Major DNS Abnormalities seen by .CN

Cindy WANG (wangxin@cnnic.cn), CNNIC

5 Nov. 2009, Beijing

Agenda

- Motivation
- What we want to learn?
- How we do it?
- Conclusions and Future work
- Discussions

Agenda

- Motivation
- What we want to learn?
- How we do it?
- Conclusions and Future work
- Discussions

This study is motivated by

- DNS is important.
- Complex system made up of thousands of service nodes interacting with each other.
- Imperfect and not error free.
- So many known and unknown vulnerabilities
- Decentralized and distributed, fully control is impossible.

However, we need to

- Know more about the operation states and situations.
- Evaluate the performance and reliability
- Predict growth
- Identify overloading and abuse
- Perceive and forecast risks and attacks.
- At some global, strategic point, which is easy to deploy.

Fortunately, we have rich data

- **Queries targeting .CN domains**
- Responses from .CN authoritative name servers
- Traffic data obtained from some DNS Cache servers
- Actively probed data
- …

Agenda

- Motivation
- What we want to learn?
- How we do it?
- Conclusions and Future work
- Discussions

Is there something unusual reflected by DNS?

- Can we identify what is happening?
  - Anything wrong or unusual use of the DNS service? Such as name servers as well as Cache servers.
  - Any DNS related attacks?
  - Important events on both Internet and Web levels. For example, the DNS collapse on May, 19, 2009. Other events that attract public to communicate on the Internet.
- Can we locate the source and prevent it from getting worse?
- Based on the queries made to .CN root servers, which are easy for collection and deployment.

Agenda

- Motivation
- What we want to learn?
- How we do it?
- Conclusions
- Discussions

Multivariate covariance analysis

- Compare the parameters of the observed query stream with the ‘normal’ DNS query stream.
- The anomaly is declared once a deviation from a normal traffic is observed.
- Does not rely upon any presumptions on the normal network packets distributions of DNS.

Rationale

- Characteristics of an information system such as DNS could be described by the correlations among its features.
- In terms of correlation, the normal patterns will be different from the abnormal ones.
- Therefore, the correlation may be selected as a change indicator, and any changes or abnormal activities will definitely change the correlation coefficients of these features in the normal situations.
- The correlations are expected to be sensitive enough to flag some changes.

OVERVIEW OF COVARIANCE ANALYSIS METHOD

- There are m features, $f_1, \cdots, f_m$. which compose a random vector $\mathbf{X} = (f_1, \cdots, f_m)'$.
- *Let* $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are the n observed vectors, $\mathbf{X}_i = (f_1^i, \cdots, f_m^i)'$ is the i^{th} observed vectors.

- $f_i^{l,j}$ is the value of f_i in the j^{th} observation during the l^{th} time interval T_l .

- We define a new variable y for time slot l :

$$y_l = \begin{pmatrix} f_1^{l,1} & \cdots & f_1^{l,n} \\ f_2^{l,1} & \cdots & f_2^{l,n} \\ \vdots & \ddots & \vdots \\ f_m^{l,1} & \cdots & f_m^{l,n} \end{pmatrix}$$

OVERVIEW OF COVARIANCE ANALYSIS METHOD

(1)

- The covariance to characterize

$$M_{y_l} = \begin{pmatrix} \sigma_{f_1^l f_1^l} & \sigma_{f_1^l f_2^l} & \cdots & \sigma_{f_1^l f_m^l} \\ \sigma_{f_2^l f_1^l} & \sigma_{f_2^l f_2^l} & \cdots & \sigma_{f_2^l f_m^l} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{f_m^l f_1^l} & \sigma_{f_m^l f_2^l} & \cdots & \sigma_{f_m^l f_m^l} \end{pmatrix}$$

- Distance measuring the change, or the anomaly

$$z_l = \left\| M_{y_l} - E\left(M_{y_l}\right) \right\|$$

Then, how to do it?

- Characterizes the behavior of the system by choosing appropriate parameters or features;
- Describe the behavior using the covariance matrix for each time interval.
- Compare the current observed patterns with the normal patterns (mean covariance matrix).
- If the distance exceeds some threshold, anomaly could be declared.

