Building a Prototype for Network Measurement Virtual Observatory

Peter Matray, Istvan Csabai, Peter Haga, Jozsef Steger, Laszlo Dobos, Gabor Vattay

Department of Physics of Complex Systems
Eotvos University, Budapest, Hungary
email:{matray, csabai, haga, steger, dobos, vattay}@complex.elte.hu

ABSTRACT

Online sharing of scientific information has accelerated the research activity in various different domains of science. This fact inspires us to initiate this kind of approach in the field of network research and review some projects pointing towards this direction. Using the experiences of similar efforts in other domains of sciences we are building a prototype node for Network Measurement Virtual Observatory. The goal of the observatory is to stimulate network research through sharing available measurement data along with analysis results and providing easy-to-use "online" network data analysis tools for network research and management purposes. We would also like to initiate discussion about standardization of network measurement data and to motivate other researchers to publish their own data and tools. In this paper we sketch the basic concept of Virtual Observatories and present a prototype system developed to share measurement data and tools associated with the ETOMIC measurement infrastructure.

Categories and Subject Descriptors

C.2.3 [**Network Operations**]: Network monitoring; H.2.8 [**Database Applications**]: Scientific databases

General Terms

Measurement, Design, Management

Keywords

Network Measurement, Virtual Observatory, Data Sharing, ETOMIC

1. INTRODUCTION

There is a saying: "information or knowledge is power". As Vint Cerf pointed out in a presentation, the saying should be refined: today it's rather the *sharing* of information that is power. Online scientific databases and their corresponding tools allow us to share our scientific achievements efficiently. Thus the rate at which science is done and discoveries are made has been accelerated.

As examples of shared online scientific databases, we can think of the Human Genome Project or recent digital sky surveys in astronomy [1]. A database or archive might also be valuable for people who did not participate in the establishment of it, and publishing the data of a large survey usually does not mean that all the work is done. For instance, different research groups are involved in revealing the genome or complex biochemical network of different species. To examine the evolution of a given protein or physiological function it might be useful to connect these separate databases and use them together. To do this efficiently researchers often need to have access to the *raw* data, not just the distilled results.

This scenario is certainly applicable for a variety of other research domains, where the actual measurement activity and the scientific analysis are separated. When aiming to understand a large-scale system, measurement groups (usually international collaborations, using some expensive infrastructure) create large surveys, then the preprocessed, calibrated, organized results of these surveys are made public through some Internet accessible archive, and the research community can analyze it or use it as reference data for decades. By combining several of the existing and evolving archives we can produce substantial new results that are impossible to achieve otherwise, using only small subsets of the available data.

One of the first large interconnected databases were introduced in the field of astronomy, hence these archives and this style of research are often called *Virtual Observatories.* Based on these experiences we are building a prototype node for Network Measurement Virtual Observatory (*nmVO*). The goal of the observatory is to stimulate network research through

a) sharing available measurement data along with analysis results

b) providing easy-to-use "online" network data analysis tools for network research/management purposes

c) initiating discussion about standardization of network measurement data

d) deploying the concept of Virtual Observatories to motivate other researchers to publish their own data/tools

MineNet'07, June 12, 2007, San Diego, California, USA.
Copyright 2007 ACM 978-1-59593-792-6/07/0006 ...$5.00.

In this paper we will elaborate on these issues through an example prototype built for the measurement data already available from the ETOMIC active probing infrastructure, on which we perform various large-scale Internet measurements (topology discovery, bandwidth estimation, queueing delay tomography etc.) since 2004.

The contribution of our work is an approach that is novel to network sciences and might fruitfully affect mining of long term, large-scale network data. In Section 3 we unfold the aims and basic concept of *nmVO*s. Then we briefly highlight the main properties of ETOMIC in Section 4, and provide a short overview of measurements performed so far, and of datasets we can offer for the community. We go on to explore the observatory database in Section 5. The database is the central component of the architecture, thus we sketch the key elements of the database schema, and provide some examples of possible queries. Finally, in Section 6 we introduce EtomicServices, a service to share the data and functions that are stored in the *nmVO*.

