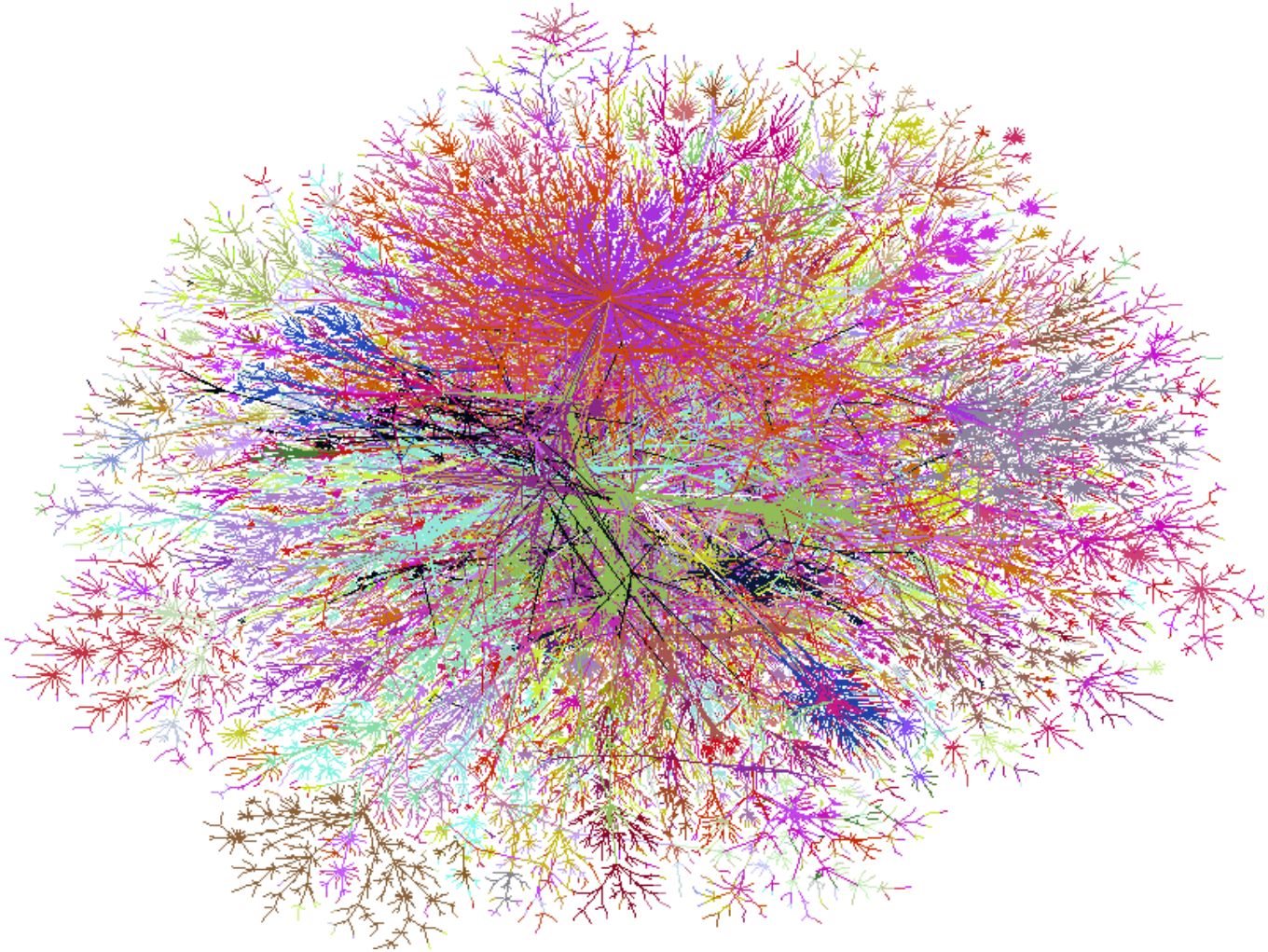
Internet Measurement and Data Analysis (1)

Kenjiro Cho

2014-09-22

introduction

how does the entire Internet look like?



lumeta internet mapping http://www.lumeta.com

http://www.cheswick.com/ches/map/

introduction (cont'd)

how does the entire Internet look like?