Background: the Baofeng Event

- Starting from 21:50 on May 19, 2009
- Internet users in Jiangsu, Anhui, Henan, Gansu, and Zhejiang, etc several provinces, reported that they suffered slow Internet speed or were unable to visit some websites.
- The network failure was due to a domain name system failure for www.Baofeng.com, the website of a Chinese music player.
- The failure further caused a surge of DNS server visits and decreased processing performance for the network.

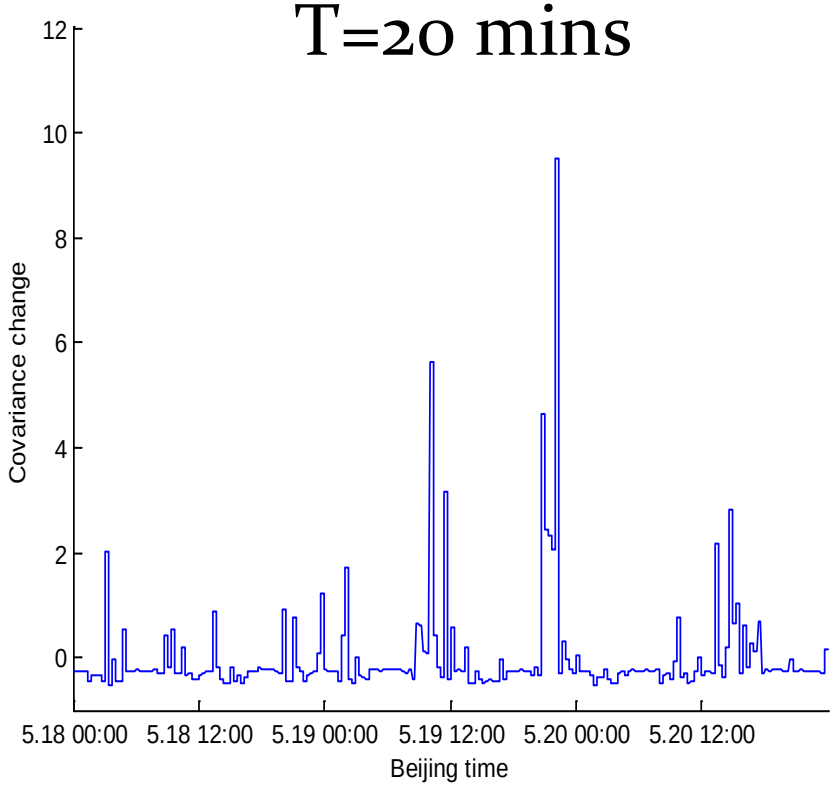
Details

- Queries made to .CN authoritative root servers during the period of 2009-5-18 00:00—2009-5-20 24:00.
- Source IP addresses are conversed to originating provinces using Maxmind™ Geolocation Database.
- Six (*m=6*) features are constructed: overall query rate, number of queries originated from Anhui, Jiangsu, Shanxi, Hebei, Zhejiang provinces (the most affected regions reported.) during each time slice.

