INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

International Computer Science Institute Activity Report 2007

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198 USA
phone: (510) 666 2900 (510) fax: 666 2956 info@icsi.berkeley.edu http://www.icsi.berkeley.edu

Principal 2007 Sponsors

Cisco
Defense Advanced Research Projects Agency (DARPA)
Disruptive Technology Office (DTO, formerly ARDA)
European Union (via University of Edinburgh)
Finnish National Technology Agency (TEKES)
German Academic Exchange Service (DAAD)
Google
IM2 National Centre of Competence in Research, Switzerland
Microsoft
National Science Foundation (NSF)
Qualcomm
Spanish Ministry of Education and Science (MEC)

Affiliated 2007 Sponsors

Appscio
Ask
AT&T
Intel
National Institutes of Health (NIH)
SAP
Volkswagen

Corporate Officers

Prof. Nelson Morgan (President and Institute Director)
Dr. Marcia Bush (Vice President and Associate Director)
Prof. Scott Shenker (Secretary and Treasurer)

Board of Trustees, January 2008

Prof. Javier Aracil, MEC and Universidad Autónoma de Madrid
Prof. Hervé Bourlard, IDIAP and EPFL
Vice Chancellor Beth Burnside, UC Berkeley
Dr. Adele Goldberg, Agile Mind, Inc. and Pharmaceutrix, Inc.
Dr. Greg Heinzinger, Qualcomm
Mr. Clifford Higgerson, Walden International
Prof. Richard Karp, ICSI and UC Berkeley
Prof. Nelson Morgan, ICSI (Director) and UC Berkeley
Dr. David Nagel, Ascona Group
Prof. Prabhakar Raghavan, Stanford and Yahoo!
Prof. Stuart Russell, UC Berkeley Computer Science Division Chair
Mr. Jouko Salo, TEKES
Prof. Shankar Sastry, UC Berkeley, Dean of the College of Engineering (Chairman)
Prof. Scott Shenker, ICSI and UC Berkeley
Dr. David Tennenhouse, New Venture Partners
Prof. Wolfgang Wahlster, DFKI GmbH

2007 Visitors in Sponsored International Programs

Name	Country	Group	Affiliation
Joan Isaac Biel	EU	Speech	AMIDA
Marijn Huijbregts	EU	Speech	AMIDA
Beatriz Trueba	EU	Speech	AMIDA
Oriol Vinyals	EU	Speech	AMIDA
Ari-Veikko Anttiroiko	Finland		TEKES
Jyri Kivinen	Finland	Algorithms	TEKES
Teemu Koponen	Finland	Algorithms	TEKES
Erno Lindfors	Finland	Algorithms	TEKES
Janne Lindquist	Finland	Networking	TEKES
Ville-Pekka Seppä	Finland		TEKES
Pekka Valkama	Finland		TEKES
Antti Vehkaoja	Finland		TEKES
Mari Zakrzewski	Finland		TEKES
Gerald Friedland	Germany	Speech	DAAD
Martin Gairing	Germany	Speech	DAAD
Martin Hilpert	Germany	Speech	DAAD
Tobias Keisling	Germany	Networking	DAAD
Thomas Kleinbauer	Germany	AI	DAAD
Christian Kreibich	Germany	Networking	DAAD
Birte Lönneker-Rodman	Germany	AI	DAAD
Andreas Maletti	Germany	AI	DAAD
Christian Müller	Germany	Speech	DAAD
Alberto Amengual	Spain	AI	MEC
Santiago Caballe	Spain	AI	MEC
Lucia Conde	Spain	Algorithms	MEC
Alberto Suarez	Spain	Algorithms	MEC
Carlos Subirats	Spain	AI	MEC
Sebastien Cuendet	Switzerland	Speech	IM2
Neha Garg	Switzerland	Speech	IM2
Kamand Kamangar	Switzerland	Speech	IM2
Mathew Magimai-Doss	Switzerland	Speech	IM2
José Millan	Switzerland	Speech	IM2
Petr Motlicek	Switzerland	Speech	IM2
Adish Singla	Switzerland	Speech	IM2

DAAD: Deutscher Akademischer Austausch Dienst TEKES: Finnish National Technology Agency
MEC: Ministerio de Educación y Ciencia AMIDA: Augmented Multi-party Interaction with Distance Access
IM2: Interactive Multimodal Information Management, National Centre of Competence in Research, Switzerland

Contents

Part I
INSTITUTE OVERVIEW

The International Computer Science Institute (ICSI) is one of the few independent, non-profit basic research institutes in the country, and is affiliated with the University of California campus in Berkeley, California. ICSI was started in 1986 and inaugurated in 1988 as a joint project of the Electrical Engineering and Computer Science Department (and particularly of the Computer Science Division) of UC Berkeley and the GMD, the Research Center for Information Technology GmbH in Germany. Since then, Institute collaborations within the university have broadened (for instance, with the Electrical Engineering Division, as well as other departments such as Linguistics). In addition, Institute support has expanded to include a range of international collaborations, US Federal grants, and direct industrial sponsorship. Throughout these changes, the Institute has maintained its commitment to a pre-competitive research program. The goal of the Institute continues to be the creation of synergy between world-leading researchers in computer science and engineering. This goal is best achieved by creating an open, international environment for both academic and industrial researchers.

The particular areas of concentration have varied over time, but are always chosen for their fundamental importance and their compatibility with the strengths of the Institute and affiliated UC Berkeley faculty. ICSI currently has a major focus on two areas: Internet Research, including Internet architecture, related theoretical questions, and network security; and Human Language Technology, including both speech and text processing. Additionally, there are efforts in theoretical computer science and algorithms for bioinformatics, and a local diversity project called the Berkeley Foundation for Opportunities in Information Technology (BFOIT). Finally, as we head into 2008, we are starting a group working on computer vision, and are returning to an earlier focus on the design and realization of computational systems, including efforts in microarchitecture.

The Institute occupies a 28,000 square foot research facility at 1947 Center Street, just off the central UC campus in downtown Berkeley. Administrative staff provide support for researchers: housing, visas, computational requirements, grants administration, etc. There are approximately one hundred scientists in residence at ICSI including permanent staff, postdoctoral Fellows, visitors, affiliated faculty, and students. Senior investigators are listed at the end of this overview, along with their current interests. The current Director of the Institute is Professor Nelson Morgan of the UC Berkeley Electrical Engineering faculty.

1 Institute Sponsorship for 2007

As noted earlier, ICSI is sponsored by a range of US Federal, international, and industrial sources. The figure below gives the relative distribution of funding among these different sponsoring mechanisms.