2. RELATED WORK

The limitations of a short paper does not allow us to review all the efforts in the field of network measurement data organization, we just highlight a few of the recent approaches.

2.1 DatCat

In 2002 US-based CAIDA began to develop the Internet Measurement Data Catalog (DatCat, [2]), a system originally designed to communicate information about their actively and passively collected Internet data. Nevertheless, DatCat also offers platform for sharing datasets stemming from researchers outside CAIDA: the project establishes a searchable index of publicly available measurement data. The catalog contains *metadata* about the measurements, the actual datasets are not stored in the database. Instead a link is stored through which the data can be reached. These include passive traffic traces, traceroute logs, BGP tables or virus propagation studies. The format of the downloadable datasets vary, there are text files and binary files too, many of them released as a compressed package.

2.2 MOME

The analogous MOME project [3] was run under the EU's Information Society Technologies (IST) Programme. Similarly to DatCat it maintains a meta-database functioning as a repository of information about accessible Internet measurement data. Nevertheless, the implementation is differing: on the one hand MOME aims to provide a certain level of standardization in order to exchange data through a unified interface. For instance, they produced a standard for storing traceroute measurements [4]. On the other hand MOME allows the users to exchange and run analysis tools (like arrival rate, jitter calculation, capacity estimation etc.) on the raw measurement data. The results of these analyses are inserted into the MOME database. Measurement data accessible through the MOME database include packet and flow traces, QoS data, routing information or HTTP traces.

2.3 MAWI

The MAWI Working Group of the WIDE Project [5] maintain a traffic data repository containing packet traces from the WIDE backbone in Japan. Traffic traces are passively collected continuously at several points within the backbone since 1999. The traces are gathered by tcpdump, and shared via the project web page. IP addresses are scrambled by a modified version of tcpdpriv. Some statistical properties, like packet size distributions, flow rates, protocol shares, etc. are automatically calculated, while the raw traces are also available for the community in gzip files.

The basic common feature of all the above, and most of the other similar projects is that in most cases they either stay on the meta data level, or gather traffic traces from a single infrastructure only. In contrast, our proposed approach goes beyond that, and tries to create an efficient scheme to handle organization and sharing of different types of raw data.

3. NETWORK MEASUREMENT VIRTUAL OBSERVATORY

The Internet is a large-scale complex system, composed of a huge number of elements with colorful features. To understand a complex system we need complex models and a lot of measurement data to be able to verify the predictions of the models. To propose new models describing traffic or topology dynamics, we have to be able to answer challenging research questions about the network's spatio-temporal properties and need large-scale and long-term observations.

Until now numerous different projects have collected a vast amount of data about various Internet traffic characteristics. However, the scope of these efforts are limited by several factors. Measurements capturing real Internet traffic are generally conducted on a dedicated measurement infrastructure. A single infrastructure usually scans only a narrow segment of the entire Internet topology. Orchestrating a single centralized project is probably not the best answer to these challenges and is not feasible because of practical reasons like financing. This calls for a self-organization similar to the development of community sites, Wikipedia, or the Internet itself. If according to this philosophy distinct datasets could be cross-matched and analyzed together, it would be possible to propose comprehensive studies dealing with the large-scale behaviour of the Internet. Furthermore, measurement archives containing historical Internet data could help data-miners to gain novel insights into the network's long-term evolution.

3.1 Data integration

The integration of different datasets in a common framework raises some major questions. These questions span from informatics-related topics to network measurement-related ones.

The idea of a framework hosting different types of network measurements, possibly produced by different research groups leads to the recognition of the need for standardization, so as to provide a unified interface for researchers and data processing programs, too. We believe that the standardization process has to specify data models instead of discussing file formats used for exchanging data. It might be reasonable to adapt the meta-database or standardization concepts from groups who already made steps in this direction. Surely, international cooperation is necessary to work out this point in details.

