Salloumi.com (Project II)

Abdullah S. Al-Salloum (0801202023)
Abdulrahman Al-Khannah (0801202021)
Ahmed O. Al-Ayyar (0801102015)
Mejbil H. Al-Shammari (0801102061)

ECO580

STATISTICAL ANALYSIS FOR BUSINESS

DR. NASRELDDEIN SADDOULLI

PROJECT II

1/23/2009

Salloumi.com Bandwidth Usage:

Salloumi Videos is a website managed to share its users' videos. Since video websites consume high volume of bandwidth (transfer traffic) due to transferring video contents within its users' ISPs, the management of the website decided to analyze its bandwidth usage over the last six months.

The data attached to this project provides statistical information about the website usage of bandwidth according to its visits, page views, and hits. We need to make sense out of this data. In particular, we are interested in the basic summary of the data (means, variations, .. etc) for each month, as well as for the whole period. We are also interested in the presence of any patterns or trends that may be viewed. For instance, are there periods during the month which bandwidth usage becomes more significant which could help in the long-term scheduling of maintenance or tweaks?

We are also interested in making some inferences regarding the usage of bandwidth. In particular, they are interested in confidence intervals of the average bandwidth usage, in addition to the average of the total of the six months. The confidence levels typically used are 90%, 95% and 99%.

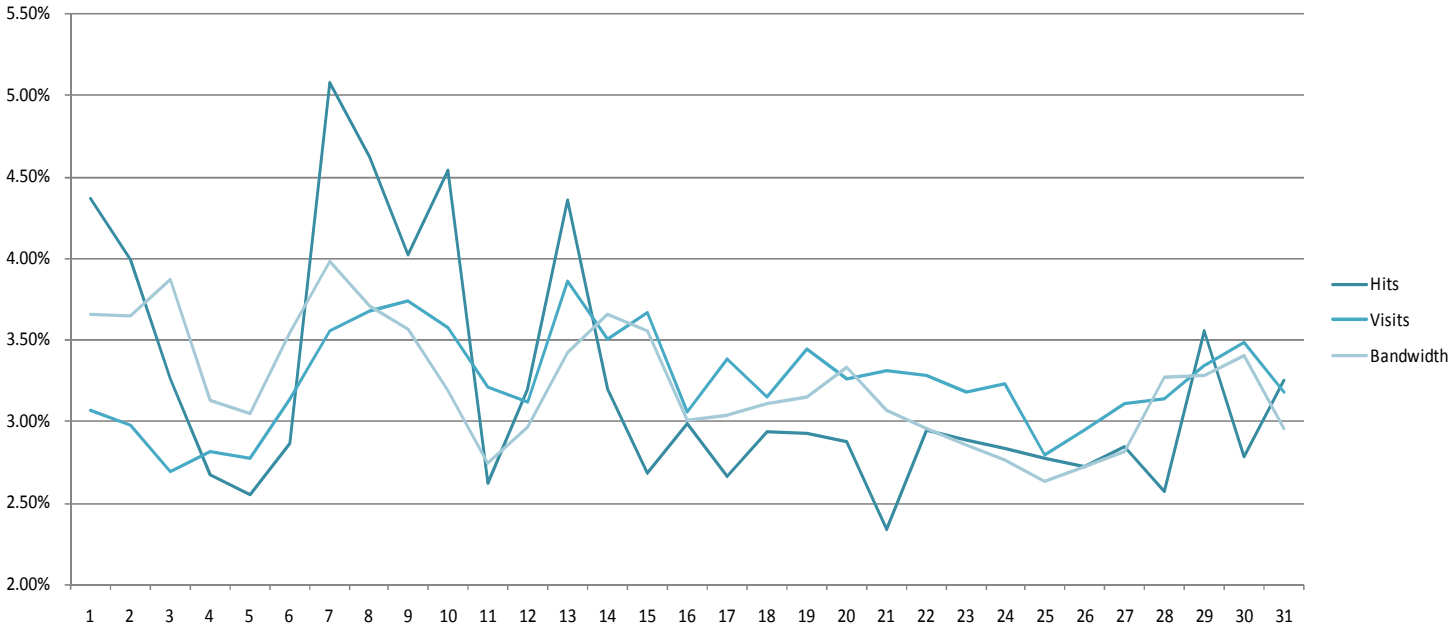
The server holding the website is located at Softlayer Inc. in Washigton DC. This data center offers free bandwidth usage up to the limit of 2,000GB per month. However, the management of Salloumi decided to cooperate with KuwaitNET Inc. to link all servers together to share the usage of bandwidth. The number of servers owned by KuwaitNET is 9 in which 10 will be the total of the servers. Each one already has 2,000GB where the total is 20TB. The servers of KuwaitNET consume the average of 10TB a month where the other 10TB is for the use of Salloumi Videos. We need to carry out the appropriate hypotheses tests to help the management make an informed decision either to take extra packages when the used bandwidth exceeds 10TB. A significance level of 5% and 1% are used since the consequences of making the wrong decision are very costly since the management will pay 0.10 USD per GB for any extra consumed bandwidth.

The bandwidth usage will depend upon hits and visits. We need to know whether hits and visits are independent so we get better measurements in limiting bandwidth usage.

Finally, we need to forecast the bandwidth usage over the next six months using regression model. Carrying out the appropriate methods to find out the forecasting equation and how each variable is responsible in affecting our dependent variable (Bandwidth).

