Morning Beacon
An Insight Company

Digital Alloy Preliminary Data Analysis

This report has two sections.
1. Network Traffic Analysis – This are simple statistics about hit rates and various frequencies.
2. Correlation Analysis – This part is about the analysis of relationship between users and their leadership status with it the social sub groups. This is the most value added section.

NETWORK TRAFFIC ANALYSIS

Summary Observations

- There is an unexpected spike in traffic each day at 12:00 EST.

File Statistics

Digital Alloy supplied two data files, Web logs and CDN logs. Here are some basic statistics from the files. The file statistics reported here are similar to what could be gathered through numerous other methods, so we don't expect any surprises or insight. It's mainly to verify some level of sanity in the data.

	Web Log	**CDN Log[1]**
Zip File Size	1,333,833,580	69,599,408 (?)
File size	9,107,169,115	61,435,871,232
Number of records in file	25,728,210	52,885,578
Start DateTime	21/Sep/2013:06:32:01 -0400	21/Sept/2013:00:00:00 Z
End DateTime	29/Sep/2013:06:25:01 -0400	22/Sep/2013:13:14:01 Z
Importable records	25,728,210	52,885,529
Imported Start	2013-09-21T10:32:01Z	2013-09-21T00:00:00Z
Imported End	2013-09-29T10:25:01Z	2013-09-22T13:14:01Z
Time Span (Seconds)	777180	54121
Hit Rate (hits/sec.)	33.1	977.2

1 The CDN log files were not used in all analysis and only a partial sample was needed. The cdnweek_sorted.log.2 file contained all of the 184,233,070 records ending 25/Sep/2013:16:07:46 Z. The cdnweek.log is a one day sample with 3,079,978 records and ends 21/Sept/2013:19:15:41 Z.

Hit Rates

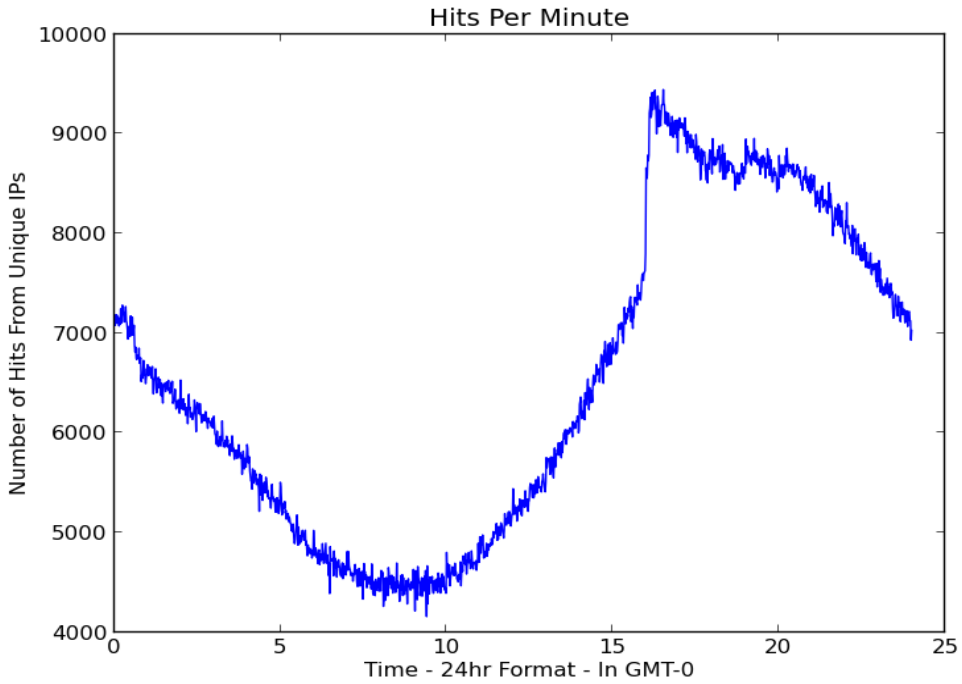
The hit rates are taken from the WebLogs only; the CDN hits are not considered. Like the file statistics, these measures should not be surprising since they can be found with a variety of commonly available tools.

