

Bibudh Lahiri

Curriculum Vitae

Revised Feb 2011

Personal Data

Address: Apt 181, Princeton Arms North II Phone: (515)451-0307
Dorchester Drive Email: bibudh@iastate.edu
Cranbury, NJ 08512 Web: http://home.eng.iastate.edu/~bibudh

Objective

Interested in a research/development position that offers the scope of designing and implementing scalable algorithms/heuristics for computing aggregates over and mining patterns from massive amount of data.

Research Interests

- Data stream algorithms for detecting exploit patterns from network packets
- Distributed algorithms for data aggregation in P2P and wireless sensor networks
- Machine learning techniques for predictive analytics and anomaly detection

Education

Ph.D. Candidate	Comp. Engg.	Iowa State University	Fall 2006 - current	GPA 3.7/4.0
Bachelor of Engineering	Comp. Sc. & Engg.	Jadavpur University	1998-2002	GPA 3.8/4.0

Professional Experience

Research Intern	Siemens Corporate Research	April 2010 - current	Princeton, NJ
Research Assistant	ECE, ISU	Fall 2006 - Spring 2010	Ames, IA
Assistant Systems Engineer	Tata Consultancy Services Limited	July 2002 - July 2006	Kolkata, Manchester

Publications

Journal Publications

- Bibudh Lahiri and Srikanta Tirthapura, “**Computing Frequent Items in a Network using Gossip**”, *Journal of Parellel and Distributed Computing (JPDC)*, 70(12), pp 1241-1253, December 2010

Refereed Conference Publications

- Bibudh Lahiri and Srikanta Tirthapura, “**Finding Correlated Heavy-Hitters over Data Streams**”, *Proc. 28th IEEE International Performance Computing and Communications Conference (IPCCC) 2009*, acceptance ratio: 43/145 = 29%
- Bibudh Lahiri and Srikanta Tirthapura, “**Computing Frequent Elements using Gossip**”, *Proc. 15th International Colloquium on Structural Information and Communication Complexity (SIROCCO) 2008*, pp. 119-130

Book Chapters (invited)

- Bibudh Lahiri and Srikanta Tirthapura, “**Stream Sampling**”, published in the *Encyclopedia of Database Systems*, by Springer Verlag GmbH

Research Competitions

- Bibudh Lahiri, “**A Generic Framework for Detecting Top- k Items from a Stream**” (extended abstract), accepted in the *ACM Student Research Contest 2010* for presentation

Papers under Review

- Bibudh Lahiri, Srikanta Tirthapura and Jaideep Chandrashekar, “**Space-efficient Tracking of Persistent Items in Massive Data Streams**”, submitted to the *5th ACM International Conference on Distributed Event-Based Systems (DEBS) 2011*

Papers under Preparation

- Bibudh Lahiri and Fabian Moerchen, “**A Divide-and-Conquer Approach for Detecting Bursts from Event Logs**”

Talks

- “Finding Correlated Heavy-Hitters over Data Streams”, **invited talk** at The DIMACS Workshop on Network Data Streaming and Compressive Sensing, Rutgers University, October 2010
- “A Generic Framework for Detecting Top- k Items from a Stream”, ACM Student Research Contest, Milwaukee, WI in March 2010
- “Finding Correlated Heavy-Hitters over Data Streams”, The 28th IEEE International Performance Computing and Communications Conference (IPCCC), Phoenix, AZ in December 2009
- “Computing Frequent Elements using Gossip”, The 15th International Colloquium on Structural Information and Communication Complexity (SIROCCO), Villars-sur-Ollon, Switzerland, 2008

Honors and Achievements

Student travel grant for ACM SRC 2010	Microsoft Research	2010
Student travel grant for IMC 2009	ACM/USENIX	2009
Student travel grant for ACM PODC 2007	ACM	2007
Hats-off award	Tata Consultancy Services Limited	2005
8th in class, graduated 1 st class Honours	Bachelor of Comp. Sc. & Engg., Jadavpur University	2002
54th among around 60,000 candidates	West Bengal Engineering Entrance Examination	1998
48th among around 300,000 candidates	West Bengal Higher Secondary Examination	1998
29th among around 500,000 candidates	West Bengal Secondary Examination	1996

Activities

- **Invited speaker at the DIMACS Workshop on Network Data Streaming and Compressive Sensing, Rutgers University, October 2010.**
- **Presented at the ACM Student Research Contest 2010 at Milwaukee, WI in March 2010.**
- **Presented at the 28th IEEE International Performance Computing and Communications Conference (IPCCC) at Phoenix, AZ in December 2009.**
- **Editorial Board Member, Springer** (since November 2009)
- **Presented at the 15th International Colloquium on Structural Information and Communication Complexity (SIROCCO 2008) at Villars-sur-Ollon, Switzerland in June 2008.**

- Reviewed papers for PAKDD 2011, INFOCOM 2010, SIROCCO 2009, ICDCN 2009, LCN 2008, DCOSS 2008 and ICDCN 2008.
- **Attended the 26th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2007) in Portland, Oregon in August 2007.**
- Attended workshops LOCALITY and DialM-POMC in Portland, Oregon in August 2007.

