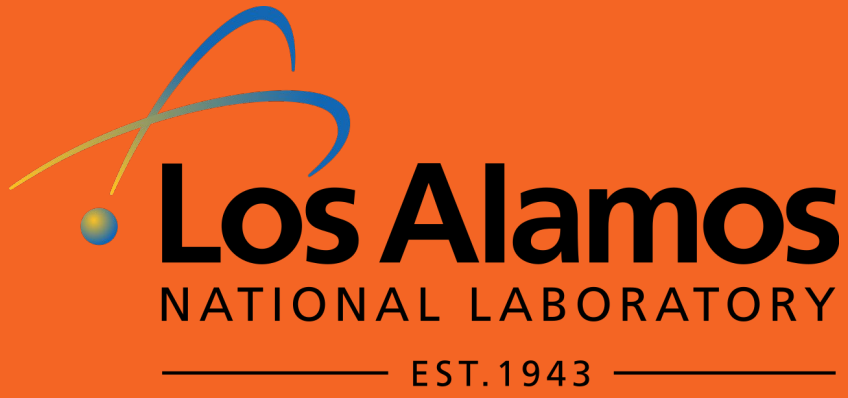
Using Network Traffic Data to Investigate Potentially Compromised Machines

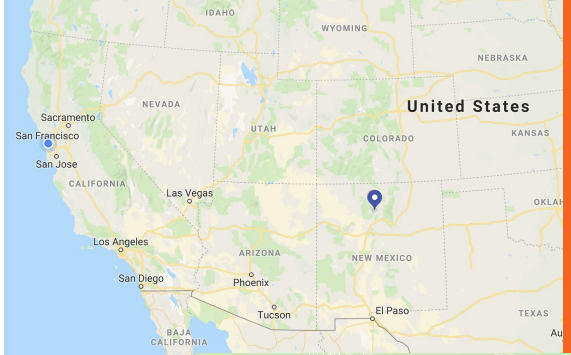
Special Thanks:

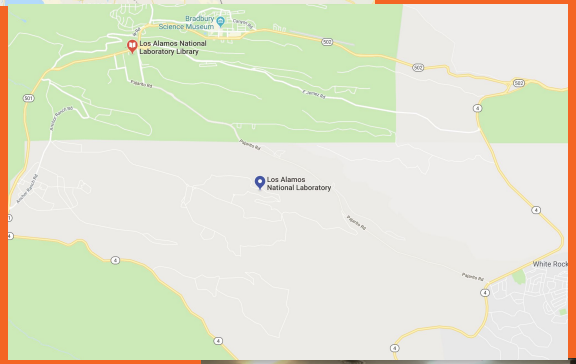


Where Is Our Data From?

This data set represents 58 consecutive days of de-identified event data collected from five sources within Los Alamos National Laboratory's corporate, internal computer network.

The Data Set Presented is collected from known human users and contains ~100 users whose computers are known to be Compromised





Network Interactions Example

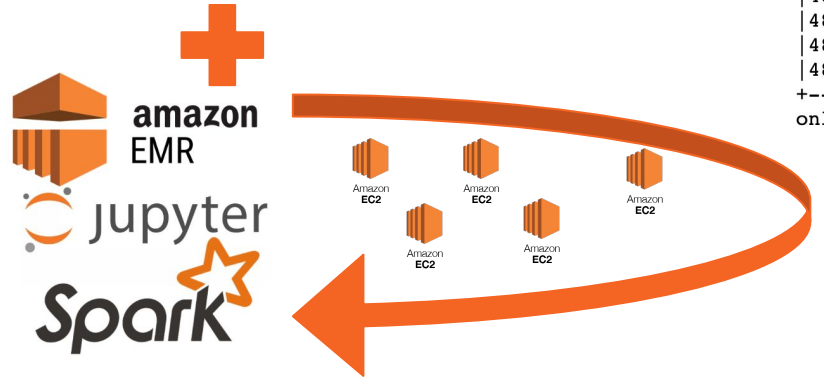
Source	Destination	Size	User	Service
Cat	Gmail	1Mb	Cat	Gmail
Gmail	Dog	1MB	Dog	Gmail

