Exploring Campus Netflow for Managing Network

Hung-Jen Yang, Miao-Kuei Ho, Lung-Hsing Kuo, Hsieh-Hua Yang, Hsueh-Chih Lin

Abstract**—**The purpose of this study was to monitoring and modeling netflows on a campus in Taiwan. There is a need to understand the netflow on a flow base for maintain and managing purpose. Random sampled campus was first identified. The netflow of the campus was monitored and collected for analysis. Based upon statistical test results, weekday flow characteristics were identified. A model of netfolw according to the protocols was also created. It is concluded that there exists significant characteristics of weekday netflow and certain predictable behavior between total amounts of flow and domain service. The characteristics revealed by this study should be consider as the fundamental information for managing campus networks.

Keywords**—**Campus network, Netflow model

I. Introduction

Since the emergence of campus data networks, the volume of data traffic carried by internet has been growing continuously due to the rapid increase in subscriber base size, communication bandwidth, and computer device capability[1]. To cope with the explosive data volume growth and best serve teachers and students, campus network operators need to design and manage campus network architectures accordingly[2-4].

The fluency of a campus networks is important for teachers' teaching and students' learning. Teachers often search matters for teaching before class teaching, teach in the classroom, and share in the Internet community after class on networks. Students also search for study materials , learn in the classroom and writing the report on networks. Campus increasingly dependent on networks, management and monitoring network traffic would be beneficial to maintain fluency on networks.

The purposes of this study were to determine if (a) understanding the usage of the campus network; and (b) creating a campus netflow model. To achieve this, the first step is to understand the spatial and temporal patterns of Internet traffic carried on campus networks. Understanding the spatial and temporal patterns of traffic and help to estimate both short- and long-term changes in network resource requirements.

II. Literature review

In this session, the flow level traffic and traditional campus networks were discussed for the research purpose

A. Flow Level Traffic

Traffic can be thought at a different level of abstraction than that of the streams of packets that actually cross the network and it can be viewed as a flow process. There are two types of flows[5]. Those are unidirectional and bidirectional flows. Nevertheless the present architecture is already carrying a mixture of traffic of different nature, that is usually classified as streaming and elastic.

This traffic mix has no bidirectional definition because of streaming flow that have no opening and closing protocol procedure, and also peer to peer flows that are a kind of multi-point-to-point elastic data transactions, therefore a uni-directional flow definition is more convenient for studying the performance of a network interface.

Instead of considering the packet level trace, a live collection of data directly at the TCP flow level was conducted.

B. Traditional Campus Networks

A traditional campus network is shown in Fig. 1. The characteristics are listed in the following:

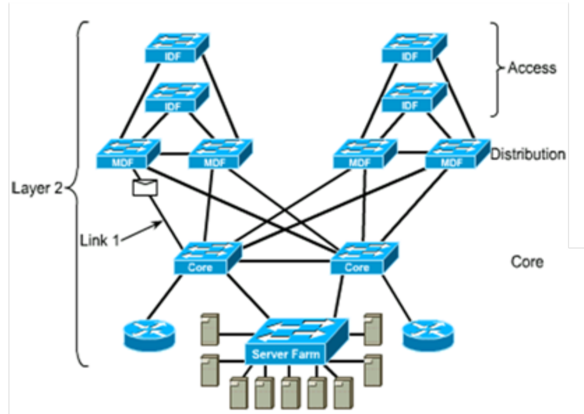

Fig. 1. Traditional Campus Networks

Manuscript received on August 1, 2013: Revised version received October 4, 2013. This work was supported in part by the Taiwan National Science Council under High Scope Project Grant 102-3113-S-017-003-GJ.

H.J. Yang is with the National Kaohsiung Normal University, 80201 Taiwan, R.O.C. (e-mail: hjyang@nknucc.nknu.edu.tw, phone: 886-7-7172930 ext. 7603,; fax: 886-7-6051206).

M.K. Ho is with the National Kaohsiung Normal University, 80201 Taiwan, R.O.C. (e-mail: hmq0402@gmail.com, phone: 886-7-7172930 ext. 7603,; fax: 886-7-6051206).

L.H. Kuo is with the National Kaohsiung Normal University, 80201 Taiwan, R.O.C. (e-mail:admi@nknucc.nknu.edu.tw)

H.H.Yang is with the Oriental Institute of Technology, 220 Taiwan, R.O.C. (e-mail:yansnow@gmail.com)

H.C. Lin is with the Taipei College of Maritime Technology, 25172, R.O.C. (e-mail: linhsiaochih@yahoo.com.tw)

- A building or group of buildings connected into one enterprise network that consists of or more LANs.
- The company usually owns the physical wires deployed in the campus.
- Generally uses LAN technologies.
- Generally deploy a campus design that is optimized for the fastest functional architecture over existing wire.

III. METHODOLOGY

For achieving the purpose of this study, a random sampled campus was identified and monitored for collecting research data. There were three group hypothesis set for testing. Based upon the statistical test results, research conclusion was reached.

