and year. To calculate octets/durations we use data in vector dockets and durations and store in vector od. When the duration is 0 it would result in an error hence they are marked with a -1 in order to skip them in the upcoming calculations.

To calculate the total number of seconds for every record from excel we do that by multiplying hours with 24, minutes with 60 and then we add them to the seconds.

```
tottime.push_back(3600 * stoi(Hrs[x]) + 60 * stoi(Min[x]) + stoi(Sec[x]));
```

Then we divide with the length of the slot which I did to categorize them into groups of specific days.

```
group.push_back(tottime[x]/slot);
```

Matrix with columns as group of the day and row numbers as dates of the month, where every excel sheet we can get the date value and group value being calculated first then adding its values of doctets/durations to the value doctets/durations being present before with date as row and group as y value. Hence, doing so, excel sheet is getting added to each row that is made that is then be used to get overall average value.

File matrix having files as rows with all groups of every day except Sunday and Saturday as columns, it also starts with the first Monday 2013-02-04. Ranks is available to every group with a value of average doctets/durations in a specific day.

To calculate Spearman's correlation coefficient, the r value is calculated using the difference between the respective ranks to get 3 correlations: r1a2a, r1a2b, r2ab2, where letter represent subject and number represent week. To get Z value, we use the 3-correlation using the formula below

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$Z = [Z_{1a2a} - Z_{1a2b}] * \frac{\sqrt{[N-3]}}{2 * [1 - r_{2a2b}] * h}$$

$$Z_{1a2a} = \frac{1}{2} log \frac{1 + r_{1a2a}}{1 - r_{1a2a}}$$

$$Z_{1a2b} = \frac{1}{2} log \frac{1 + r_{1a2b}}{1 - r_{1a2b}}$$

$$h = \frac{1 - [f * rm^2]}{1 - rm^2}$$

$$f = \frac{1 - r_{2a2b}}{2 * [1 - rm^2]}$$

$$rm^2 = \frac{r_{1a2a}^2 + r_{1a2b}^2}{2}$$

To calculate P of these z values we used the function below

```
static double PFunction(double z)
{
    double p = 0.3275911;
    double a1 = 0.254829592;
    double a2 = -0.284496736;
    double a3 = 1.421413741;
    double a4 = -1.453152027;
    double a5 = 1.061405429;

    int sign;
    if (z < 0.0)
        sign = -1;
    else
        sign = 1;

    double x = Math.Abs(z) / Math.Sqrt(2.0);
    double t = 1.0 / (1.0 + p * x);
    double erf = 1.0 - (((((a5 * t + a4) * t) + a3)
      * t + a2) * t + a1) * t * Math.Exp(-x * x);
    return 0.5 * (1.0 + sign * erf);
}
```

Therefore, when p <= 0.05: correlation coefficient calculated for internet usage (e.g.; b) is smaller than the known subject (e.g.; a), as subject b will be pointed as subject fat from subject a.

On the other hand, when p > 0.05: correlation coefficient calculated for internet usage (e.g.; b) is *not significantly* smaller than the known subject (e.g.; a), as subject b will be

pointed as subject fat from subject a, where then subject b will be pointed as indistinguishable from subject a.

Indistinguishable is when p value > 0.05, when distinguishable is when p value <= 0.05

