FUZZYBASE: AN INFORMATION – INTELLIGENT RETRIEVAL SYSTEM

Michael H. Smith
EECS Department
University of California
Berkeley, CA, USA
mhs@mining.ubc.ca

Ljiljana Trajković
School of Engineering Science
Simon Fraser University
Burnaby, BC, Canada
ljilja@cs.sfu.ca

Stuart Rubin
Department of Computer Science
Central Michigan University
Mt. Pleasant, MI, USA
stuart.rubin@cmich.edu

ABSTRACT

We describe the conceptual design of **FuzzyBase**, an information – intelligent system. The purpose of the proposed system is to facilitate intelligent and fast retrieval of information that is of interest to scientific research communities with specific needs, such as getting *relevant* technical information *fast*. It involves the intelligent retrieval of information (both crisp and fuzzy), a subsystem for collecting usage statistics and traffic data, and a high end computer and communications infrastructure.

1. INTRODUCTION

A scientific research community has many electronic needs, such as digital libraries [7], [8], [10], electronic journals [9], mailing lists, and bulletin boards. In this paper, the authors propose the design of an information – intelligent system called **FuzzyBase** to meet these needs [15]. The **FuzzyBase** system offers a specific type of digital information: it is not a digital library, nor an electronic journal, and it is not as simple as a mailing list, nor as impersonal as a bulletin board. Instead, it is a customized information system for a scientific society that includes the above mentioned components. It facilitates intelligent and fast retrieval of information that is of interest to a scientific research community with specific needs. One such need is getting *relevant* technical information *fast*. This system would involve:

- the intelligent retrieval of information, both crisp and fuzzy,
- a system for collecting usage statistics and traffic data, and
- a high end computer and communications infrastructure.

One of the major problems that researchers face today in keeping informed of current scientific results is the need for the search of *relevant* information from a wide variety of sources. **FuzzyBase**'s solution to this problem consists of three components:

- A user's profile database detailing the types of information relevant and important to the users of the system,
- Electronic databases and links to WWW sites that contain relevant information, and
- An interactive search engine that can integrate a user's profile with possible relevant databases and WWW sites in order to retrieve useful and significant information.

This prototype would be one of the first broad-based push technology electronic research systems for a major scientific society, and has the potential to serve as a prototype for other societies. We believe that **FuzzyBase** could be a testbed for future research in databases, retrieval algorithms, and networking.

Our paper consists of two parts. Sections 2 and 3 describe the overall proposed design of our **FuzzyBase** prototype, as well as certain issues dealing with building such a complex system. These issues include designing the architecture of the system in a distributed environment, building the information retrieval and data mining structure, and recognizing the importance of traffic collection and characterizations. The second part, Section 4, focuses in greater detail on the issue of traffic collection and characterization and how they relate to the design of **FuzzyBase**. This is because the development of high-speed networks, intended to carry diverse traffic resulting from statistically multiplexed data, voice, video, and image applications, such as those to be built in **FuzzyBase**, is difficult to accomplish without knowing and understanding their traffic characteristics.

2. THE DESIGN OF FUZZYBASE

The overall design of **FuzzyBase** is shown in Figure 1. It would consist of a number of components:

- **CVs and the profile-builder**: This component automatically builds a profile for each user based on their curriculum vitae (CV). It is used to query and retrieve relevant information based on an understanding of who the user is and what their needs are.