A. Sample

The sampled campus is located in middle Taiwan. The source of this data was from the traffic statistics of a room in Taiwan Education Network Center. There are 342 organizations in the room, including 236 elementary schools, 72 junior schools, 30 high schools, two special schools and two teaching network center. In this study, the campus was random sampling in order to create a campus netflow model of the elementary schools.

In Table 1, detail information of the organizations of a room is listed in Taiwan Education Network Center.

Table 1. Analysis of organizations information in the room

Number of elementary schools	Number of junior schools	Number of high schools	Number of special schools	Number of teaching network center	Total number of organizations
236	72	30	2	2	342

There are 101 classes, 176 faculties, and 2999 students on the campus. The total numbers of people are 3175 on the campus.

In Table 2, detail information of school is listed.

Table 2. Analysis of school information

Number of classes	Number of staff	Number of students	Total number of people
101	176	2999	3175

B. The Data Collection

The data collected for this study were including Flows, amounts in Mbytes, and twenty sub-items of protocols in percentage. The Flows are presented the number of times, the Mbytes are presented the Mbytes and the twenty sub-items of protocols are presented the percentage of all organizations.

Data was collected from 2013/02/15 to 2013/03/28, the first six weeks in a semester. The sources of netflow Statistics are downloaded from The Education Network Center in Taiwan.

C. Data Analysis & Hypotheses

For revealing researched evidences, the mean, standard deviation were used. For verifying the hypothesis, t-test, one-way anova and regression were applied.

A independent t-test was applied to verify the data sets and the difference was verified by ANOVA procedure. A regression procedure was applied to explore collected data and to creat a model of campus netflows.

The null hypotheses of this study are listed in followings.

1.1. There exists no significant difference between the FLOWs mean values of weekday and weekend.

1.2. There exists no significant difference between the MBytes mean values of weekday and weekend.

2.1There exists no significant difference between the FLOWs mean values of each day.

2.2. There exists no significant difference between the MBytes mean values of each day

3.1. There exists no significant variables which could be used for predicting Mbytes.

The first and second group of hypothesis were applied to explore characteristics of campus netflows. The third hypothesis was applied for creating netflows model for predicting the behavior of netflows.

IV. FINDINGS

The purpose of this study was to exploring the campus netflows. In the first session, descriptive statistics of netflows were presented. In Table 3, the dependent variable of FLOWs was described. The mean, standard deviation, and N were listed according to the day of week. The highest value is on Monday with the mean of 560724.67 times. The lowest value is on Saturday with the mean of 264463.83 times. The lowest value of the standard deviation is on Monday (48875.580) and the highest value of the standard deviation is on Sunday (287224.071), they show the times of FLOWs are most concentrated on Monday and most scattered on Sunday.

Table 3. Descriptive Statistics of FLOWs

week	Mean	Std. Deviation	N
1	560724.67	48875.580	6
2	550216.00	65211.063	6
3	536775.83	54540.306	6
4	427973.83	156098.289	6
5	433116.00	137912.396	6
6	264463.83	163848.496	6
7	314966.83	287224.071	6
Total	441176.71	178614.524	42

In Table 4, the dependent variable of MBytes was described. The mean, standard deviation, and N were listed according to the day of week. The highest value is on Monday with the mean of 15276.4783 MBytes. The lowest value is on Sunday with the mean of 2247.9000 MBytes. The lowest value of the standard deviation is on Wednesday (2395.02902) and the highest value of the standard deviation is on Thursday (5478.88773), they show the MBytes are most concentrated on Wednesday and most scattered on Thursday.

Table 4. Descriptive Statistics of MBytes

week	Mean	Std. Deviation	N
1	15276.4783	2925.02902	6
2	14414.9767	4403.97061	6
3	15074.5700	2395.41020	6
4	10936.5817	5478.88773	6
5	12453.2567	4123.43313	6
6	4648.8150	4826.14580	6
7	2247.9000	2662.73554	6
Total	10721.7969	6140.30711	42

In Table 5, the dependent variables of each protocol were described on FLOWs. The N, Minimum, Maximum, Mean, and standard deviation were listed according to the protocols of icmp, www, https, proxy, domain, smtp, pop3, ftp, ssh, telnet, port137, port138, port445, port139, irc7000, irc6667, irc6660, and other protocol. The highest mean value is port445 with the mean of 579597.0145. The lowest mean value is pop3 with the mean of 2011.7709. The lowest value of the standard deviation is pop3 (5152.7922) and the highest value of the standard deviation is port139 (527046.8179), they show the protocols are most concentrated on the pop3 protocol and most scattered on the proxy protocol.