Computer skills

Embedded Systems: MICA2 platform for sensor networks, nesC, TinyOS 1.1

Programming Languages: C, C++, Java, nesC, Unix Shell Script, Latex, PL/SQL

Intrusion Detection Systems: Bro, Snort

Operating Systems: FreeBSD 7.1, Red Hat Enterprise Linux 4.0, Windows XP

Markup Languages: HTML, XML

Middleware: Java RMI, Socket API

Database: Oracle 9i, SQL Server 2008

IDE: Oracle 9i JDeveloper 9.0.2, Eclipse 3.1

J2EE frameworks: BC4J, Struts, Maverick

Statistical packages: R 2.7.2, MatLab 7.4.0

Research and Development Projects

1. **Space-efficient Tracking of Persistent Items in Massive Data Streams (with Srikanta Tirthapura and Jaideep Chandrashekar from Intel Labs Berkeley (Fall 2009-Fall 2010):** Network exploits like botnets and port scans typically exhibit a stealthy but temporally persistent communication pattern. In contrast with heavy hitters, persistent items do not necessarily contribute significantly to the volume of a stream, and may escape detection by traditional volume-based anomaly detectors. We first show that any online algorithm that tracks persistent items exactly must necessarily use a large workspace, and is infeasible to run on a traffic monitoring node. Motivated by this lower bound, we introduce an approximate formulation of the problem and present a small-space algorithm to approximately track persistent items over a large data stream. Our experiments on a real traffic dataset shows that in typical cases, the algorithm achieves a space compression of upto factor of 5x-20x, while incurring very few false positives ($< 1\%$) and false negatives ($< 4\%$). To our knowledge, this is the first systematic study of the problem of detecting persistent items in a data stream, and our work can help detect anomalies in a data stream that are not volume-based, but show a regular temporal pattern of behavior.

Roles played:

- Identified the problem by collaborating with Intel labs
 - Worked on the design, analyses and simulation of the algorithms (with Srikanta Tirthapura and Jaideep Chandrashekar)
2. **Designing Efficient Heuristics for Detecting Bursts from Event Logs (Fall 2010, with Fabian Moerchen, Alexis Motto Legbedji and Ioannis Akrotirianakis):** While mining patterns from timeseries data generated by healthcare equipment, we observed a bursty pattern in the occurrence of the events, i.e., there are “bursty windows” in time where at least k events occur within an interval of length at most t . Our goal is to search for a critical threshold pair (k^*, t^*) such that the number of bursty windows meeting this threshold pair is significantly higher than that for any other combination of threshold values. An exhaustive search over the entire possible range of k -values and t -values is impractical, hence we are currently working on the design of efficient heuristics for the search.

Roles played:

- Formulated the problem
 - Prepared the ground truth by writing programs in Java and plotting results in MatLab
 - Currently designing different heuristics and experimenting with them
3. **Analysis of CT Machine Logs for Predictive Maintenance (Summer 2010, with Dmitriy Fradkin and Fabian Moerchen):** The X-ray tube is one of the most important and expensive components of the Computed Tomography (CT) machines. The tubes have to be replaced and serviced regularly for routine maintenance. Predictive maintenance of the CT machines becomes easier if it can be predicted when the tube should be replaced next time. Since we only knew when in the past tube replacements had taken place, we had to rely on unsupervised learning techniques. We trained a multidimensional Gaussian Mixture Model using data (on temperature, current, voltage etc) from time-windows much before the replacement, and then derived the likelihood values of the data points based on these models as the dates approached the replacement date. For a significant number of machines, we noticed a steady decline in the likelihood values as we approached the replacement dates.

Roles played:

- Worked on training and testing the GMM on event logs using R
4. **Finding Correlated Heavy-Hitters over Data Streams (Fall 2008-Summer 2009, with Srikanta Tirthapura):** We designed a small-space, deterministic approximation algorithm to answer queries of the following form: “On a stream of (x, y) tuples, maintain the y -values that occur frequently alongwith the x -values that appear frequently in the stream”. Previous techniques for correlated aggregates cannot handle queries of this form, since even for a one-dimensional stream, heavy-hitters cannot be maintained exactly using small space. Our online data stream algorithm is easy to implement and uses workspace which is orders of magnitude smaller than the stream itself. We present provable guarantees on the maximum error estimates, as well as experimental results, that demonstrate the space-accuracy trade-off, on more than one billion packet headers from a backbone network link. This work was published in IPCCC 2009.

Roles played:

- Identified the problem
 - Worked on the design, analyses and simulation of the algorithms (with Srikanta Tirthapura)
5. **A Generic Framework for Detecting Top- k Items from a Stream (Summer 2009):** We propose a generic framework for identifying the k most frequent items from a network data stream in a single pass, using a workspace much smaller than the stream itself. We observed that any small-space sketch that approximately maintains the frequencies of the items appearing in a stream can be utilized to extract the top- k items. We show theoretically how the Misra-Gries sketch and the count-min sketch can fit into this framework, analyze the space-complexities of both, and present experimental results on packet traces from a backbone network link. With the realistic assumption that the frequencies of the items follow a Zipfian distribution, we fully present an algorithm, based on the Misra-Gries sketch, that takes only $O(k^{O(1)}/\epsilon)$ space and guarantees that the solution provided is ϵ -approximate. An extended abstract of this work got accepted in the ACM Student Research Competition, 2010.