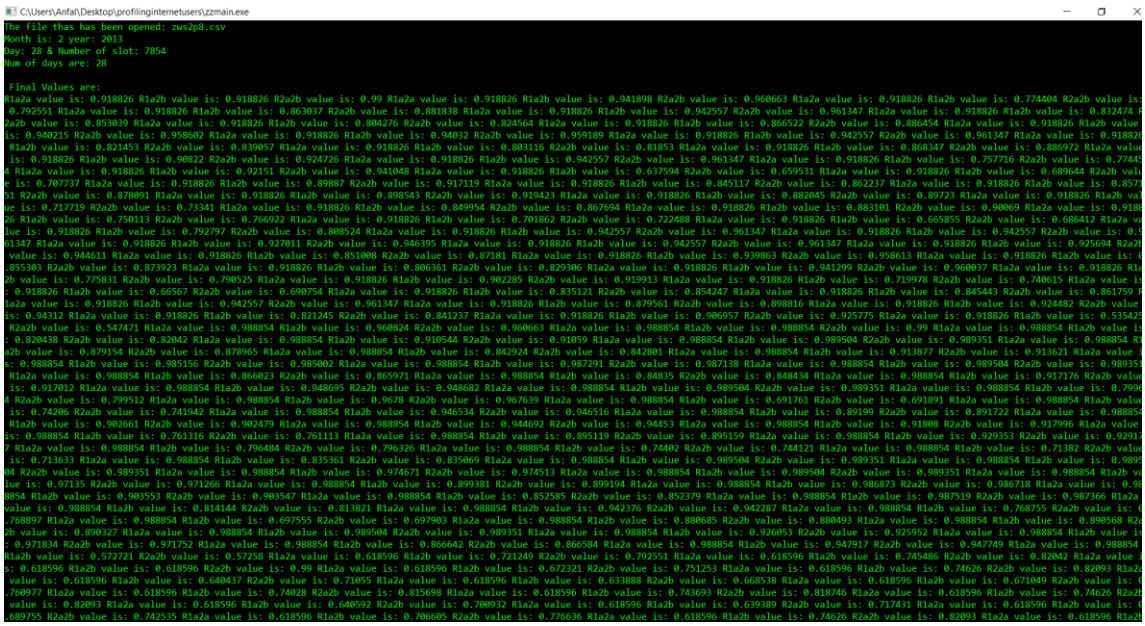
Code output issues

> I tried fixing my issues with outputting nan “not a number” whenever the value p falls in the same user, such as when its diagonal. I checked all my codes, and everything seems to be doing the right calculations. Hence, I am not sure what exactly is it doing in between.

