# NetCube: a comprehensive network traffic analysis model based on multidimensional OLAP data cube

Daihee Park , Jaehak Yu , Jun-Sang Park and Myung-Sup Kim

Joobin Jin

Dept. of AI and Bigdata, SCH Univ.

jjb0821@sch.ac.kr

# Contents

1. Introduction
2. Requirement analysis and related work
3. NETCUBE: A comprehensive traffic analysis model
4. Experiments
5. Conclusion
6. How can I apply it?

# Introduction

- Due to the Internet utilization and data traffic spikes Network overloads, bottlenecks occur and struggling to run a stable service.

  ->To address these issues, network traffic monitoring and analysis is critical.

- Network monitoring and analytics are broadly divided into “active” and “passive” methods.
- Active methods provides slightly limited performance information.
- Passive methods measuring the performance of a target network by collecting and analyzing traffic data directly from specific sections of the network.

## Introduction – passive methods

- Passive methods can provide more information, but are limited in their ability to handle large volumes of traffic on high-speed networks.
- Current passive traffic monitoring systems don't provide comprehensive analytics

  -> Need to **multidimensional analysis**
- But, three reason why current manual analysis systems fail to perform multidimensional analysis.
  - 1. Lack of effective storage management methods for large, long-term traffic data
  - 2. Lack of effective integration methods for various separately built analytics methodologies
  - 3. Lack of effective ways to extract useful information from large amounts of traffic data accumulated over time

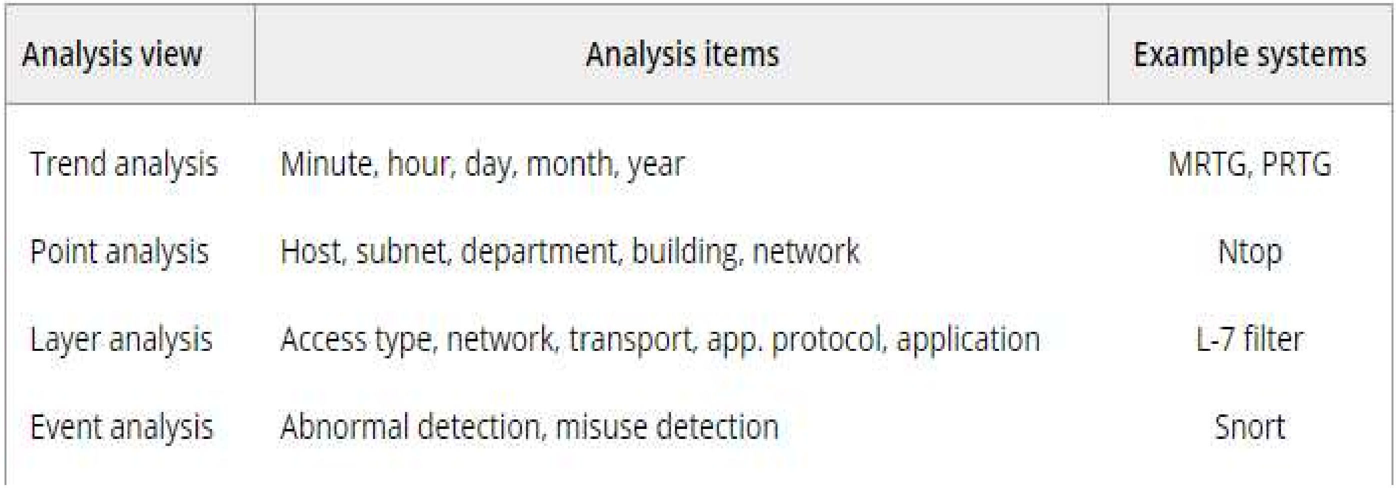
# Introduction – OLAP(On-line Analytic Processing)

- Propose a new design methodology for effective traffic analysis system that can overcome the limitations of the current passive analysis methods.
- A new design methodology

  -> Using long-term accumulated network traffic data in a data warehouse to build data cube models for multi-dimensional analytics required by network operators.
- Based on the built data cube model, perform multidimensional traffic analysis through OLAP operations with different levels of abstraction for different traffic analysis purposes

## Requirement analysis and related work

- Trend analysis : Analyze changes in network traffic over time
- Point analysis : Analyze the host or location from which the traffic was generated
- Layer analysis : Analyze traffic characteristics based on layers of the network protocol structure
- Event analysis : Analyze specific events in network traffic, such as unusual behavior of traffic



| Analysis view | Analysis items | Example systems |
|---|---|---|
| Trend analysis | Minute, hour, day, month, year | MRTG, PRTG |
| Point analysis | Host, subnet, department, building, network | Ntop |
| Layer analysis | Access type, network, transport, app. protocol, application | L-7 filter |
| Event analysis | Abnormal detection, misuse detection | Snort |

# Requirement analysis and related work - Limit

- **Trend analysis** is limited to analyzing traffic by host and subnet while displaying various graphs over time.
- **Point analysis** makes it difficult to analyze which applications generated the traffic or if the traffic was generated normally
- **Layer analysis** does not provide enough data to characterize traffic generated by a host or section over a long period of time and may not fully reflect traffic generated by anomalies
- **Event analysis** is difficult to effectively obtain trends and specialized information about event occurrences

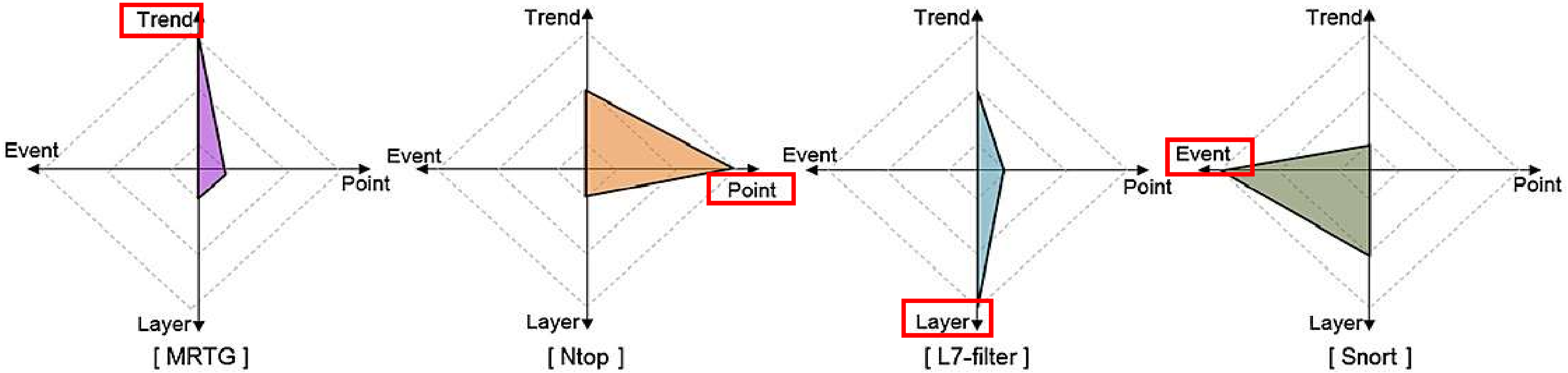


Figure1 Performance evaluation based on traffic analysis viewpoints with radar charts

# Requirement analysis and related work

- Trend analysis
- Point analysis
- Layer analysis
- Event analysis

→

- Limitations of network traffic analysis systems that use multiple perspectives, focusing on individual perspectives and failing to provide comprehensive analysis.