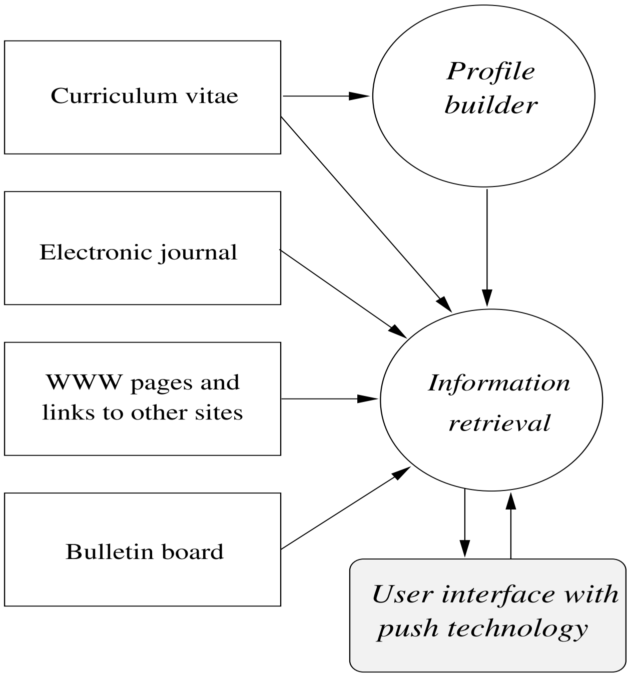

Figure 1: **FuzzyBase** architecture and components.

- **Electronic journal**: An electronic journal incorporated into **FuzzyBase** can be designed to be fully interactive with the **CVs and the profile-builder** so that information can be stored efficiently and relevant information can be retrieved quickly.

- **WWW pages and links to other sites**: They consist of links to relevant scientific information and research results, pointers to electronic abstracts and table of contents of other journals, links to various conferences, calls for papers, abstracts, and conference papers.

- **Bulletin board**: This is a place where members can post and exchange questions and ideas.

- **Information retrieval**: Fuzzy retrieval techniques with a natural language front end are used.

- **User interface with push technology**: This component is designed to send automatically over time requested information.

We now discuss these components in greater detail.

2.1. CVs and the profile-builder

One of the problems of information retrieval is retrieving information that is *relevant* to the user. As the retrieval system usually does not have an understanding of who the user is or how sophisticated he/she may be, the information retrieved is often irrelevant to the needs of the user. Hence, profile-builder is designed to automatically build a profile for each user from their CV's. This profile would be used to query and retrieve relevant information to the user. For example, North American Fuzzy Information Processing (NAFIPS) members' CVs would be electronically added to our database. A profile would be automatically created for each member. This profile would consist of a member's research interests and would enable them to discover other researchers with similar research interests and manuscripts published in their area of interest. Since each CV would include a large number of published references, even those in obscure journals, a significant database can be built quickly. Thus, we would have a real-world electronic experts database with electronic resumes and cross references.

This system can be used as a testbed to examine and solve a number of research issues dealing with automatic profile building. The idea here is essentially to recognize that a sequence of words is embedded in context. This context can yield a picture of the individual that can contain far more information than does only the set of relevant keywords. For example, the CV might contain the phrase: "research interests include knowledge-based systems, fuzzy logic, and data-based systems." In addition to the stated interests, it may be concluded on the basis of the contextual sequence that the person is more interested in AI than in databases. This hypothesis would be useful to an intelligent agent for purposes of retrieval. Of course, a plethora of additional inductions are possible in general.

2.2. Electronic journal

In order to create a specific, controllable database for research on querying and information retrieval of relevant information, a new *NAFIPS Electronic Journal* (with a separate editorial board) is to be established. Manuscripts would be electronically submitted and peer reviewed. **FuzzyBase** would allow us to design an electronic journal to go beyond being only an electronic version of a hardcopy journal, but to have embedded in it features unique to an electronic journal. These features include demonstrations, animations, data links, links to source codes and simulations, agents for cross linkings to members CVs, and to other papers. The electronic journal would be interlinked with other sites in our digital library.

The *NAFIPS Electronic Journal* would consist of electronic submission of research manuscripts that would be peer reviewed. The electronic nature of the journal allows for the fast publication of new research ideas. The manuscripts would be also published in a hardcopy format, either as a book/proceedings once a year, or as a special issue of an already existing journal. These hardcopy publications would be suitable for libraries, and would be accessible to researchers and students who would want archival records of current research. We plan to develop links to tutorials with relevant scientific information and research (such as information on fuzzy logic and related areas), as well as pointers to electronic abstracts of other journal submissions. Full text of each manuscript would be available to individual and library subscribers.

2.3. WWW pages and links to other sites

Our prototype would also have WWW page support with links to other sites with relevant scientific information. We would also establish pointers to electronic table of contents and abstracts of other journals and tutorials, WWW pages and links to other relevant sites would be designed.