- no one knows
- but, everyone is interested

the theme of the class

- looking at the Internet from different views
  - how to measure what is difficult to measure
  - how to extract useful information from huge data sets

this kind of approach will be increasingly important in the future information society

Internet measurement and data analysis

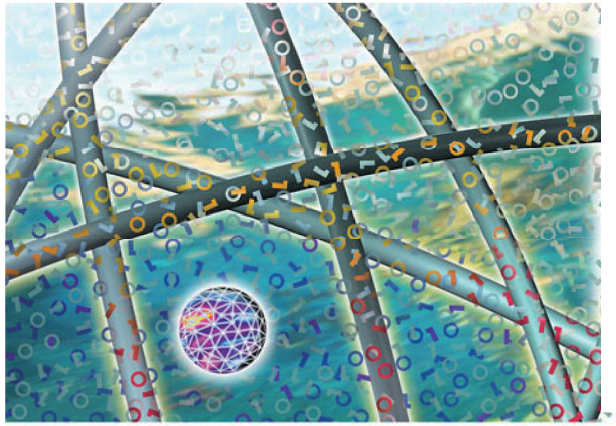
- Faculty: Kenjiro Cho <kjc@sfc.keio.ac.jp>
- TA: Yohei Kuga <sora@sfc.wide.ad.jp>
- SA: TBA
- URL: http://web.sfc.keio.ac.jp/~kjc/classes/sfc2014f-measurement/
- support email (facaulty, TA, SA): <imda2014f@sfc.wide.ad.jp>
- textbooks, references: the lecture slide materials will be provided online.
- programming: data processing exercises by Ruby
- evaluation: 2 assignments (20% each) and a final report (60%)

what you will learn in the class

- how to understand statistical aspects of data, and how to process and visualize data
  - which should be useful for writing thesis and other reports
- programming skills to process a large amount of data
  - beyound what the existing package software provides
- ability to suspect statistical results
  - the world is full of dubious statistical results and infomation manipulations
  - (improving literacy on online privacy)

- programming and hands-on data analysis
  - just reading textbooks isn't enough
  - certain skills can be learned only through first hand experiences