Table 5. Descriptive Statistics of each Protocol on FLOWs

	N	Minimum	Maximum	Mean	Std. Deviation
icmp	42	8564.3633	463143.5588	111060.7053	98301.5029
other	42	10355.4154	560030.6838	168457.8097	131312.5080
www	42	119494.4790	2381745.5420	562861.4002	467270.6763
https	42	33922.0500	1051111.2840	435791.7392	208909.0931
proxy	42	0	2407426.4090	318830.1069	446917.7025
domain	42	8480.5125	200211.6732	103670.5288	61344.9621
smtp	42	0	98925.4260	4351.3042	15797.1746
pop3	42	0	16719.4736	2011.7709	5152.7922
ftp	42	16719.4736	291196.9260	91727.2912	67106.3626
ssh	42	0	2089934.2000	204587.7653	346465.2361
telnet	42	8188.5252	1675108.7140	551580.7342	440475.5471
port137	42	19785.0852	864259.4937	179188.2575	147728.4893
port138	42	0	263543.8512	17706.4327	47588.8468
port139	42	0	3074015.2660	205942.8718	527046.8179
port445	42	281163.4800	759670.4392	579597.0145	146569.0899
irc7000	42	0	292576.1408	29573.8463	56686.1454
irc6667	42	0	1470853.2650	159014.5936	262757.5890
irc6660	42	0	434706.3136	86022.0558	126793.2284

Valid N (listwise)	42

In Table 6, the dependent variables of each protocol were described on MBytes. The N, Minimum, Maximum, Mean, and standard deviation were listed according to the protocols of icmp, www, https, proxy, domain, smtp, pop3, ftp, ssh, telnet, port137, port138, port445, port139, irc7000, irc6667, irc6660, and other protocol. The highest mean value is telnet with the mean of 15874.7038. The lowest mean value is pop3 with the mean of 41.2361. The lowest value of the standard deviation is pop3 (127.56322) and the highest value of the standard deviation is port139 (16971.70025), they show the protocols are most concentrated on the pop3 protocol and most scattered on the proxy protocol.

Table 6. Descriptive Statistics of each Protocol on MBytes

	N	Minimum	Maximum	Mean	Std. Deviation
icmp	42	45.16	14864.10	3218.5041	3254.24571
other	42	96.31	16505.89	4570.8441	4077.21083
www	42	901.84	23393.04	12662.1418	7068.49437
https	42	208.25	31675.26	11533.4792	7768.01273
proxy	42	.00	36499.42	6709.1009	8657.43677
domain	42	52.06	6033.38	3034.6167	2087.62772
smtp	42	.00	837.31	68.7223	185.50964
pop3	42	.00	493.38	41.2361	127.56322
ftp	42	115.24	9452.71	2372.6939	2269.53298
ssh	42	.00	61307.00	5248.2487	10271.26238
telnet	42	59.17	49286.71	15874.7038	13844.12930
port137	42	167.46	26149.16	4997.9193	4951.40879
port138	42	.00	7767.48	484.0502	1418.93331
port139	42	.00	103027.64	5774.5821	16971.70025
port445	42	1806.39	25460.85	14979.6231	8294.28242
irc7000	42	.00	8335.84	820.4397	1661.69972
irc6667	42	.00	49952.22	4244.1415	8711.98733
irc6660	42	.00	13149.02	2172.9916	3822.60207
Valid N (listwise)	42				

A. Difference between Weekday and Weekend

In this session, the difference between weekday and weekend is presented. A one-way ANOVA procedure was applied to verify the difference.

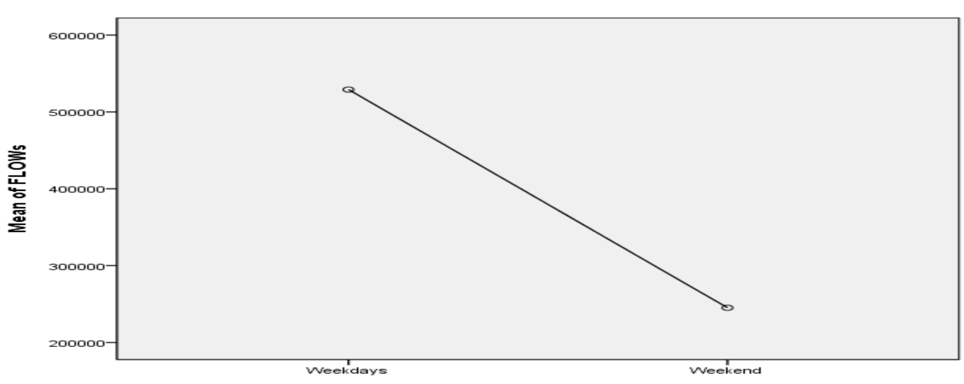

Fig. 2. Mean plot of FLOWs

In Fig. 2, the mean plot was shown the FLOWs mean of weekday and weekend.

Table7. Group Statistics of FLOWs

items	Group	N	Mean	Std. Deviation
FLOWs	weekdays	29	529028.69	57605.225
	weekend	13	245199.23	202484.684

In Table 7, the values of mean plot Fig. 2. are listed. The FLOWs mean of weekdays is 529028.69 times and the FLOWs mean of weekend is 245199.23 times. The FLOWs standard deviation of weekdays is 57605.225 and the FLOWs standard deviation of weekend is 202484.648.

