Data Retention
- an ISP's view -

Presented to the

Forum for the Prevention of Organized Crime

Klaus Landefeld
EuroISPA

Brussels, 14 June 2004

Scope

This presentation is aimed to provide an operational view on how ISPs will be affected by a general, undifferentiated retention proposal.

It will not judge, but merely educate.

Content

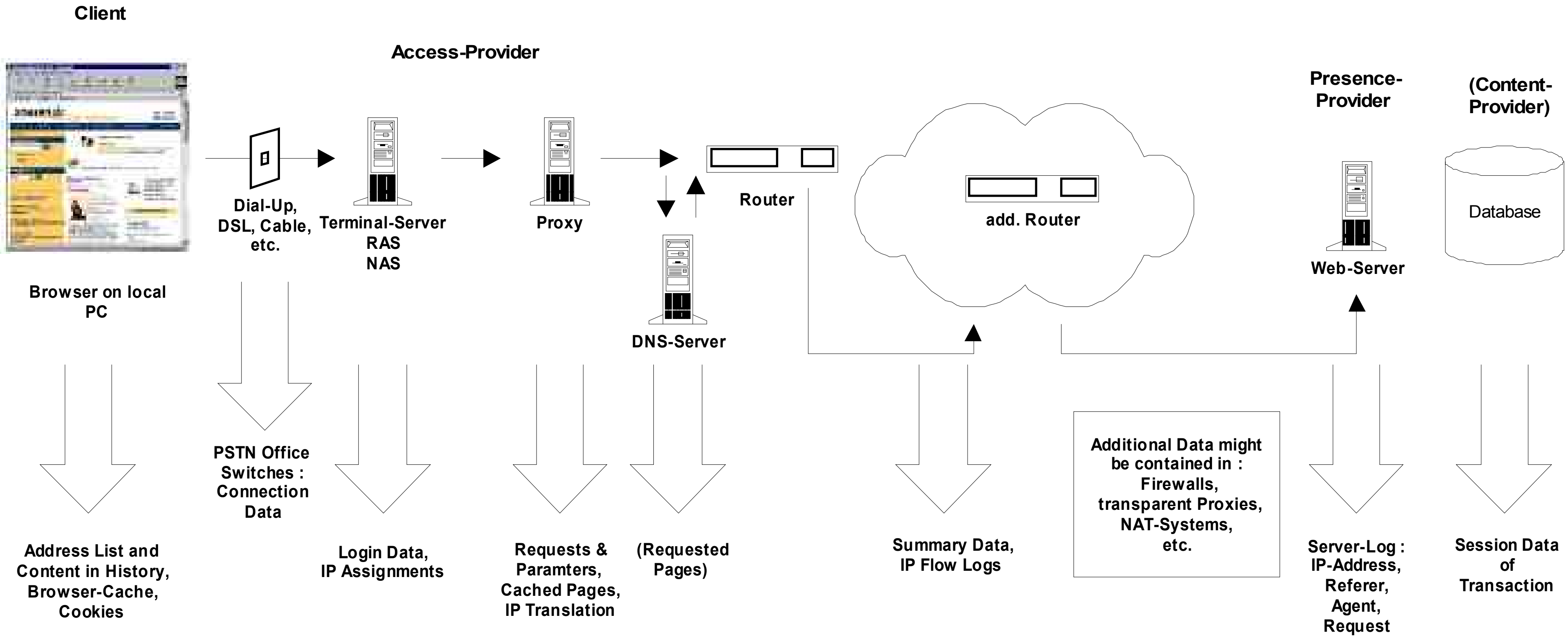
I. Data Collection
II. Data Storage
III. Data Retrieval
IV. Considerations

I. – Data Collection

Sources of Data

- Data Collected from Forwarding Equipment
  - i.e. Routers, Switches, Network Elements
- Data Collected from Individual Services
  - i.e. Servers, Service Gateways
- Data Collected from Individual Databases
  - i.e. User Data, Billing information, etc.

Data Traces: World Wide Web



Client

Access-Provider

Presence-Provider

(Content-Provider)

Browser on local PC

Dial-Up, DSL, Cable, etc.

Terminal-Server
RAS
NAS

Proxy

DNS-Server

Router

add. Router

Web-Server

Database

PSTN Office Switches : Connection Data

Address List and Content in History, Browser-Cache, Cookies

Login Data, IP Assignments

Requests & Paramters, Cached Pages, IP Translation

(Requested Pages)

Summary Data, IP Flow Logs

Additional Data might be contained in : Firewalls, transparent Proxies, NAT-Systems, etc.