**※ Need to understand traffic from a multi-dimensional perspective to extract useful information for efficient Qos(Quality of Service) Provision and resource management**

# Requirement analysis and related work

- Propose new analytics models to store, combine, and analyze large-scale and long-term traffic data
- This model centers on building data cubes within a data warehouse and

  **online analytical processing(OLAP)** operations for multidimensional traffic analysis
- Data mining techniques can be applied to extract useful information for network operators

# NETCUBE: A comprehensive traffic analysis model

## -Overall NetCube traffic analysis model

- NetCube multidimensional traffic analysis model comprehensively reflects trend, point and layer perspective
- NetCube can analyze large amounts of traffic data over long periods of time through a variety of OLAP operations at multiple levels of abstraction

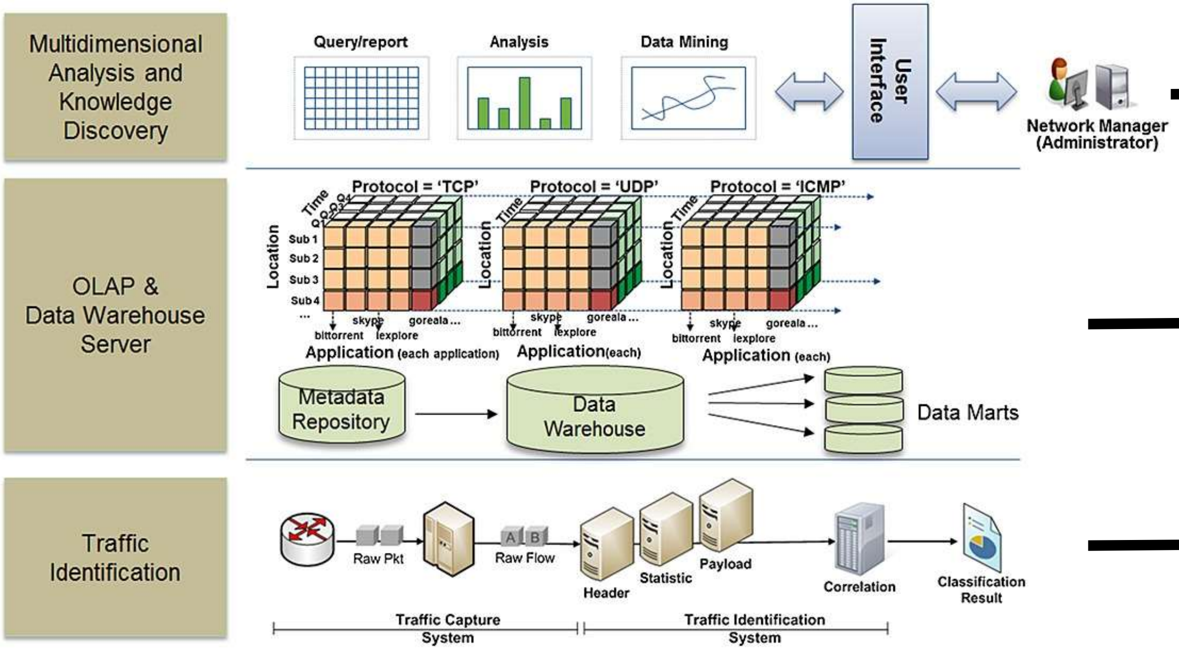


Figure2. overall NetCube traffic analysis model

→ Perform multidimensional analysis using OLAP and data-mining technique

→ The flow information is stored in the data

→ NetCube traffic analysis model performs application layer traffic identification process

# NETCUBE: A comprehensive traffic analysis model

## -Application traffic identification

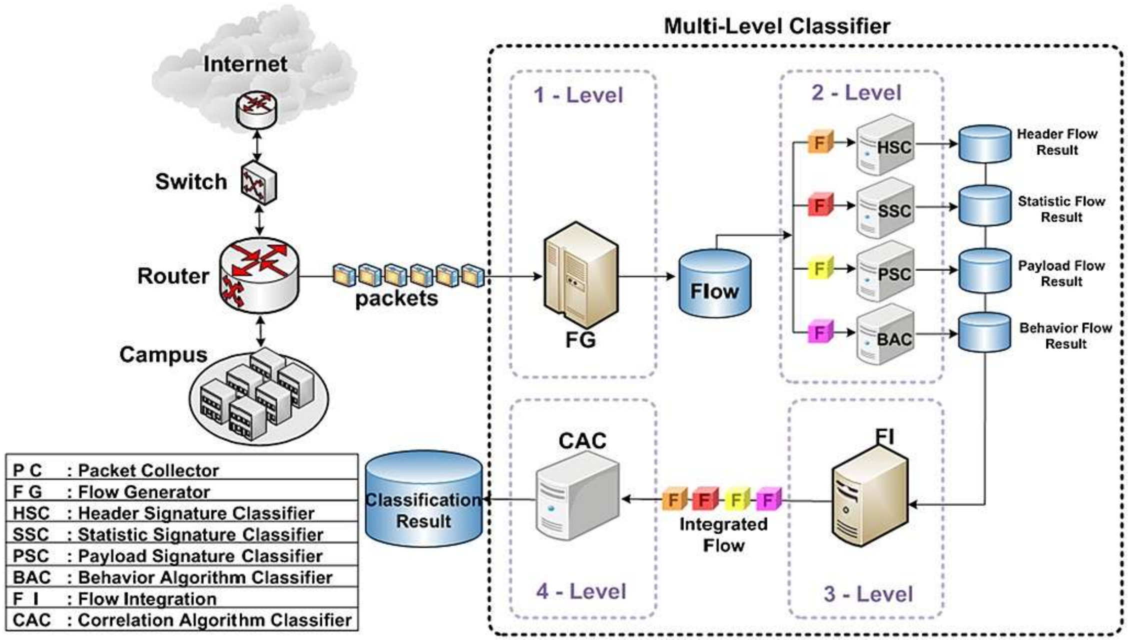


Figure2. overall NetCube traffic analysis model

- In the first level, the flow generator(FG) captures all the raw packets from a target network and aggregates them into a group
- The flow data are delivered to the second level, where several individual identification modules work in order to identify the application name of each flow separately
- The identification results of each flow are integrated in the third level: flow integration(FI)
- In the final level, the correlation algorithm classifier(CAC) additionally determines the application name of unknown flows which are not identified in the second level

# NETCUBE: A comprehensive traffic analysis model
## -Multidimensional data cube

- The model proposed in this study represented by a multidimensional data cube, which can model

  and display data from different perspectives

- It is defined by dimensions(4) and fact(1)

The perspectives or entities with respect to which a user wants to keep record

four dimensions: time for a trend analysis
location for a point analysis
application and protocol for a layer analysis

The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Used the bandwidth and flow count of traffic as measurements in a fact table

- The dimension in the data cube can be easily added if a new analysis viewpoint is required

# NETCUBE: A comprehensive traffic analysis model

## -Star schema

- 4 dimensions table and 1 fact table

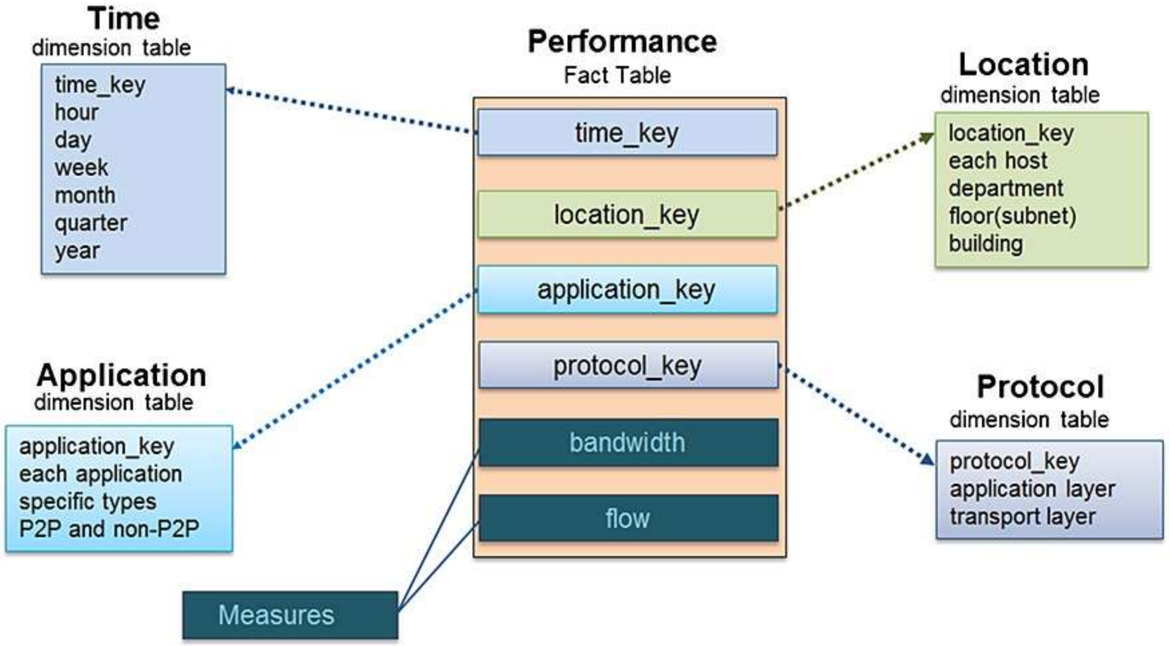


Figure4. NetCube star schema

# NETCUBE: A comprehensive traffic analysis model

## -Concept hierarchy

- A Concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level and more general concepts
- Concept hierarchies allow data to be handled at varying levels of abstraction