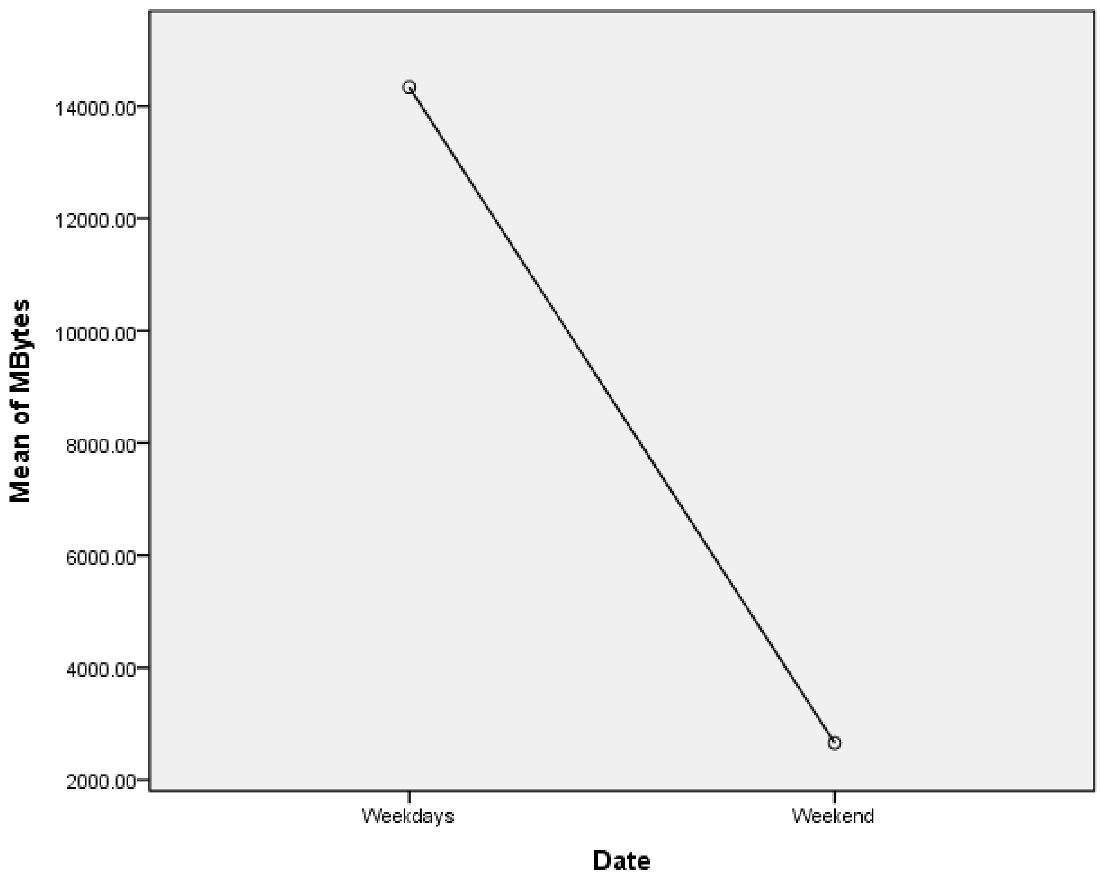

Fig. 3. Mean Plot of Mbytes

In Fig. 3, the mean plot was shown the MBytes mean of weekday and weekend.

Table 8. Group Statistics

items	Group	N	Mean	Std. Deviation
MBytes	weekdays	29	14336.0838	2977.22239
	weekend	13	2659.1569	2478.93106

In Table 8, the values of mean plot Fig. 3. are listed. The MBytes mean of weekdays is 14336.0838 MBytes and the

MBytes mean of weekend is 2659.1569 MBytes. The Mbytes standard deviation of weekdays is 2977.22239 and the MBytes standard deviation of weekend is 2478.93106.

Table 9. ANOVA of FLOWS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	723114380304.057	1	723114380304.057	49.451	.000
Within Groups	584914702450.515	40	14622867561.263		
Total	1308029082754.571	41			

In Table 9, an ANOVA table of FLOWs on weekdays and weekend is presented. According to the significant level less than 0.05, it is concluded that there exists significant difference between weekday and weekend. The analysis showed that the FLOWs on weekdays greater than the FLOWs on weekend.

Table 10. ANOVA of MBytes

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1223909146.791	1	1223909146.791	152.072	.000
Within Groups	321929079.119	40	8048226.978		
Total	1545838225.910	41			

In Table 10, an ANOVA table of MBytes on weekdays and weekend is presented. According to the significant level less than 0.05, it is concluded that there exists significant difference between weekday and weekend. The analysis showed that the MBytes on weekdays greater than the MBytes on weekend.