Server-Log : IP-Address, Referer, Agent, Request

Session Data of Transaction

Forwarding Equipment - Sources

Routers, Switches

Other Network Elements (RAS, NAS, Firewall)

- Authentication Data
- IP Assignments
- SNMP Data
- Syslog Data
- Flow Logs (IP Flow Data)
- Routing Tables/Structural Data

Forwarding Equipment - Network

- Only edge devices logged today
- Only authentication&volume data logged today
- Flow Logs not logged today, but biggest source
- Requested Data >> Data logged today
- Factor: 50 to 100+ times volume logged today
- Roughly 10-20 times the cost for collection today due to man power, network utilization and the level of pre-processing required

Forwarding Equipment - Issues

- Same data logged several times in Network
- Operator typically knows only one side (source/destination) of individual flows
- Data provided is raw data, pre-processing required to produce meaningful, readable data
- Information required to produce meaningful logs will often require outside source (i.e. DNS, other Providers)

Individual Services - Sources

Servers, Service Gateways

- Multiple Log File formats, plain text
- Huge amounts of data, dependant on log detail level
- Log files designed for humans, not easily machine readable
- Data not systematically logged, but biggest source
- Potential Data >> Data logged today
- Factor: 200+
- Roughly 5-10 times cost for collection today, but collection not the critical issue

Individual Services - Issues

- Not in-line with other data (i.e. User Names might differ from service to service)
- Anonymous use often possible
- High probability of source being intermediary (Proxy, NAT, Load Balancing)

Individual Databases

User Data, Billing information, etc.

- Requires Software originally used to create Database Structure
- Personal, sensitive Information
- Processing needs to conform to Data Protection standards, i.e. deletion might be mandatory
- No additional data to be retained, no additional collection cost

II. – Data Storage Procedures

Keeping Data - Situation

- Storage Periods not harmonized
- Most logs deleted after 30 days, no long-term storage
- Secure Storage of Logs not industry standard
- Separate Storage required for different types of data, some are off-, some are on-line
- Acceptable legal framework for operators missing, data protection and retention directives almost mutually exclusive

Keeping Data - Issues

- Proposal enhances requirement for clear legislation
- Dependant on country (existing period 0 to 12 month), anywhere from 24 to 36 month of additional data requires storage
- Combined with the increase in Data to be Collected, realistic assumptions for a mid-sized ISP requires 300-600 times the storage capacity used today
- Systems storing several TB of data not in common use, disproportionate cost (x2) will lead to 600 to 1200 times the cost budgeted for storage today (!)

Keeping Data - Face-Off

- Storage of raw data will not lead to useful information, if not stored in correlation
- Establishing correlation will violate individual Data Protection
- Establishing correlation requires significant processing & will be costly, but not covered by case based cost recovery

III. – Data Retrieval Procedures

How to respond if data is requested?

- Man Power problem
- Processing Power problem
- Know-How problem
- Availability of software

Man power

- Personnel needs to be readily available
- Turn-around time of tech staff degrades retrieval capabilities, training required
- Inaccessibility for staff without court order requires special procedures

Processing Power

- Systems need to be available and updated regularly
- Dependant on type of request, several TB of data need to be processed per case
- Time frames should be defined to scale systems
- Significant cost even without actual cases, cost recovery uncertain

Know-How

- Systems change radically on a regular base, knowledge of how data was stored is lost over time
- Correlating data not normal procedure for ISP, might be exclusive for law enforcement
- ISP not necessarily data-mining experts, quality might fluctuate wildly

Availability of software

- Data retrieval Software required
- Database and/or Enterprise software required from time of creation, needs to be maintained
- Data Structures need to be available for all recent and past logs, structures evolve over time
- Individual code might be required dependant on type of request

The Data Mining exercise

- Correlations might be wrong, quality degrades over time – it might not mean what you think
- Clear definitions required on what data to extract, currently not provided by law enforcement
- Significant cost associated with processing TB of data, cases will cost several EUR 1.000 each
- If new code/correlation required, several EUR 10.000 realistic to complete
- Exercise only complete if data from various sources is correlated

IV. – Considerations

Reality check: Collection today

- On relative terms, ISP's retain less and less data
- 2/3 of all connections are unmetered connections today, steeply rising
- DSL, Cable, Dial-up – mostly unmetered connections
- Email, Web, P2P, Chat – mostly unmetered services
- VoIP, Video – only partially metered
- 65%+ of all data stored today is already for legal and not operational purposes