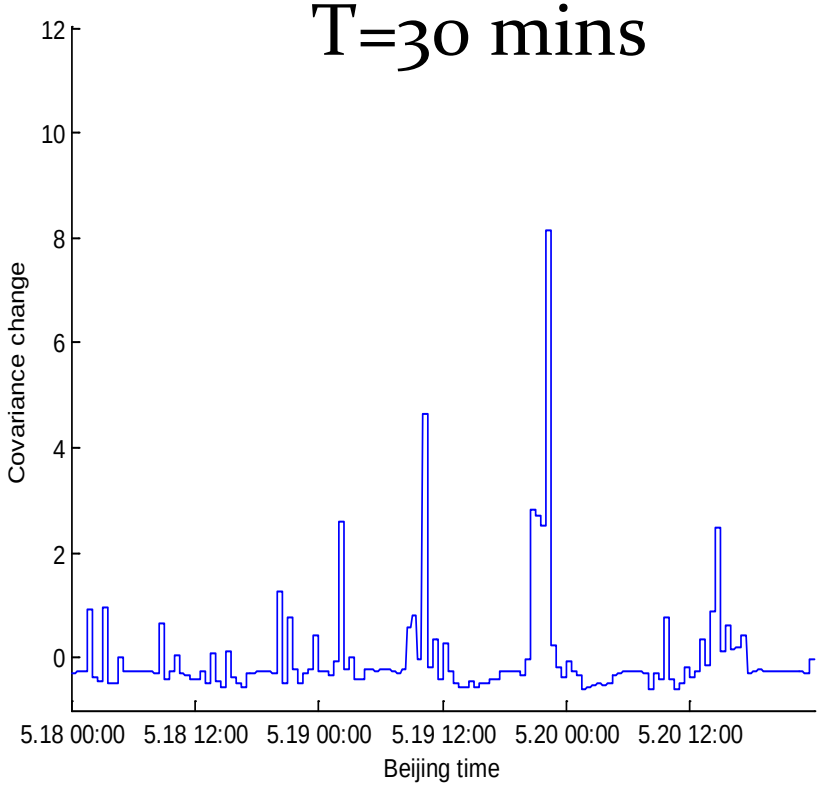
Covariance Matrix Distance calculated

- Distances between the covariance matrix and the mean of all covariance matrices under normal situations
- Change points (Peaks) coincide with the time of the event reported . Around 10pm on 19 May.
- Major anomalies are identified by both settings of *T*.