FLOW.txt

```
1,9,C3090,N10471,C3420,N46,6,3,144
1,9,C3538,N2600,C3371,N46,6,3,144
2,0,C4316,N10199,C5030,443,6,2,92
```

AUTH.txt

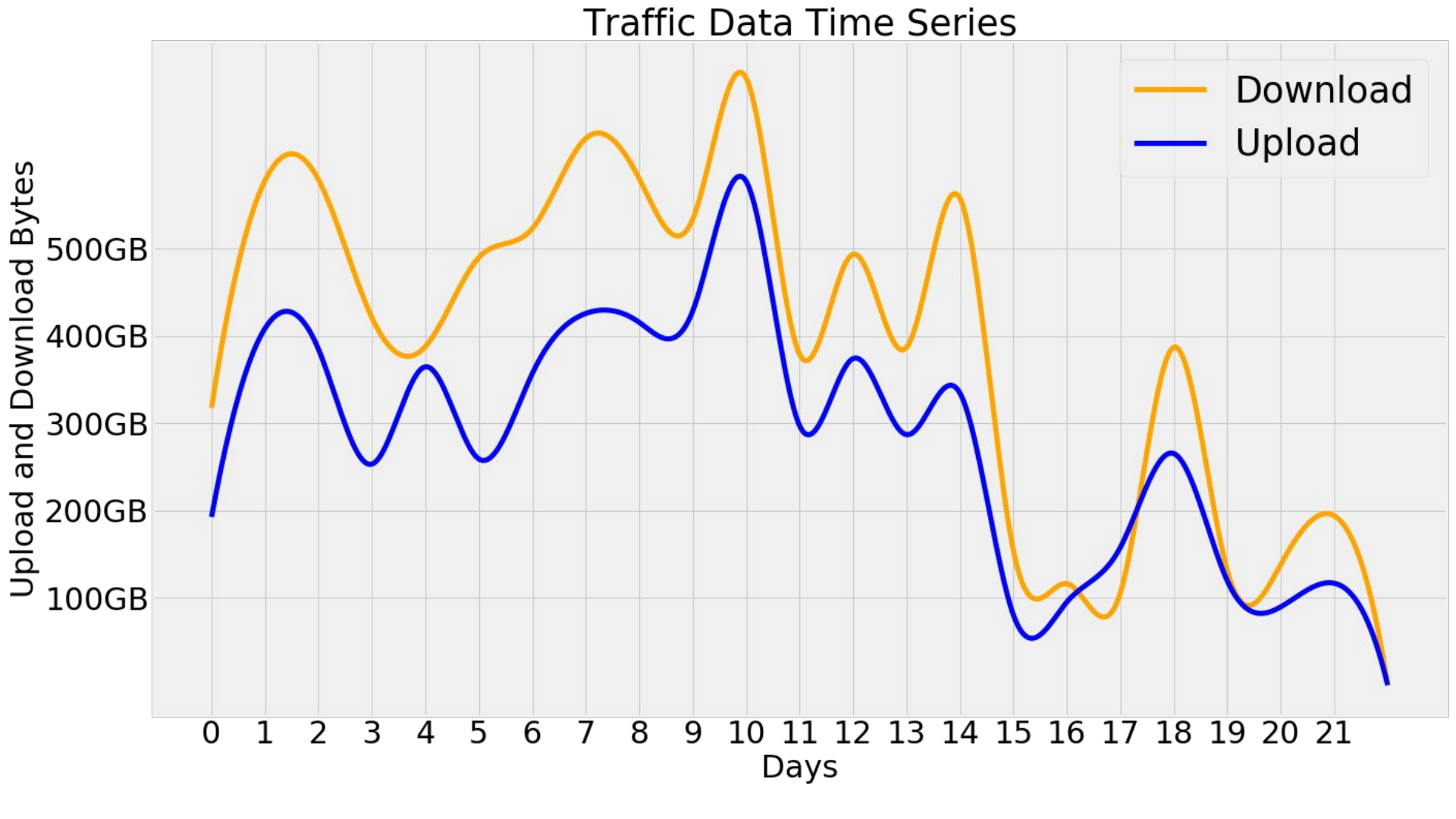
dst_user_domain	src_comp	dst_comp	auth_type	logon_type	auth_orientation	Success
U113@DOM1	C1710	C1710	Negotiate	Interactive	LogOn	Success
U75@DOM1	C1710	C1710	Negotiate	Interactive	LogOn	Success
U75@DOM1	C1710	C1710	?	Interactive	LogOff	Success
U113@DOM1	C1710	C1710	?	Interactive	LogOff	Success
U6801@DOM1	C15205	C15205	Negotiate	Interactive	LogOn	Success