Incompatibility

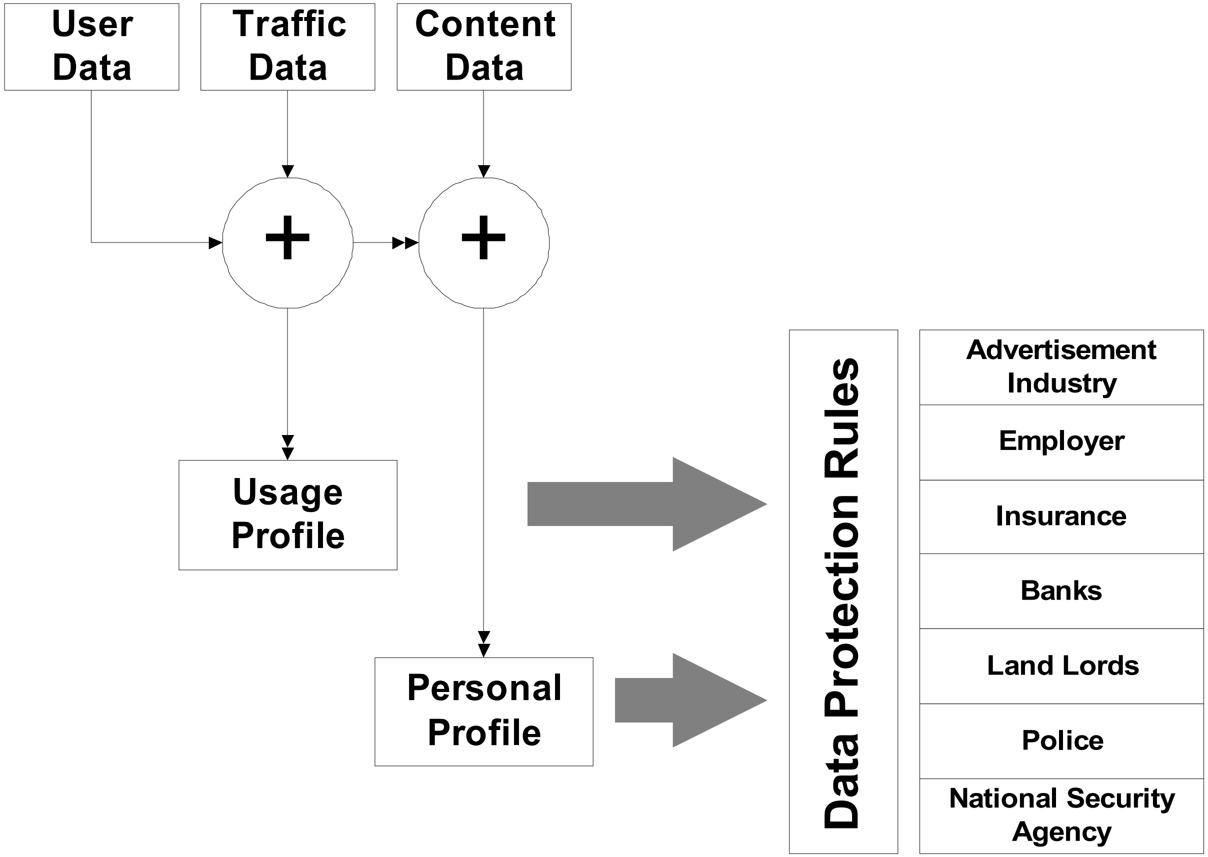
- Advent of unmetered services requires less and less data to be stored & processed by ISP’s
- Law enforcement requires more and more data to be stored & processed by ISP’s
- Collection & Storage “exclusive” for official use?

- Total cost of proposal will increase cost of services, but Europe already top of the list
- In order to maintain the competitive environment, a harmonization of cost recovery might be required

Topics to consider

- New technologies (i.e. AOL 6pipes) will mislead investigations
- Transfer of legacy services (i.e. Telephony) to IP will prompt requirement to retain content as well
- Law enforcement might be required to retrieve data, need to avoid responsibility if data is available but not used
- "Available" data will be a target for civil cases as well

Profiles and Interested Parties



Thank you for your attention

EuroISPA

European Internet Services Providers Association

klaus.landefeld@eco.de