Roles played:

- Identified the problem
- Worked independently on the design, analyses and simulation of the algorithms

6. **Design of Gossip-based protocols for identifying frequent items in a distributed database: ECE, ISU. Summer 2007 - Summer 2008 (with Srikanta Tirthapura):** The objective was to design distributed Gossip-based algorithms for identifying the frequently occurring data items in large-scale distributed systems, with probabilistic guarantees on accuracy. The nodes exchange small-sized “sketches” or synopses of their individual data sets through an underlying gossip mechanism. These scalable, fault-tolerant and self-stabilizing algorithms can be applied to identify the most popularly accessed resources in large P2P networks. This work was published in SIROCCO 2008 and in JPDC, volume 70, 2010.

Roles played:

- Identified the problem
- Worked on the design and the probabilistic analyses of the algorithms (with Srikanta Tirthapura)
- Simulated the algorithm on Java

7. **Design and implementation of a distributed, lightweight sensor network directory: ECE, ISU. Fall 2006 - Spring 2007 (with Srikanta Tirthapura and Bojian Xu):** The goal was to design and implement an efficient and scalable distributed sensor network directory to track mobile objects in a wireless sensor network. The algorithm adopted was the distributed Arrow protocol, in which local change in the object’s position does not usually result in a global change in the network. This was deployed in the first system demo with a fixed spanning-tree-based network of MICA2 motes.

Roles played:

- Designed the initial system demo (with Srikanta Tirthapura)
- Implemented the initial system demo with nesC on TinyOS and CrossBow motes (with Bojian Xu)
- Demonstrated the application in the Graduate Research Workshop organized by the Department of Electrical & Computer Engineering, ISU in April, 2007.

8. **A peer-to-peer file-sharing system: ECE, ISU. Fall 2007:** This application, developed using socket programming in Java, supported file sharing across multiple machines. The design was adopted from the Gnutella specification and the application was successfully tested on a platform involving multiple Linux machines.

Roles played:

- Worked on the object-oriented design of the protocol
- Implemented the protocol over a network of Linux machines

Industrial Projects

1. **Data Extraction from Text Logs of CT Machines (at Siemens Corporate Research, Princeton, Summer 2010):** We worked with logs obtained from Computed Tomography (CT) machines where each machine had a number of photomultiplier tubes (PMTs) and two detectors. The service history of these PMTs and detectors were available as free-format text in the log files. The goal was to convert this unstructured textual history into structured data, which gives accurate description of when each PMT was serviced/replaced, and which detector(s) was/were serviced alongwith it.

Roles played:

- Implemented a text-parser using the regular expression API of Java.

2. **Trailblazer (Acuity)(TCS, Kolkata, Feb 2006 - July 2006):** Acuity is a web-based application developed for McGraw-Hill Digital Learning to enhance the functionality of an existing application named TrailBlazer. It is used to create test, exercise and assignments for students; score the students' response and generate predictive, diagnostic and summary reports. The system also monitors the structure of state curriculum and tracks the progress of a student in class.

Roles played:

- Interacted with the customer for requirement collection
- Designed the class and sequence diagrams using UML tool Enterprise Architect 6.0
- Developed the application on Maverick framework using Eclipse 3.1

3. **DUoS and Associated Distribution Systems (TCS, Kolkata and Manchester (UK), Nov 2002 - Jan 2006):** The goal of the DADS program was to meet the regulatory requirement to separate electricity distribution and supply businesses of United Utilities. The application developed by TCS provided a common web-based interface across all the distribution functions and it replaced a number of different legacy applications that were used for distribution billing previously.

Roles played:

- Interacted with the customer for requirement collection
- Developed prototypes and specification documents for the middleware and database components
- Developed web components on BC4J (Business Component for Java) framework using Oracle 9i JDeveloper 9.0.2
- Developed and maintained programs for data migration from legacy applications
- Joined the project as a developer and later took the responsibility as a process owner of Payment Management and Revenue Protection Service modules

References

Prof. Srikanta Tirthapura (Major professor)
Dept. of Electrical & Computer Engineering
Iowa State University
Phone: (515)294-3546
Email: snt@iastate.edu
Web: <http://home.eng.iastate.edu/~snt>

Prof. Daji Qiao (Research Collaborator)
Dept. of Electrical & Computer Engineering
Iowa State University
Phone: (515)294-2390
Email: daji@iastate.edu
Web: <http://home.eng.iastate.edu/~daji/>

Prof. Yong Guan (Co-advisor)
Dept. of Electrical & Computer Engineering
Iowa State University
Phone: (515)294-8378
Email: guan@iastate.edu
Web: <http://home.eng.iastate.edu/~guan>