Hit Rates per day

Date	Hours	Minutes	Count	Hits/Minute
Saturday / Sept 21	13:28 (13.47)	797	1839864	2308
Sunday / Sept 22	24:00	1440	2709081	1881
Monday / Sept 23	24:00	1440	3136865	2178
Tuesday / Sept 24	24:00	1440	3591921	2494
Wednesday / Sept 25	24:00	1440	4060703	2820
Thursday / Sept 26	24:00	1440	3453893	2399
Friday / Sept 27	24:00	1440	3261886	2265
Saturday / Sept 28	24:00	1440	2768672	1923
Sunday / Sept 29	10:25 (10.42)	625	905325	1449

The graph below shows the hit rate of unique IP address requesters each minute over the course of a day. The count is the sum of the number of hits for each of the days in the sample. There is a notable spike at 16:00Z (12:00 EST)

Morning Beacon
An Insight Company



Data Statistics

These statistics are based on the web log data only. CDN data is not included in this.

Verb Frequency

Verb	Count
GET	24,592,501
POST	1,016,244
HEAD	119,293
PROPFIND	42
OPTIONS	118
PUT	6
Undefined	6
TOTAL	**25,728,210**

Status Code Frequency

Status Code	Count
200	23,140,540
206	93,157
301	1,508,513
302	745,402
304	126,934
400	35
401	491
403	14,488
404	94,368
405	1
406	2
416	1
500	4278
Total	**25728210**

Unique Requestors and Referrers

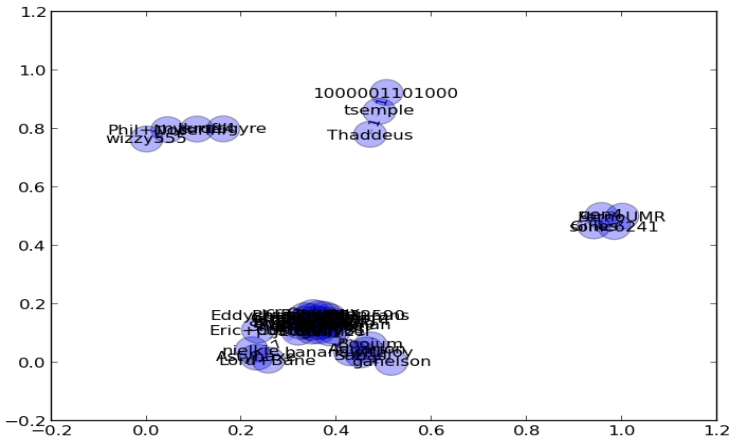
IP Type	Count
Unique Requesting IP	1,253,908
Average number of requests per Requester	20.5
Unique Referring IP	494,001
Escapist referrals	396,463