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

July 2008 Overview:



In July, 2008, as from the beginning, we see that in percentage, bandwidth usage is about 3.60% compared to hits which are 4.50% and also compared to visits which are 3.10%. Since a visitor which is counted as 1 in visits can interpret as much hits as he stays on the website. As long as he hits files, the bandwidth usage increases.

In July, we see that hits, bandwidth, and visits are fair enough as of the beginning of the month. However, from 6 to 11 of July, we see that hits and bandwidth are higher in percentage than bandwidth. This can be caused due to the friendly interface on the website where users are interested in text or images rather than watching videos. Hitting a video is counted as 1 and so hitting an image or text file. However, videos are using more bandwidth than texts. That is why leading visitors to spend more hits on texts is our aim. The more hits, the high rank we gain.

From 15 to 31 of July, we see that hits are less in percentage than bandwidth where we figure out that users during this period are interested in videos more than texts or images. It can also refer to a problem in protecting videos from being downloaded, the more videos downloaded, the more bandwidth used. Assume downloading protection was broken, a user watching a video will, meanwhile, download another video or maybe two. Here, less visits, less hits, but more bandwidth.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 3, 7, 14, 20, 28, 30 of July ▼ 5, 11, 16, 25 of July

Tweaks are needed on (first view):

3_1, 7_1, 14_1, 20_1, 28_1, 30_1.

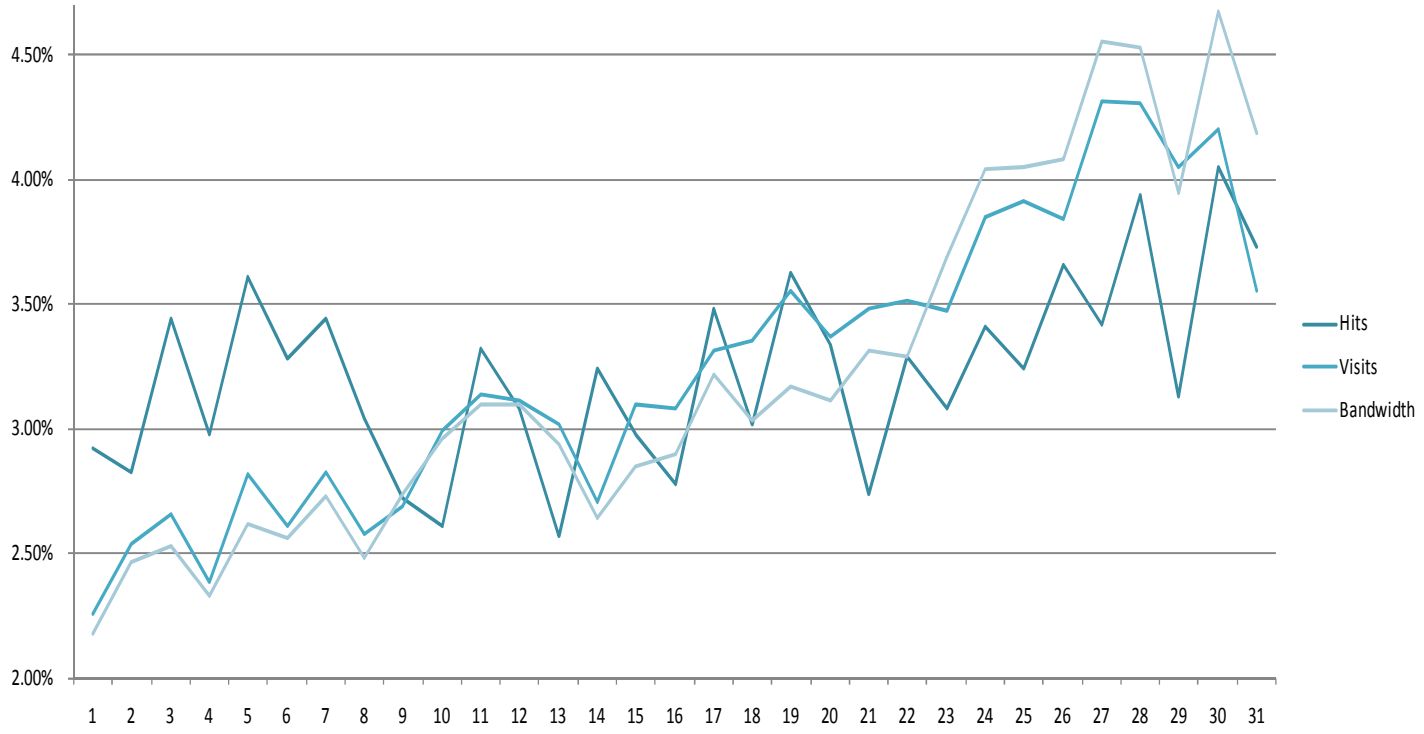
Tweaks are needed on (incrementally):

3_1, 7_1, 14_1, 20_1, 28_1, 30_1.

Mean	116987546.2
Standard Error	2357548.808
Median	114260350
Mode	#N/A
Standard Deviation	13126276.24
Sample Variance	1.72299E+14
Kurtosis	-0.819808026
Skewness	0.304448764
Range	48979309

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

August 2008 Overview:



What we see in August is a good process. Bandwidth usage was low compared to visits and hits which can be due to the community activities in Salloumi Videos such as "new profile's services" like instant messaging, commenting and so on. The overall process, then, goes smoothly higher which indicates that the website is going well. Bandwidth, hits, and visits are moving relatively higher.