Beyond the question of common data formats we have to consider several other aspects of manipulating large datasets. Access to publicly available measurement datasets seems to

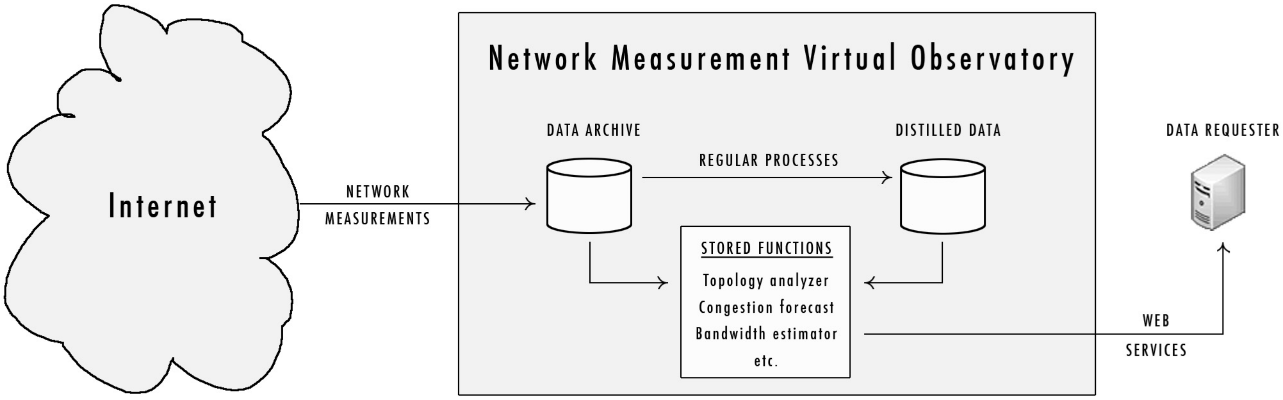

Figure 1: Architecture of the observatory

be an easy task: network trace files can be easily shared through the web. Nonetheless, downloading terabytes of data through the Internet might become prohibitive. (DatCat itself offers approximately 8,5 TB of data as of April of 2007.) The offline transportation of measurement data is unlikely due to several reasons, and to make the problem even worse, the frequent upgrades and curation of the data makes it hopeless to have an always up to date version.

If we would like to merge different catalogs in order to perform joint analyses, we need to solve data conversion and storage problems. For network measurements the traditional way of scientific research is to store the collected raw data in files in some standard (like traceroute dump, tcpdump) or custom format, and process it according to the research questions to be answered. Detailed analysis of a complex network requires large statistical samples, therefore measurements in high bandwidth networks will produce traces with substantial size, even if just a few parameters of the packets are recorded (like IP addresses, arrival time, protocol, size or delay). Practically, measurements can produce dozens of megabytes at each monitoring node that sums up to hundreds of megabytes or even terabytes in multi-node measurements. Focusing on the data analysis results exclusively by discarding the raw data itself is not a solution: measurement data gathered today cannot be reproduced in the future. We have to store the original datasets to allow further re-analysis (applying the various different statistical methods to be developed in the coming years), and to support the study of the long-term evolution of the network.

Particular network measurements (like network tomography) pose another problem: the data we have to cope with is not just large in size but it is also complex. The file-based storage of large and complex datasets might become easily unmanageable (think of nontrivial search tasks), so the application of a *well structured operational database* seems to be necessary. Besides the efficient data storage and access properties, the advantage of using a database management system is that it obliges the users to create a schema for the data. The emergent data models may serve as a starting point for a future standardization, which is necessary if we want to combine data from various measurements.

Another cornerstone of the implementation is the system's *scalability.* It is difficult to imagine a central database that contains long-term traces and is widely used and scales with the number of users and the amount of contributed network data. Building on the practice of astrophysical Virtual Observatories we propose a distributed architecture, where loosely connected observatories share their catalogs about available measurement data and analysis tools.

3.2 Web services

The role of an *nmVO* goes beyond the simple data collecting and archiving functions. On the consumer side there are the researchers or network managers who want to poll the archive to get either a (usually small) subset of the archive or some composite information. To sketch the principle through a possible application, consider a peer-to-peer overlay network that needs management information in order to optimize routing between peers. It would be unthinkable to use gzipped measurement data for similar purposes. On the contrary, a scenario is feasible where one turns to an *nmVO* to get, let's say the average loss of a link on Mondays between 2 and 3 o'clock or the fastest path between two nodes. This means that beyond the data itself, easy-to-use tools are also needed to perform such data transformation queries efficiently.