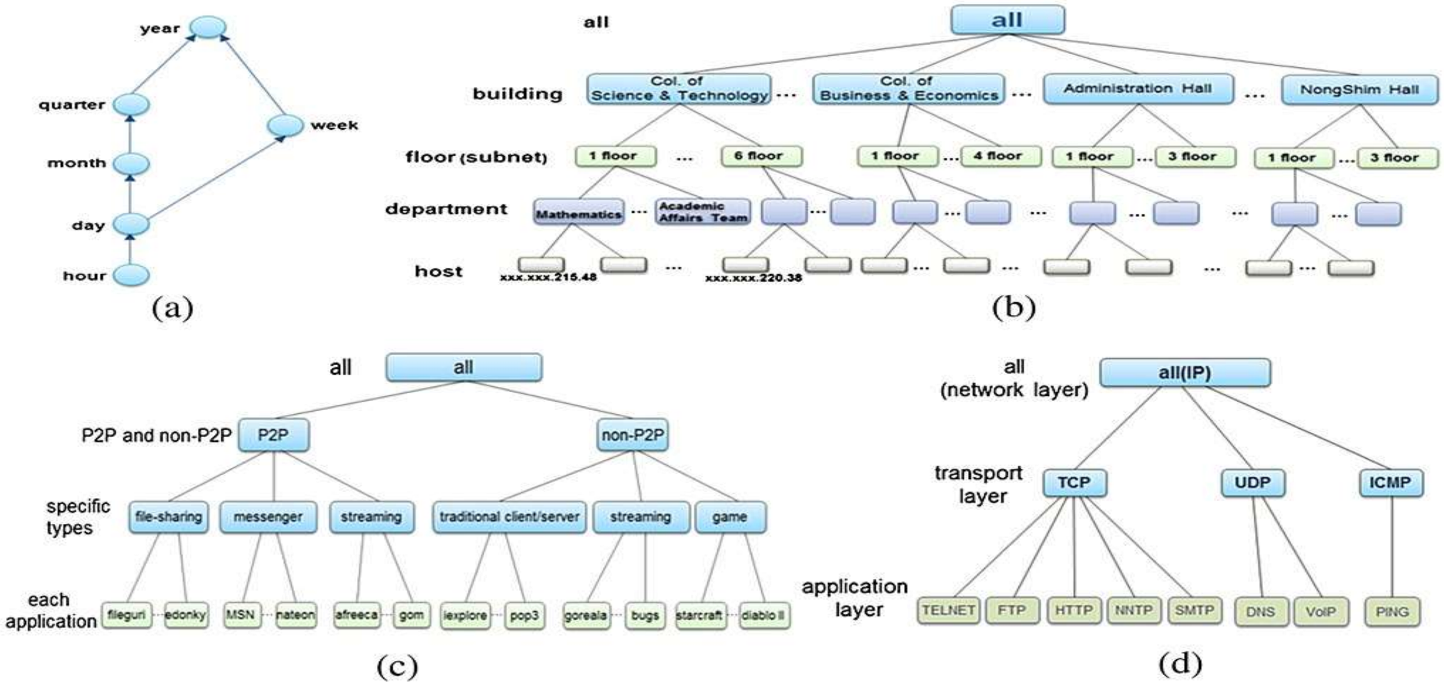


Figure5. Concept hierarchy

- (a) the dimension time
- (b)the dimension location
- (c) the dimension application
- (d) the dimension protocol

# NETCUBE: A comprehensive traffic analysis model

## -OLAP operation

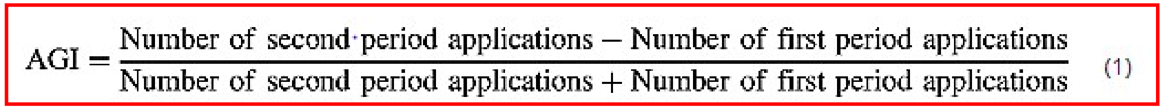
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies
- A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand
- Several OLAP operations such as **roll-up**, **drill-down**, **slice** and **dice** are used to retrieve useful information from the data warehouse of traffic data

## NETCUBE: A comprehensive traffic analysis model
### -OLAP operation

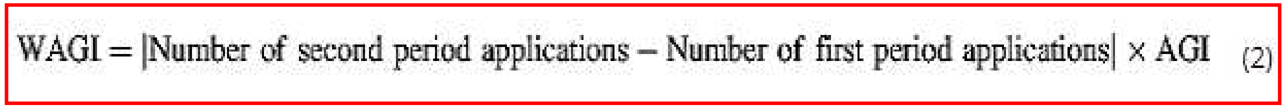
- The **roll-up** operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction
- The **drill-down** operation is the reverse of the roll-up operation
- The **slice** operation selects one dimension of the given cube and this results in a subcube
- The **dice** operation defines a subcube by performing a selection on two or more dimensions

# NETCUBE: A comprehensive traffic analysis model

## -Traffic analysis using data-mining techniques



$$\mathrm{AGI} = \frac{\text{Number of second period applications} - \text{Number of first period applications}}{\text{Number of second period applications} + \text{Number of first period applications}} \quad (1)$$



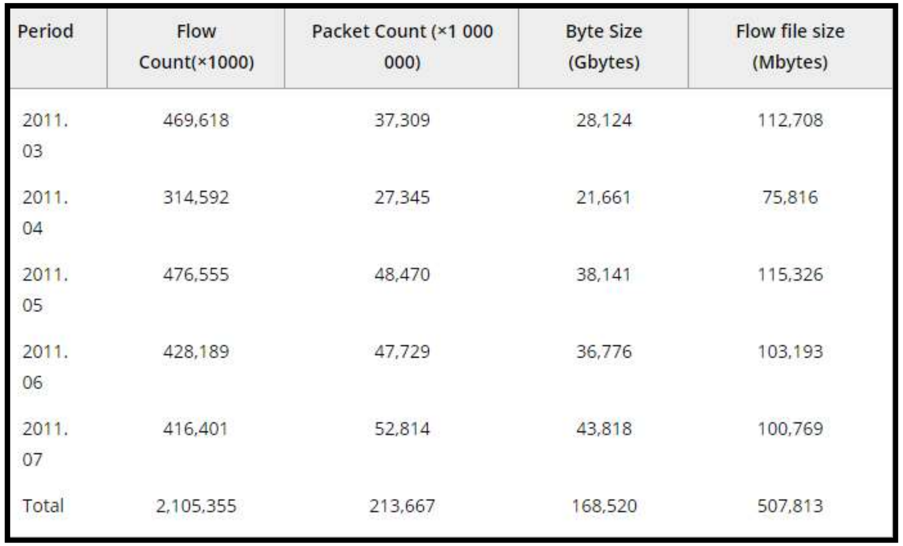
$$\mathrm{WAGI} = |\text{Number of second period applications} - \text{Number of first period applications}| \times \mathrm{AGI} \quad (2)$$

→ In this study, we utilize WAGI rather than AGI to analyze the traffic evolution of each application over the target period

# Experimets

- The primary goal of the experiments is to evaluate the performance, scalability, and utility of the NetCube model
- Datasets:

  - The authors use a real-world network traffic dataset for a realistic assessment



| Period | Flow Count(×1000) | Packet Count (×1 000 000) | Byte Size (Gbytes) | Flow file size (Mbytes) |
|---|---|---|---|---|
| 2011. 03 | 469,618 | 37,309 | 28,124 | 112,708 |
| 2011. 04 | 314,592 | 27,345 | 21,661 | 75,816 |
| 2011. 05 | 476,555 | 48,470 | 38,141 | 115,326 |
| 2011. 06 | 428,189 | 47,729 | 36,776 | 103,193 |
| 2011. 07 | 416,401 | 52,814 | 43,818 | 100,769 |
| Total | 2,105,355 | 213,667 | 168,520 | 507,813 |