B. Difference among weekdays

In this session, the difference among days is presented and verified. A independent t-test was applied to verify the data sets and the difference was verified by ANOVA procedure.

Table 11. Test of Homogeneity of Variances

items	Levene Statistic	df1	df2	Sig.
FLOWs	1.74698372	6	35	0.139
MBytes	0.730161071	6	35	0.628

In Table 11, tests of homogeneity of FLOWs and MBytes were listed. According to the significant levels are both greater than 0.05, it is concluded that both variables are qualified for ANOVA test procedure.

Table 12. ANOVA summary table

		df	Mean Square	Sig.
FLOWs	Between Groups	6	82716137834.98	0.007**
	Within Groups	35	23192350164.13	
	Total	41		
MBytes	Between Groups	6	165063456.66	0.000**
	Within Groups	35	15870213.88	
	Total	41		

In Table 12, ANOVA summary information of degree of freedom, mean square, and significant level were presented. According to the significant levels, it is concluded that both variables exist significant difference among weekdays (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday) .

In the Fig. 4., based upon statistical processing, an estimated marginal means of FLOWs were drawn according to weekdays. The estimated marginal means of FLOWs were highest on Monday and lowest on Saturday. The Figure shows the FLOWs are more than 400000 times from Monday to Friday.

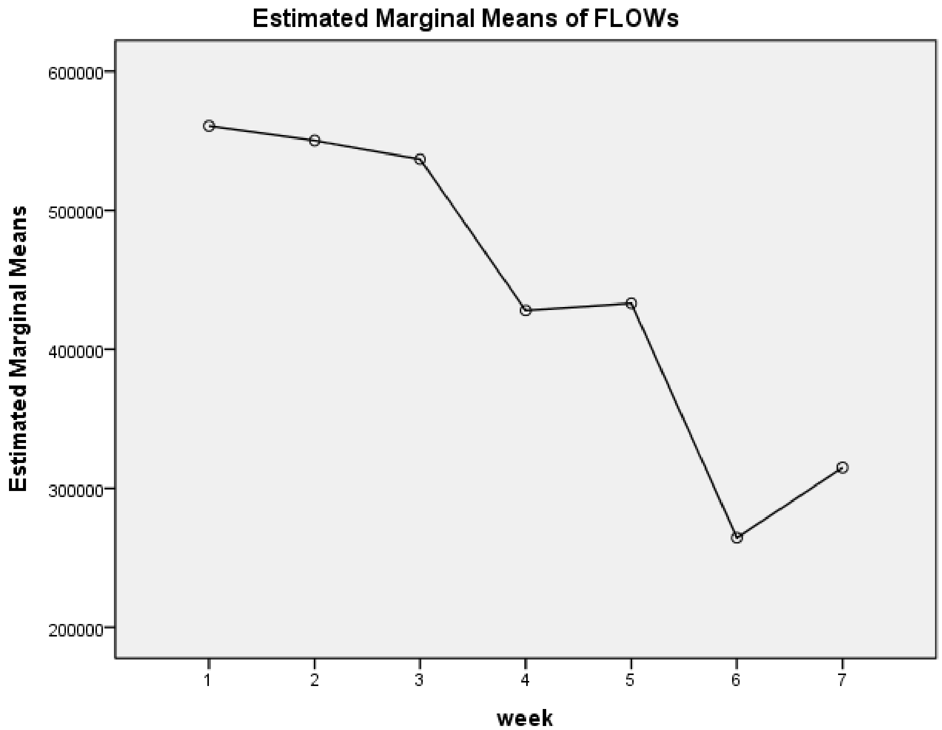

Fig. 4. Estimated Mean of FLOWs

In the Fig. 5, based upon statistical processing, an estimated marginal means of MBytes were drawn according to weekdays. The estimated marginal means of MBytes were highest on Monday and lowest on Sunday. The estimated marginal means of MBytes were highest on Monday and lowest on Saturday. The Figure shows the MBytes are more than 10000 MBytes from Monday to Friday.