US Federal funding in 2007 came from a range of grants to support research Institute-wide. Most of this funding comes from the National Science Foundation, DARPA and the Disruptive Technology Office (DTO, now IARPA). International support in 2007 came from

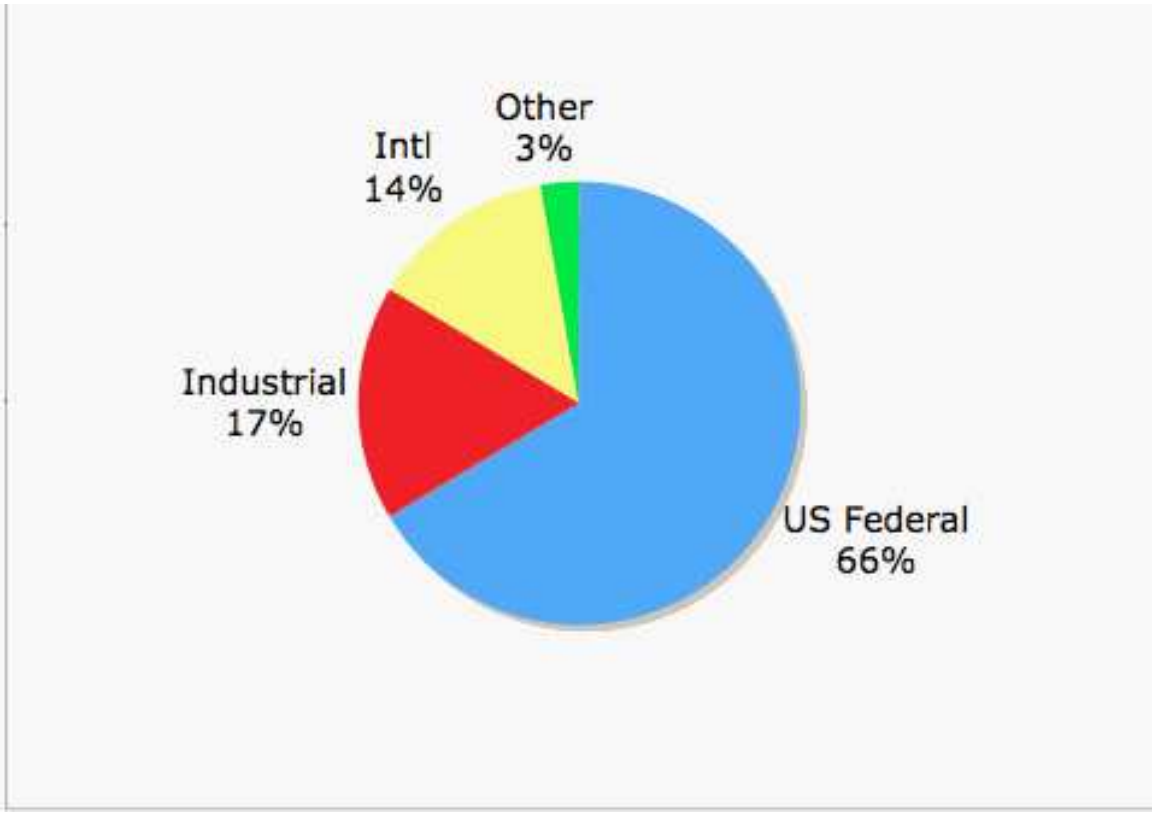

Figure 1: Distribution of sources of ICSI revenue for 2007.

government and industrial programs in Germany, the Ministry of Education and Science in Spain, the National Technology Agency of Finland, and the Swiss National Science Foundation (through the Swiss Research Network IM2). Additional support came from the European Union (as a partner in the Integrated Project, AMIDA). Industrial support in 2007 was provided by Qualcomm, Ask, Google, Cisco, SAP, AT&T, Intel, Volkswagen, AppScio, and Vyatta. Total ICSI revenue was $8.2M in 2007.

2 Institutional Structure of ICSI

ICSI is a nonprofit California corporation with an organizational structure and bylaws consistent with that classification and with the institutional goals described in this document. In the following sections we describe the two major components of the Institute's structure: the Administrative and Research organizations.

2.1 Management and Administration

The corporate responsibility for ICSI is ultimately vested in the person of the Board of Trustees, listed in the first part of this document. The current Chair of that organization is Professor Shankar Sastry of the EECS Department. Ongoing operation of the Institute is the responsibility of Corporation Officers, namely the President, Vice President, and

Secretary-Treasurer. The President also serves as the Director of the Institute, and as such, takes responsibility for day-to-day Institute operations.

Internal support functions are provided by three departments: Computer Systems, Finance, and Administrative Operations. Computer Systems provides support for the ICSI computational infrastructure, and is led by the Systems Manager. Finance is responsible for payroll, grants administration, benefits, human resources, and generally all Institute financial matters; it is led by the Controller. All other support activities come under the general heading of Administrative Operations, and are supervised by the Operations Manager; these activities include office assignments, housing, visas, grant proposal administration, and support functions for ongoing operations and special events.

2.2 Research

Research at ICSI is overwhelmingly investigator-driven, and themes change over time as they would in an academic department. Consequently, the interests of the senior research staff are a more reliable guide to future research directions than any particular structural formalism. Nonetheless, ICSI research has been organized into Groups: the Networking Group (internet research), the Algorithms Group, the AI Group, and the Speech Group. Consistent with this organization, the bulk of this report is organized along these lines, with one sub-report for each of the four groups. New efforts in computer architecture and computer vision are beginning in earnest in 2008.

Across these activities, there is a theme: scientific studies based on the growing ubiquity of connected computational devices. In the case of Networking studies, the focus is on the Internet; in the case of Speech and AI, it is on the interfaces to the distributed computational devices. The Algorithms group continues to develop methods that are employed in a range of computational problems, but recently has focused on problems in computational biology.

Senior Research Staff: The previous paragraphs briefly described the clustering of ICSI research into major research themes and working groups. Future work could be extended to new major areas based on strategic Institutional decisions and on the availability of funding to support the development of the necessary infrastructure. At any given time, though, ICSI research is best seen as a set of topics that are consistent with the interests of the Research Staff. In this section, we give the names of the current (April 2008) senior research staff members at ICSI, along with a brief description of their current interests and the Research Group that the researcher is most closely associated with. This is probably the best snapshot of research directions for potential visitors or collaborators. Not shown here are the postdoctoral Fellows, visitors, and graduate students who are also key contributors to the intellectual environment at ICSI.

Mark Allman (Networking): congestion control, network measurement, network dynamics, transport protocols and network security;

Krste Asanovic (Architecture): computer architecture, parallel programming, VLSI design, new device technologies for computing systems;

Collin Baker (AI): developing semantic frames for a large portion of the common English lexicon, and studying the extent to which these frames are applicable to other lan-

guages, - including Spanish, German and Japanese - all of which have ongoing FrameNet-related projects. Also investigating the extent to which a currently manual semantic annotation process can be automated and accelerated, using automatic semantic role labeling and computer-assisted frame discovery;

Jerome Feldman (AI): neural plausible (connectionist) models of language, perception and learning and their applications;

Charles Fillmore (AI): building a lexical database for English (and the basis for multilingual expansion) which records facts about semantic and syntactic combinatorial possibilities for lexical items, capable of functioning in various applications: word sense disambiguation, computer-assisted translation, information extraction, etc.;

Sally Floyd (Networking): congestion control, transport protocols, queue management, and network simulation;

Dilek Hakkani-Tur (Speech): spoken language understanding, spoken dialog systems, active and unsupervised learning for spoken language processing;

Eran Halperin (Algorithms): computational biology, computational aspects of population genetics, combinatorial optimization, algorithm design;

Adam Janin (Speech): statistical machine learning, particularly for speech recognition, speaker recognition, and language understanding; use of higher level information (e.g., semantics) in speech recognition;

Richard Karp (Algorithms and Networking): mathematics of computer networking, computational molecular biology, computational complexity, combinatorial optimization;