As you may see, the there is nothing wrong with August usage of bandwidth, everything is relatively going well. However, we see at the end of August, hits percentage is less than visits percentage which indicates that page views per visitor is less within the ending period of this month.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 28, 30 of Aug. ▼ 1, 4, 8, 15 of Aug.

Tweaks are needed on (first view):

28$_2$, 30$_2$.

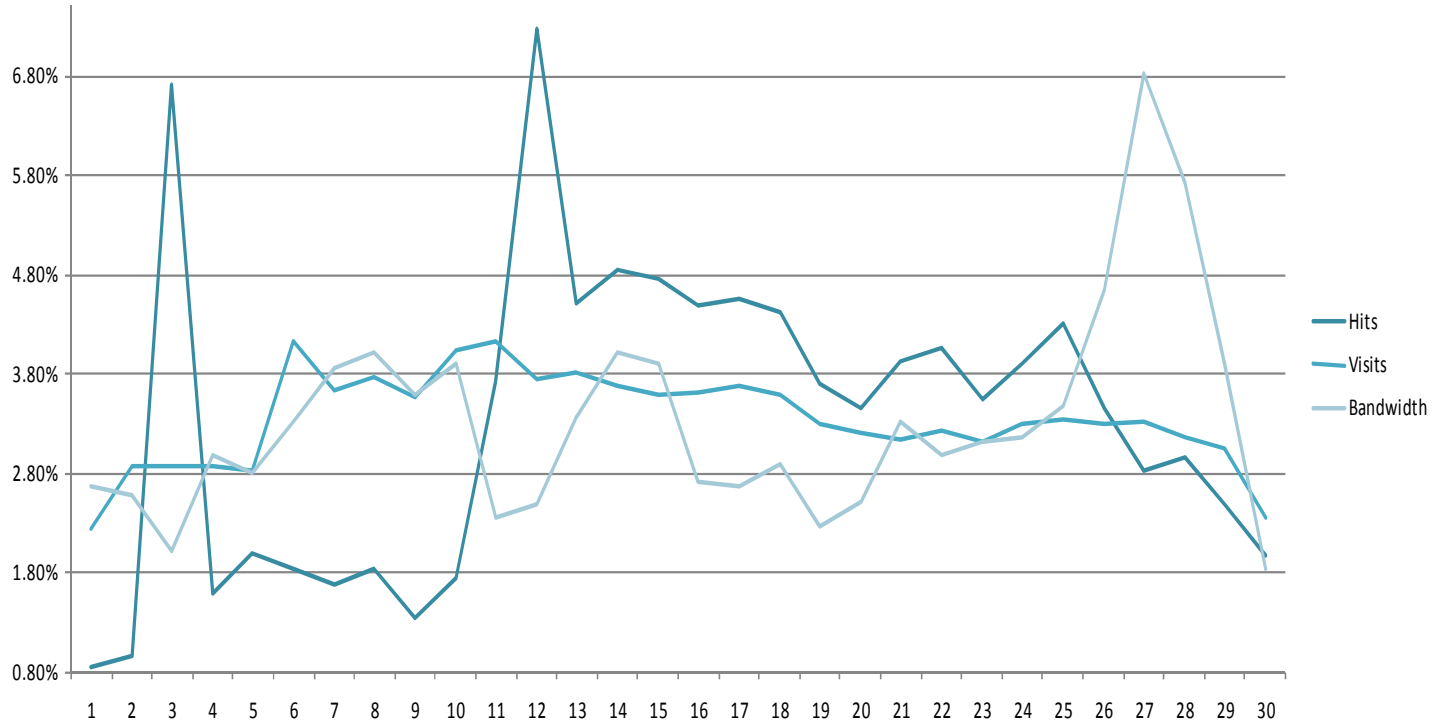
Tweaks are needed on (incrementally):

3$_1$, 7$_1$, 14$_1$, 20$_1$, 28$_2$, 30$_2$.

Mean	165646254.8
Standard Error	6493687.7
Median	159268683
Mode	#N/A
Standard Deviation	36155322.96
Sample Variance	1.30721E+15
Kurtosis	-0.645686169
Skewness	0.660636689
Range	127801365

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

September 2008 Overview:



What happens in September is quite different. We see that visits curve is stable and always between 1.8% and 3.9%. However, the variation of hits and bandwidth changes in high ratio. There are three noticeable actions occurring this month.

The first one on the 3rd of September where we notice that Hits are higher in percentage than bandwidth and visits which is caused by accessing more files other than videos (small sized files). The second one was on the 12th of September with the same action of the 3rd of September. The third one is different where we see our bandwidth is high in percentage and visits and hits are low. This can be caused due to watching long videos with low ratio of hits.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 7, 14, 20, 28 of Sep. ▼ 3, 11, 19, 30 of Sep.

Tweaks are needed on (first view):

7_2, 14_2, 20_2, 28_3.