T=20 mins

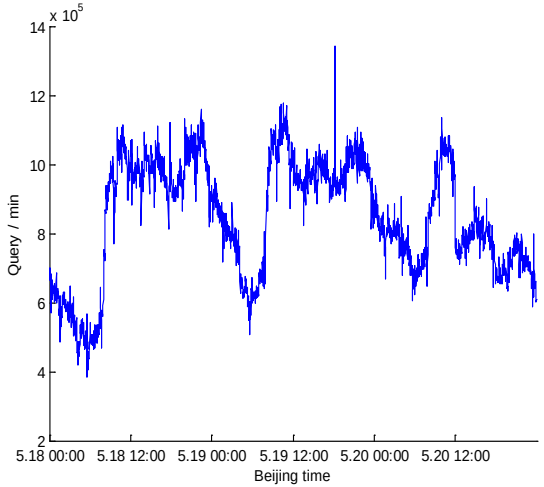


T=30 mins

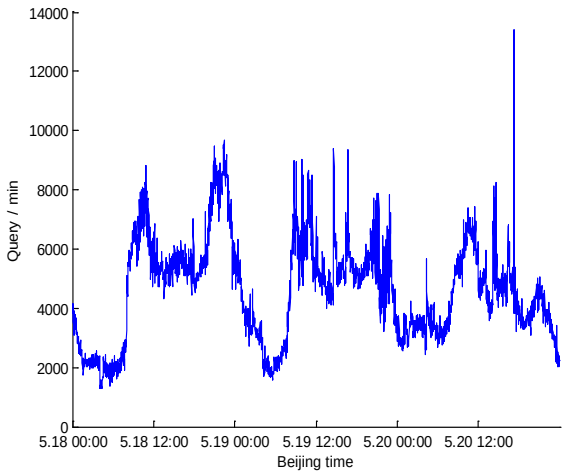


The query rates of the six provinces

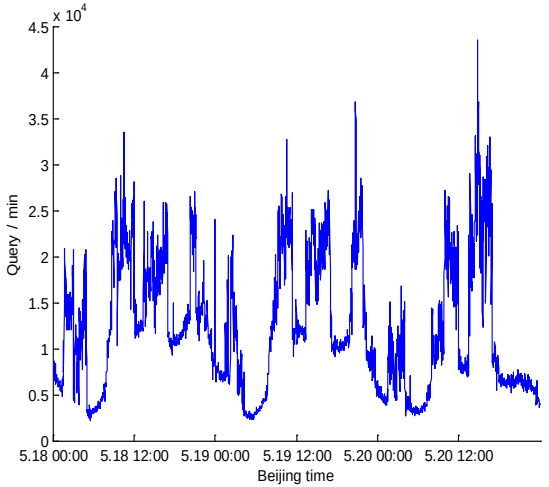
Overall



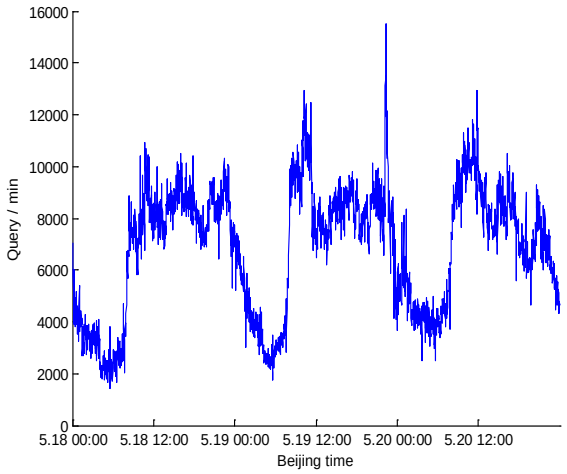
Anhui



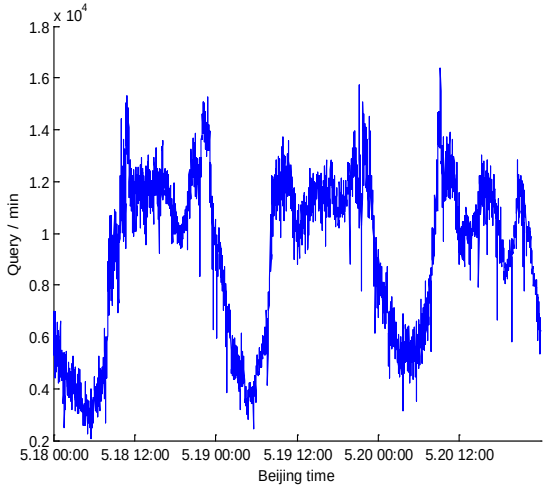
Jiangsu



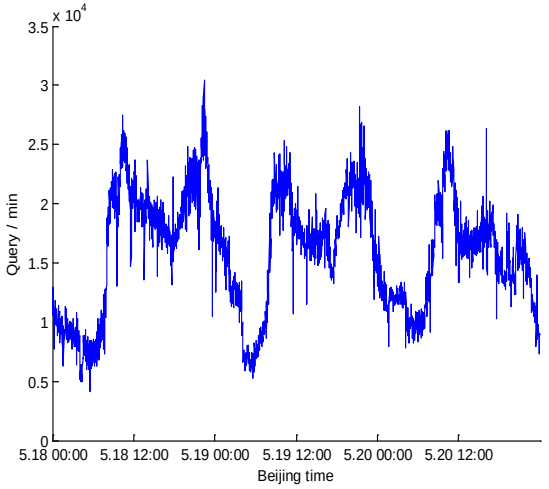
Shanxi



Hebei



Zhejiang

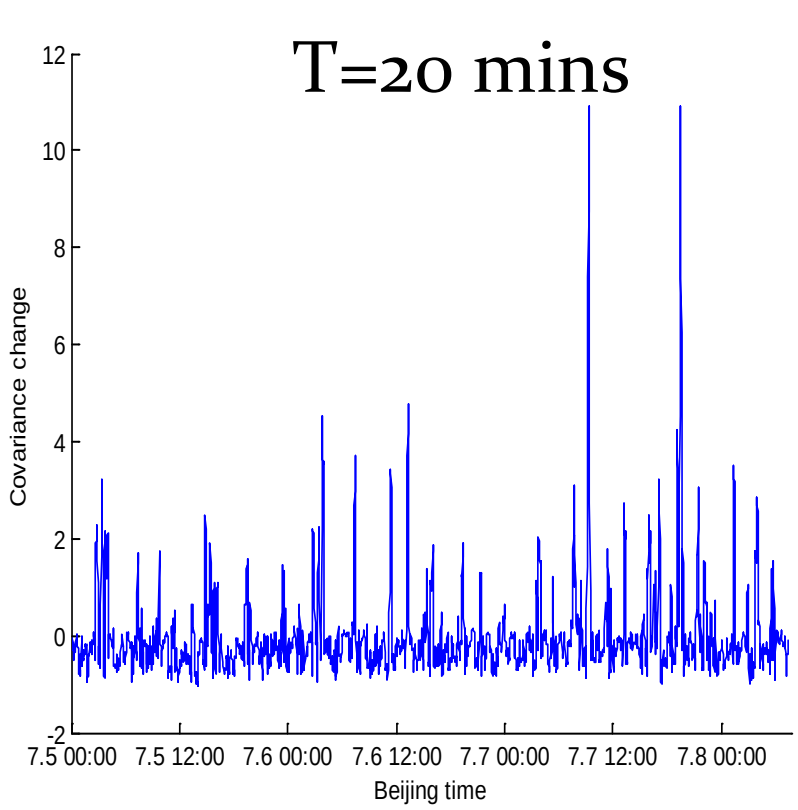


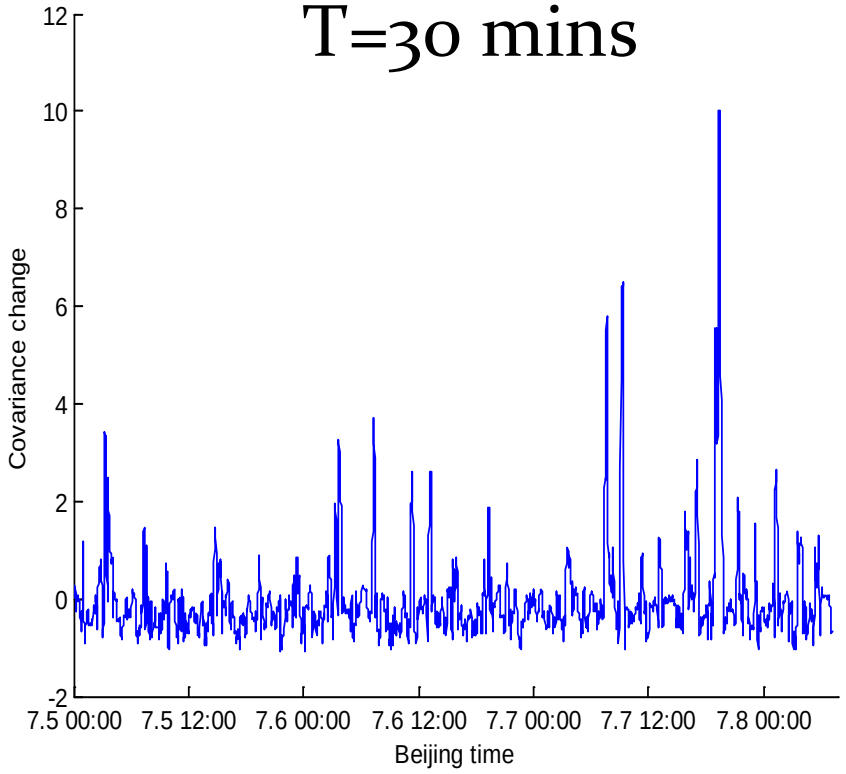
Another event on 5th July,2009

- Queries made to .CN authoritative root during the period of 2009-7-5 00:00—2009-7-7 24:00.