Paul Kay (AI): analyzing the data from the World Color Survey, which gathered color naming data in situ from 25 speakers each of 110 unwritten languages from 45 distinct language families, in order to (1) assess whether cross-language statistical universals in color naming can be observed and (2) measure the degree to which the boundaries of color categories in individual languages can be predicted from universal focal colors;

Nelson Morgan (Speech): signal processing and pattern recognition, particularly for speech classification tasks;

Nikki Mirghafori (Speech): speech processing, particularly speech and speaker recognition;

John Moody (Algorithms): machine learning, multi-agent systems, statistical computing, time series analysis, computational finance;

Srini Narayanan (AI): probabilistic models of language interpretation, graphical models of linguistic aspect, graphical models of stochastic grammars, semantics of linguistic aspect, on-line metaphor interpretation, and embodied rationality; more recently models of the role of sub-cortical structures (like basal ganglia-cortex loops) in attentional control;

Vern Paxson (Networking): intrusion detection; internet measurement; measurement infrastructure; packet dynamics; self-similarity;

Scott Shenker (Networking): congestion control, internet topology, game theory and mechanism design, scalable content distribution architectures, and quality of service;

Elizabeth Shriberg (Speech): modeling spontaneous conversation, disfluencies and repair, prosody modeling, dialog modeling, automatic speech recognition, utterance and topic segmentation, psycholinguistics, computational psycholinguistics (also with SRI International);

Andreas Stolcke (Speech): probabilistic methods for modeling and learning natural languages, in particular in connection with automatic speech recognition and understanding (also with SRI International);

Nicholas Weaver (Networking): worms and related malcode; automatic intrusion detection and response; hardware accelerated network processing;

Part II
Research Group Reports

1 Research Group Highlights

The following are a selection of key achievements in our research groups for the year 2007, both in group development and in research per se. Although not a complete listing and, by necessity, quite varied given the different approaches and topics of each group, it should nonetheless give the flavor of the efforts in the ICSI community for the last year. Not listed are efforts in computer architecture and computer vision (which are starting in earnest in 2008), or the continuing community effort that is the Berkeley Foundation for Opportunities in Information Technology (BFOIT),which assists underrepresented students in computer science and engineering in getting college acceptances and scholarship dollars.

1.1 Networking

- **New Awards:** Three members of the networking group have received significant awards during this past year. Sally Floyd was awarded the ACM SIGCOMM Award, which has been awarded annually since 1989 to an individual for lifetime achievement in and contributions to the field of data communications. Vern Paxson was awarded the 2008 Grace Murray Hopper Award; this award honors the outstanding young computer professional of the year, selected on the basis of a single recent major technical or service contribution. Scott Shenker was awarded an Honorary Doctorate of Science from The University of Chicago.

- The founding of the Finland-ICSI Center for Novel Internet Architectures (FICNIA) was celebrated on September 27th, 2007. The mission of the Center is to conduct fundamental research in novel Internet architectures, aiming at a significant contribution towards the future development of the Internet. The Center will combine and coordinate the research efforts at ICSI, led by Professor Scott Shenker, with those of HIIT and other Finnish reseachers, led by Professor Kimmo Raatikainen. It will operate through joint research activities, research visits, and events.

- High-performance hardware architectures for detecting network attacks: we have pursued designs and implementations for three different parallel architectures for intrusion detection, one based on constructing a cluster of PCs, one on using a

custom FPGA front-end unit, and one on exploiting the burgeoning area of multicore processors.

- Modern Internet systems often combine different applications (e.g., DNS, web, and database), span different administrative domains, and function in the context of network mechanisms like tunnels, VPNs, NATs, and overlays. Diagnosing these complex systems is a daunting challenge. Although many diagnostic tools exist, they are typically designed for a specific layer (e.g., traceroute) or application, and there is currently no tool for reconstructing a comprehensive view of service behavior. This past year researchers at ICSI and UCB developed X-Trace, a tracing framework that provides such a comprehensive view. It does so by attaching metadata to requests and then using this metadata to reconstruct the paths traversed by those requests, even across multiple network layers.

1.2 Algorithms

- We have developed an efficient computational method, based on dynamic programming and importance sampling, for avoiding the confounding effects of population structure on measuring the significance of SNP-based association tests.

- The genomes of individuals of mixed ancestry consist of a mosaic of short segments that have originated from the different ancestral populations. We have devised methods of inferring the structure of an individual's mosaic through an analysis of single-nucleotide polymorphism.

1.3 Artificial Intelligence

- Graduate Student Joseph Makin and Srini Narayanan completed the first computational simulation of the entire mammalian coagulation pathway. The model uses techniques from hybrid state control theory and continues to attract considerable interest from clinicians and from the pharmaceutical industry.

- In 2007, Frame analysis was adopted as part of a new community wide evaluation task, Semeval, which signals a move in computational linguistics from word sense disambiguation to richer text level semantics. FrameNet members were instrumental in designing, organizing, and implementing the frame and semantic role recognition tasks within Semeval. FrameNet annotations were used as gold standard texts for training and evaluation. The first Semeval evaluation workshop was conducted during the ACL conference in Prague.

- Srini Narayanan won a Google research award to investigate NLP techniques for multilingual information access in developing countries.

- The FrameNet group won four NSF awards in 2007, for research on (1) creating annotation on the American National Corpus, (2) aligning WordNet and FrameNet, (3) creating extended paraphrases using the FrameNet data, and (4) grammatical constructions on which standard parsers fail.

1.4 Speech

- Group members have extended representation of sentences for distillation to syntactic and semantic graphs, investigated blind-feedback methods and cross-lingual information distillation, all of which resulted in significant (over 5% absolute for overall F-measure) improvements in our distillation performance.

- The Speech Group diarization system, which continues to score extremely well in competitive evaluations, has now been sped up by roughly a factor of four using a combination of fast-match methods and the new ICSI implementation of a fast logarithm.

- For the first time, we were able to achieve significant improvements in the diarization of meetings by detecting speech segments in which the speakers overlapped.

- ICSI developed and evaluated its first language recognition system.

- ICSI speech researchers worked with a group of others on campus to successfully win a very competitive award from Microsoft and Intel to fund a Parallel Lab, which will include the ICSI application area of speech recognition.

2 Networking

2.1 Measurements and Modeling

Community use of network traces: In this effort we endeavor to frame for the community issues surrounding how to provide and use network measurement data made available for sharing among researchers. While previous work has focused on the technical details of enabling sharing via traffic anonymization, our concern here is on higher-level aspects of the process, such as potential harm to the provider (e.g., by de-anonymizing a shared dataset) or interactions to strengthen subsequent research (e.g., helping to establish ground truth). We believe the community would benefit from a dialog regarding expectations and responsibilities of data providers, and the etiquette involved with using other's measurement data. To this end, we have developed a set of guidelines that aim to aid the process of sharing measurement data. We consider these not as specific rules, but rather a framework under which providers and users can better attain a mutual understanding about how to treat particular datasets [8].