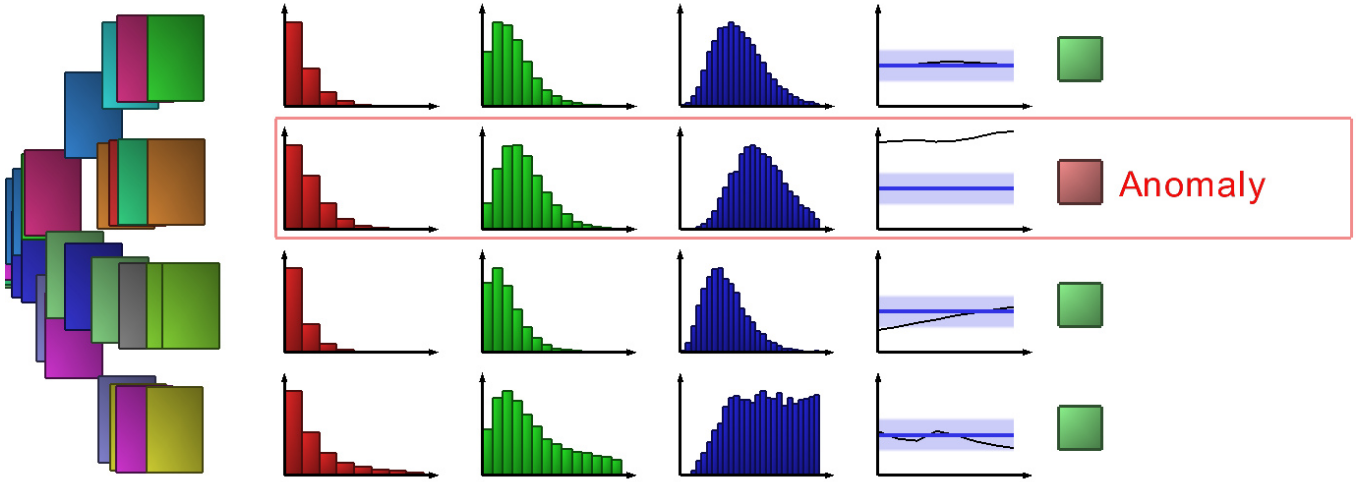
Big Data everywhere

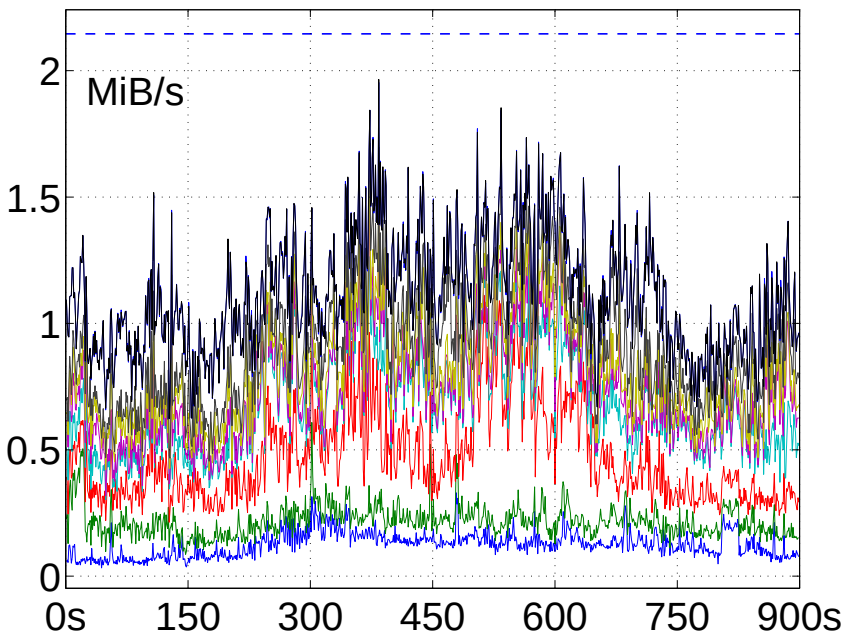


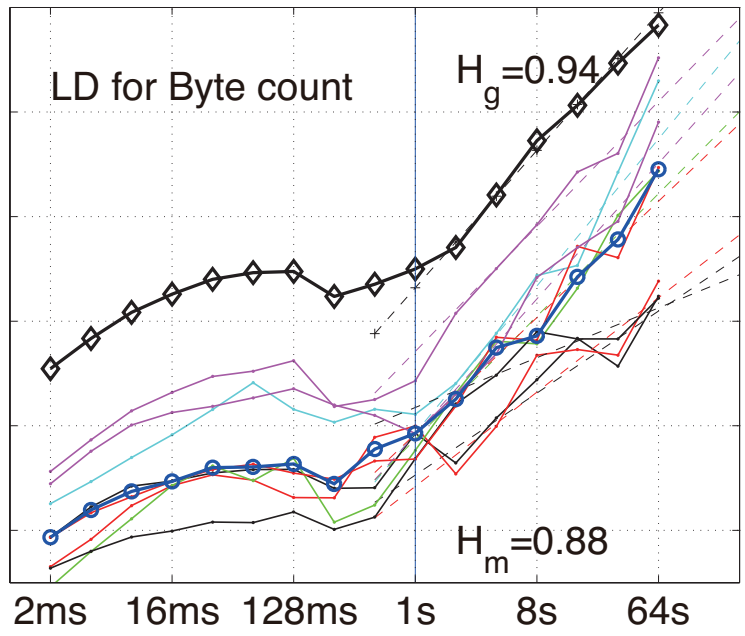
big data and Internet measurement

- big data: broadly, technologies for extracting valuable information hidden in a large volume of unstructured data
  - often aiming at constructing new service or business models
- most technologies have been around
  - search ranking, online recommender systems, etc.
- Internet measurement: efforts to understand the Internet from huge but incomplete data
  - need to use inferences by statistical methods