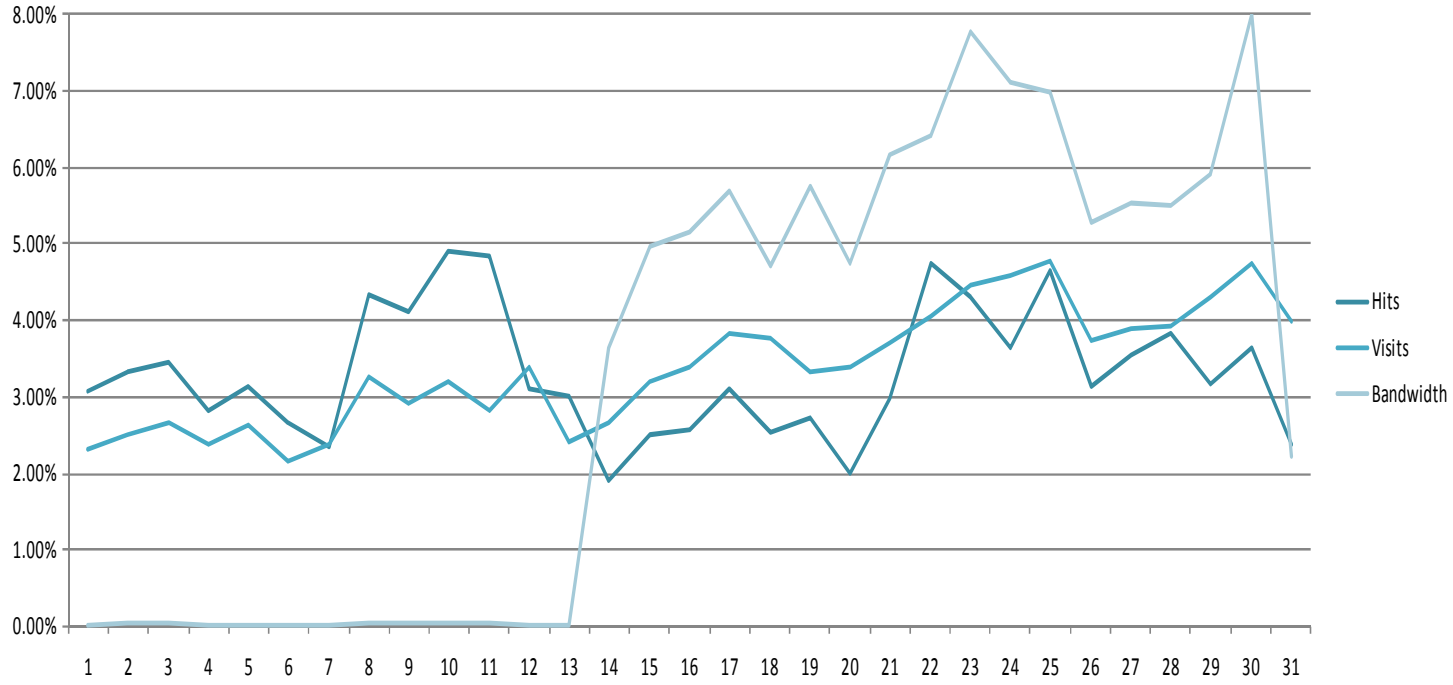
Tweaks are needed on (incrementally):

3_1, 7_2, 14_2, 20_2, 28_3, 30_2.

Mean	285286637.5
Standard Error	16364095.16
Median	268578583
Mode	#N/A
Standard Deviation	89629840.51
Sample Variance	8.03351E+15
Kurtosis	3.565861167
Skewness	1.581843556
Range	425730818

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

October 2008 Overview:



In this analysis there will be something different. As you see in the graph, the bandwidth usage percentage from the 1st to the 13th of October is about 0%. That was due to a problem in the Statistics software of Salloumi server. The software did not count the bandwidth as it should. Therefore, this month will be excluded from our 6-month analysis because the wrong information will give us wrong estimation and would affect our regression model.

However, what we can notice in this month according to the second half is that the bandwidth is always above visits and hits. This month would be considered as a good month for watching than visiting and hitting. But, on the 31st of October we see than hits, visits and bandwidth all fall. This might be caused due to a holiday day.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 17, 19, 23, 30 of Oct. ▼ 18, 20, 31 of Oct.

Tweaks are needed on (first view):

17_1, 19_1, 23_1, 30_3.