Reactive measurement: Reactive measurement (REM) is a measurement technique in which one measurement's results are used to decide what (if any) additional measurements are required to further understand some observed phenomenon. While reactive measurement has been used on occasion in measurement studies, what has been lacking is *(i)* an examination of its general power, and *(ii)* a generic framework for facilitating fluid use of this approach. We believe that by enabling the coupling of disparate measurement tools, REM holds great promise for assisting researchers and operators in determining the root causes of network problems and enabling measurement targeted for specific conditions. We are currently exploring reactive measurement using two approaches. First, we have a prototype REM system that a researcher can use to connect measurement tools together and we have been using this to explore HTTP failures. In addition, we have built a plug-in for the Firefox web browser that takes ancillary measurements when an error in loading a web page is found. These measurements are then sent back to a central database for analysis. While these efforts are in fairly early stages, the overall architecture is more mature [10].

A longitudinal study of scanning: The literature available for analyzing network activity—much less the specialized form of activity due to network *attacks*—is very sparse in terms of long-term, longitudinal studies examining the evolution of activity over time. In this project, we analyze 20+ billion fine-grained TCP connection records recorded over a 12-year period at the Lawrence Berkeley Laboratory to assess how the activity of "scanning" (i.e., attackers probing for vulnerabilities) evolves over more than a decade. Incessant scanning of hosts by attackers looking for vulnerable servers has become a fact of Internet life. We have studied the onset of scanning in the late 1990s and its evolution in terms of characteristics such as the number of scanners, targets and probing patterns. While our study is preliminary at this point, it has provided the first longitudinal examination of a now ubiquitous Internet phenomenon [11].

An open measurement platform: In this project we pursue the development of a measurement framework that researchers can use to abstract away the mundane logistical details that tend to dog every measurement project [6]. The measurement community has outlined the need for better ways to gather assessments from a multitude of vantage points, and our system is designed to be an *open community-oriented* response to this desire. While many previous efforts have approached this problem with heavyweight systems that ultimately fizzle due to logistical issues (e.g., hosts breaking and no money to replace them), we take the opposite approach and attempt to use the lightest-weight possible framework that allows researchers to get their work done. In particular, we take the approach of designing a system without any sort of central "core" component, and therefore the system has no single point of failure. In addition, our proposed system is *community-oriented* in that there is no central control; we build just enough mechanism for the community to get their work done and police the infrastructure. In addition, our proposed system works in an open fashion, such that results from the community's infrastructure are immediately provided to the community through publicly available "live feeds".

Measuring OpenDHT: A number of applications have proposed using distributed hash tables (DHTs) for a variety of database tasks (e.g., new naming schemes or aggregating RSS feeds). DHTs present logistical challenges to set up and deploy. Therefore, OpenDHT was developed as a general-purpose DHT service that provides a simple *get()/put()* interface to developers and is advertised as useful across many applications. While in principle various proposed databases fit nicely within the DHT abstraction, and OpenDHT is attractive for its ease of use, each application will ultimately have its own performance requirements—which the DHT system may or may not be able to meet. Therefore, in practice the suitability of using a DHT data structure may not be readily apparent. In this project we have captured an initial set of simple measurements to assess the performance of the OpenDHT service. In addition, we are assessing the reliability and responsiveness for various operations from an application's viewpoint.

2.2 Security, Malware, and Intrusion Detection

Exploiting multi-core processors to parallelize network intrusion prevention: It is becoming increasingly difficult to implement effective systems for preventing network attacks, due to the combination of (1) the rising sophistication of attacks requiring more complex analysis to detect, (2) the relentless growth in the volume of network traffic that we must analyze, and, critically, (3) the failure in recent years for uniprocessor performance to sustain the exponential gains that for so many years CPUs enjoyed ("Moore's Law"). For commodity hardware, tomorrow's performance gains will instead come from *multicore* architectures in which a whole set of CPUs executes concurrently.

Taking advantage of the full power of multi-core processors for network intrusion prevention requires an in-depth approach. In this project we work towards developing an architecture customized for parallel execution of network attack analysis. At the lowest layer of the architecture is an "Active Network Interface" (ANI), a custom device based on an inexpensive FPGA platform. The ANI provides the in-line interface to the network, reading in packets and forwarding them after they are approved. It also serves as the

front-end for dispatching copies of the packets to a set of analysis threads. The analysis itself is structured as an event-based system, which allows us to find many opportunities for concurrent execution, since events introduce a natural, decoupled asynchrony into the flow of analysis while still maintaining good cache locality. Finally, by associating events with the packets that ultimately stimulated them, we can determine when all analysis for a given packet has completed, and thus that it is safe to forward the pending packet—providing none of the analysis elements previously signaled that the packet should instead be discarded [41].

FPGA-based acceleration of network intrusion detection/prevention: As discussed above, with ever-increasing network speeds and traffic volumes, there is a growing need to support network intrusion detection using custom hardware. Yet designing such hardware in a fashion that is both robust and sufficiently flexible takes great care. The *Shunting* project explores developing an architecture that combines the power of high speed network elements with the flexibility of highly programmable network intrusion detection systems (NIDSs). The core of the architecture is a network forwarding element (the "Shunt") that works in conjunction with a NIDS by diverting a subset of the traffic stream through the NIDS. Because the NIDS receives the actual traffic itself rather than a copy, the architecture enables the NIDS to instantly block attack traffic (i.e., "intrusion prevention").

The key insight leveraged by the architecture is that in many environments, the vast majority of the high-volume traffic is confined to a small fraction of the connections. Furthermore, these high-volume connections are generally of little interest from an intrusion-detection perspective *after* they have been initially established. That is, it is important to analyze the connections' surrounding context (control session, initial authentication dialog, concurrent logins, etc.), but, once established, the connections themselves can be safely skipped. The core of the hardware support is based on extending the usual packet-filter model with lookup tables. Such tables can be indexed at a variety of granularities: individual connections (source address/source port/destination address/destination port tuples); specific source or destination addresses or pairs; and source/destination prefixes. The key is allocating sufficient memory in the network element so that these tables (particularly the first, per-connection) can be large. The element then looks up incoming packets in the given tables to find if they match flavors of traffic specified by the tables. If so, the element executes the action associated with the table element, where the actions are one of: pass-through, shunt, or drop. Such tables allow the NIDS to communicate fine-grained go/no-go/inspect decisions to the network element in a concise manner: it simply sends over new table entries and their associated actions as it makes decisions concerning whether a given connection, pair of hosts, source, or destination is deemed trustworthy, malicious, or undecided.

To date we have implemented a prototype Shunt hardware design using the NetFPGA 2 platform, capable of Gigabit Ethernet operation. We have also adapted the Bro intrusion detection system to utilize the Shunt framework to offload less-interesting traffic. We evaluated the effectiveness of the resulting system using traces from three sites, finding that the IDS can use this mechanism to offload 55%–90% of the traffic, as well as gaining

intrusion prevention functionality [48, 31].

Bro cluster: While the task of parallelizing network intrusion analysis might at first blush seem fairly simple—split the traffic among multiple CPUs on a per-connection basis, and we're done—in reality, such a division becomes significantly more subtle when we must consider higher-level analyses that require coordination of information *across* connections or hosts. This project has developed a *clusterizable* version of the Bro intrusion detection system, with a focus on an approach whereby if one can dedicate N commodity PCs to the task of executing Bro, then the execution of Bro will be approximately N times more efficient.