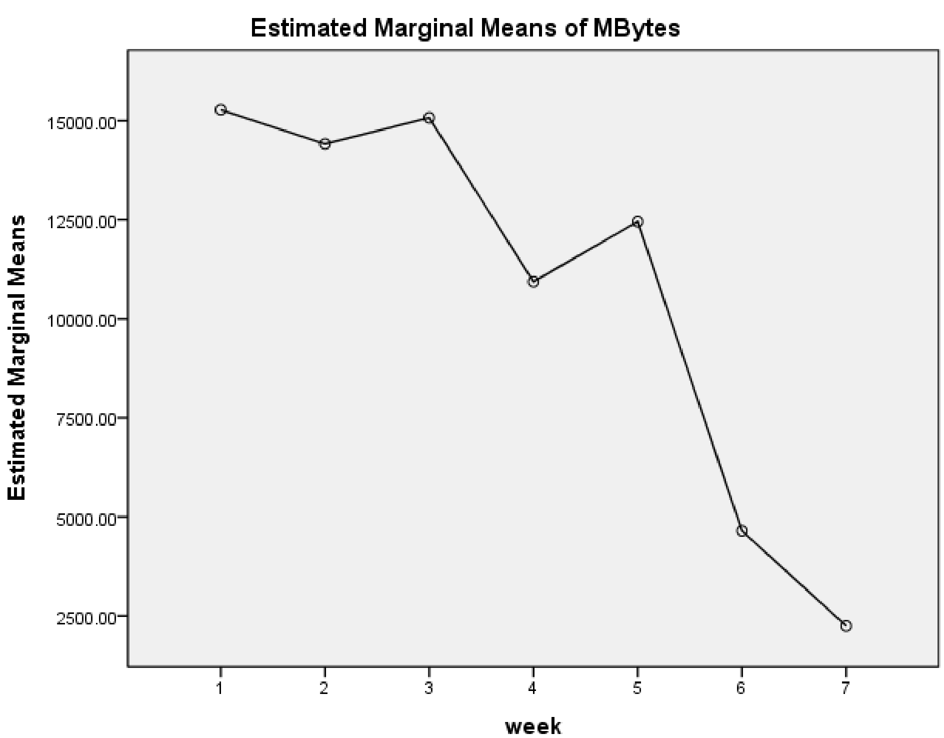

Fig. 5. Estimated Mean of Mbytes

C. Netflow Model

In this session, a model of campus netflows was created. A regression procedure was applied to explore collected data.

1. The prediction model on MBytes

A model of campus netflows was created on FLOWs.

In Table 13, the model with domain, other and irc6660 variables was entered. The criteria were probability-of-F-to-enter <= .050 and probability-of-F-to-remove >= .100 with stepwise.

Table 13. The model with variables entered

Model	Variables Entered	Variables Removed	Method
1	domain	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	other	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	irc6660	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: FLOWs

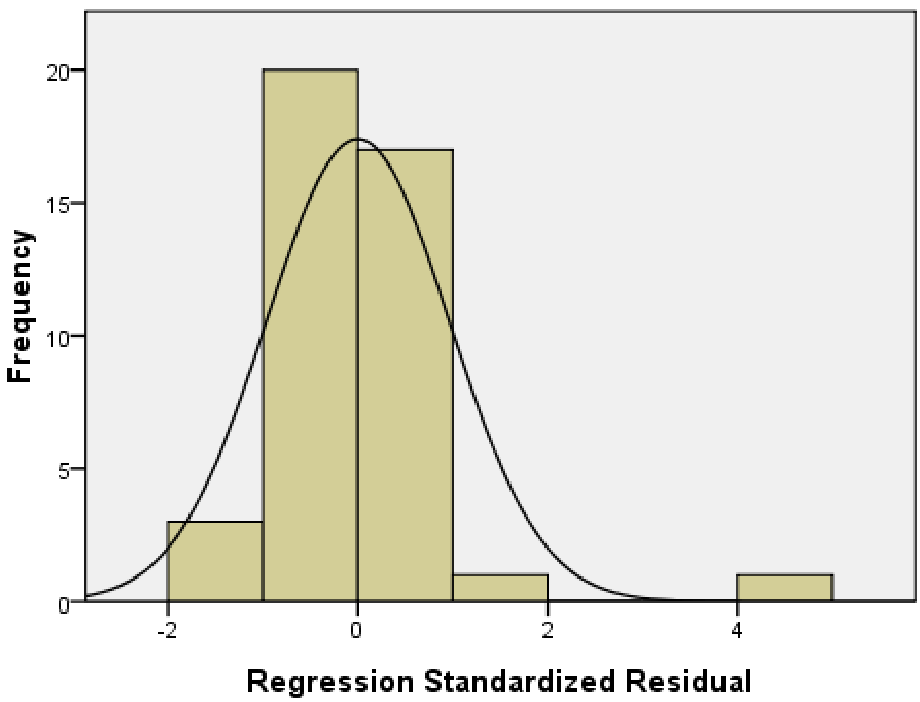

Fig. 6. Residuals of the model

In Fig. 6, the residuals of the model were presented. The histogram of Studentized residuals compares the distribution of the residuals to a normal distribution. The smooth line represents the normal distribution. The closer the frequencies of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