Morning Beacon
An Insight Company

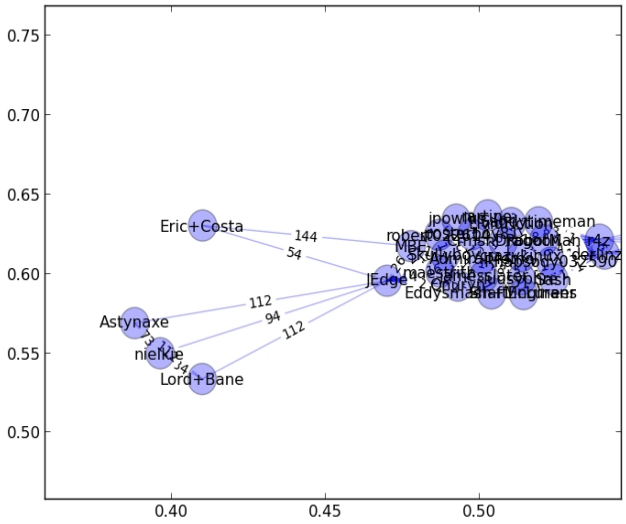
CORRELATION ANALYSIS

User Relationships

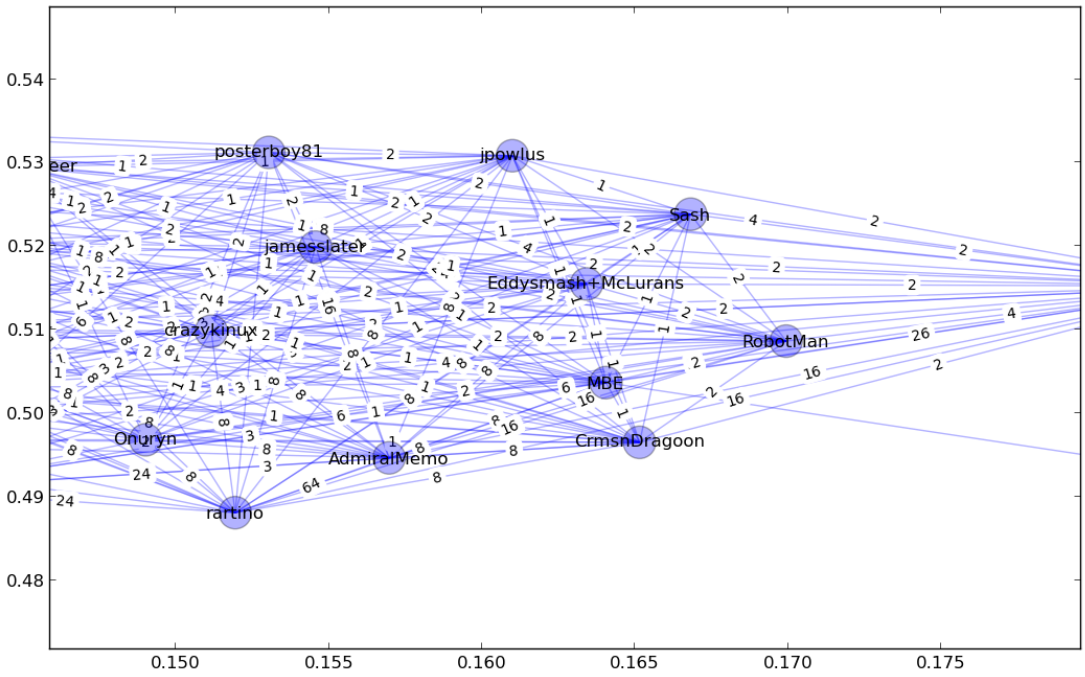
The image below shows a simple clustering of users for a a small 2 hour period. The analysis can be done for longer periods, but the graph gets so cluttered that it has no visual value. The graph only shows users that are in cliques, not the users that are in a clique of one.



The next image is a zoom-in of one portion of the graph to show the details. The numbers on the edges show the weight of the relationship.



This last network shows some of the connections inside of the large cluster so that we can make the connections with video watches.



When we run some correlation analysis, we found that (in a limited time interval),

- jamesslater watched 6 movies
- AdmiralMemo watched 2 of the same movies
- In all cases, AdmiralMemo watched the movie first
- MBE watched 4 movies
- AdmiralMemo watched 3 of the same movies
- Again, in all cases, AdmiralMemo watched the movie first.

Next Steps

We have extracted each video description and be able to find which users watched which type of movie reviews. Since we were able to identify the community leaders as show above next step would be to do correlation analysis between sub social groups and the types of movie reviews that they have watched and which leader have influenced the sub social group by watching what type of movie reviews.

Of course having access to user demographic information through digitalalloy's user database would allow us to generate better results since now we can use this information in our correlation analysis and be able to give more insight about the user interaction.