example: anomaly detectio by sketch and statistical feature comparison

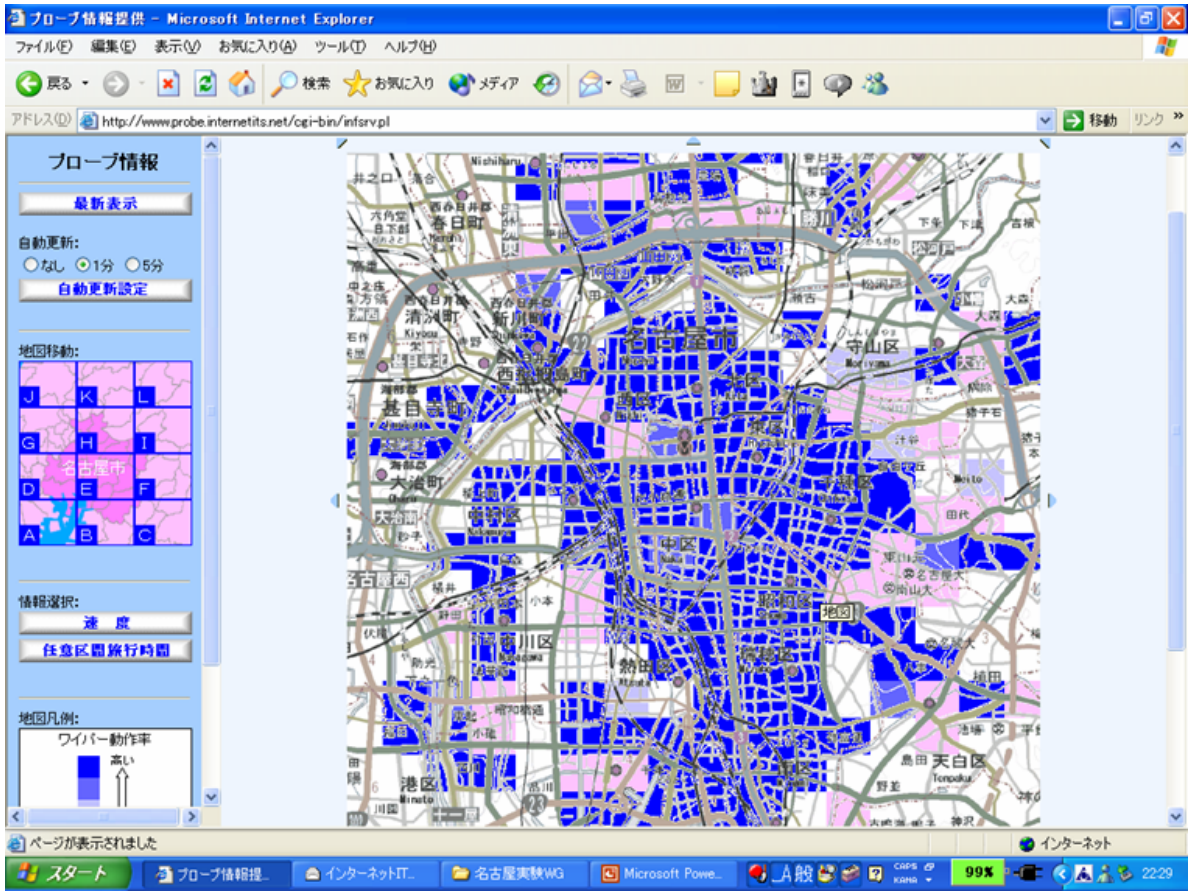






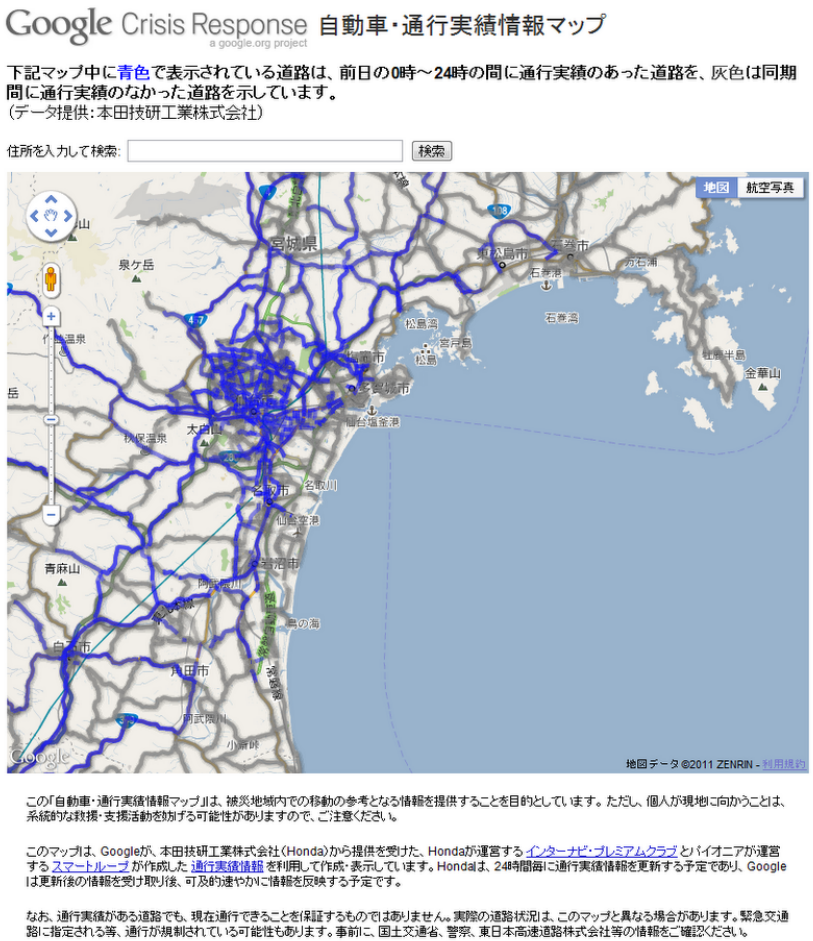
example: Internet vehicle experiments

- by WIDE Project In Nagoya in 2001
  - location, speed and wiper usage data from 1,570 taxis
  - blue areas indicate high ratio of wiper usage, showing rainfall in detail