Table1. Traffic trace statistics

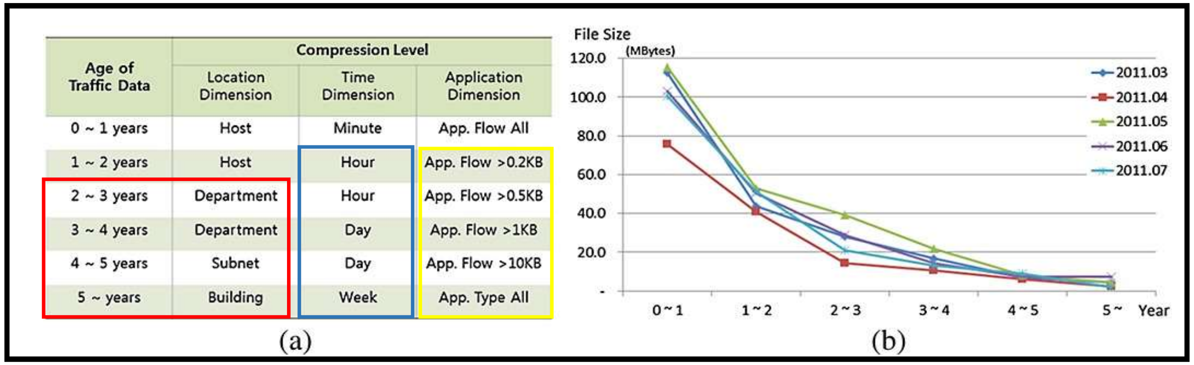
The amount of traffic collected over the five months was about 168TB

When converted to flow format, this amounts to about 507GB

# Experimets

- The level of compression applied based on the age of the traffic data

The elapsed time since the data was collected or recorded



| Age of Traffic Data | Compression Level | | |
|---|---|---|---|
| | Location Dimension | Time Dimension | Application Dimension |
| 0 ~ 1 years | Host | Minute | App. Flow All |
| 1 ~ 2 years | Host | Hour | App. Flow >0.2KB |
| 2 ~ 3 years | Department | Hour | App. Flow >0.5KB |
| 3 ~ 4 years | Department | Day | App. Flow >1KB |
| 4 ~ 5 years | Subnet | Day | App. Flow >10KB |
| 5 ~ years | Building | Week | App. Type All |

Figure 6.

Traffic data older than 2,4 and 5 years aggregated to the department, subnet, and building level, respectively

In the time dimension, data older than 1 years is compressed at the hour level

As the data's age reaches 3 or 5 years, it's summarized into the day or week levels

In the application dimension, flow data below certain size thresholds are aggregated into application types

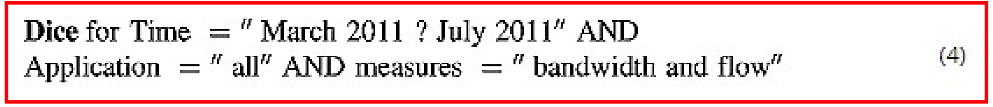
These thresholds increase over time, ranging from 0.2kbytes to 10kbytes

- (b) shows that only 2 MB of storage is need to store flow from July 2011, which is five years old.

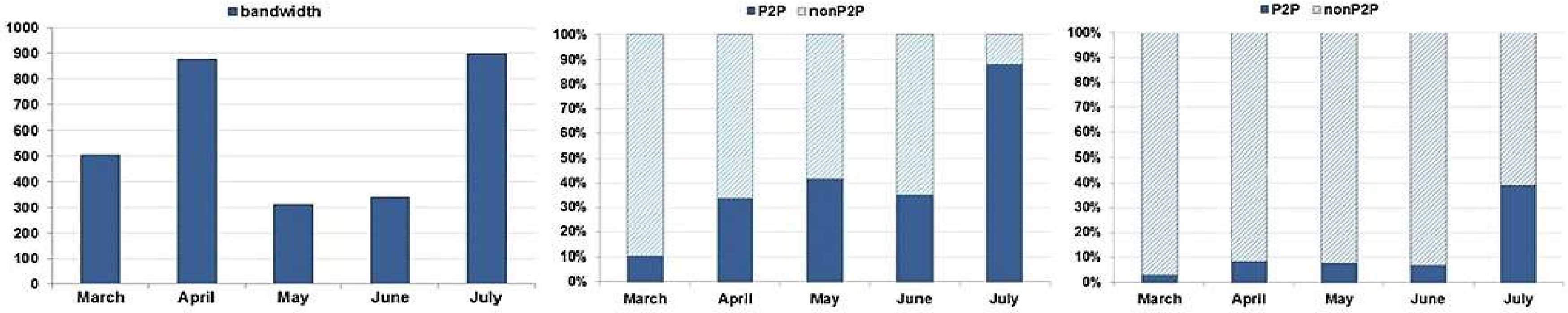
# Experimets

## -Traffic analysis results using OLAP

- Introducing multidimensional analysis that can be achieved by adjusting the level of abstraction using the NetCube model



Dice for Time = " March 2011 ? July 2011" AND
Application = " all" AND measures = " bandwidth and flow" (4)



(a) (b) (c)

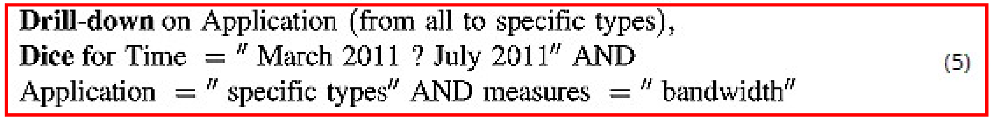
Figure7. The group of graphs obtained from formula (4)
(a) Monthly bandwidth (b) Monthly application bandwidth (c) Monthly application flow

- In particular, P2P bandwidth occupies more than 80% of the entire bandwidth in a campus network during the July vacation period(Figure7a,b)
- Thus, Need a Qos management policy for P2P applications related to summer vacation, as excessive usage of P2P interferes with the stable services for daily business traffic

# 실험

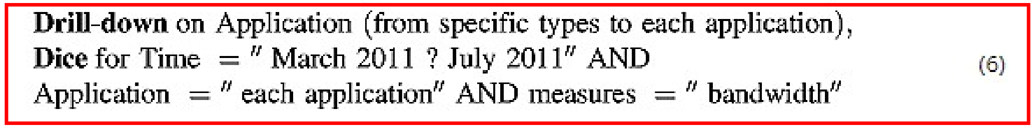
## -OLAP를 이용한 트래픽 분석 결과

- To understand the results of the detailed analysis, conducted a **drill-down** operation for the application dimension, as shown in Formula (5)



**Drill-down** on Application (from all to specific types),
**Dice** for Time = ″ March 2011 ? July 2011″ AND
Application = ″ specific types″ AND measures = ″ bandwidth″ (5)

- Conduct a **drill-down** operation, as shown in Formula (6). down to the lowest level to confirm in detail what file-sharing application programs are being used



**Drill-down** on Application (from specific types to each application),
**Dice** for Time = ″ March 2011 ? July 2011″ AND
Application = ″ each application″ AND measures = ″ bandwidth″ (6)