To date, this effort has developed prototypes successfully operating at the Lawrence Berkeley National Laboratory and the University of California at Berkeley [46], and evaluated the implementation using stress-testing [49]. The next step to pursue is a more efficient architecture for distributing events between the multiple nodes than the current mesh/broadcast model, which burdens large clusters with excessive communication overhead.

Investigating the underground economy: One of the most disturbing recent shifts in Internet attacks has been the change from attackers motivated by glory or vanity to attackers motivated by commercial (criminal) gain. This shift threatens to greatly accelerate the "arms race" between defenders developing effective counters to attacks and attackers finding ways to circumvent these innovations. A major driving force behind the shift to criminalized malware has been the development of *marketplaces* that criminals use to foster a specialized economy of buyers and sellers of specialized products and services. This project, joint with UC San Diego, aims to explore these marketplaces in an attempt to characterize their constituencies, impact, and sundry elements, in the hope that such an analysis might shed light on bottlenecks/weakspots present in the underground economy that can then be targeted to provide maximal benefit for defenders [30].

Visibility into network activity across space and time: The premise of this project is that for key operational networking tasks—in particular troubleshooting and defending against attacks—there is great utility in attaining views of network activity that are *unified across time and space*. By this we mean that procedures applied to analyzing past activity match those applied for detecting future instances, and that these procedures can seamlessly incorporate data acquired from a wide range of devices and systems. To this end, we have pursued development of *VAST* (Visibility Across Space and Time), a system that can process network activity logs comprehensively, coherently, and collaboratively [5]. The VAST system archives data from a multitude of sources and provides a query interface that can answer questions about what happened in the past, as well as notifying operators when certain activity occurs in the future. Its policy-neutral structure allows a site to specify custom procedures to coalesce, age, sanitize, and delete data.

In addition, the VAST system can facilitate operationally viable, cross-institutional information sharing. In contrast to today's inefficient and cumbersome operational practices—phone calls, emails, manual coordination via IM—we envision a framework that enables

operators to leverage each others' VAST systems. To address the important trust and privacy constraints of a such a setting, we in addition introduce the notion of a per-site *Clearing House* component that provides operators with fine-grained control over the flow of information, enabling them to deploy the full spectrum from automated sending and receiving of descriptions of activity, to holding all requests for explicit, manual approval.

Building a large-scale honeyfarm: A key tool for detecting new worm outbreaks in their early stages is the *honeyfarm*, a large collection of honeypots fed Internet traffic by a "network telescope". However, actual operation of a honeyfarm in a large-scale environment presents difficult scaling challenges. We have designed and implemented *GQ*, a honeyfarm system capable of analyzing in real-time the scanning probes seen on a quarter million Internet addresses. GQ's architecture emphasizes high fidelity, scalability, isolation, stringent control, and wide coverage. In operational use, GQ has captured 66 distinct types of worms over the course of four months. Current efforts focus on improving the system's ability to prefilter the incoming stream of Internet probes to more effectively find probes of likely interest, and on increasing the performance of the VM system to more efficiently support VM instantiation.

Opportunistic personas: Cryptographic security mechanisms often assume that keys or certificates are strongly tied to a party's identity. This requirement can in practice impose a high bar on making effective use of the cryptographic protections, because securing the coupling between credentials and actual identity can prove to be an arduous process. In this project we explore a more relaxed form of identity, termed *opportunistic personas*, that works by *(i)* generating cryptographic credentials on an as-needed basis, *(ii)* associating credentials not with a user per se but instead as a link to past behavior by the same actor, as a means to inform future interactions, and *(iii)* managing these credentials *implicitly* in an opportunistic fashion [4].

Internet situational awareness: Effective network security administration depends to a great extent on having accurate, concise, high-quality information about malicious activity in one's network. "Honeynets"—collections of sacrificial hosts ("honeypots") fed traffic seen on an unused region of a network—can potentially provide such detailed information, but the volume and diversity of this data can prove overwhelming. In this project we explore ways to analyze the probes seen by honeynet data in order to assess whether a given "event" present in the honeynet reflects the onset of a new Internet worm, a benign misconfiguration, or a concerted effort to scan the site. For this latter (the most common), we then attempt to refine the analysis to assess whether the scanning *targeted* the site in particular, or was merely part of a much broader, indiscriminate scan.

Building a time machine: Insight into past network traffic can have enormous value, both for forensics when analyzing a problem detected belatedly, and to augment real-time decision-making, both to inform *reactive measurement* (see above) and to give additional pinpoint context to a network intrusion detection system (NIDS). This project aims to develop a network *time machine*, which works by passively bulk-recording as much network

traffic as possible. The time machine maintains a ring buffer of recent network traffic that matches a given criteria. This criteria needn't be a simple static filter—the decision of what to capture and for how long could be much richer and incorporate more context. This buffer resides in RAM for fast access, with the decision of what traffic to record in the buffer, and how to filter it (e.g., retaining the first N bytes of each connection), being driven off of a collection of policies describing retention for different types of activity. In addition, recorded traffic migrates from RAM to a given allocation of disk space, which is also managed per a collection of policies that again determine which traffic to migrate, how to filter it, and how to expire it as the disk allocation fills up.

Recent efforts on this project have focused on integrating our Time Machine implementation with a real-time NIDS by providing an API by which a NIDS (in our case, the Bro system) can query activity seen in the recent past for given connections or hosts. This coupling has the potential to greatly offload the NIDS, allowing it to process only lighter-weight request streams and not response streams, unless it sees a problematic request, in which case it can at that point ask the time machine for a copy of the reply to that particular request.

Robust TCP stream normalization: One of our previous efforts investigated algorithms by which hardware devices can reassemble TCP bytestreams even in the presence of adversaries who will attempt to subvert the hardware's operation by overwhelming its state management. This follow-on project looks at the next step: how to *normalize* the byte stream in order to assure that we can remove any ambiguities in terms of inconsistent TCP retransmissions. Hardware vendors have asserted that simply hashing the contents of previously seen packets suffices to provide such normalization. We have found that by itself, this approach renders a great deal of "collateral damage" in terms of retransmitted traffic that must be discarded because it does not align with previously recorded hashes, or in *evasion* opportunities, if such traffic is allowed to proceed in the absence of a previous hash against which we can check it. The approach we have developed in this regard, however, is robust to such variations, as well as to attackers who deliberately target the state we must manage to provide such normalization [47].

Rate-based scan detection: This project explores developing light-weight worm detection algorithms that offer significant advantages over fixed-threshold methods. The first algorithm, RBS (rate-based sequential hypothesis testing), aimed at the large class of worms that attempt to quickly propagate, thus exhibiting abnormal levels of the rate at which hosts initiate connections to new destinations. The foundation of RBS derives from the theory of sequential hypothesis testing, the use of which for detecting randomly scanning hosts was first introduced by our previous work with the *Threshold Random Walk* algorithm [33]. The sequential hypothesis testing methodology enables engineering the detectors to meet false positives and false negatives targets, rather than triggering when fixed thresholds are crossed. In this sense, the detectors that we introduce are truly adaptive.