Japan Earthquake

- the system is now part of ITS
- usable roads info released 3 days after the quake
  - data provided by HONDA (TOYOTA, NISSAN)



the age of data

- big data is not just for marketing
- technological innovations known as the data revolution are occurring in every field
- previously difficult applications become possible
  - access to huge amount of data, analysis of data constantly being updated, and applications to non-linear models
- big data analysis becomes an indispensable research method in all areas of science and technology

example: impact to science

e-science: paradaigm shift?

- theory
- experiment
- simulations (enabled by computer)
- data-driven discovery (enabled by big data)

Google's Chief Economist Hal Varian on Statistics

The McKinsey Quarterly, January 2009

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

data analysis is merely a tool

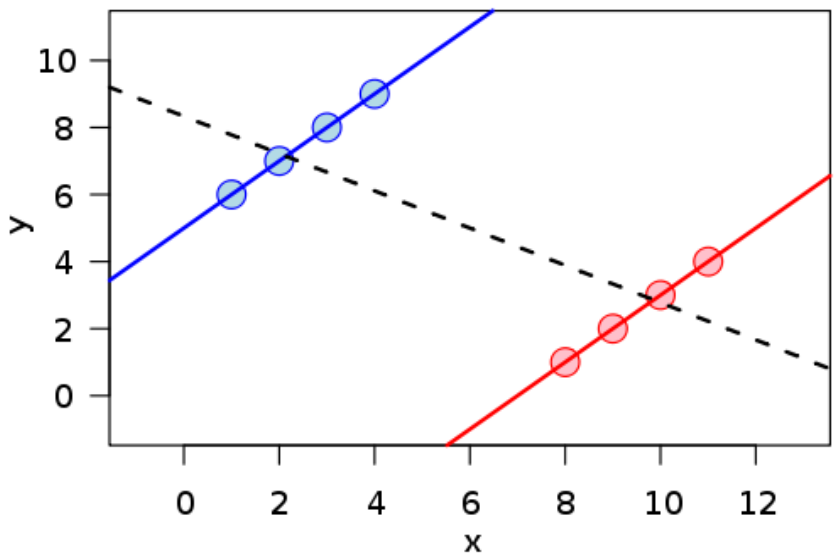
- recent big data trends focus too much on tools and methods but data analysis is merely a tool
- data analysis is an iterative process
  - forming a hypothesis, verifying it with data
  - if the results are unexpected, you find new questions
  - repeating the process will uncover interesting facts
- analysis without purpose ends up with useless numbers
- if you identify what to get from data, you will see a path forward

fundamental change to creative thinking process?

- data-driven decision making has been always important
- but, ICT pushes it to a completely different level (in quality, quantity, expressions)
- now, we can literally interact with data (data-human interaction)

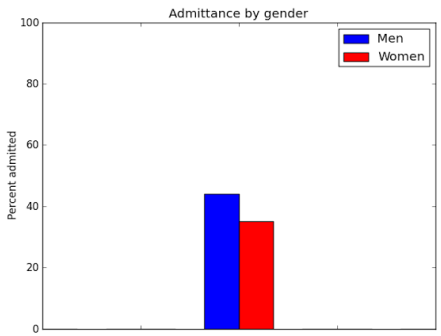
Simpson's paradox

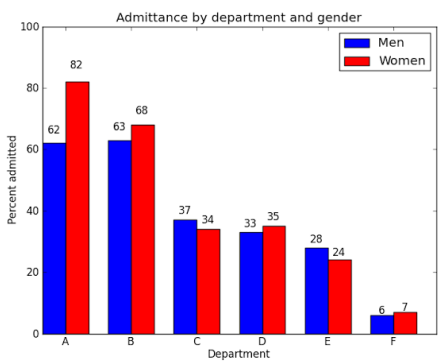
- one of paradoxes in statistics
  - a trend observed for aggregated groups is different from that observed for each groups