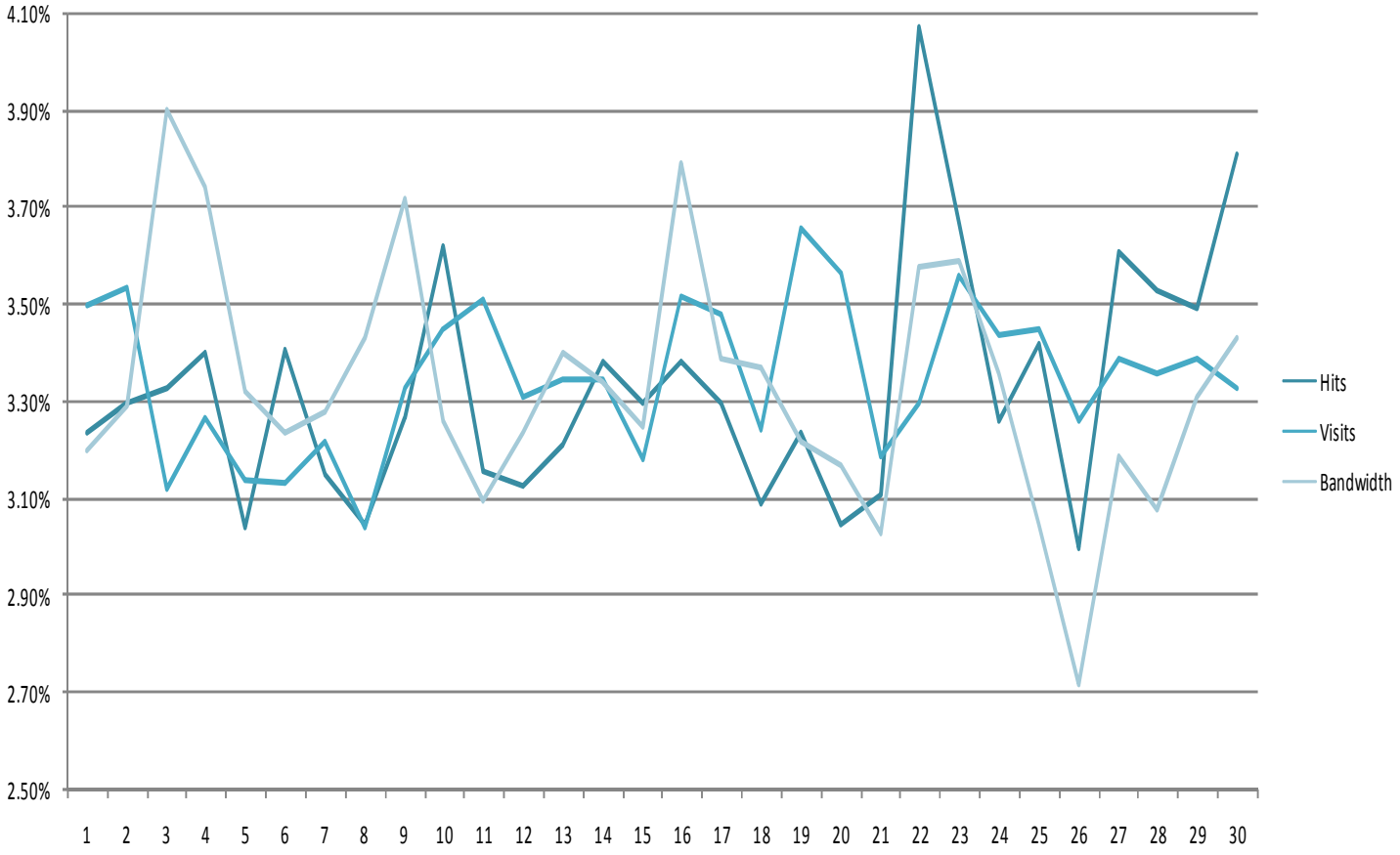
Tweaks are needed on (incrementally):

3_1, 7_2, 14_2, 17_1, 19_1, 20_2, 23_1, 28_3, 30_2.

Mean	59200059.45
Standard Error	10079240.26
Median	86276348
Mode	#N/A
Standard Deviation	56118834.72
Sample Variance	3.14932E+15
Kurtosis	-1.8092145
Skewness	0.031515941
Range	146445370

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

November 2008 Overview:



Although we see a huge variation on the 26th of November we, in a wider look, see that Salloumi is stable and the variations between hits, visits and bandwidth are relative. The website is processing very well during this month as should be expected.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 3, 9, 16, 23, 30 of Nov. ▼ 11, 21, 26 of Nov.

Tweaks are needed on (first view):

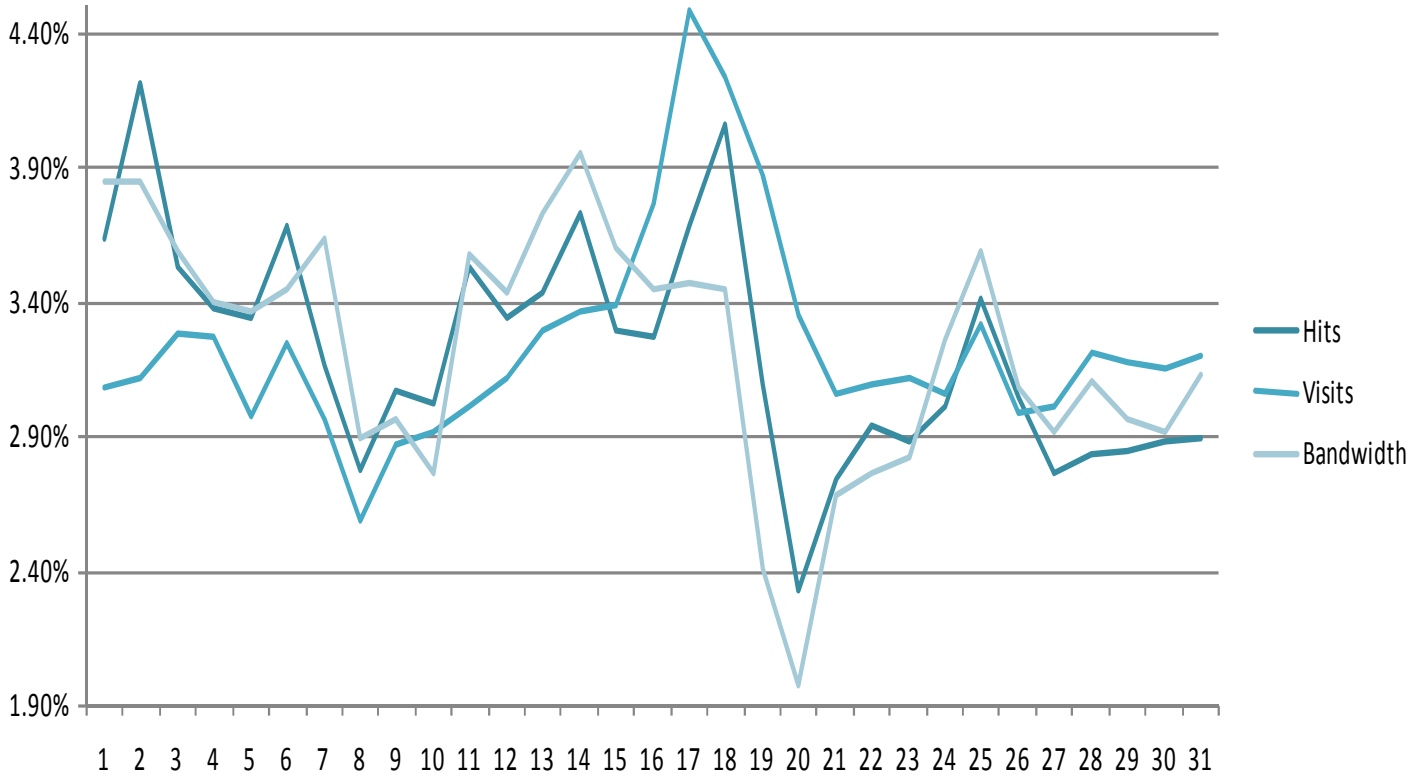
3_2, 9_1, 16_1, 23_3, 30_3.