In addition, our design calls for the creation of a database of electronic experts with electronic resumes and cross references. For example, NAFIPS members' resumes would be electronically added to the database and a profile would be automatically created for each member, including their research interests. This would enable members to find other researchers and published references in their area of interest. The implementation would be based on push technology, where requested information is automatically sent over time to the subscriber. Information would be used to periodically inform members with a similar profile of the latest abstracts, paper titles, new researchers, and of new research areas that had been added to our database or, possibly, to other linked sites.

2.4. Bulletin board

We would establish a bulletin board where members can post questions and ideas, with links to various conferences such as NAFIPS, calls for papers, abstracts, and conference papers.

2.5. Information retrieval

Another component of **FuzzyBase** is the querying and fuzzy information retrieval of relevant information. Preliminary investigation of fuzzy retrieval and data mining techniques planned to be used for extracting useful information from the database of electronic manuscripts, tutorials, resumes, and links to other relevant sites was described in more detail in [15]. One approach involves the automatic and dynamic acquisition of fuzzy membership functions and the use of transformation rules to map an input, such as curriculum vitae, into an output, such as a key(s) for hashing to an appropriate response.

2.6. User interface with push technology

A front end to the fuzzy retrieval component is a specification subsystem that uses natural language. Rather than scan a list of keywords, use boolean query searches, or successive menu windowing, we propose to use a natural language front end. This front end would be fully adaptive and can learn to translate domain-specific natural language from a trusted user. The generated query is then passed on to a fuzzy retrieval system.

Push technology, where requested information is automatically distributed over time to the subscriber, would be used to periodically (weekly or monthly) and automatically via email inform members (or anyone else with an existing profile), of the latest new abstracts, paper titles, relevant conferences, new tutorial information, new researchers and research, that have been added to our database or, possibly, to other linked sites. Appropriate search engines and cross-references would be developed. This subsystem would be broader than the ACM Digital Library's proposed member profile to support automated notification of the appearance of relevant ACM articles because **FuzzyBase** would have a broad base of information such as articles, conference calls for papers, tutorial information, a database of other CVs, links to other societies and journals, conference papers, and links to university and industrial projects. Furthermore, **FuzzyBase** would use member profiles generated from their CVs (containing detailed information on research interests, and papers they have published), and would have an extensive understanding of who the user is and what their interests are. Thus, we can perform analysis on the information available in **FuzzyBase** to retrieved relevant information for that user. We can choose to do shallow or deep searches depending on how deep a profile the system develops from each CV.

3. RESEARCH AND DEVELOPMENT ISSUES

A number of research and development issues need to be considered when building **FuzzyBase**. These research issues are difficult to address because of a lack of suitable real-world testbeds. Computer simulations of such systems are often inaccurate in absence of realistic models. Thus, the authors propose to develop a small experimental research prototype of **FuzzyBase** for the electronic querying and retrieval of research archival information.

We plan to build the system with the help and involvement of the NAFIPS and possibly with the IEEE Systems, Man, and Cybernetics (SMC) Societies. By having scientific societies involved, it is possible to build a working, real-world prototype of an information – intelligent system suitable for the retrieval of research archival information. Another advantage is that our information domain would be much smaller but more focused than general systems that are often too complex. Consequently, the research questions arising from our system are more specific and hence more solvable.

The following issues would need to be addressed during the design of our system and systems of similar complexity:

3.1. FuzzyBase architecture in distributed environment

We are investigating suitable architectures and system components that are necessary for making an efficient and reliable system. An important issue is the selection of the best suitable architecture for the high-speed data network to support fast retrieval of the information [2]. This architecture also needs to support a scalable database systems containing the electronic information. Finally, the prototype system would be built on the SUN Ultra 30 platform, under Solaris 2.6 operating system.

3.2. Information retrieval and data mining

We are exploring a design of a system to automatically build a user's profile from supplied CVs as part of the design of a querying and information retrieval system of technical research data. We are also investigating suitable search engines to be implemented in **FuzzyBase** that can function across distributed environments, and we plan to address the problem of synchronization of query results across distributed environments.

We are investigating the choice of an appropriate database structure that could support a fast information retrieval. It is planned to investigate data mining techniques [3], [16] for extracting useful information for the proposed electronic publishing system. We plan to use fuzzy techniques for extracting desired information from the database of electronic manuscripts, tutorials, resumes, and links to sites with relevant research information. A more detailed analysis of our approach to the data mining techniques was presented in [15].