The XML based Web service technology [6] combined with database management systems is a good candidate to solve these issues. They allow to run either simple queries or more complicated functions that are stored on the server side, where the data is. Thus there's no need to transfer unprocessed bulk data through the network. The developer can create a client application that invokes a series of Web services through remote procedure calls. This is done seamlessly, which means that Web services can be called as if they were local functions on the client side. Data conversions into/from XML are done automatically, messages are carried through the HTTP protocol. See Figure 1 for the Web service enabled *nmVO* architecture.

3.3 Privacy issues

Passive Internet measurements raise important privacy and anonymity concerns. The problem is even worse when traces are gathered from different measurement platforms, since the ability of linking raw data to correlate detailed logs implies the possibility of drawing a general picture of the end-user network behaviour. Besides the application of known anonymization and cryptographic methods ([7, 8]), the proposed database/Web service-based approach might facilitate to overcome these issues via role-based access control.

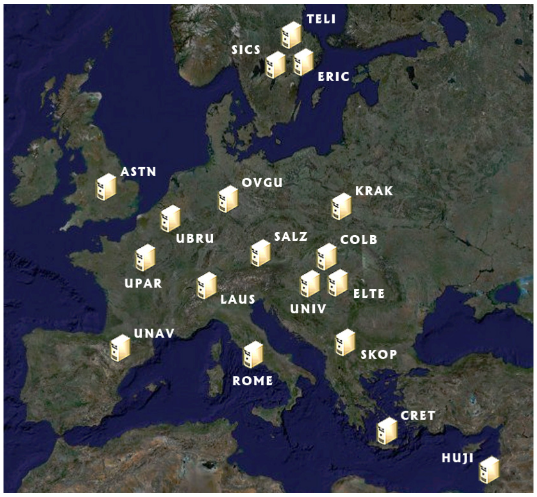

Figure 2: Deployment of ETOMIC nodes

4. MEASUREMENT DATA

Our main guideline during the establishment of the virtual observatory is to provide a suitable framework for network data and analysis tools associated with the ETOMIC infrastructure. ETOMIC is a multi-purpose high-precision active measurement platform deployed across Europe. To help the analysis of spatio-temporal network dynamics, a number of measurements are executed periodically (every day, or even several times a day), generating a vast amount of network data. It is reasonable to integrate all ETOMIC data and analysis tools under an *nmVO*, both for better research efficiency, and to allow others using our data.

4.1 The ETOMIC infrastructure

The ETOMIC (European Traffic Observatory Measurement InfrastruCture, [9]) project is focusing on realizing a Pan-european measurement infrastructure consisting of measurement nodes deployed at selected European locations (see Figure 2). ETOMIC is targeted to provide the scientific community with a measurement platform that is fully open, reconfigurable, extremely accurate and globally synchronized. The system has been designed to allow researchers to perform any kind of measurement experiments: users are provided with a web based interface to manage their softwares, schedule experiments and reserve nodes in the infrastructure. Currently the infrastructure contains 18 GPS synchronized active probing nodes that generate timestamps with an overall accuracy of 0.1 μs. (www.etomic.org)

4.2 Measurements realized so far

Most of the periodic ETOMIC measurements were launched in the spring of 2005. As a major instance, there were approximately 1200 distinct queueing delay tomography measurements performed since April 2005 [10]. To be brief, we list our most significant network measurements in keywords: periodic experiments to trace temporal changes in the inter-ETOMIC *network topology*; regular *one way delay* measurements; *queueing delay tomography* to draw congestion maps of the internal network; *available bandwidth* measurements; end-to-end *loss probability* experiments; *router fingerprinting* to match IP addresses to router ports; *joint experiments* in cooperation with Dimes and PlanetLab ([12], [13]).