# Experimets

## -Traffic analysis results using OLAP

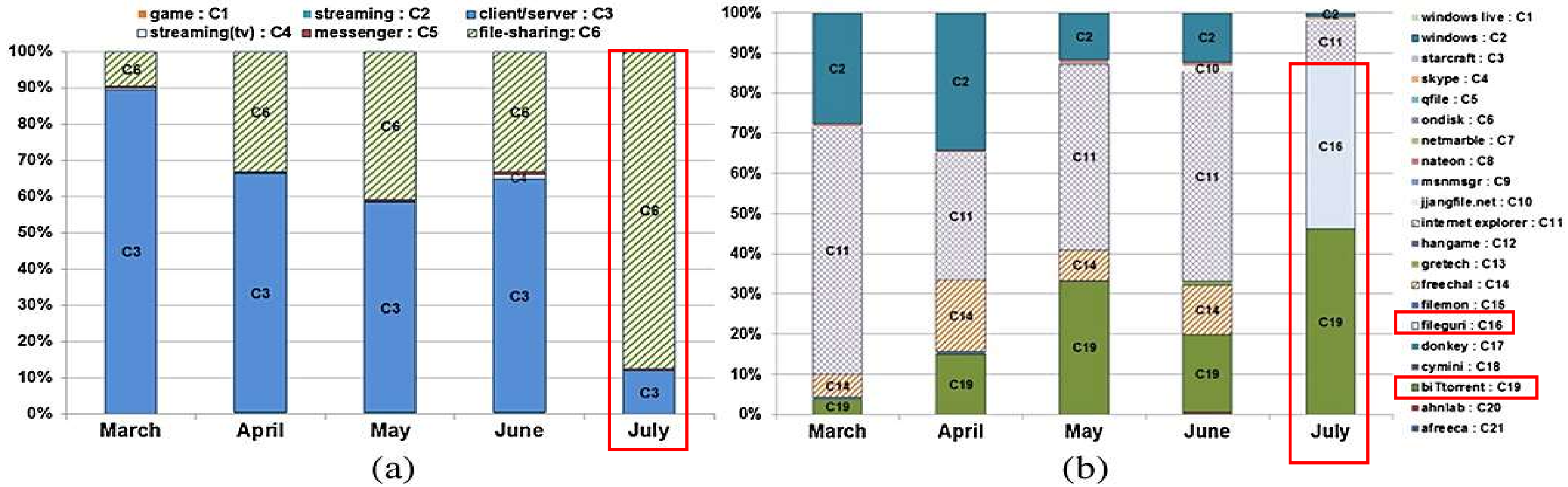


Figure8. (a) Monthly application(specific types) bandwidth (b) Monthly application(each application) bandwidth

- Figure 8(a) reveals that the type P2P application which used a large amount of bandwidth in July was a file-sharing application
- Figure 8(b) shows that the file-sharing programs used were Fileguri, a domestic Korean program, and BitTorrent, which has many domestic and international users

# Experimets

## -Traffic analysis results using OLAP

- Chose the time and location dimensions and analyzed changes in the bandwidth usage of each department according to the time slots

Dice for Time = " March 2011 ? July 2011" AND
Location = " department" AND measures = " bandwidth" (7)

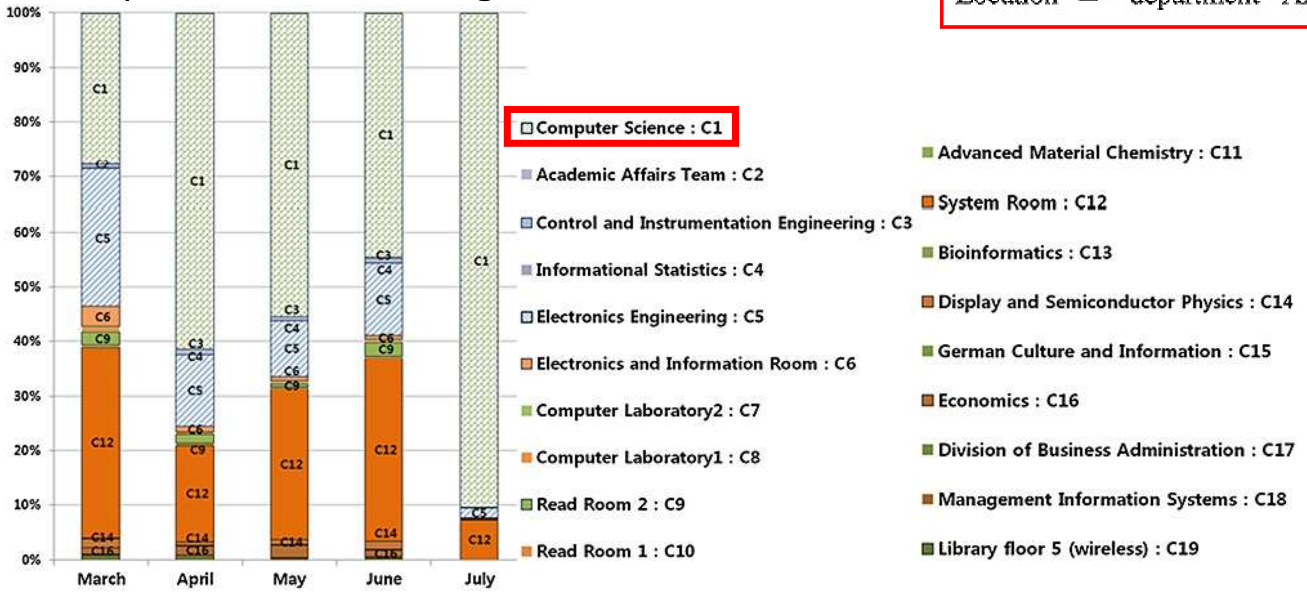


Figure9. Usage rate of traffic for each department by time slot

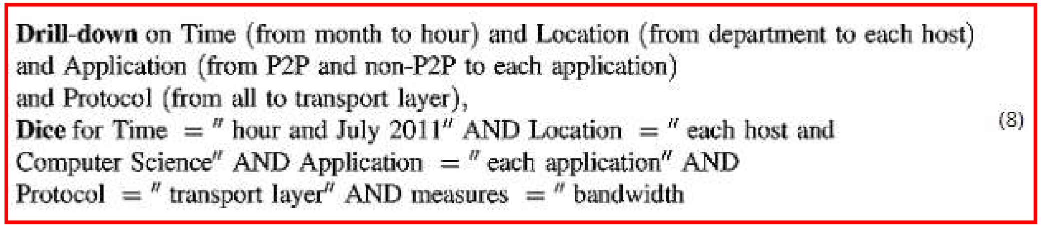
- From formula (7), Figure 9 presents the change in bandwidth usage according to university department from March to July 2011

- In particular, the bandwidth used in the Department of Computer Science in July covered more than 90% of the total bandwidth of the campus

# Experimets

## -Traffic analysis results using OLAP

- In order to carry out a detailed analysis for the Department of Computer Science, **drill-down** as illustrated in formula (8)



**Drill-down** on Time (from month to hour) and Location (from department to each host)
and Application (from P2P and non-P2P to each application)
and Protocol (from all to transport layer),
**Dice** for Time = ″ hour and July 2011″ AND Location = ″ each host and
Computer Science″ AND Application = ″ each application″ AND
Protocol = ″ transport layer″ AND measures = ″ bandwidth (8)