Tweaks are needed on (incrementally):

3_2, 9_1, 7_2, 14_2, 16_1, 17_1, 19_1, 20_2, 23_2, 28_3, 30_3.

Mean	467018536.2
Standard Error	6350625.648
Median	462614359
Mode	#N/A
Standard Deviation	34783809.22
Sample Variance	1.20991E+15
Kurtosis	0.802215017
Skewness	0.276216285
Range	165071535

December 2008 Overview:



All changes in December are normal. As hits and visits change, bandwidth changes in similar percentage. According to the last two months we notice that the statistical information is getting quite accurate and so the website performance doing. The bandwidth usage falls on the 20th of December due to the fall of visits and hits and that is regular. The highest bandwidth usage of this month was on the 15th of December due to the high percentage of hits and visits which is normal too.

SUMMARY OF THIS MONTH

Bandwidth Average Summary, in days:

▲ 1, 2, 7, 12, 15, 25 of Dec. ▼ 10, 20 of Dec.

Tweaks are needed on (first view):

1_1, 2_1, 7_3, 12_1, 15_1, 25_1.

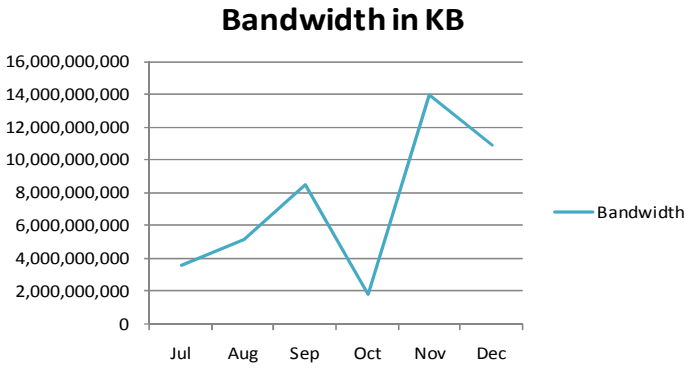
Tweaks are needed on (incrementally):

1_1, 2_1, 3_2, 9_1, 7_3, 12_1, 14_2, 15_1, 16_1, 17_1, 19_1, 20_2, 23_2, 25_1, 28_3, 30_3.

Mean	354480947.6
Standard Error	9200220.624
Median	370967879.5
Mode	#N/A
Standard Deviation	50391683.7
Sample Variance	2.53932E+15
Kurtosis	0.369284103
Skewness	-0.691837245
Range	216445886

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

6-month Overview:

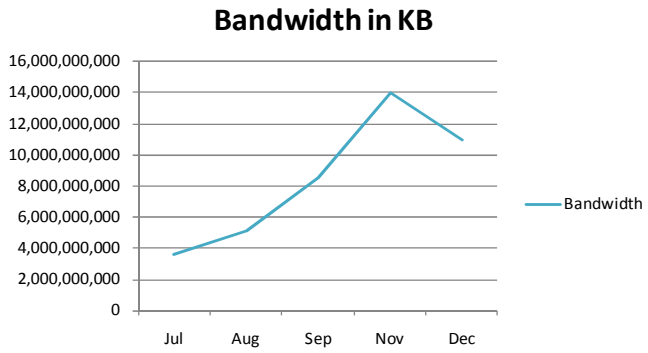


As you can see, the bandwidth increased from July to September. Then it falls on October. It increases in November in high volume and then falls again on December.

The fall on October should be counted since it isn't real and it was due to software issues. So we need to eliminate counting October to see how this graph should be.

That is how it should appear. The bandwidth in increasingly getting higher month after month due to the raise in website popularity which is normal. The usage of bandwidth started from 4TB, and increased up to 14TB on November, then declined to 11TB on December. This declining could be due to Christmas holiday.