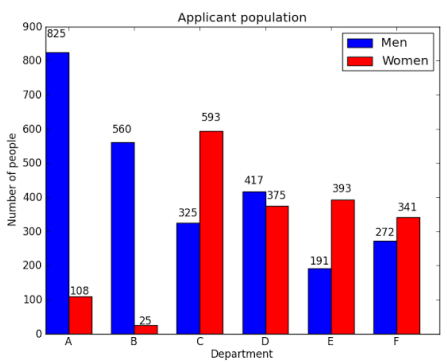


example: UC Berkeley gender bias case in 1973

- the university was sued for bias against women as acceptance ratio to graduate school is lower for women
- investigation revealed that most departments had statistically significant bias in favor of women
- the reason: women tend to apply to competitive departments with low acceptance ratio







source:http://mathemathinking.blogspot.jp/2012/06/simpsons-paradox.html

self-introduction

Kenjiro Cho

- positions
  - Research Director, IIJ Research Lab
  - Guest Professor, Keio SFC
  - Adjunct Professor, JAIST
  - Board member, WIDE Project
- bio
  - BE in electronics from Kobe University in 1984.
    - started as a hardware engineer at Canon, Inc, then became interested in operating systems
  - M.Eng in computer science from Cornell University in 1993
    - studied computer science and distributed systems
  - Researcher at Sony Computer Science Labs from 1996
    - research on the Internet
  - Ph.D. (Media and Governance) from Keio University in 2001
  - Researcher at IIJ from 2004
- research topics
  - Internet measurement and management
  - networking support in operating systems
  - datacenter management and cloud services

class overview

It becomes possible to access a huge amount of diverse data through the Internet. It allows us to obtain new knowledge and create new services, leading to an innovation called "Big Data" or "Collective Intelligence". In order to understand such data and use it as a tool, one needs to have a good understanding of the technical background in statistics, machine learning, and computer network systems.

In this class, you will learn about the overview of large-scale data analysis on the Internet, and basic skills to obtain new knowledge from massive information for the forthcoming information society.

class overview (cont'd)

Theme, Goals, Methods

In this class, you will learn about data collection and data analysis methods on the Internet, to obtain knowledge and understanding of networking technologies and large-scale data analysis.

Each class will provide specific topics where you will learn the technologies and the theories behind the technologies. In addition to the lectures, each class includes programming exercises to obtain data analysis skills through the exercises.

Prerequisites

The prerequisites for the class are basic programming skills and basic knowledge about statistics.

In the exercises and assignments, you will need to write programs to process large data sets, using the Ruby scripting language and the Gnuplot plotting tool. To understand the theoretical aspects, you will need basic knowledge about algebra and statistics. However, the focus of the class is to understand how mathematics is used for engineering applications.

class schedule (1/4)

- Class 1 Introduction (9/22)
  - Big Data and Collective Intelligence
  - Internet measurement
  - Large-scale data analysis
  - exercise: introduction of Ruby scripting language
- Class 2 Data and variability (9/29)
  - Summary statistics
  - Sampling
  - How to make good graphs
  - exercise: graph plotting by Gnuplot
- Class 3 Data recording and log analysis (10/6)
  - Network management tools
  - Data format
  - Log analysis methods
  - exercise: log data and regular expression

class schedule (2/4)

- Class 4 Distribution and confidence intervals (10/20)
  - Normal distribution
  - Confidence intervals and statistical tests
  - Distribution generation
  - exercise: confidence intervals
  - **assignment 1**
- Class 5 Diversity and complexity (10/27)
  - Long tail
  - Web access and content distribution
  - Power-law and complex systems
  - exercise: power-law analysis
- Class 6 Correlation (11/3)
  - Online recommendation systems
  - Distance
  - Correlation coefficient
  - exercise: correlation analysis

class schedule (3/4)

- Class 7 Multivariate analysis (11/10)
  - Data sensing and GeoLocation
  - Linear regression
  - Principal Component Analysis
  - exercise: linear regression
- Class 8 Time-series analysis (11/17)
  - Internet and time
  - Network Time Protocol
  - Time series analysis
  - exercise: time-series analysis
  - **assignment 2**
- Class 9 Topology and graph (12/1)
  - Routing protocols
  - Graph theory
  - exercise: shortest-path algorithm
- Class 10 Anomaly detection and machine learning (12/8)
  - Anomaly detection
  - Machine Learning
  - SPAM filtering and Bayes theorem
  - exercise: naive Bayesian filter

class schedule (4/4)

- Class 11 Data Mining (12/15)
  - Pattern extraction
  - Classification
  - Clustering
  - exercise: clustering
- Class 12 Search and Ranking (12/22)
  - Search systems
  - PageRank
  - exercise: PageRank algorithm
- Class 13 Scalable measurement and analysis (1/14)
  - Distributed parallel processing
  - Cloud computing technology
  - MapReduce
  - exercise: MapReduce algorithm
- Class 14 Privacy Issues (1/19)
  - Internet data analysis and privacy issues
  - Summary of the class