- IP addresses are conversed to the originating regions using Maxmind™ Geolocation Database.
- Three (m=3) features are constructed: number of queries originated from recursive name servers distributed in Xinjiang, China and world regions outside China.
- Covariance matrices for each fixed time slice T=20 and T=30 are constructed respectively.

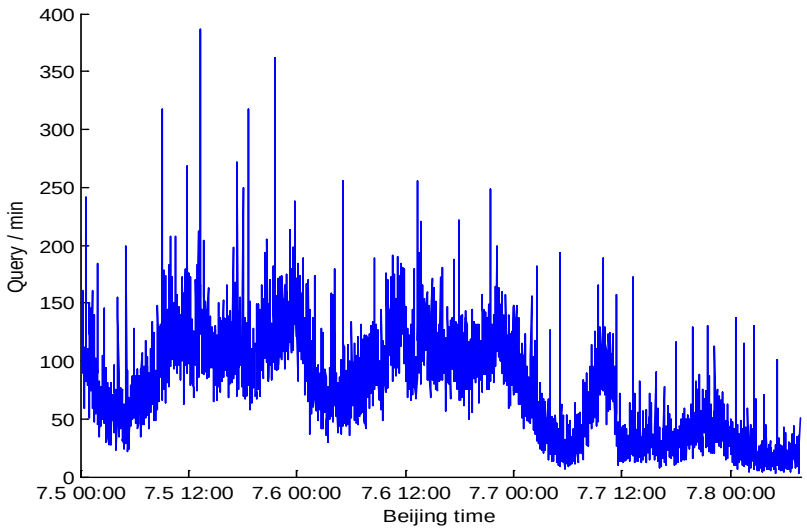
Covariance Matrix Distance

- Distances between the covariance matrix and the mean of all covariance matrices under normal situations
- Change points (Peaks) occurred with a delay instead of simultaneously.
- Major anomalies are identified by both settings of *T*.