3.3. Traffic collection and characterization

As discussed in Section 4, our real-world prototype system would serve as a testbed for collecting usage statistics and traffic data. Collected traffic data traces from **FuzzyBase** would be used to predict its network performance and utilization. They would enable us to learn about suitable architectures of high-speed data networks needed to support such electronic systems. Past measurements of WWW traffic [1], [6] indicate complex traffic patterns that exhibit self-similar behavior [11], [17]. The impact of such traffic, rather different from Poisson models used to design telephone networks, on the system architecture that supports data, voice, image, and video applications is an open research problem.

4. IMPACT OF TRAFFIC CHARACTERIZATION ON SYSTEM DESIGN

As discussed earlier, the development of existing and future high-speed networks, intended to carry diverse traffic resulting from statistically multiplexed data, voice, video, and image applications, is difficult to accomplish without knowing and understanding their traffic characteristics. In this section, we discuss this issue and how they relate to and influence the design of **FuzzyBase**.

The **FuzzyBase** prototype would be used as a testbed for collecting usage and traffic statistics. The controllable size of the system and the variety of applications that would be implemented would provide a rich source of traffic data. Meaningful Internet and WWW traffic measurements are not readily available. Performing traffic measurement is usually difficult because of the large number of users, variability of a typical Internet connection, and variety of applications [6]. The few Internet Service Providers (ISP) that collect statistics are either unable to release them due to customer privacy concerns, or are unwilling to share the data for proprietary reasons. Hence, there is a growing need in the researchers community for not-intrusive traffic measurements and collection. In the absence of well understood and tractable traffic models, these traffic data are proving to be an invaluable tool for quantifying network performance. They can be expected to provide invaluable answers to a number of open traffic engineering issues. For example, it is to be expected that by using a prototype of **FuzzyBase** to explore call admission control policies, congestion control algorithms, and optimal routing algorithms using realistic traffic traces, this would provide reliable and trustworthy results leading to a better design of **FuzzyBase**.

4.1. Failure of traditional traffic models

In the past, traffic measurements and characterization have been essential in designing telephone networks. For these networks Poisson traffic model proved adequate for their design, provisioning, and planning. Traditional mathematical approaches to traffic modeling, such as queuing theory, have met with limited success in today's networking environment. That is, at least in part, due to the complexity of the traffic supported by today's networks, which carry traffic from data, video, image, and voice applications. Inevitably, simulation and empirical techniques would need to play a larger role in designing today's and future networks than the traditional mathematical techniques, which were essential for the development of telephone networks. For example, recent traffic measurements and detailed empirical studies of traffic traces revealed that traditional traffic models cannot capture the actual network traffic and have led to the development of self-similar or fractal models [5], [12], [13] for high-speed network traffic [11], [14], [17]. The essential difference between traditional traffic modeling and the self-similar models has practical implications for **FuzzyBase**'s network and systems engineering and Quality of Service (QoS) requirements because they will affect the design of the system's elements (links, routers, WWW servers).

Traffic characterization work has recently been shown to be promising due to the presence of traffic "invariants" (e.g., self-similarity) [6], [11], [17], and traffic measurement tools and techniques are slowly emerging. Measured traces of actual **FuzzyBase** traffic will be made available to the research community.

4.2. Traffic measuring tools

We plan to deploy in **FuzzyBase** hardware and software traffic measurement tools and techniques that are currently being used by researchers and are just becoming available to the research community. Obtaining a reasonable insight into the performance of WWW based systems such as **FuzzyBase** can be greatly facilitated with a traffic measurement infrastructure. Such infrastructure requires deployment of traffic collection and performance evaluation tools, and development of techniques for a non-intrusive measuring and analyzing traffic congestion levels, queuing delays, and packet loss.

These measurement tools may collect traffic statistics or capture the part or the whole packet. These statistics can be gathered per packet basis (inter-arrival times and packet sizes), or per window basis. While collecting only traffic statistics for **FuzzyBase** is valuable for establishing certain parameters of network performance, collection of actual packets is essential for careful traffic modeling and characterization of **FuzzyBase**. Capturing individual packets allows for post-processing of information and modeling of traffic based on various traffic sources, applications, and priority.