- from month to hour for the time dimension

- from the department to each host for the location dimension

- from P2P and non-P2P to each application for the application dimension

# Experimets

## -Traffic analysis results using OLAP

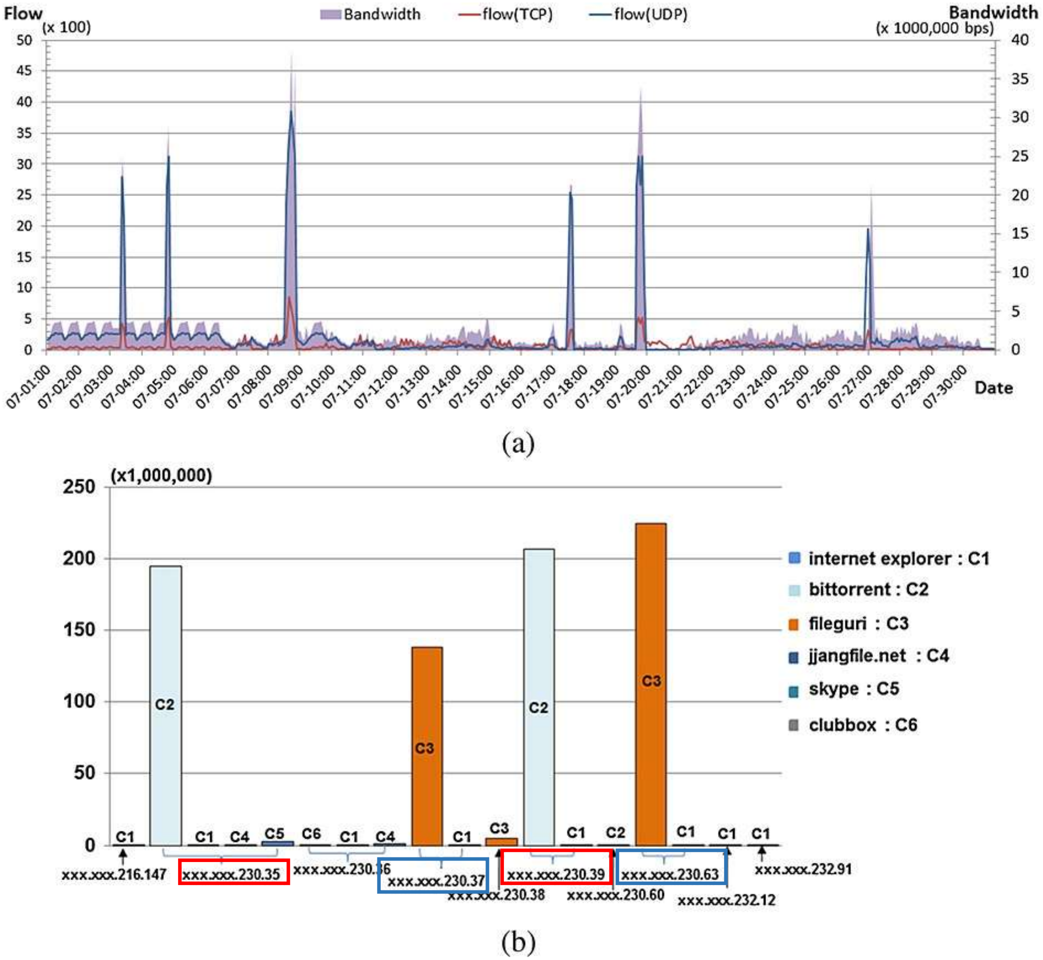


Figure10. (a) Bandwidth and flow usage of each protocol in the Department of Computer Science in July
(b) An analysis of the application bandwidth used by the host

- The analysis shows that large amount of traffic usages appear six times over several days in July
- These unusual high traffic usages are caused by high increase of UDP flows, most of which were data transfer flows by P2P applications
- xxx.xxx.230.35, 230.39 -> bittorrennt
- xxx.xxx.230.37, 230.63 -> fileguri

# Experimets

## -Traffic analysis results using data mining

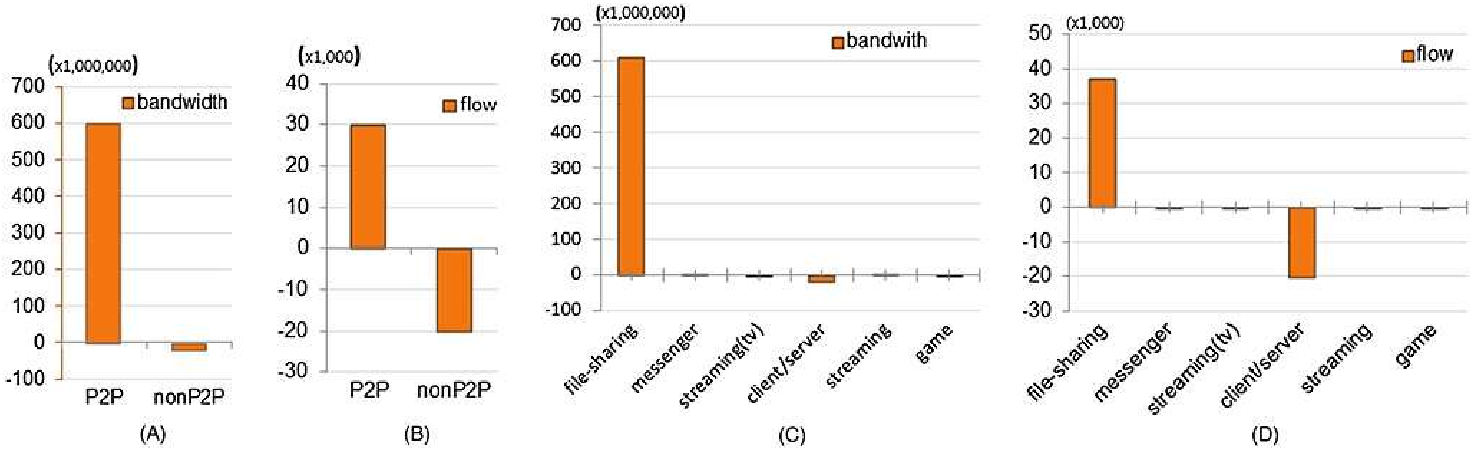


Figure11. The WAGI of P2P and non-P2P applications in (a) bandwidth and (b) flow
Lower application level WAGI in (c) bandwidth and (d) flow

- Describes the WAGI calculated between two periods:

  the first period = June,

  the second period = July

- In (A),(B) P2P is positive value -> means that July usage is higher than June usage

  non-P2P is negative value -> means that July usage is lower than June usage

- In order to examine this in more detail, performed a **drill-down** operation into the application dimension
- In (C),(D) found out that the file-sharing application has significantly increased in the second period over other applications

# Experimets

## -Traffic analysis results using data mining

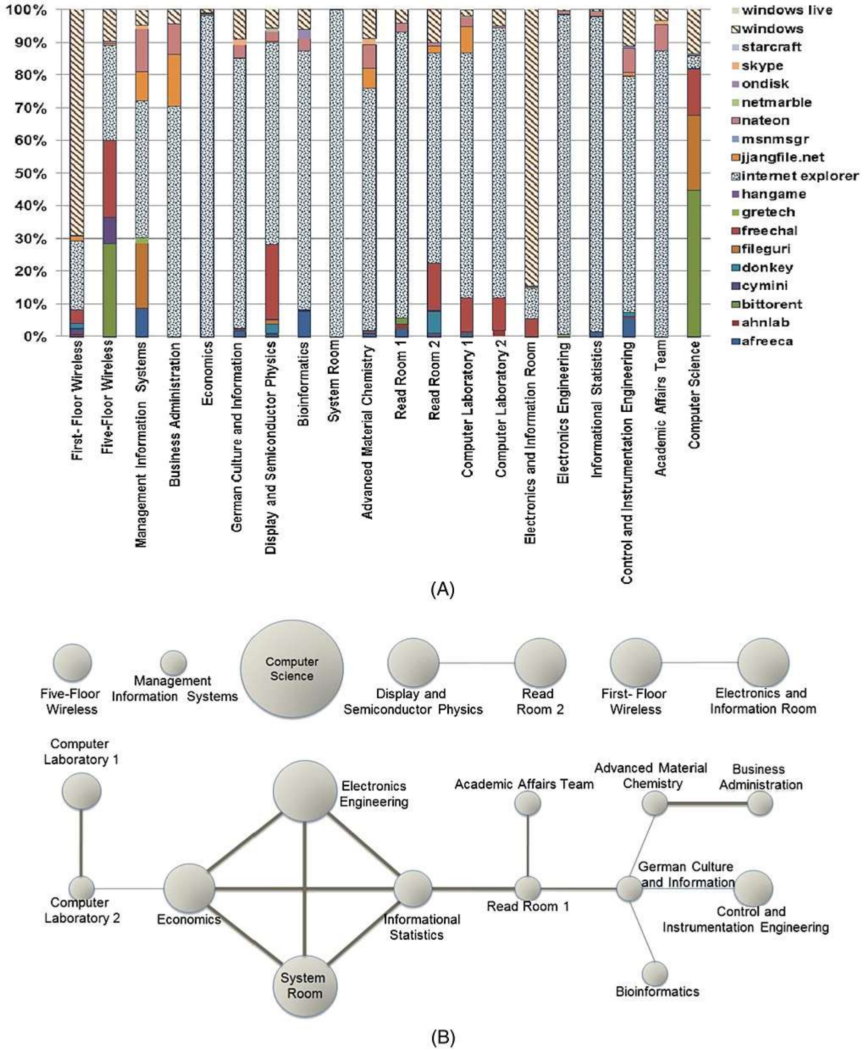
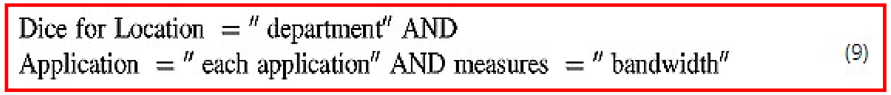