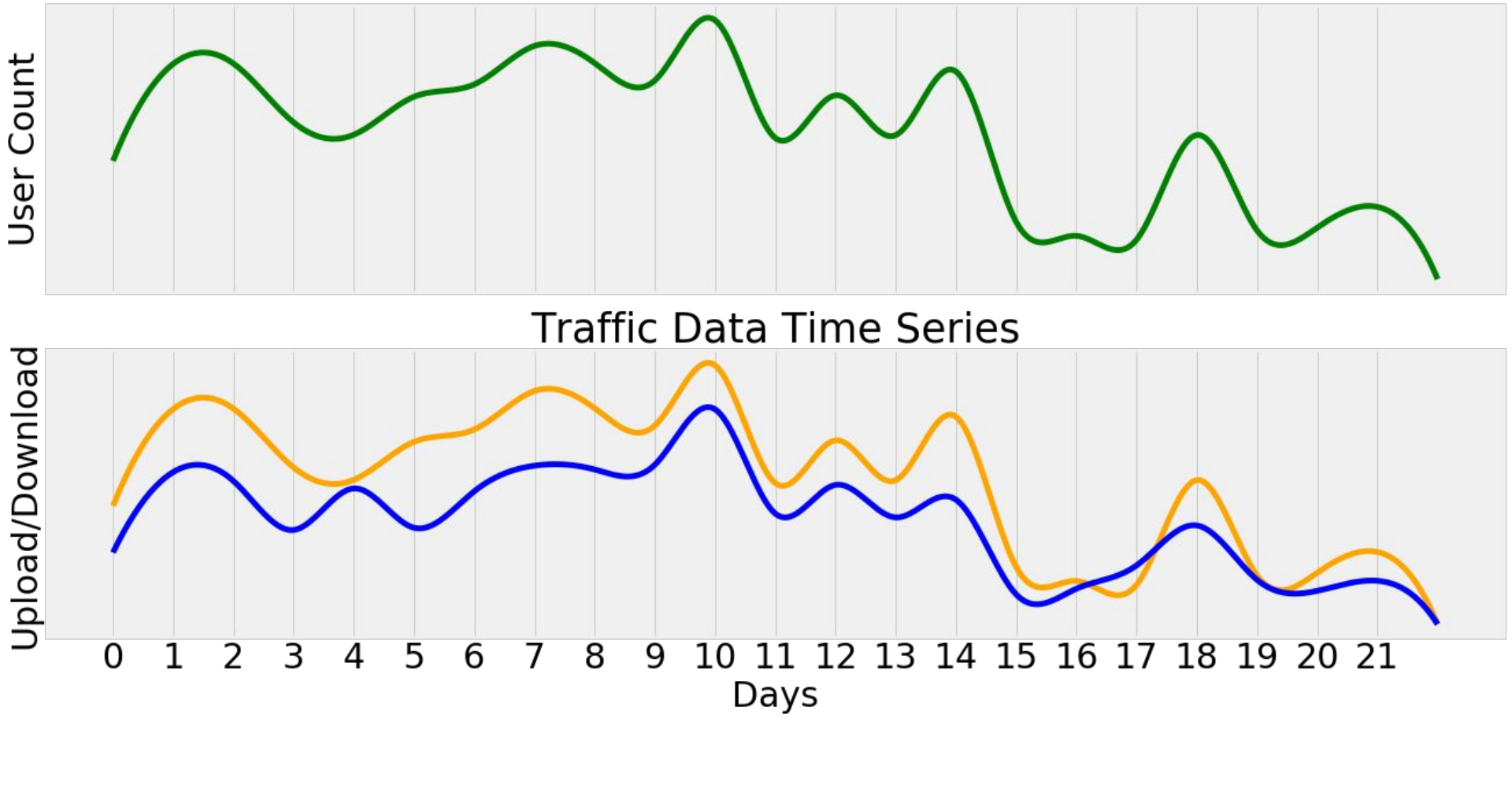


Filled DataFrame

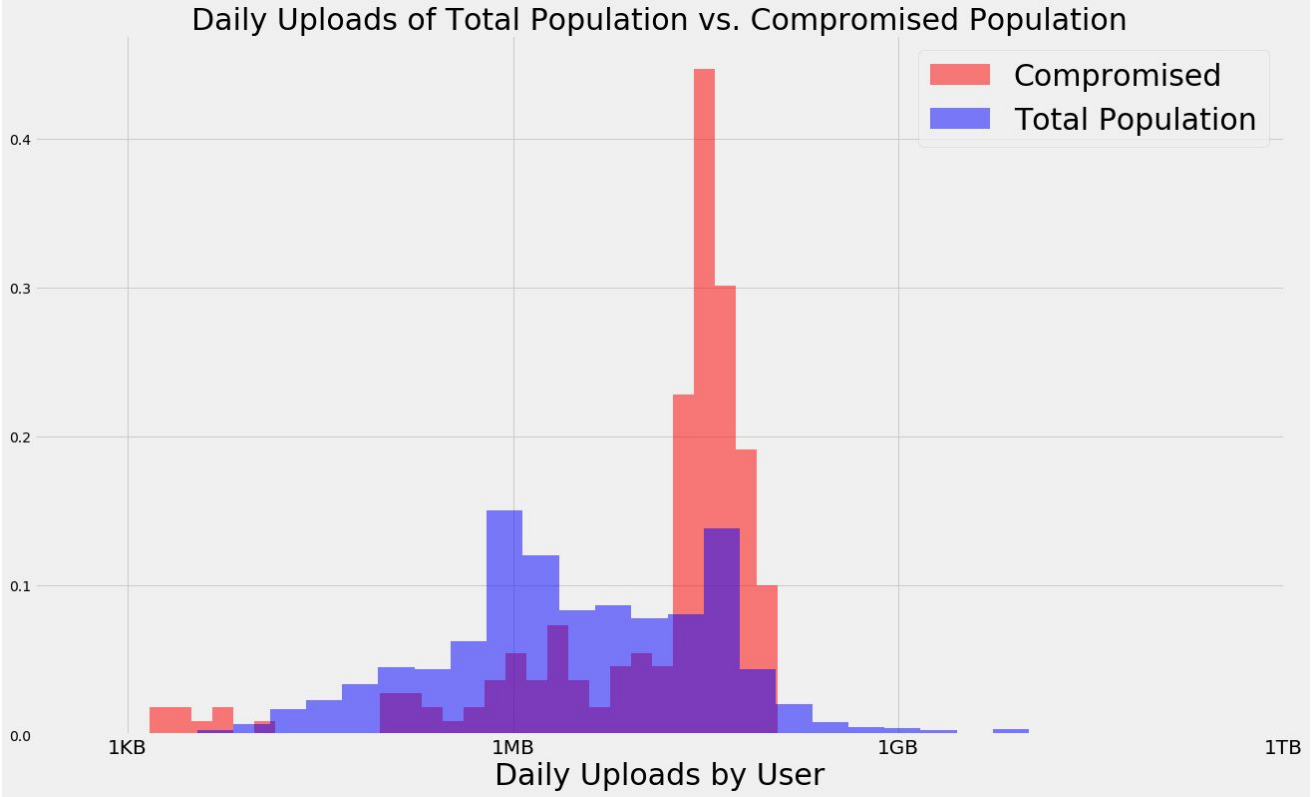
time	src_comp	dst_comp	Upload	Download	user	service	upload_bytes	download_bytes
47746	C15	C5721	1	0	C15	C5721	46	0
47776	C5720	C8751	0	1	C8751	C5720	0	18848
47776	C8751	C5720	1	0	C8751	C5720	5166	0
47784	C1015	C4773	0	1	C4773	C1015	0	92
47836	C8751	C5720	1	0	C8751	C5720	5055	0
47841	C5720	C8751	0	1	C8751	C5720	0	18848
47859	C1015	C4773	0	1	C4773	C1015	0	92
47866	C15	C5721	1	0	C15	C5721	46	0
47901	C8751	C5720	1	0	C8751	C5720	5212	0
47906	C5720	C8751	0	1	C8751	C5720	0	18848
47935	C1015	C4773	0	1	C4773	C1015	0	92
47966	C8751	C5720	1	0	C8751	C5720	5212	0
47968	C5720	C8751	0	1	C8751	C5720	0	18848
47986	C15	C5721	1	0	C15	C5721	46	0
48011	C1015	C4773	0	1	C4773	C1015	0	92
48031	C8751	C5720	1	0	C8751	C5720	5212	0
48032	C5720	C8751	0	1	C8751	C5720	0	23820
48087	C1015	C4773	0	1	C4773	C1015	0	92
48093	C8751	C5720	1	0	C8751	C5720	6305	0
48101	C5720	C8751	0	1	C8751	C5720	0	18848