Internet measurement

- network measurement (engineering)
  - measurement in limited environment
  - snapshot at a time
- Internet measurement
  - measurement of the Internet as a large-scale open system
  - open system (keep changing with undefined inputs)
  - behaviors of people on the Internet (social science)

Internet measurement – measuring unmeasurable Internet

- need for generic measurement data for the Internet
  - example: typical traffic usage of Internet users
- the Internet is an open system continuously changing, evolving, and expanding
  - no central point, representative locations, different behaviors are observed depending on observing location and time
  - seeking for generality of the Internet: measuring unmeasurables
- for operation of the Internet, for development of protocols, equipment and services
  - seeking for the best estimates, predicting the future, and revisiting the existing knowledge
- user behavior: need to consider not only from technical aspects but also from social, political and economical aspects

characteristics of network data and behavior

- skewed distributions with large variance
  - inherent mechanism to make burst transfer
  - skewed utilization: e.g., a handful users generate most traffic
- anomalies everywhere
  - bugs, mis-configurations, spec mismatches, accidents, maintenance's
- interferences among various mechanisms
- aggregation
  - complex behavior as a whole (more than the sum of the individual components)

why measurement of Internet is so hard?

- traditional measurement is mainly to improve measurement accuracy according to some engineeringly defined metrics
- Internet measuement is to exploer Internet from massive and noisy data, using statistical methods
  - massive, diverse, skewed, dynamic data
  - complex behavior of open distributed systems
  - resilient and fault-tolerant mechanisms
  - anomalies are the norm

massive volume

- unprecedented scale with unprecedented growth
- far more data than we can analyze
  - techniques needed to reduce data size
    - filtering: e.g., record only TCP SYN packets
    - aggregation: e.g., flow-based accounting
    - sampling: e.g., record 1 in n packets
  - also, techniques needed to reduce dimensionality
- still, details matter
  - a big impact often comes from a small and minor fraction
  - look at the whole while paying attention to details

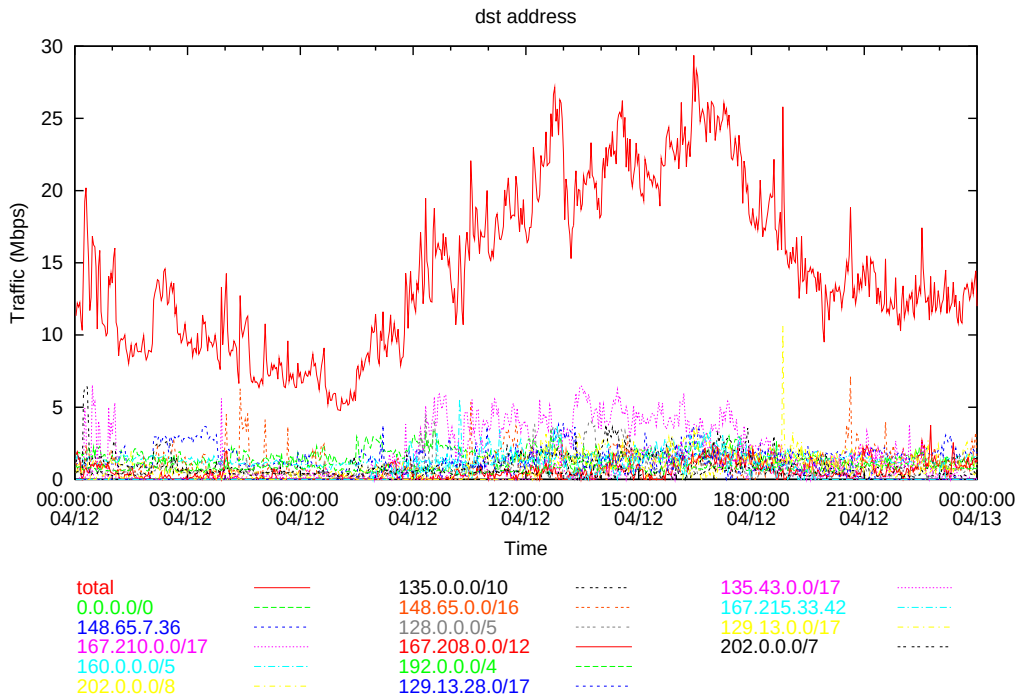
diverse data

- different behaviors are observed from different locations and time
  - country, region, time
  - industry/university/home, backbone/access networks
- different services use different technologies and have different users

typical data doesn't exist!