Figure12. (a) Application traffic distribution by subnets
(b) Traffic pattern similarity map by subnets

- (A)-> An application traffic distribution for each subnet analyzed using formula (9)



Dice for Location = " department" AND
Application = " each application" AND measures = " bandwidth" (9)

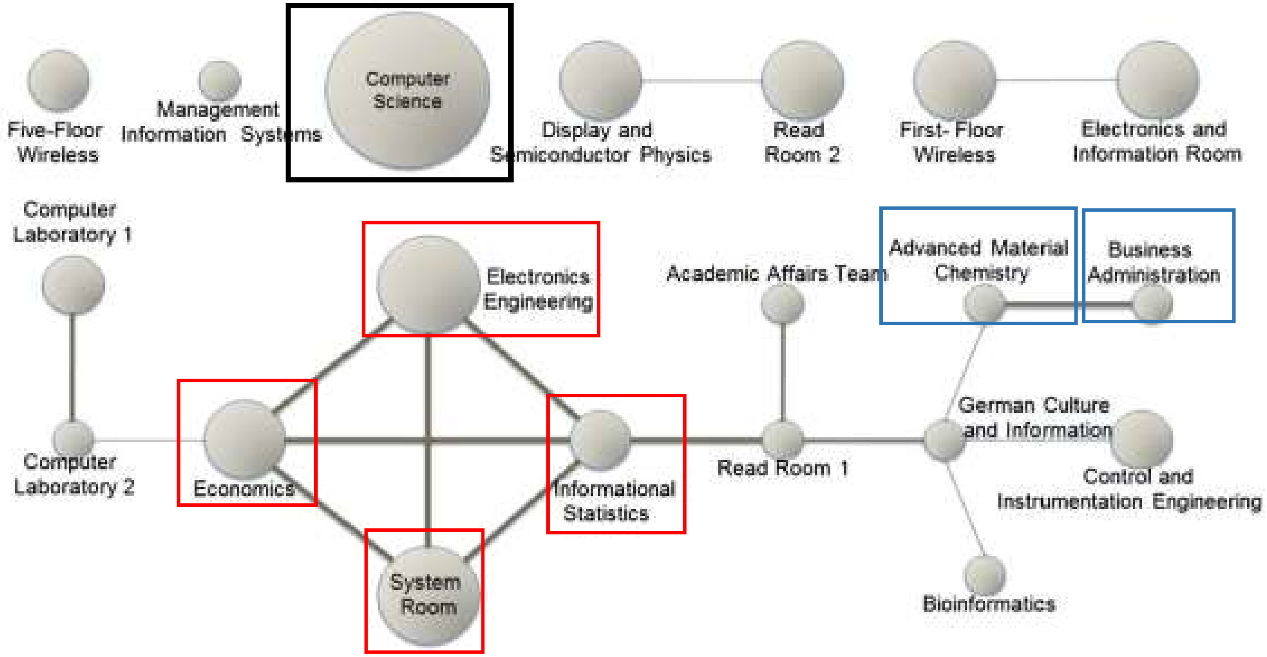
- (B) -> A subnet similarity map that was drawn by applying the cosine similarity function

  The size of the circle indicates the amount of traffic generated by each subnet

  The thickness of the links among the nodes represents the degree of relationship between the nodes

# Experimets

## -Traffic analysis results using data mining



## Conclusion

- Identifies limitations in existing traffic analysis approaches and introduces NetCube, a novel multidimensional traffic analysis model
- NetCube's key feature include:

  -Four dimensions for analysis: time, location, application type, protocol

  -An incremental compression storage model for efficient data handling

  -Capability for multidimensional analysis using OLAP operations

  -Detailed insights into traffic growth and inter-subnet relationship
- By offering detailed insights into ever-changing network environments, NetCube aids in making timely decisions and formulating Qos policies

# How can I apply it?

- OPT_DT : 운행 일자
- TRANSCO_CD : 운송회사 코드
- CAR_REG_NO : 운행차량 번호
- TRIP_ID : 운행ID
- OPT_HHMISSSS : 운행 시 분 초
- GPS_X : 경도
- GPS_Y : 위도
- OPT_SPD : 차량의 속도
- OVERSPD_IS_20KM_EXC : 20km과속 초과 유무
- OVERSPD_TM_20KM_EXC : 20km과속 초과 시간
- OVERSPD_IS_40KM_EXC : 40km과속 초과 유무
- OVERSPD_TM_40KM_EXC : 40km과속 초과 시간
- OVERSPD_IS_60KM_EXC : 60km과속 초과 유무
- OVERSPD_TM_60KM_EXC : 60km과속 초과 시간
- LTM_OVERSPD_IS : 장기 과속 유무
- LTM_OVERSPD_TM : 장기 과속 시간
- SDN_ACCEL_IS : 급 가속 유무
- SDN_START_IS : 급 출발 유무
- SDN_DECEL_IS : 급 감속 유무
- SDN_STOP_IS : 급 정지 유무
- SDN_LTURNL_IS : 급 좌회전 유무
- SDN_RTURN_IS : 급 우회전 유무
- SDN_UTURN_IS : 급 유턴 유무
- SDN_OVERTKG_IS : 급 앞지르기 유무
- SDN_COURSE_CHG_IS : 급 차선변경 유무
- IDLE_STD_TM_EXC_IS : 공회전기준시간 초과 유무
- DATA_PRST_DTTM : 무시(전처리 예정)

# How can I apply it?

1. 먼저 같은 운송차 별로 사고위험성이 높은 급 가속, 급 감속, 급 정지와 같은 운전 패턴 분석.
   -> 그 후 운전 패턴이 위험한 운송차들이 많이 존재하는 운송회사 추출

2. 20km,40km,60km과속 변수가 존재하므로 각각의 과속 여부를 판단한 후 얼마나 장기 과속을 했는지 여부 판단.

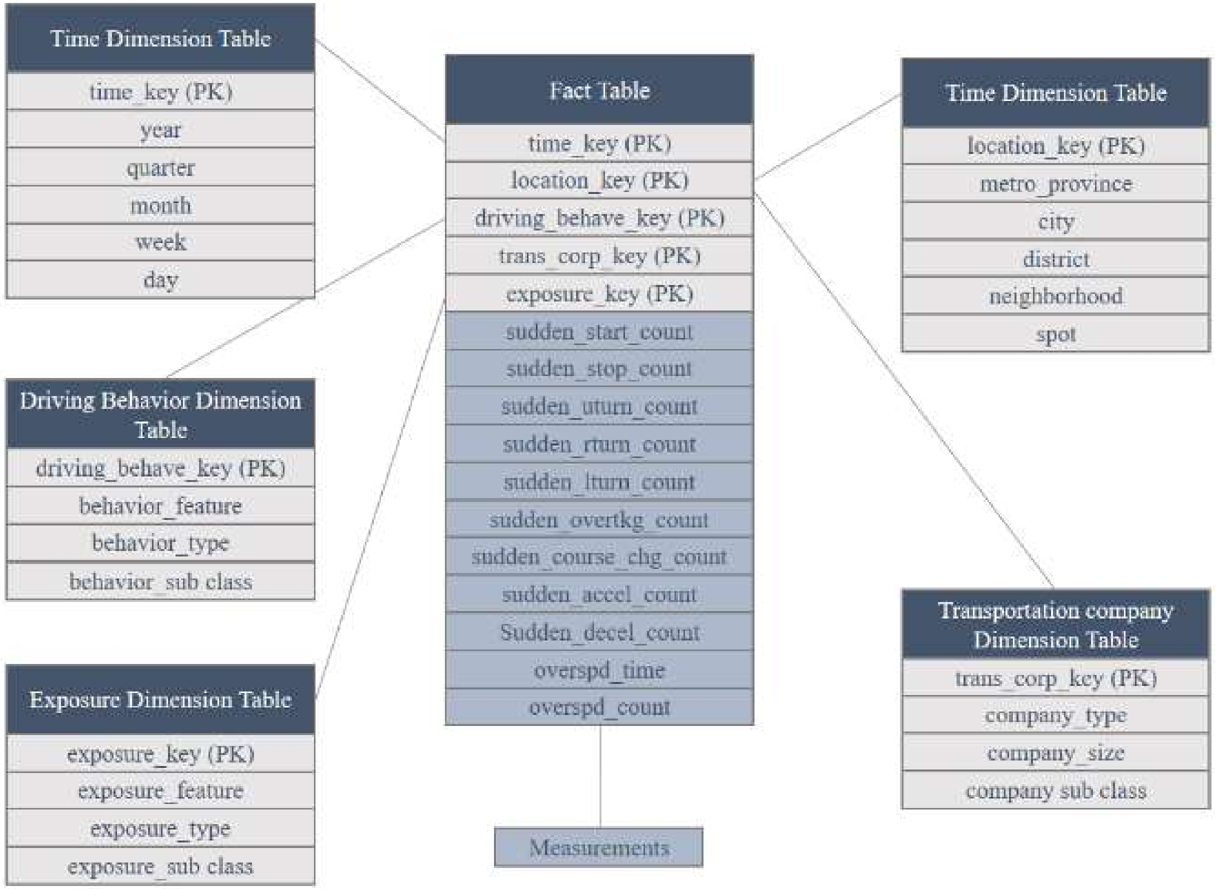
3. 과속 여부 판단한 데이터로 특정 지역에서 과속이 많이 일어나는지 GPS데이터를 활용하여 운행 지역 분석.