We then developed RBS+TRW, an algorithm that combines fan-out rate (RBS) and probability of failure (TRW) of connections to new destinations. RBS+TRW provides a unified framework that at one end acts as a pure RBS and at the other end as pure TRW,

and extends RBS's power in detecting worms that scan randomly selected IP addresses. Using four traces from three qualitatively different sites, we evaluated RBS and RBS+TRW in terms of false positives, false negatives, and detection speed, finding that RBS+TRW provides good detection of high-profile worms, internal Web crawlers, and a network monitoring tool that we used as proxies for targeting worms. In doing so, RBS+TRW generates fewer than 1 false alarm per hour for wide range of parameter choices [35, 36].

Detecting hidden changes introduced to Web pages: While Web pages sent over HTTP have no integrity guarantees, it is commonly assumed that such pages are not modified in transit. In this project we investigate this question, finding evidence of surprisingly widespread and diverse changes made to web pages between the server and client. Over 1% of web clients receive altered pages, often changes with undesirable consequences for web publishers or end users. Such changes include popup blocking scripts inserted by client software, advertisements injected by ISPs, and malicious code inserted by malware using ARP poisoning. Additionally, we found that even changes introduced by client software can inadvertently cause harm, such as introducing cross-site scripting vulnerabilities into every page a client visits. To help publishers understand and react appropriately to such changes, we develop the concept of *Web tripwires*–client-side JavaScript code that can detect in-flight modifications to a web page. We have investigated several Web tripwire designs and show that they are more flexible and less expensive than switching to HTTPS, without requiring changes to current browsers [42].

Predicting resource consumption of network intrusion detection systems: When installing network intrusion detection systems (NIDSs), operators are faced with a large number of parameters and analysis options for tuning trade-offs between detection accuracy versus resource requirements. In this effort we set out to assist this process by understanding and predicting the CPU and memory consumption of such systems. We started towards this goal by devising a general NIDS resource model to capture the ways in which CPU and memory usage scale with changes in network traffic. We then used this model to predict the resource demands of different configurations for specific environments, leading to an approach to derive site-specific NIDS configurations that maximize the depth of analysis given predefined resource constraints.

We have validated our approach by applying it to the open-source Bro NIDS, testing the methodology using real network data, and developing a corresponding tool that automatically derives a set of configurations suitable for a given environment based on a *sample* of the site's traffic. While no automatically generated configuration can ever be optimal, these configurations provide sound starting points, with promise to significantly reduce the traditional trial-and-error NIDS installation cycle.

Testing evasion resilience of network intrusion detection systems: Network intrusion detection systems (NIDS) face a difficult, fundamental problem in the degree to which attackers can exploit ambiguities present when monitoring network traffic in order to undermine the correctness of the NIDS's analysis to evade detection. However, many of today's NIDSs lack the additional mechanisms required to resist different forms of evasion,

because the underlying problems are subtle and—critically—*not visible* to the customers who purchase these systems.

Remedying this common shortcoming of modern NIDS functionality requires the widespread availability of *test suites* oriented towards probing the degree to which a NIDS exhibits evasion vulnerabilities. In this project we undertake the creation of a framework to facilitate the development of such test suites. Our prototype system takes as input a packet trace and from it constructs a configurable set of variant traces that introduce different forms of ambiguities that can lead to evasions. Our test harness then uses these variant traces in either an *offline* configuration, in which the NIDS under test reads traffic from the traces directly, or a *live* setup, in which we employ replay technology to feed traffic over a physical network past a NIDS reading directly from a network interface, and to potentially live victim machines. Summary reports of the differences in NIDS output tell the analyst to what degree the NIDS's results vary, reflecting sensitivities to (and possible detections of) different evasions. We have used it to test the open-source *Snort* and *Bro* systems [34].

Packet symmetry: Distributed denial-of-service attacks remain one of the biggest challenges the Internet is facing today. The vast majority of countermeasures proposed to date focus on the victim, such as enabling "pushback" of filters that are intended to drop traffic from the many attackers. In this project, we take the view that in a "well-tempered" Internet no actor should need to send vastly more traffic than it receives. We explore the notion of *packet symmetry* as a metric that steers adaptive throttling mechanisms at the sender side, based on the ratio of the numbers of packets sent vs. received over a given period of time. Using a combination of trace-based analysis and live deployment, we are investigating the feasibility of the idea and its resilience to attackers having large numbers of clients at their disposal.

Measuring the Storm botnet: Malware authors have discovered that their creations can not only be used for destructive purposes, but can also generate money: by organizing large numbers of subverted machines into so-called *botnets*, these Internet miscreants are able to extort money via denial-of-service attacks, undertake identity theft or mass marketing by sending phishing and other unsolicited email, exploit the growing Internet advertising business through fraudulent ad-clicking activity, and heighten the resilience of malware-serving infrastructures through increasing levels of distribution and indirection.

In 2007, the "Storm" botnet rose to prominence as one of the biggest and technically advanced botnets ever created. While press articles were quick to report ever-larger numbers of infected hosts (reaching well into the hundreds of thousands of infected machines), in reality few scientifically sound measurement techniques are available to confirm such reports. We are actively engaged in analysis and infiltration of the Storm botnet, to improve our understanding of the scope of the threat as well as the technical details of the malicious activities such large-scale botnets undertake, and—in the longer term—to enable network operators to prevent, identify, and eradicate future botnet activity on their networks.

2.3 Internet Protocols

Reacting to spurious retransmissions: TCP and SCTP both provide reliability by retransmitting lost data. In addition, losses are taken as an indication of network congestion and used to trigger congestion control in the data sender (i.e., a reduction in the sending rate). Spurious retransmissions are generally caused by a transport's lack of ability to cope with the dynamics of a network path (e.g., a widely varying round-trip time). Several schemes have been devised to detect spurious retransmissions. In this effort, we investigate a response to spurious timeouts whereby we alter the calculation of the retransmission timer in an attempt to account for more variation in the round-trip time and prevent further spurious retransmissions [15].

Quick-Start: A fundamental aspect of communication in general-purpose, best-effort, packet-switched networks is determining an appropriate sending rate. The appropriate sending rate depends on the characteristics of the network path between the two peers (such as bandwidth and propagation delay), as well as the amount of load placed on the network by others at the given time. Traditionally, TCP has used a set of congestion control algorithms for determining this rate. The problem addressed by the *Quick-Start* project is how a particular connection on an under-utilized network path can increase its sending rate to take advantage of the available capacity more rapidly than allowed by TCP's traditional congestion control algorithms. Quick-Start is a proposed mechanism for end nodes to request permission from routers along the path to use a higher sending rate. We note that Quick-Start is not in fact a congestion control mechanism, in that it doesn't detect or respond to congestion, and does not replace the traditional congestion control mechanisms of the transport protocol; rather, Quick-Start is an optional mechanism that flows could use to get approval from routers to send at a high sending rate on a significantly-underutilized path. We have both evaluated [43] and specified [27] Quick-Start.

Jump Start: While with Quick-Start we begin from the premise that starting transmission at a high sending rate requires approval from each node along a network path, with the Jump Start project we explore the opposite view: what are the implications if hosts can simply start sending at whatever rate they deem appropriate? Other than this change, traditional congestion control remains in place, so that inappropriate bursts of traffic will not be sustained over lengthy periods of time. Clearly, such an approach is fraught with potential problems. Our initial work in this area involves both developing ways to deal with such problems, and assessing how much of a concern these envisioned problems might be in real networks [40].