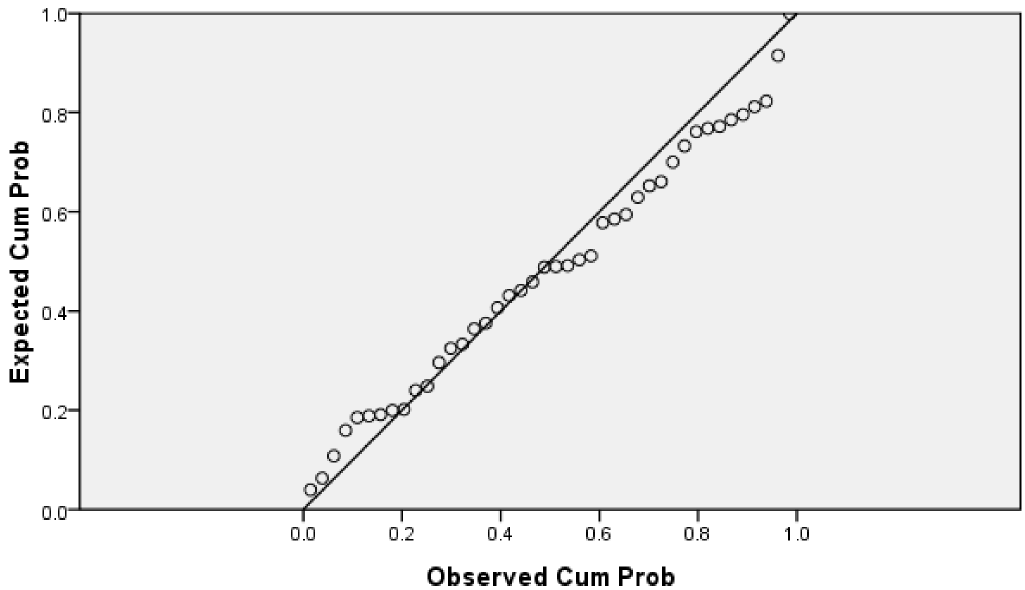

Fig. 7. The P-P plot of residuals

In Fig. 7, the residuals of the model were presented in a P-P plot. This Studentized residuals compares the distribution of the residuals to a normal distribution. The diagonal line represents the normal distribution. The closer the observed cumulative probabilities of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

Table 14. Model summary table

Sequence selected variables	R	R Square	R Square Change	F	F Change	Beta
domain	.719	.518	.518	42.922	42.922	.560
other	.780	.608	.091	30.290	9.035	.324
irc6660	.805	.648	.040	23.344	4.310	.201

In Table 14, the final model with domain, other and irc6660 variables was created. The overall contribution is 64.8%. The model is listed in the following formula.

FLOWs =0.560* domain+0.324* other+0.201* irc6660

2. The prediction model on MBytes

A model of campus netflows was created on Mbytes.

In Table 15, the model with www variable was entered. The criteria were probability-of-F-to-enter <= .050 and probability-of-F-to-remove >= .100 with stepwise.

Table 15. The model with variables entered

Model	Variables Entered	Variables Removed	Method
1	www	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: MBytes

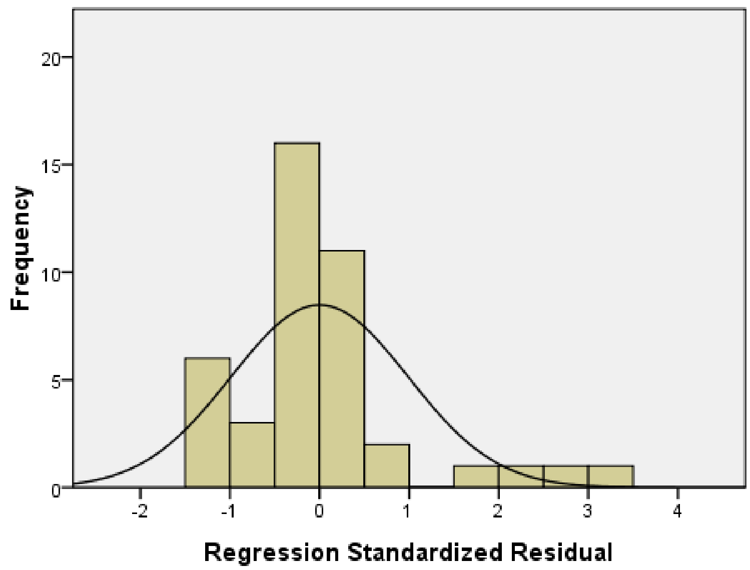

Fig. 8. Residuals of the model

In Fig. 8, the residuals of the model were presented. The histogram of Studentized residuals compares the distribution of the residuals to a normal distribution. The smooth line represents the normal distribution. The closer the frequencies of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