4. 11대 위험 운전을 활용하여 급 우회전, 급 좌회전, 급 유턴, 급 차선변경, 급 앞지르기 여부 판단
   -> 추후에 무엇을 할 수 있는지 고민할 예정

5. 운행 일자(계절과 주말 구별),변수와 시간대(밤과 낮 구별)변수를 이용해 주행 패턴 분석

6. 다음 법과 같이 공회전 제한시간 초과 유무 여부 판단
   ->초과한 차량이 많은 운송회사 선별

# How can I apply it?



| Time Dimension Table |
|---|
| time_key (PK) |
| year |
| quarter |
| month |
| week |
| day |

| Driving Behavior Dimension Table |
|---|
| driving_behave_key (PK) |
| behavior_feature |
| behavior_type |
| behavior_sub class |

| Exposure Dimension Table |
|---|
| exposure_key (PK) |
| exposure_feature |
| exposure_type |
| exposure_sub class |

| Fact Table |
|---|
| time_key (PK) |
| location_key (PK) |
| driving_behave_key (PK) |
| trans_corp_key (PK) |
| exposure_key (PK) |
| sudden_start_count |
| sudden_stop_count |
| sudden_uturn_count |
| sudden_rturn_count |
| sudden_lturn_count |
| sudden_overtkg_count |
| sudden_course_chg_count |
| sudden_accel_count |
| Sudden_decel_count |
| overspd_time |
| overspd_count |

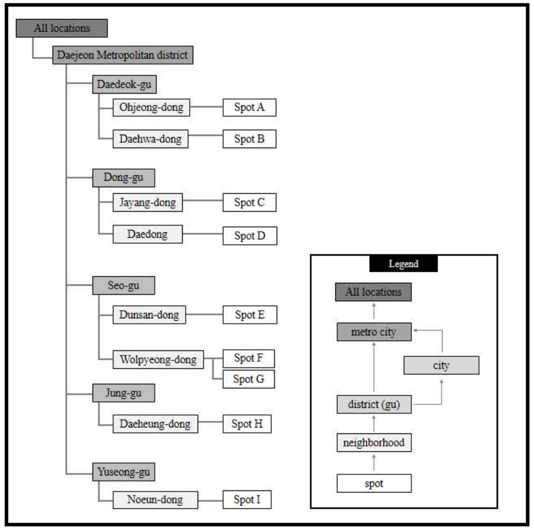
Measurements

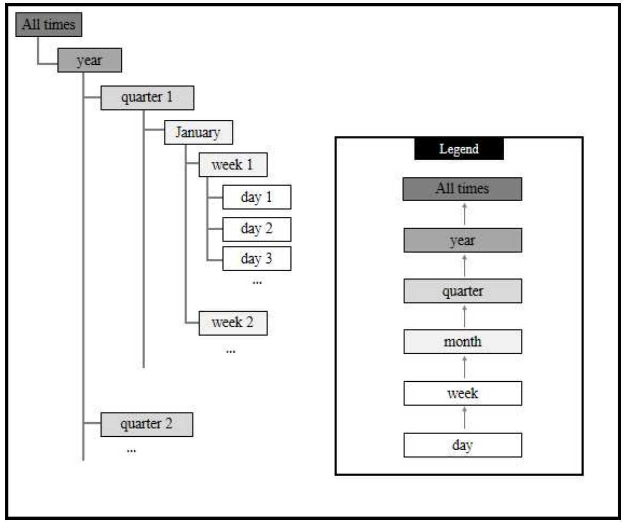
| Time Dimension Table |
|---|
| location_key (PK) |
| metro_province |
| city |
| district |
| neighborhood |
| spot |

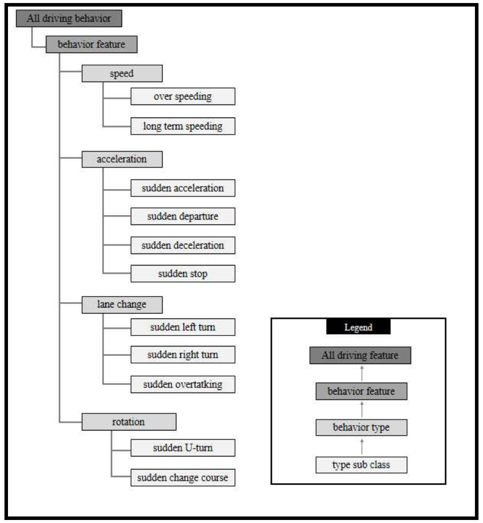
| Transportation company Dimension Table |
|---|
| trans_corp_key (PK) |
| company_type |
| company_size |
| company sub class |

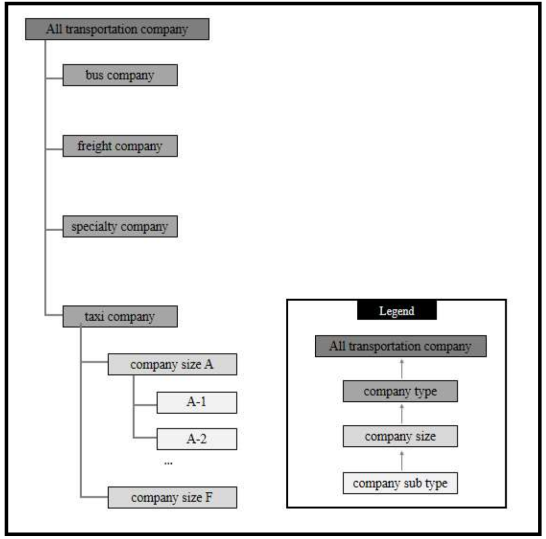
- 11대 위험 운전(진한 파란색)을 고려하여 작성된 스타스키마로 분석 진행

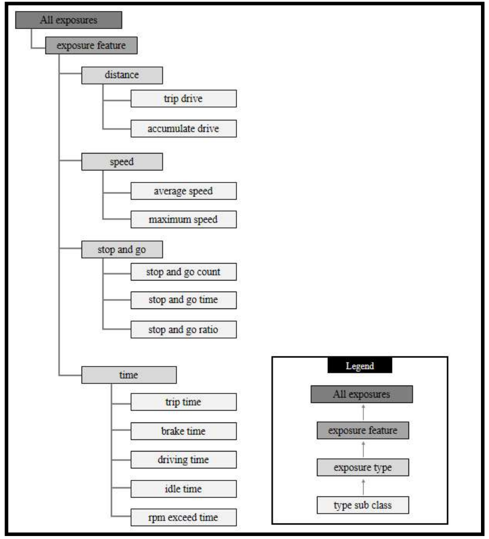
# How can I apply it?











- 데이터에 맞게 작성된 개념 계층도를 활용하여 OLAP연산을 통한 추상화 수준을 달리하여 구체화를 진행하면서 분석을 실시할 예