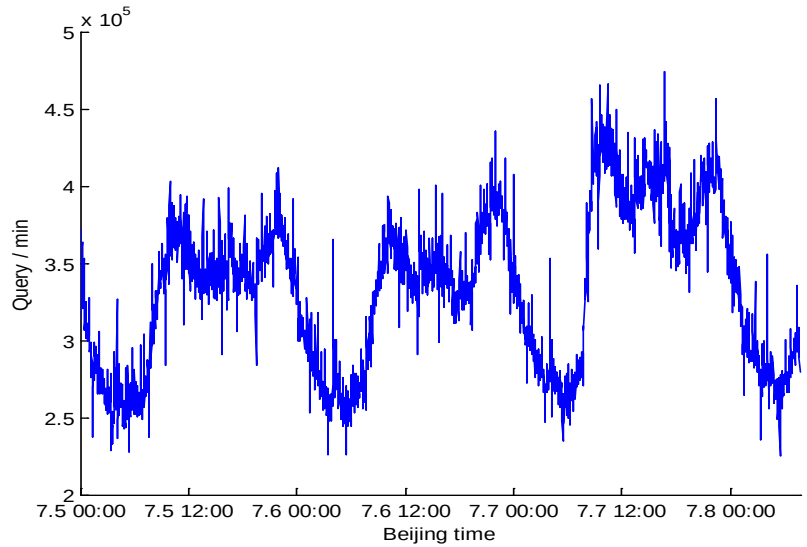
Query load by Cache Name servers from

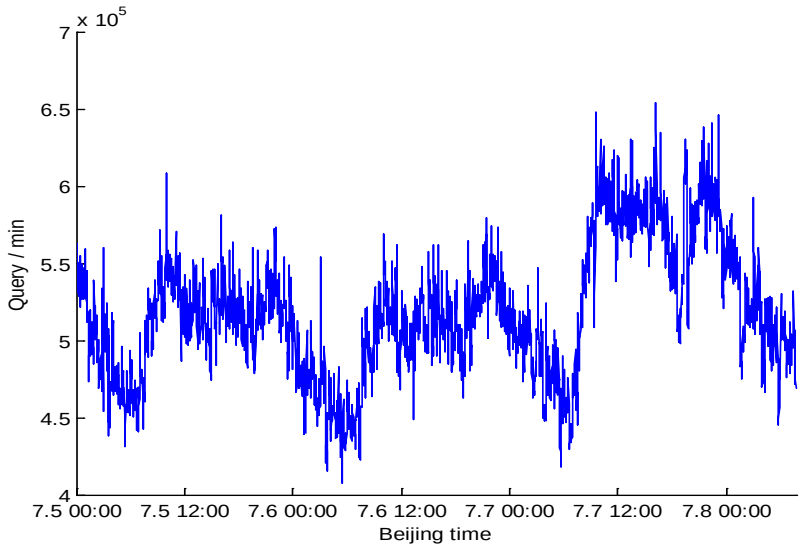


Xinjiang

China

Outside China





What can be done if the anomalies were detected in real-time?

- DNS
  - Important recursive servers located in the affected regions could be traced, monitored and anomalies found.
  - Further more, backup DNS services could be launched in time.
- Web level
  - Web advertisers would be happy by attracting much more eyeballs during the flush period.
  - For Web governance …

Agenda

- Motivation
- What we want to learn?
- How we do it?
- Conclusions and Future work
- Discussions

Conclusions

- Multivariate correlation analysis on the detection of anomalies.
- In terms of correlation, the normal patterns will be different from the abnormal ones.
- Detecting the correlation changes among different features could determine the occurrence of the anomalies.
- The complexity of the method is linear and makes the on-line detection practical.
- Advantage of independence from packet distribution assumption.

Future work

- Further investigations on which features are valid or sufficient for DNS DDoS / Anomaly detection.
- Extensive simulation experiments needed for evaluating the detection rate of various known abnormalities or DNS-related DDos attacks .
- On the selection of the appropriate observed time interval.

Discussions