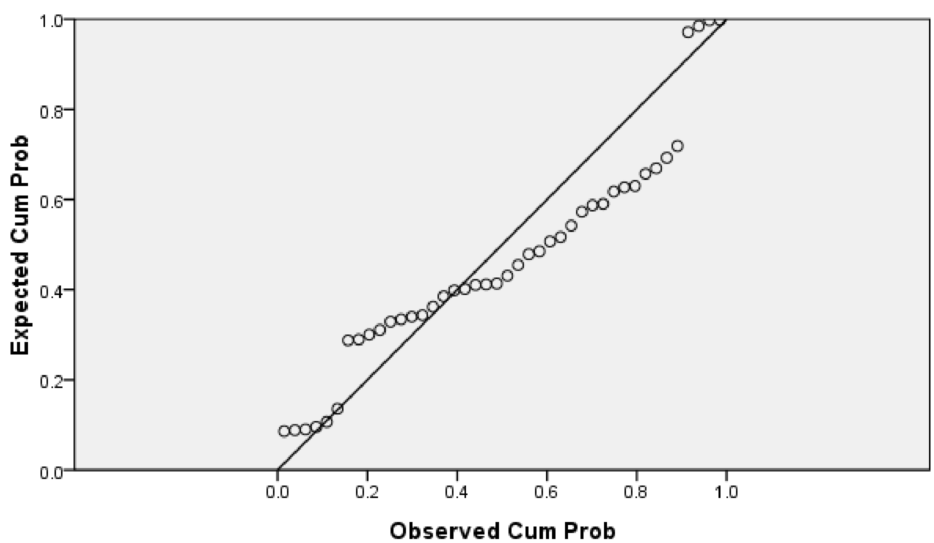

Fig. 9. The P-P plot of residuals

In Fig. 9, the residuals of the model were presented in a P-P plot. This Studentized residuals compares the distribution of the residuals to a normal distribution. The diagonal line represents the normal distribution. The closer the observed cumulative probabilities of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

Table 16. Model summary table

Sequence selected variables	R	R Square	R Square Change	F	F Change	Beta
www	.515	.265	.265	14.405	14.405	-.515

In Table 16, the final model with www variables was created. The overall contribution is 26.5%. The model is listed in the following formula.

Mbytes=-0.515* www

V. Conclusion

The purpose of this study was to monitoring and modeling netflows on a campus in Taiwan. There is a need to understand the netflow on a flow base for maintain and managing purpose. Random sampled campus was first identified.

The netflow of the campus was monitored and collected for analysis. Based upon statistical test results, weekday flow characteristics was identified. A model of netfolw according to the protocols was also created. It is concluded that there exists significant characteristics of weekday netflow and certain predictable behavior between total amounts of flow and domain service.

It is suggested that the model could be applied as the criteria for monitoring campus network. The characteristics revealed by this study should be consider as the fundamental information for managing campus networks.

References

[1] H.-J. Yang and H.-H. Yang, "Developing IT Knowledge Transfer Model by Using SEM," *WSEAS Transactions on Mathematics,* vol. 5, pp. 174-181, 2006.

[2] H. H. Yang, J. C. Yu, and H. J. Yang, "A Study of Verifying Knowledge Reuse Path Using SEM," in *the European Computing Conference 2007 Proceedings, September 25-27, 2007*, E. C. Conference, Ed., ed Vouliagmeni Beach, Athens, Greece,: WSEAS, 2007.

[3] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analysis and modelling of YouTube traffic," *Transactions on Emerging Telecommunications Technologies,* vol. 23, pp. 360-377, June 2012 2012.

[4] L. H. Kuo, H. M. Wei, L. M. Chen, M. C. Wang, M. K. Ho, and H. J. Yang, "An Evaluation Model of Integrating Emerging Technology into Formal Curriculum," *INTERNATIONAL JOURNAL OF EDUCATION AND INFORMATION TECHNOLOGIES*, vol. 6, pp. 250-259, 2012.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," presented at the SIGMETRICS'11, 2011.

Hung-Jen Yang obtained a Master (M.S.) in Technology Education from University of North Dakota and a Ph.D. in Industrial Technology Education from Iowa State University. He is currently conducting research on knowledge transfer, and knowledge reuse via information technology. His research has appeared in a variety of journals including those published by the WSEAS.

Miao-Kuei Ho received her Master (M.S.) in Industrial Technology Education (2003~2005) from National Kaohsiung Normal University. She is pursuing her Ph.D. in Industrial Technology Education from National Kaohsiung Normal University now. She has been teaching elementary school since 1990. As a senior teacher, she has had ten years field research experience with NSC projects. Her research interests include technology education, educational technology, and learning theory. She also has experts in technology process skills.

Lung-Hsing Kuo received his Master (M.E.) in Education (1990~1993) and Ph.D. in Education from (1993~1997) National Kaohsiung Normal University. He is the director of the center for teacher career and professional development in National Kaohsiung Normal University. His research interests include social Science Research Methodology, Continuing Education, Human and social, Youth Study, Emotion development and management, Counseling and Education Issues.

Hsieh-Hua Yang received her Master (M.S.) in Health Promotion and Health Education from National Taiwan Normal University and Ph.D. in Health Policy and Management from National Taiwan University. She is a professor in Health Care Administration at Oriental Institute of Technology. Her research focuses are health behavior, social network and organizational behavior.

Hsueh-Chih Lin received a Master (M.S.) in Industrial Technology Education from National Kaohsiung Normal University on the year of 2006. He received a Ph.D. in the department Industrial Technology Education at National Kaohsiung Normal University on the year of 2013. His research is focus on Design and technology, technology development and technology education.