Time windows and R values

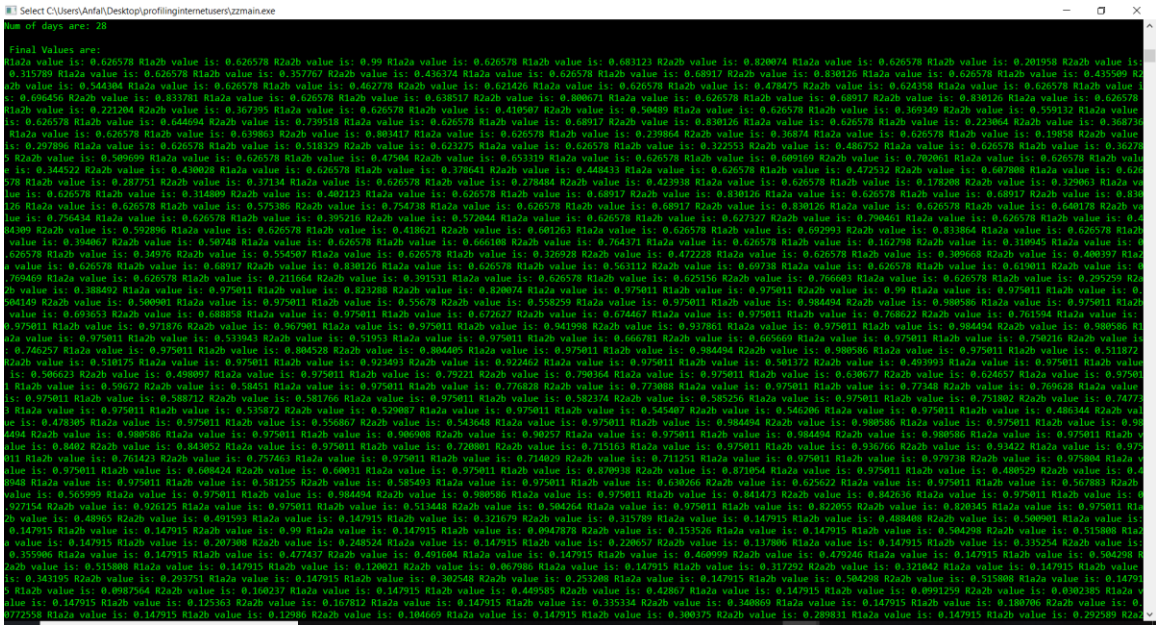
10-second interval

Week 1 &2	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	User 11	User 12	User 13	User 14	User 15	User 16	User 17
User 1	0.5	0	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1
User 2	1	0.5	1	1	7.7716E-16	1	1	1	1	1	7.77156E-16	1	1	1	1	7.77156E-16	1
User 3	0	0	0.5	0	0	5.5511E-17	1.5412E-07	0	0	0	0	3.33067E-16	4.996E-16	0	0	0	1
User 4	0	0	1	0.5	0	1	1	0.0934323	0	0	0	1	1	6.01301E-08	0	0	1
User 5	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 6	0	0	1	0	0	0.5	1	0	0	0	0	1	1	0	0	0	1
User 7	0	0	1	0	0	1	0.5	0	0	0	0	1	1	0	0	0	1
User 8	0	0	1	1	0	1	1	0.5	0	0	0	1	1	1	0	0	1
User 9	1	0	1	1	0	1	1	1	0.5	0	0	1	1	1	1	0	1
User 10	1	0	1	1	0	1	1	1	1	0.5	0	1	1	1	1	0	1
User 11	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 12	0	0	1	0	0	0.999733	1	0	0	0	0	0.5	1	0	0	0	1
User 13	0	0	1	0	0	0	0.992357	0	0	0	0	0.000024979	0.5	0	0	0	1
User 14	0	0	1	1	0	1	1	0.995862	0	0	0	1	1	0.5	0	0	1
User 15	0	0	1	1	0	1	1	1	0	0	0	1	1	1	0.5	0	1
User 16	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 17	0	0	0	0	0	0	8.9675E-08	0	0	0	0	0	0	0	0	0	0.5
User 18	1	0	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1
User 19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
User 20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
User 21	0	0	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1
User 22	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	0	1
User 23	0	0	1	7.6571E-10	0	1	1	0	0	0	0	1	1	0	0	0	1
User 24	0	0	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1
User 25	0	0	1	1	0	1	1	1	0	0	0	1	1	1	0	0	1
User 26	0	0	1	0	0	0	0.00094299	0	0	0	0	5.05151E-15	0	0	0	0	1
User 27	0	0	1	1.0381E-14	0	1	1	0	0	0	0	1	1	0	0	0	1
User 28	0	0	1	1	0	1	1	1	0	0	0	1	1	1	0	0	1
User 29	0	0	1.057E-09	0	0	0	0	0	0	0	0	0	0	0	0	0	1