The increase in such percentage could be due to the popularity of some video clips which are distributed over the web. It could be also the high number of websites linking to Salloumi.



Mean	277313218.3
Standard Error	11083983.99
Median	260681067
Higher C.I. 95%	295546372
Lower C.I. 95%	259080064.6
Higher C.I. 99%	303083481.1
Lower C.I. 99%	251542955.5
Standard Deviation	136652532.3
Sample Variance	1.86739E+16
Kurtosis	-1.272880554
Skewness	0.279597515
Range	487791743
Minimum	95481159
Maximum	583272902
Sum	42151609180
Count	152
Confidence Level(95.0%)	21899723.27

Here we have the descriptive statistical information. We see that the average usage of bandwidth is about 0.25TB per day which is good for a video website actually. We got also the higher C. 95% which is 295546372KB, and the lower C. 95% which is 259080064.6KB.

Bandwidth in KB	
Jul	3626613927
Aug	5135033899
Sep	8558599123
Oct	1833700535
Nov	14010556088
Dec	10978348223

Abdullah S. Al-Salloum
Abdulrahman Al-Khannah
Ahmed O. Al-Ayyar
Mejbil H. Al-Shammari

Hypotheses Testing for Packaging Bandwidth:

✓ Formulating the hypotheses test: $\begin{cases} H_0: \mu \leq 10TB/n \\ H_1: \mu > 10TB/n \end{cases}$: One Tail;

July 2008:

(M_0=10TB/n, n=31, Standard Deviation= 13126276.23, Mean= 116987546.2)

T-Test: t = -87.2062959

When Alpha: 0.01:
Area = 0.5 – 0.01 = 0.49
t = 2.457
-87.2062959 ≤ 2.457? Yes! → Do Not Reject H_0

When Alpha: 0.05:
Area = 0.5 – 0.05 = 0.450
t = 1.697
-87.2062959 ≤ 1.697? Yes! → Do Not Reject H_0

Conclusion:
There is no evidence that the monthly average of the bandwidth usage exceeds the 10TB where it would not require bandwidth packaging.

August 2008:

(M_0=10TB/n, n=31, Standard Deviation= 36155322.96, Mean= 165646254.8)

T-Test: t = -24.1672217

When Alpha: 0.01:
Area = 0.5 – 0.01 = 0.49
t = 2.457
-24.1672217 ≤ 2.457? Yes! → Do Not Reject H_0

When Alpha: 0.05:
Area = 0.5 – 0.05 = 0.450
t = 1.697
-24.1672217 ≤ 1.697? Yes! → Do Not Reject H_0

Conclusion:
There is no evidence that the monthly average of the bandwidth usage exceeds the 10TB where it would not require bandwidth packaging.

September 2008:

(M_0=10TB/n, n=30, Standard Deviation= 89629840.51, Mean= 285286637.5)

T-Test: t = -2.936104649

When Alpha: 0.01:
Area = 0.5 – 0.01 = 0.49
t = 2.457
-2.936104649 ≤ 2.457? Yes! → Do Not Reject H_0

When Alpha: 0.05:
Area = 0.5 – 0.05 = 0.450
t = 1.697
-2.936104649 ≤ 1.697? Yes! → Do Not Reject H_0

Conclusion:
There is no evidence that the monthly average of the bandwidth usage exceeds the 10TB where it not would require bandwidth packaging.

November 2008 (Note we skipped October):

(M_0=10TB/n, n=30, Standard Deviation= 34783809.22, Mean= 467018536.2)

T-Test: t = 21.05071378

When Alpha: 0.01:
Area = 0.5 – 0.01 = 0.49
t = 2.457
21.05071378 ≤ 2.457? No! → Reject H_0

When Alpha: 0.05:
Area = 0.5 – 0.05 = 0.450
t = 1.697
21.05071378 ≤ 1.697? No! → Reject H_0

Conclusion:
There is evidence that the monthly average of the bandwidth usage exceeds the 10TB where it would require bandwidth packaging.

December 2008:

(M_0=10TB/n, n=31, Standard Deviation= 50391683.7, Mean= 354480947.6)

T-Test: t = 3.467341025

When Alpha: 0.01:
Area = 0.5 – 0.01 = 0.49
t = 2.457
3.467341025 ≤ 2.457? No! → Reject H_0

When Alpha: 0.05:
Area = 0.5 – 0.05 = 0.450
t = 1.697
3.467341025 ≤ 1.697? No! → Reject H_0

Conclusion:
There is evidence that the monthly average of the bandwidth usage exceeds the 10TB where it would require bandwidth packaging.

Accordingly, bandwidth packages should be taken into consideration in November and December. Therefore, we need to take a look at the hits visits of both months to know the amount of hits and visits causing bandwidth to be higher than the limit of 10TB.

November:

Total Hits	13301462
Total Visits	46119

December:

Total Hits	11736732
Total Visit	48675

As you see, we should take bandwidth into consideration when hits are above 1,000,000 hits, and when visits are above the 45,000 per month which are reasonable.

Hits vs. Visits:

We agreed that we take bandwidth usage into consideration when total hits are above 1,000,000 and visits are above 45,000. The problem is that we have to check both visits and hits. We, yet, do not know if hits are dependent upon visits and, accordingly, we are going to do a test.