Traffic collection usually requires consensus from the service providers for performing the measurements. It can be intrusive or non-intrusive, where the non-intrusive measurements do not interfere with a network normal operation. In addition, we may also differentiate between passive or active collection. Passive collection is performed by monitoring network links and often requires special purpose hardware, while active measurements rely on software probes that collect traffic information by injecting additional packets into the network. Having ownership of **FuzzyBase** would enable us to overcome difficulties associated with traffic measurements from public networks and WWW sites.

4.3. Traffic analysis

We plan to apply statistical tools and methods to explore the underlying nature of the collected traffic traces (e.g., long-range dependencies) from **FuzzyBase**. Researchers have already detected a consistent thread of self-similarity in traffic traces collected from a variety of networks, including WWW servers [1], [6]. In all these traces, self-similarity implied that the means and variances were insufficient to characterize the statistical process and that a certain underlying correlation structure was retained over a range of time scales. Such a new concept, quite different from the Poisson traffic models that greatly helped characterize traffic in telephone networks, requires new traffic models that rely on traffic measurements from a wide variety of network sources and locations.

The self-similar nature of aggregate traffic [11], [17] is closely related to the infinite variance phenomena (i.e., heavy-tailed distributions with infinite variance) at the level of individual sources and/or applications. Such distributions are qualitatively different from the exponential distributions commonly encountered in traffic modeling in the past and are the major cause of the observed long-range dependence property in measured network traffic at the at the aggregate level. The existence of long-range dependency in **FuzzyBase**'s traffic trace may be quantified by estimating its Hurst parameter H, which is related to the decay rate of the autocorrelation function of the time series representing packet arrivals. Once known, the Hurst parameter can be used to generate synthetic traffic traces with the desired long-tailed distribution properties.

4.4. Network control mechanisms

One of the problems of the Internet is controlling the admission of users and avoiding network congestion. Available traffic information will influence the choice and the performance of adopted connection admission control (CAC) policy, and, ultimately, the achieved QoS parameters in the network supporting the **FuzzyBase**. A natural traffic description is the random amounts of traffic generated in a certain time interval (time scale). As an example of possible use of the benchmark suite of traffic traces, we propose to investigate the performance of several CAC policies in packet networks using the collected traffic data. Also, one aspect of **FuzzyBase** design would be to quantify how much more efficiently a network can be run if one knows the traffic description over several time scales. By using **FuzzyBase** we can examine the use of traffic statistics obtained from measured traffic traces (e.g., connection time, traffic volume, burstiness, packet destinations) and investigate how these statistics can be used to control the admission of users to the **FuzzyBase** WWW servers and for avoiding their congestion.

4.5. Establishing billing and pricing policies

The total cost of the Internet is estimated to $200 million a year. While originally subsidized by NSF, Internet is today managed by commercial and nonprofit carriers. It is, therefore, to be expected that the cost of using the network would be recovered through some form of billing and pricing policies. Not all Internet users generate similar amounts of traffic, nor do they send packets over similar distances, nor do they require identical service guarantees. As a result, a variety of pricing policies are currently discussed in the literature. While some network providers have implemented a connection-based billing scheme, others charge users by the number of packets sent internationally (public networks in Chile and New Zealand); yet others have designed different usage-based systems (Compuserve). In a usage-based system, users are billed individually based on the characteristics of the traffic they generate. Such systems, hence, require exact traffic measuring and monitoring. Several recently suggested Internet pricing policies based on various traffic parameters, such as the economic of congestible resources, the effective bandwidth concept, or collected traffic traces from TCP users, have been recently considered. It is planned to use collected statistics from **FuzzyBase** to model various billing mechanisms and explore better pricing policies.

Other important areas that may benefit from reliable and meaningful traffic traces of **FuzzyBase** are novel pricing policies and tariffing strategies for existing and future Internet services. Such services might well rely upon certain measures of traffic sent to and received from Internet users.