5. DATABASE DESIGN

Regular files and regular file-systems don't support efficient searching and ordering of measurement data, since they are based on a sequential processing principle. The application of a database management system, as these are designed for storing and indexing large sets of data resolves these problems.

To enable efficient query execution it is insufficient to store the data in a database as it is. We have to define *indices* to allow rapid response to more complex search challenges. To carefully design indices we need to know exactly how the database is going to be used in the future, which is not the case. Still, we are gaining experience to find optimal indexing, primarily for the raw data tables, where the bulk of the packet-level information is stored.

In the next section we highlight the essential database elements and contexts, but refrain from going into deep details. We also remark that we are putting effort into launching an object oriented database simultaneously to see the performance differences between the two paradigms.

5.1 Schema in brief

The core of the database layout deals with the organization of measurement data. The schema is built on a few essential relations like `Node` containing the discovered nodes/IP addresses, `RawPacketData` incorporating packet-level datasets or `MeasurementCollector` serving as a key to join different submeasurements and analysis results belonging to the same logical unit. To illustrate the schema, we give some example queries too.

`Node`. Describes all IP addresses involved in the measurements. We use a `nodeID` for each IP address, that is unique across the database. The `type` of the node is also stored, if known (ETOMIC, Dimes, PlanetLab etc.).

`MeasurementCollector`. This is a central relation in the database: it's role is to integrate the different measurements, submeasurements and analyzed results that should be handled together. As an example, in an ETOMIC queueing delay tomography measurement every host is sending probe packets to all the other hosts. This action itself generates as much one way delay *timeseries* as the number of ETOMIC hosts participating in the measurement. These timeseries are handled separately, as atomic submeasurements of the tomography experiment. Along with the delay data collection, we are continuously scanning the underlying topology by means of *traceroute* measurements. After the timeseries and the traceroutes are inserted into the database, data evaluation functions deliver the simplified *topology segments* and the calculated *queueing delay distributions*. All these elements (timeseries, traceroutes, topology segments, delay distributions) belong to the very same tomography experiment. It is also possible that a single tomography measurement has many different evaluations, using different statistical methods or modified parameters. These also should be handled together with the earlier analyses. The function of `MeasurementCollector` is to hold these seemingly detached parts together.

`Measurement`. This is a meta-descriptor of atomic, one source-one destination measurements. From this table, usually many different records are linked to the very same record in the MeasurementCollector table, namely to which they all belong. Along with other crucial fields, the measurement's sourceID, destinationID and temporal information (startTime, endTime) is stored here.

By running SQL queries on the MeasurementCollector and Measurement tables we can easily summarize meta-measurement information like the total number/frequency of a given type of experiment. It is also becoming trivial to select measurements fitting to given temporal patterns (like weekday vs. holiday experiments).

RawPacketData. One of the main reasons for using a database instead of storing network data in files, is to be able to complete possibly complex search functions. Thus packet-level representation of raw data seems to be unavoidable. Relevant packet information (previously stored in raw data files) is handled by RawPacketData.

Based on the main ETOMIC measurements, we determined a minimal set of packet attributes to be used in the prototype, to map our raw files. As two key elements, we store destination timestamps in the field sent, and delay values in delay. We mention that saving received timestamps instead of the delays is a theoretically equivalent solution. We chose the latter, since the great majority of our data is delay oriented. See section 5.2 for an example, how to use simple aggregate queries to determine delay distributions.

Traceroute & TraceDetail. Topology discovery is an important type of network measurements. It might be used for several purposes, like tracking down changes in the network's structural map in the time domain. We execute traceroute measurements producing the well known logs containing a listing of hops through which the packets traverse towards the destination station. In Traceroute we give a meta-description of traceroute experiments. The relation is very similar to Measurement, with some extensions like success (indicating the success of the experiment) and flags (containing auxiliary information to help other measurements). The TraceDetail table contains the detailed hop sequences and RTT values.

With the help of these relations, responding to topology related questions is heavily simplified compared to the file-based approach. Besides the regularly executed topology evaluations we can answer questions about the number of discovered paths between two given nodes, or the nature of route changes in a given network segment and time interval.