Early Retransmit: In this effort we introduce a new mechanism for TCP and SCTP for recovering lost segments when a connection's congestion window is small. The "Early Retransmit" mechanism [2] allows the transport to reduce (in certain special circumstances) the number of duplicate acknowledgments required to trigger a fast retransmission. This allows the transport to use Fast Retransmit to recover packet losses that would otherwise require a lengthy retransmission timeout.

Congestion control for small-packet flows: TCP-Friendly Rate Control (TFRC) is a congestion control mechanism for unicast flows operating in a best-effort Internet environment. TFRC is intended for applications that use a fixed packet size, and is designed to be reasonably fair when competing for bandwidth with TCP connections using the same packet size. For congestion control for applications that send small packets, we have designed TFRC-SP, a Small-Packet (SP) variant of TFRC [28]. This design involves a fundamental question about whether the congested points in the network are limited by their sending rate in packets per second (e.g., CPU cycles at routers), or by their sending rate in bytes per second (e.g., bandwidth); the development of TFRC-SP is based on the assuption that the limitation today is generally one of bandwidth rather than of CPU cycles. The design goal for TFRC-SP is to achieve the same bandwidth in bytes per second as a TCP flow with 1500-byte packets experiencing the same packet drop rates. TFRC-SP enforces a maximum sending rate, to prevent a single flow from sending small packets arbitrarily frequently.

The usefulness of best-effort traffic: In this effort we develop a number of observations on the capabilities and limitations of "simple best-effort" traffic, defined loosely as Internet traffic that is not covered by Quality of Service mechanisms, congestion-based pricing, cost-based fairness, admissions control, or the like [26]. One core observation is that simple best-effort traffic serves a useful role in the Internet, and is worth keeping. While differential treatment of traffic can clearly be useful, we believe such mechanisms have utility primarily as *adjuncts* to simple best-effort traffic, not as *replacements*. A second observation is that for simple best-effort traffic, some form of rough "flow rate fairness" is a useful goal for resource allocation, by which we mean attaining equal flow rates for different flows over the same path.

Specifying new congestion control algorithms: The IETF's standard congestion control schemes have been widely shown to perform inadequately for various environments (e.g., high-speed or wireless networks). Recent research has yielded many alternate congestion control schemes. However, using these new congestion control schemes in the global Internet has possible ramifications to both the network and to traffic using the currently standardized congestion control. In this effort we developed guidelines for the IETF to use when evaluating suggested congestion control algorithms that significantly differ from the general congestion control principles outlined in current standards [25]. The guidance is intended to be useful to both authors proposing alternate congestion control, and for the IETF community when evaluating whether a proposal is appropriate for publication in the RFC series.

Updating standard TCP congestion control. ICSI researchers have been instrumental in codifying algorithms for TCP congestion control (previously developed by V. Jacobson) as Internet standards [7]. This effort focuses on revising this previous work to clarify issues and ambiguities that have been identified since its publication [9].

2.4 Novel Internet Architectures

Architectural support for network trouble-shooting: Troubleshooting is an inherent part of network operation: no matter how well networks are designed, something eventually fails, and in large networks, failures are ever-present. In the past, troubleshooting has mostly relied on *ad hoc* techniques cobbled together as afterthoughts. However, both the importance and difficulty of troubleshooting has intensified as networks have become crucial, ubiquitous components of modern life, while at the same time their size and complexity continues to grow. These twin pressures highlight the urgent need to integrate troubleshooting as a first-class citizen when developing a network architecture.

This project pursues a key set of building blocks for developing networks that are much more amenable to troubleshooting. *Annotations* provide a means for associating meta-information with network activity. One use of annotations is to *track causality* in terms of how instances of network activity relate to previous activity. We envision much more powerful forms of *logging*, enhanced by notions of *distillation* of logged information into more abstract forms over time, and *dialog* between system components that generate log entries and the logger itself, which can call back to the component to support highly flexible distillation as well as interactive debuggers. Finally, we feed logs from multiple observation points into *repositories* that construct aggregated views of activity and mediate the ways in which sites share information for cooperative trouble-shooting.

Selective connectivity and architecting for energy efficiency: The Internet's architecture largely and implicitly assumes full-time connectivity, a notion that is embodied in key networking principles including fate sharing, soft state, and the end-to-end principle. In contrast, efforts to allow for more graceful operation in the presence of forced disconnectedness have recently been undertaken that change the underlying style of networking used by applications to accommodate both host-level and hop-by-hop disconnectedness (e.g., for deep space networks where connectivity depends on orbital mechanics). In this project we explore architectural constructs to support *selective connectivity*, whereby a host can *choose* whether to be "connected" or "disconnected". While we keep our notion of selective connectivity general, the driver behind our thinking is to allow hosts to go to sleep to realize energy savings while not sacrificing their standing in the network. Studies show that enabling such sleeping offers large potential energy savings. In this context, we explore approaches focusing on assistants, soft state, host-based control, and application primitives [3].

Selective sharing: We envision the blossoming of sensing applications in an urban context, enabled by increasingly affordable and portable sensing hardware, and ubiquitous wireless access to communication infrastructure. In this effort, joint with UC Los Angeles, we pursue development of the *Partisans* architecture, featuring infrastructure-supported selective data sharing and verification services. Sharing such information in a manner that aptly balances utility of the provided data with the necessary privacy and security assurances raises a number of novel issues.

Personal namespaces: This effort explores the introduction of an over-arching namespace that serves to abstract away the Internet's current and obscure naming schemes from *users*. We argue for users to have *personal namespaces* that are not concerned with unique naming of resources, but rather focused on aiding user's interactions with the system. This additional namespace does not replace any of our current (or future) naming systems. Rather, our vision calls for adding a naming layer that provides the ability for users to meaningfully alias network resources (especially their own). These aliases become context-sensitive, provider independent names for objects that can be easily shared among people. We have designed a strawman system—called *pnames*—in high level terms as a starting point in the discussion of how such a system might be built [1].

A new addressing scheme for the Internet: Todays IP network layer provides little to no protection against misconfiguration or malice. Despite some progress in improving the robustness and security of the IP layer, misconfigurations and attacks still occur frequently. In [13] we show how a network layer that provides accountability, i.e., the ability to associate each action with the responsible entity, provides a firm foundation for defenses against misconfiguration and malice. We present the design of a network layer that incorporates accountability called AIP (Accountable Internet Protocol) and show how its features — notably, its use of self-certifying addresses — can improve both source accountability (the ability to trace actions to a particular end host and stop that host from misbehaving) and control-plane accountability (the ability to pinpoint and prevent attacks on routing).

A new communications API: We contend that a new networking API could better serve the needs of data- and service-oriented applications, and could more easily map to heterogeneous environments, than the pervasive Sockets API does. In [20], we present an initial design of a networking API based on the publish/subscribe paradigm, along with an exploration of its security implications, examples to demonstrate several common use cases, and a discussion of how the implementation of such an API could leverage a wide range of networking technologies. We propose this model not as a final design but as the first step towards a wider community discussion of the need for a modern communications API.

A protection architecture for enterprise networks: Ethane [16] is a new architecture for enterprise networks which provides a powerful yet simple management model and strong security guarantees. Ethane allows network managers to define a single, network-wide, fine-grain policy, and then enforces it at every switch. Ethane policy is defined over human-friendly names (such as "bob, "payroll-server", or "http-proxy) and dictates who can talk to who and in which manner. For example, a policy rule may specify that all guest users who have not authenticated can only use HTTP and that all of their traffic must traverse a local web proxy.