4.6. Relevance to a QoS concept

FuzzyBase traffic measurements will provide invaluable information necessary for addressing various QoS issues such as: available network capacity, network availability and outage, or predicting response times. In the absence of a clear definition of these QoS issues, traffic measurements might provide insight into QoS parameters important both to the ISP and to the Internet users of **FuzzyBase**. Establishing a metrics for measuring **FuzzyBase**'s performance might well be driven by the quality of benchmark traffic traces.

Furthermore, these measurements would influence the design of the network elements and the provisioning of **FuzzyBase**'s resources. For instance, some vendors of high-speed WWW servers might relay on traffic characteristics (e.g., flow duration statistics) to assist them in optimizing router performance. For instance, by recognizing traffic suitable for fast switching (caching), one can greatly increase effective switching capacity. Further development and establishing the validity of such technology inevitably requires high quality traffic traces.

5. CONCLUDING REMARKS

One of the major problems that researchers face today in keeping informed of current scientific results is the need for searching for *relevant* information from a wide variety of sources. In this paper, we proposed the design of **FuzzyBase**, an information – intelligent system, to meet these needs. **FuzzyBase** system would facilitate intelligent and fast retrieval of information that is of interest to scientific research communities with specific needs, such as getting *relevant* technical information *fast*.

The **FuzzyBase** prototype consists of subsystems for automatic profile-building from members' CVs, a new *NAFIPS Electronic Journal*, WWW pages and links to other sites, a bulletin board, an information retrieval subsystem, and a user interface with push technology. It would also employ a system for traffic collection and characterization in order to better understand and design **FuzzyBase**. It has a potential to be a testbed for future research in databases, retrieval algorithms, and networking.

6. REFERENCES

[1] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 5, pp. 631–645, Oct. 1997.

[2] I. Ben-Shaul and G. Kaiser, "Coordinating Distributed Components over the Internet," *IEEE Internet Computing*, Vol. 2, pp. 83–86, 1998.

[3] M. Berry and G. Linoff, *Data Mining Techniques*, New York, NY: John Wiley & Sons, Inc., 1997.

[4] G. J. Chaitin, "Randomness and Mathematical Proof," *Scientific American*, Vol. 232, pp. 47–52, 1975.

[5] D. R. Cox, "Long-Range Dependence: A Review," *Statistics and Appraisal, Proc. 50th Anniversary Conf.*, H. A. David and H. T. David, Eds., The Iowa State University Press, pp. 55–74, 1984.

[6] M. E. Crovella and A. Bestavros, "Self-Similarity in World-Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp. 835–847, Dec. 1997.

[7] Special Issue on Digital Libraries., *Communications of the ACM*, Vol. 39, No. 4, pp. 22–78, April 1995.

[8] Special Issue on Digital Library Initiative., *IEEE Computer*, Vol. 29, No. 5, pp. 22–76, May 1996.

[9] P. J. Denning, "The ACM Digital Library Goes Alive," *Communications of the ACM*, Vol. 40, No. 7, pp. 28–29, July 1997.

[10] E. A. Fox, R. M. Akscyn, R. K. Fruta, and J. J. Leggett, "Introduction," Special Issue on Digital Libraries, *Communications of the ACM*, Vol. 39, No. 4, pp. 22–28, April 1995.

[11] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. on Networking*, Vol. 2, No. 1, pp. 1–15, Feb. 1994.

[12] B. B. Mandelbrot and M. S. Taqqu, "Robust R/S Analysis of Long Run Serial Correlation," *42nd Session of the International Statistical Institute*, Manila, pp. 1–37, Dec. 1979.

[13] I. Norros, "A Storage Model with Self-Similar Input," *Queueing Systems*, Vol. 16, pp. 387–396, 1994.

[14] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *Proc. ACM SIGCOMM '94*, London, England, pp. 257–268, Sept. 1994.

[15] M. H. Smith, S. Rubin and Lj. Trajković, "Fuzzy Data Mining for Querying and Retrieval of Research Archival Information," to be presented at *17th NAFIPS Int. Conf.*, Pensacola Beach, Florida, Aug. 1998.

[16] S. M. Weiss and N. Indurkhyaa, *Predictive Data Mining*, San Francisco, CA: Morgan Kaufmann Pub., 1998.

[17] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *IEEE/ACM Trans. on Networking*, Vol. 5, No. 1, pp. 71–86, Feb. 1997.