Distilled results. Due to space limitations, we just mention here that relations exist to store distilled, evaluated data too. These include Topology, describing simplified topologies; TomographyResult, that integrates different results in relation with tomography measurements; QDDistribution storing resolved queueing delay distributions, and other result descriptors.

5.2 Example query

To give an impression how the database might support data mining, scientific research and traffic engineering we provide an example query. Suppose that one needs the delay distribution on a specific network segment aggregated for a given time interval, for instance. To simplify the query we choose two ETOMIC nodes, let's say the one in Paris (nodeID=12) and the one located at our university in Budapest (nodeID=9). To provide a time frame, we aggregate data on this segment for one month (November 2006). To retrieve the distribution we execute the query below.

```
SELECT delay, count(*) as number_of_packets
FROM Measurement m, RawPacketData r
WHERE m.sourceID=12 and m.destinationID=9 and
      m.measID=r.measID and
      m.startTime between
      '11/1/2006 13:02:51' and '12/1/2006 13:02:51'
GROUP BY delay
ORDER BY delay
```

In spite of the huge number of records in RawPacketData (few billion rows at the moment) the query is executed in a few seconds, thanks to indexing techniques. In Figure 3 we plotted the output of the example query executed for three different 1-month periods. The plot shows temporal changes in the delay distributions: the delay mean value decreases in the latter experiments.

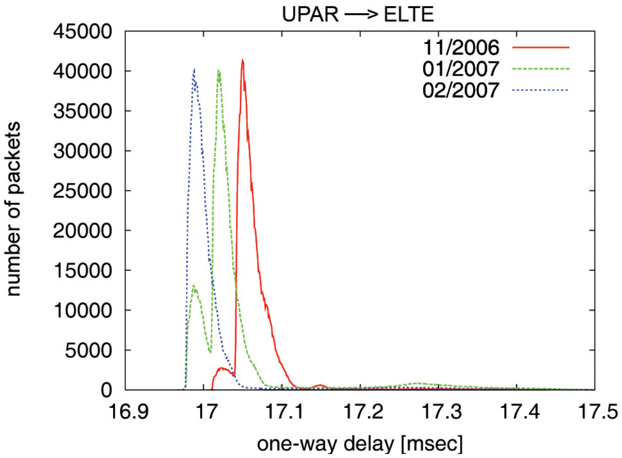

Figure 3: Results of three delay distribution queries

6. *EtomicServices* FOR THE NMVO

Our prototype *nmVO* is based on the technology developed for sharing astronomical data [11]. Though it is not finalized yet, the service (*EtomicService*) is already working: the archive contains 3.5 billion records at the moment, the earliest data is from the January of 2006. All the following services can be accessed both manually through a web interface (http://amd1.colbud.hu/casjobs), and also via Web services (for which the proxy function definition WSDL can be requested from the server) for client applications.

- **Registration:** Users first have to register and get an ID. Among other data they provide their email address to be able to get management information, like server maintenance, etc.
- **Login:** With the user ID and password (or with the unique Web service ID generated for each user) users can log in, and out.
- **Schema browser:** Just by clicking and typing in the browser window, users can explore the data schema, both for the provided standard tables and MyDB tables (see below).

- **Query:** Users can execute queries formulated in SQL.

- **Batch queues:** Since we expect several users using the services simultaneously, and complex data-mining queries on a huge archive can take a long time, we need techniques to avoid the overloading of the server. In the current version there are two queues, one fast (limited to 1 minute and maximum of 1MB returned data) and a long (maximum 500 minutes, and up to the size of free space in the user's MyDB, which is 500MB at registration). The status of the batch queues can be monitored on a separate page.

- **Download:** The results of a query can be directed to the browser, or data can be stored in MyDB (see below). MyDB tables can be extracted into standard CSV or XML format and downloaded to the user's computer. When using Web services, the result is converted to XML, and the client side seamlessly parses it and converts it to suitable data structure (e.g. to an array).

- **MyDB:** MyDB is a fully functional standalone database created for each user. Users can store resulting or temporary datasets here, run joint queries against these or the main data tables, delete or rename MyDB tables and extract data to files.