constant change

- daily, weekly traffic pattern
- trend changes over time
  - web in 90s and p2p/video in 2000s completely changed traffic pattern
- hard to predict future!



limitations of Internet measurement

- problems often occur at boundaries of different networks
  - cooperation needed but not easy
- need understanding and help from operators
- cost: measurement doesn't come free
  - limitations to measure high-end routers with a PC
- privacy and confidential information in data
  - barriers for researchers to access commercial data

measurement targets

- network infrastructure
  - traffic, CDN, packet loss, delay, jitter, topology, routes, DNS
- applications
  - Web, email, messaging, P2P, gaming, SNS, videos
- security and anomaly detection
  - anomalies, attacks, flash crowd

broader targets

- connections among SNS users, popular keyword extraction, online privacy
- SPAM/virus, MapReduce, GeoLocation services, Web server log analysis
- search ranking (PageRank), online recommender systems (collaborative filtering)

possible topics to be studied in the class

- online recommender systems(collaborative filtering)
- search ranking (PageRank)
- SPAM filtering
- Web server log analysis
- MapReduce and other big data technologies
- Internet topology and packet routing
- how users are connected in social network services
- GeoLocation services
- packet analysis
- Internet traffic analysis

summary

Internet measurement and data analysis

- measurement is basis for all technologies
- for networking, it is an attempt to observe invisible networks
- need to consider not only from technical aspects but also from social, political and economical aspects

theme of the class

- Internet measurement and data analysis as case studies
- learn how to measure what is difficult to measure
- learn how to extract useful information from huge data sets

Introduction to Ruby

Ruby

- a scripting language for object-oriented programming
- supports wide range of functions for text processing and system management
- free software started in 1993
- original author: Yukihiro Matsumoto
- became popular for Ruby on Rails (a web application framework)

Ruby information
Ruby official site: `http://www.ruby-lang.org/`
Ruby reference manual: `http://www.ruby-lang.org/en/documentation/`
Ruby の歩き方: `http://jp.rubyist.net/magazine/?FirstStepRuby`

Ruby characteristics

- interpreter language: no need to compile for execution
- highly portable: runs on most platforms
- simple syntax
  - no predefined data type for variables, variables can store any data and are dynamically typed
  - no need to declare variables, variable types (local variables, global variables, instance variables) can be inferred from variable names
- garbage collection: users do not need to manage memory
- object-oriented
  - everything is an object
  - class, inheritance, methods
  - iterator and closure
    - control structures and procedures can be written in object-oriented manner
- powerful string operations/regular expressions
- built-in support for large integers
- Ruby's shortcomings: a bit slower than its competitors

Ruby commands

- irb: Ruby's interactive interface

```
$ irb --simple-prompt
>> puts "Hello"
Hello
```

- ruby: Ruby main program

```
$ ruby test.rb
```

or,

```
$ ruby -e 'puts "Hello".reverse'
olleH
```

exercise: a program to count text lines

count the number of text lines in a file given by the argument

```
filename = ARGV[0]
count = 0
file = open(filename)
while text = file.gets
  count += 1
end
file.close
puts count
```

write to “count.rb” and then run it

```
$ ruby count.rb foo.txt
```

rewrite it in a more rubyish way

```
#!/usr/bin/env ruby
count = 0
ARGF.each_line do |line|
  count += 1
end
puts count
```

next class

Class 2 Data and variability (9/29)

- Summary statistics
- Sampling
- How to make good graphs
- exercise: graph plotting by Gnuplot

references

[1] Ruby official site. `http://www.ruby-lang.org/`

[2] gnuplot official site. `http://gnuplot.info/`

[3] Mark Crovella and Balachander Krishnamurthy. *Internet measurement: infrastructure, traffic, and applications*. Wiley, 2006.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[5] Raj Jain. *The art of computer systems performance analysis*. Wiley, 1991.

[6] Toby Segaran. *Programming Collective Intelligence*. O'Reilly Media. 2007.

[7] Allen B. Downey. *Think Stats: Probability and Statistics for Programmers*. O'Reilly Media. 2011. `http://thinkstats.com/`

[8] Chris Sanders. *Practical Packet Analysis, 2nd Edition*. No Starch Press. 2011.

[9] あきみち、空閑洋平. インターネットのカタチ. オーム社, 2011.

[10] 井上洋, 野澤昌弘. 例題で学ぶ統計的方法. 創成社, 2010.

[11] 平岡和幸, 掘玄. プログラミングのための確率統計. オーム社, 2009.