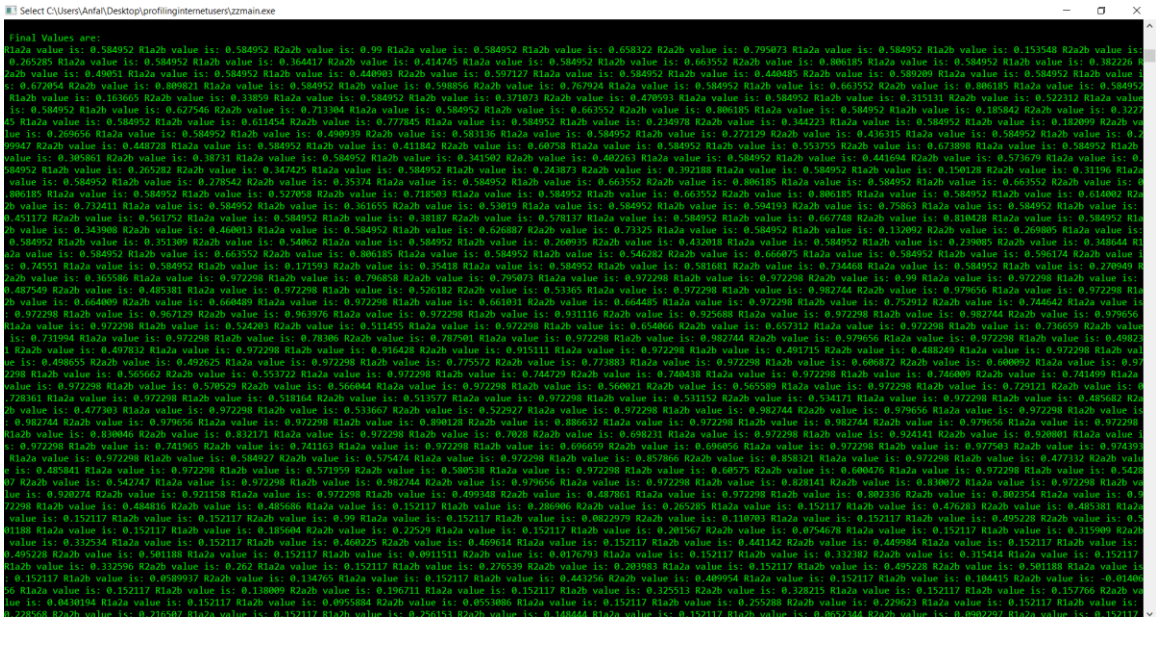
227-second interval

Week 1 & 2	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	User 11	User 12	User 13	User 14	User 15	User 16	User 17	
User 1	0.5	1.0375E-13	1	1	0	1	1	1	0	0.0811558	0	1	1	1	0.0483217	0	1	
User 2	1	0.5	1	1	0	1	1	1	0.998648	1	0	1	1	1	1	0	1	
User 3	8.8555E-09	0	0.5	0.915481	0	0.0407433	0.0310794	5.6556E-11	0	0	0	0.744517	1.59678E-08	3.91462E-10	2.16077E-06	0	0.900248	
User 4	0	0	0.98787	0.5	0	1.2108E-06	0.00453214	6.6613E-16	0	0	0	0.755691	1.13389E-05	1.55431E-14	0	0	0.929333	
User 5	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1	
User 6	0	0	1	1	0	0.5	0.831183	1.3015E-06	0	0	0	1	0.907398	0.000465622	4.94049E-15	0	1	
User 7	0.15586	0	1	1	0	1	0.5	1	0	1.62592E-08	0	1	1	1	0.973835	0	1	
User 8	0.997903	0	1	1	0	0.999483	1	0.5	0	2.27472E-10	0	1	1	1	0.999255	0	1	
User 9	1	0.00038908	1	1	0	1	1	1	0.5	1	0	1	1	1	1	0	1	
User 10	1	0	1	1	0	1	1	1	0	0.5	0	1	1	1	1	0	1	
User 11	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1	
User 12	0.991579	0.00016228	1	1	2.9548E-06	1	1	0.942319	1.8457E-05	0.0199254	2.95477E-06	0.5	1	0.999988	0.999741	2.95477E-06	1	
User 13	0.00099821	0	1	1	0	0.305197	0.362982	7.6615E-05	0	0	0	1	0.5	0.993057	0.0022698	0	0.999986	
User 14	5.1585E-08	0	1	1	0	0.999339	0.997578	0.0662141	0	0	0	1	0.999997	0.5	3.42898E-09	0	1	
User 15	0.999761	0	1	1	0	1	1	1	0	5.66755E-05	0	1	1	1	0.5	0	1	
User 16	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1	
User 17	0.00084436	0	0.999952	0.675933	0	0.0558084	0.567902	3.8017E-11	0	0	0	0.972918	0.00107374	0.964575	5.97304E-07	0	0.5	
User 18	1	0	1	1	0	1	1	1	0	6.00872E-09	0	1	1	1	1	0	1	
User 19	2.0567E-05	0	1	0.999969	0	0.399098	0.00486972	0.0165054	0	0	0	1	0.723192	0.785449	4.76968E-07	0	0.993598	
User 20	0.00704605	0	0.997841	0.999148	0	0.00122823	0.938188	0.00239958	0	0	0	0.999991	0.00068376	0.623079	0.00813168	0	0.964238	
User 21	0.999999	0	1	1	0	1	1	1	0	2.1043E-11	0	1	1	1	1	0	1	
User 22	0	0	1	0.998425	0	0.280845	0.00023797	2.0689E-06	0	0	0	0.943553	0.0416231	8.09813E-05	1.02141E-14	0	0.999994	
User 23	0	0	0.991534	0.656555	0	1.6613E-05	3.8258E-05	0	0	0	0	0.997888	0.00391712	1.20471E-08	0	0	0.992368	
User 24	0.0924678	0	1	1	0	0.999957	0.999078	0.988957	0	0	0	1	1	0.95432	0.810256	0	1	
User 25	2.0184E-08	0	1	1	0	0.999728	0.993625	0.00876723	0	0	0	1	0.999306	0.96313	1.08524E-13	0	1	
User 26	0.745241	9.5208E-12	1	1	7.605E-15	0.999602	0.996729	0.661005	5.8303E-13	3.07145E-05	7.60503E-15	1	0.998264	0.999916	0.575128	7.60503E-15	1	
User 27	0	0	0.206882	0.659087	0	2.1205E-14	4.7238E-11	0	0	0	0	0.791433	7.23854E-12	1.30584E-06	0	0	0.000137243	
User 28	3.9084E-05	0	1	1	0	0.999999	0.998086	0.521455	0	0	0	1	1	0.997149	0.0231741	0	1	
User 29	0.00042918	0	0.990648	0.999705	0	0.0773031	0.0230927	1.1292E-09	0	0	0	1	0.0246285	0.379629	1.45581E-05	0	0.999917	