✓ Formulating the hypotheses test: $\begin{cases} H_0: \text{Hits are independent of visits} \\ H_1: \text{Hits are not independent of visits} \end{cases}$

	Visits	Hits	Total
Jul	25497	4046358	4071855
Aug	34158	4392299	4426457
Sept	52538	14715895	14768433
Oct	16825	3639265	3656090
Nov	46119	13301462	13347581
Dec	48675	11736732	11785407
Total	223812	51832011	52055823

$\chi^2 = 20016.10317$ where p-value < 0.005
When Alpha is 0.05 and 0.01 we still reject H null.
So, there is evidence that hits are not independent of visits.

		fij	eij	fij - eij	((fij - eij)^2)/eij
Jul	Hits	4046358	4054348	-7990.22	15.74693637
	Visits	25497	17506.78	7990.217	3646.790069
Aug	Hits	4392299	4407426	-15126.6	51.91570455
	Visits	34158	19031.38	15126.62	12023.0165
Sept	Hits	14715895	14704937	10958.31	8.166270844
	Visits	52538	63496.31	-10958.3	1891.204405
Oct	Hits	3639265	3640371	-1105.78	0.335887341
	Visits	16825	15719.22	1105.782	77.78723366
Nov	Hits	13301462	13290194	11268.41	9.554190289
	Visits	46119	57387.41	-11268.4	2212.628886
Dec	Hits	11736732	11734736	1995.902	0.339472929
	Visits	48675	50670.9	-1995.9	78.61761024
				X^2 =	20016.10317

Bandwidth vs. Hits and Visits:

We thought about connecting these variables together according to our statistical analysis. We ran out a regression to figure out the equation that gives us the bandwidth usage in KB.

Regression Statistics								
Multiple R	0.714333501							
R Square	0.510272351							
Adjusted R Square	0.503698826							
Standard Error	96269858.88							
Observations	152							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	1.43885E+18	7.19423E+17	77.62537051	7.97296E-24			
Residual	149	1.38091E+18	9.26789E+15					
Total	151	2.81976E+18						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-29564513.64	31381508.48	-0.942099825	0.347666501	-91574786.91	32445759.62	-91574786.91	32445759.62
Hits	165.8647383	65.45584381	2.533994348	0.012310216	36.5231305	295.2063461	36.5231305	295.2063461
Visits	186987.4283	31691.97338	5.900150998	2.34939E-08	124363.6724	249611.1843	124363.6724	249611.1843

Although the variation explained by this regression is bad since R-square is 51% but it helps. Accordingly, our equation would be as the following:

$$Bandwidth\ in\ KB = -29564513.64 + 165.8647383H + 186987.4283V$$
$$Where\ H: hits, V: visits$$

Bandwidth Usage over Time:

Regression Statistics								
Multiple R	0.798689361							
R Square	0.637904695							
Adjusted R Square	0.635490726							
Standard Error	82503424.42							
Observations	152							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	1.79874E+18	1.79874E+18	264.2555782	6.62068E-35			
Residual	150	1.02102E+18	6.80682E+15					
Total	151	2.81976E+18						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	87651303.02	13450134.42	6.516760372	1.02725E-09	61075111.16	114227494.9	61075111.16	114227494.9
Time Period	2479240.722	152512.9126	16.25593978	6.62068E-35	2177889.651	2780591.794	2177889.651	2780591.794

Although the variation explained by this regression is bad since R-square is 63.5% but it helps. Accordingly, our equation would be as the following:

$$Bandwidth\ in\ KB = 87651303.02 + 2479240.722T$$
$$Where\ T: Time\ Period$$
$$Time\ period\ of\ 1st\ of\ January, 2009: 152$$

Conclusion:

Bandwidth goes high on..

Salloumi Videos consume bandwidth the most on the following days of each month: 3,7,14,20,23,28 and 30. During these days, management should be concerned about attracting visitors to use texted files more than videos so less bandwidth is consumed.

Hits and visits..

We do not have evidence that hits are dependent upon visits which means that in order to take bandwidth into consideration we take a look at visits or hits. We may limit hits to 1,000,000 or visits to 45,000 to eliminate bandwidth over usage.

Packages and the future..

We believe as per the forecasting equation given above that Salloumi Videos will need packaging of bandwidth from now on since the monthly bandwidth will exceed 10TB.

How this project benefited us!

Abdullah S. Al-Salloum

"Working on such cases gave me the needed confidence to work on such large sized data observations in the future career where I could get the needed information from semi-dummy numbers and visualize them using Excel. This project gave me the power to understand how Excel is one of the applications that leads to a successful work. Group-wise, I found it interesting and achievable although we faced many conflicts in opinions. I loved working on regression and figuring out the forecasting equations."

Abdulrahman M. Al-Khannah

"The main benefit of the project that I gained is the proper way to extract the needed information from the survey – observations to summarize and, then, analyze them where we finally get the appropriate conclusion. Regarding the group, I did not face any issue; however, the benefits were the group discussion and dividing the work effort to increase the output efficiency."

Ahmed O. Al-Ayyar

"I gained much information on how it works while getting observations and define the needed analysis. Working as a group member helped me a lot while distributing the work between us. It is a good experience."

Mejbil H. Al-Shammari

"The benefit that I got from this experiment is how to take serious decisions based on analysis which would assist me in my future career. However, regarding the group work we have experienced, I liked the work cooperation between all group members."