only showing top 20 rows


Traffic Data Time Series
Download
Upload
Upload and Download Bytes
500GB
400GB
300GB
200GB
100GB
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Days


User Count
Traffic Data Time Series
Upload/Download
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Days

Compare Distributions of Daily Uploads

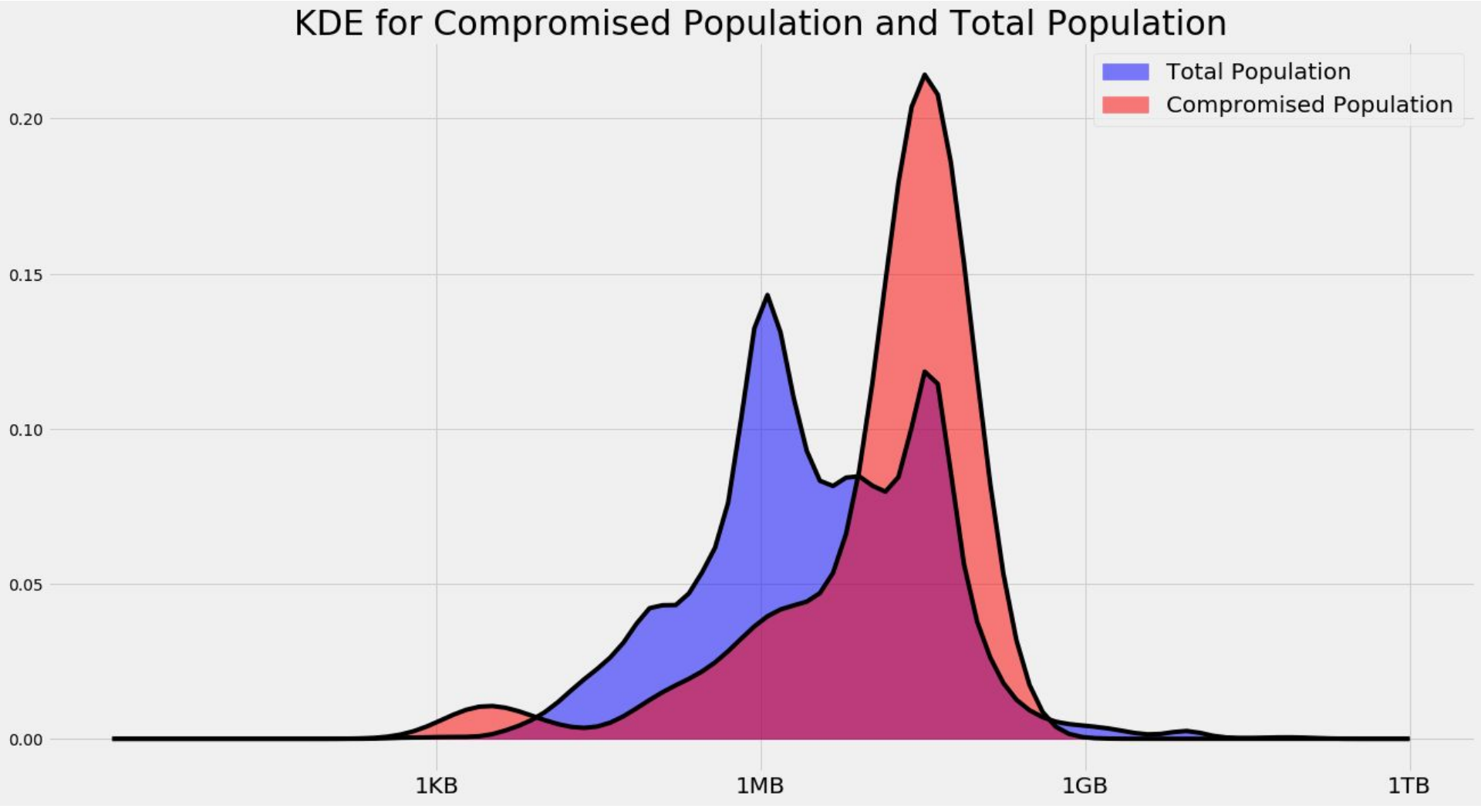


H_0=50% chance any random item from Compromised exceeds any random item from total