- **Import:** Users can also import data to MyDB, for example the user can import a set of dates or IP addresses to find the matching records in the archive. In the future we would like to enhance this feature as a more flexible tool for data sharing, and also allow users to upload not just data but data-mining procedures and functions too.

- **Groups:** Similarly to Unix file system groups, it is possible to create groups and share tables among members. This can be very useful for workgroups working on a research project that requires frequent exchange of data.

- **History:** When an SQL query is submitted, not just the results, but the query itself is stored, too. Users can share it with group members, modify and rerun it with new data, etc.

All the user information, the structure of the archive, the history, queue limitation settings, etc. are stored in a separate database. This makes it possible to search among the queries in the history, for example.

7. SUMMARY & PLANS

In this paper we presented an efficient way to organize and share different types of measurement data and corresponding analysis tools to help network researchers in proposing new traffic models, protocol designs etc. Beyond the concept itself, we also introduced a prototype node implementing the architecture. The standard data provider of the prototype observatory is the ETOMIC active probing infrastructure, what we also described briefly, along with EtomicServices, that provide easy access to server side data manipulation tools. To see the concept working, we sketched a simple network research example implemented via the prototype database.

We plan to enable various non-ETOMIC network measurements to be included in the framework. To do so, international collaboration is necessary. (We have active discussions with Dimes, Mome and Lobster.) We are hoping that our prototype might serve as a first link in the chain of Network Measurement Virtual Observatories.

8. ACKNOWLEDGMENTS

The authors thank the partial support of the National Office for Research and Technology (NKFP 02/032/2004 and NAP 2005/ KCKHA005) and the EU IST FET Complexity EVERGROW Integrated Project.

9. REFERENCES

[1] D.G. York et al. The Sloan Digital Sky Survey: Technical Summary, *The Astronomical Journal* 120, p.1579-1587 (2000).

[2] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker and k claffy. The internet measurement data catalog, *SIGCOMM Computer Communication Review* Vol.35. No.5. p.97-100 (2005).

[3] P. A. Gutierrez, A. Bulanza, M. Dabrowski, B. Kaskina, J. Quittek, C. Schmoll, F. Strohmeier, A. Vidacs and S. Zs. Kardos. A Scalable System for Sharing Internet Measurements. *IPS-MoMe2005 Workshop*, Warsaw, Poland (March 2005).

[4] S. Niccolini, S. Tartarelli, J. Quittek, M. Swany. How to store traceroute measurements and related metrics, *http://tools.ietf.org/html/draft-niccolini-ippm-storetraceroutes-03*

[5] K. Cho, K. Mitsuya and A. Kato. Traffic Data Repository at the WIDE Project *USENIX 2000 FREENIX Track*, San Diego, CA (June 2000).

[6] The W3C's Web Service Activity page. *http://www.w3.org/2002/ws/*

[7] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization, *SIGCOMM Computer Communication Review* Vol.36. No.1. (2006).

[8] M. Roughan and Y. Zhang. Secure Distributed Data Mining and its Application in Large-Scale Network Measurements, *SIGCOMM Computer Communication Review* Vol.36. No.1. (2006).

[9] D. Morato, E. Magana, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Haga, G. Simon, J. Steger and G. Vattay. ETOMIC: A testbed for universal active and passive measurements, *TRIDENTCOM 2005*, Best Testbed Award, p283-289, Trento, Italy (February 2005).

[10] P. Mátray, G. Simon, J. Stéger, I. Csabai and G. Vattay. Results of Large-Scale Queueing Delay Tomography Performed in the ETOMIC Infrastructure, *IEEE Global Internet Symposium 2006*, Barcelona, Spain (April 2006).

[11] A.S. Szalay, T. Budavari, T. Malik, J. Gray and A. Thakar. Web Services for the Virtual Observatory *Proc. SPIE Conference on Advanced Telescope Technologies*, 4846, 124 (2002).

[12] The Dimes project, *http://www.netdimes.org*

[13] The PlanetLab project, *http://www.planet-lab.org/*