5-minute interval

Week 1 & 2	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	User 11	User 12	User 13	User 14	User 15	User 16	User 17
User 1	0.5	1.9518E-13	1	1	1.1102E-16	1	1	1	0	0.112883	1.1102E-16	1	1	1	0.00106107	1.1102E-16	1
User 2	1	0.5	1	1	0	1	1	1	0.99993	1	0	1	1	1	1	0	1
User 3	0.00020036	0	0.5	0.932804	0	0.203656	0.151731	1.0956E-06	0	0	0	0.881977	1.7885E-07	1.0856E-06	0.00127098	0	0.97977
User 4	0	0	0.907359	0.5	0	1.5998E-05	0.00323917	1.7174E-12	0	0	0	0.430966	4.24E-06	4.1E-12	0	0	0.643008
User 5	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 6	1.9323E-13	0	1	0.999994	0	0.5	0.516898	8.9289E-06	0	0	0	0.999988	0.632738	0.00051119	5.5562E-11	0	1
User 7	0.577198	1.0895E-12	1	1	5.5511E-17	1	0.5	1	1.6653E-16	0.00011893	5.5511E-17	1	1	1	0.992241	5.5511E-17	1
User 8	0.993643	0	1	1	0	0.993464	1	0.5	0	9.8197E-08	0	1	1	1	0.995028	0	1
User 9	1	6.5295E-06	1	1	0	1	1	1	0.5	1	0	1	1	1	1	0	1
User 10	1	0	1	1	0	1	1	1	0	0.5	0	1	1	1	0.999998	0	1
User 11	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 12	0.999961	0.0422205	1	1	0.00463137	1	1	0.99367	0.00740648	0.451675	0.00463137	0.5	1	1	0.999983	0.00463137	1
User 13	0.0304042	0	0.999993	1	0	0.520452	0.353975	0.0007467	0	0	0	0.999992	0.5	0.992175	0.00981138	0	0.999788
User 14	0.00025541	0	1	1	0	0.99932	0.979411	0.140836	0	0	0	1	0.999868	0.5	4.0211E-06	0	1
User 15	0.999935	1.4822E-14	1	1	0	1	0.998424	1	0	0.0113967	0	1	1	1	0.5	0	1
User 16	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	0.5	1
User 17	0.0204031	0	0.999934	0.809018	0	0.0939866	0.7259	1.9937E-06	0	0	0	0.98199	0.00784562	0.979851	0.00014498	0	0.5
User 18	1	0	1	1	0	1	1	1	0	0.00155574	0	1	1	1	1	0	1
User 19	0.0529939	5.5511E-17	1	0.999991	0	0.828928	0.0582236	0.296602	1.6653E-16	1.0483E-09	0	1	0.952305	0.947249	0.0016304	0	0.995559
User 20	0.167288	0	0.992265	0.999981	0	0.0789582	0.978607	0.115672	0	4.5032E-12	0	0.999984	0.015019	0.890457	0.0925981	0	0.997154
User 21	1	0	1	1	0	1	1	1	0	8.9547E-06	0	1	1	1	1	0	1
User 22	0	0	0.999997	0.989367	0	0.495832	0.00013875	0.00014698	0	0	0	0.853288	0.0689804	0.00182357	9.3918E-11	0	0.999828
User 23	0	0	0.885058	0.343561	0	0.0001134	2.5528E-05	5.5511E-17	0	0	0	0.959194	0.00027358	2.9926E-07	0	0	0.932294
User 24	0.43171	0	1	1	0	0.99924	0.956257	0.850904	0	0	0	1	0.999877	0.825354	0.753163	0	1
User 25	2.601E-05	0	1	1	0	0.9993	0.811947	0.00726346	0	0	0	1	0.962597	0.869144	6.3566E-12	0	1
User 26	0.9453	3.0557E-08	1	1	1.4758E-10	0.999521	0.990378	0.799994	3.7334E-09	0.00607667	1.4758E-10	1	0.996428	0.999582	0.790364	1.4758E-10	1
User 27	0	0	0.059426	0.570607	0	3.3646E-11	1.1608E-09	0	0	0	0	0.646239	2.4565E-10	3.1957E-06	0	0	4.2504E-05
User 28	0.00012391	0	1	1	0	0.999997	0.907735	0.275892	0	0	0	1	1	0.990998	0.0438783	0	1
User 29	0.0527626	0	0.988713	0.999958	0	0.441714	0.0734625	1.0007E-05	0	3.3523E-13	0	0.999997	0.104686	0.64966	0.00334414	0	0.999874



Results

Results for each time window is in a different excel spread sheet. Below report are for complete p-value