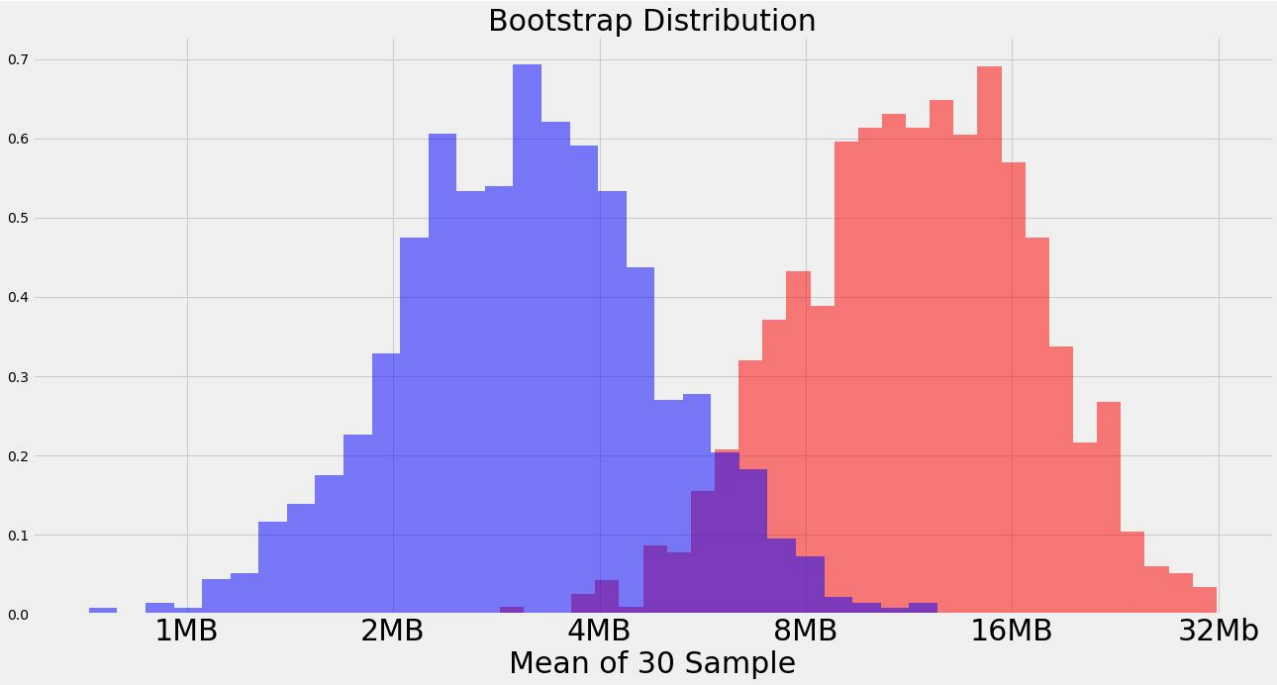
α = .05

Run Mann Whitney U Test on the two distributions

P=1.2472e-20


KDE for Compromised Population and Total Population
Total Population
Compromised Population
0.20
0.15
0.10
0.05
0.00
1KB
1MB
1GB
1TB

Further Questions



Given a group of 50 users is it possible to tell whether they were likely drawn from the Total Population of from the distribution of compromised users?