Ethane has a number of salient properties difficult to achieve with network technologies today. First, the global security policy is enforced at each switch in a manner that is resistent to spoofing. Second, all packets on an Ethane network can be attributed back to the sending host and the physical location in which the packet entered the network. In

fact, packets collected in the past can also be attributed to the sending host at the time the packets were sent – a feature that can be used to aid in auditing and forensics. Finally, all the functionality within Ethane is provided by very simple hardware switches.

The trick behind the Ethane design is that all complex functionality, including routing, naming, policy declaration and security checks are performed by a central controller (rather than in the switches as is done today). Each flow on the network must first get permission from the controller which verifies that the communicate is permissible by the network policy. If the controller allows a flow, it computes a route for the flow to take, and adds an entry for that flow in each of the switches along the path.

With all complex function subsumed by the controller, switches in Ethane are reduced to managed flow tables whose entries can only be populated by the controller (which it does after each succesful permission check). This allows a very simple design for Ethane switches using only SRAM (no power-hungry TCAMS) and a little bit of logic.

Resolving inter-domain policy disputes: The Border Gateway Protocol (BGP) allows each autonomous system (AS) to select routes to destinations based on semantically-rich and locally-determined policies. This autonomously exercised policy-freedom can cause instability, where unresolvable policy-based disputes in the network result in interdomain route oscillations. Moreover, several recent works have established that such instabilities can only be eliminated by enforcing a globally accepted preference ordering on routes (such as shortest path). To resolve this conflict between policy autonomy and system stability, we propose [22] a distributed mechanism that enforces a preference ordering only when oscillations due to these disputes occur. This preserves policy freedom when possible, and imposes stability when required.

Convergence-free routing: Current distributed routing paradigms (such as link-state, distance-vector, and path-vector) involve a convergence process consisting of an iterative exploration of intermediate routes triggered by certain events such as link failures. The convergence process increases router load, introduces outages and transient loops, and slows reaction to failures. We propose a new routing paradigm where the goal is not to reduce the convergence times but rather to eliminate the convergence process completely. To this end, we propose [39] a technique called Failure-Carrying Packets (FCP) that allows data packets to autonomously discover a working path without requiring completely up-to-date state in routers. Our simulations, performed using real-world failure traces and Rocketfuel topologies, show that: (a) the overhead of FCP is very low, (b) unlike traditional link-state routing (such as OSPF), FCP can provide both low lossrate as well as low control overhead, (c) compared to prior work in backup path precomputations, FCP provides better routing guarantees under failures despite maintaining lesser state at the routers.

Data-oriented network architecture: The Internet has evolved greatly from its original incarnation. For instance, the vast majority of current Internet usage is data retrieval and service access, whereas the architecture was designed around host-to-host applications such as telnet and ftp. Moreover, the original Internet was a purely transparent carrier of packets, but now the various network stakeholders use middleboxes to improve security and

accelerate applications. To adapt to these changes, we propose the Data-Oriented Network Architecture (DONA) [38], which involves a clean-slate redesign of Internet naming and name resolution.

Loss and Delay Accountability The current Internet provides no information on the fate of transmitted packets. As a result, when packets get lost or delayed, there is no clean way for the affected parties to localize the problem and fix it (if it is local), ask for compensation (if a service-level agreement has been violated), or simply learn from it (e.g., re-assess a peering agreement with an under-performing neighbor). Probing tools like traceroute can help localize network failures, however, they draw their conclusions based on the fate of probes, not actual traffic, which makes them susceptible to manipulation by transit networks. Moreover, such probing tools often reveal the internal structure and routing policies of ISPs, giving the latter an incentive to render their networks opaque to probing.

The goal of this project is a way to change this lack of accountability in the Internet: a clean, yet practical solution that tells network entities what they need to know (who is responsible for losing or delaying their packets), but not what they shouldn't (the internal structure and policies of other networks or ISPs). The key idea behind our work is that this information need not (and should not) be extricated by ad-hoc probing tools that treat the Internet as a black box and try to reverse-engineer its structure and failures. Rather, it should be provided by a cooperative, incentive-based framework, where networks provide verifiable information on their own performance and, in exchange, learn how their own traffic is being treated by their neighbors.

To this end, we propose AudIt [14], an explicit accountability interface, through which ISPs can pro-actively supply feedback to traffic sources on loss and delay, at administrative-domain granularity. Notably, our interface is resistant to ISP lies and can be implemented with a modest NetFlow modification. On our Click-based prototype, playback of real traces from a Tier-1 ISP reveals less than 2% bandwidth overhead. Finally, our proposal benefits not only end systems, but also ISPs, who can now control the amount and quality of information revealed about their internals.

Global Environment for Network Innovations (GENI): GENI is a facility for network experimentation being planned by the NSF, in collaboration with the research community. It's goal is to enable the research community to invent and demonstrate a global communications network and related services that will be qualitatively better than today's Internet.

2.5 Distributed Systems

IRIS Project: The Infrastructure for Resilient Internet Services (IRIS) project combines the efforts of 12 PIs from five institutions (ICSI, UCB, MIT, Rice, NYU). The IRIS project is developing a novel decentralized infrastructure, based on distributed hash tables (DHTs), that will enable a new generation of large-scale distributed applications. DHTs are robust in the face of failures, attacks and unexpectedly high loads. They are scalable, achieving large

system sizes without incurring undue overhead. They are self-configuring, automatically incorporating new nodes without manual intervention or oversight. They provide a simple and flexible interface and are simultaneously usable by many applications.

Replay debugging for distributed applications: We have developed a new replay debugging tool, liblog, for distributed C/C++ applications. It logs the execution of deployed application processes and replays them deterministically, faithfully reproducing race conditions and non-deterministic failures, enabling careful offline analysis. To our knowledge, liblog is the first replay tool to address the requirements of large distributed systems: lightweight support for long-running programs, consistent replay of arbitrary subsets of application nodes, and operation in a mixed environment of logging and non-logging processes. In addition, it requires no special hardware or kernel patches, supports unmodified application executables, and integrates GDB into the replay mechanism for simultaneous source-level debugging of multiple processes.

Attested append-only memory: Researchers have made great strides in improving the fault tolerance of both centralized and replicated systems against arbitrary (Byzantine) faults. However, there are hard limits to how much can be done with entirely untrusted components; for example, replicated state machines cannot tolerate more than a third of their replica population being Byzantine. In [18], we investigate how minimal trusted abstractions can push through these hard limits in practical ways. We propose Attested Append-Only Memory (A2M), a trusted system facility that is small, easy to implement and easy to verify formally. A2M provides the programming abstraction of a trusted log, which leads to protocol designs immune to equivocation the ability of a faulty host to lie in different ways to different clients or servers which is a common source of Byzantine headaches. Using A2M, we improve upon the state of the art in Byzantine-fault tolerant replicated state machines, producing A2M-enabled protocols (variants of Castro and Liskov's PBFT) that remain correct (linearizable) and keep making progress (live) even when half the replicas are faulty, in contrast to the previous upper bound. We also present an A2M-enabled single-server shared storage protocol that guarantees linearizability despite server faults. We implement A2M and our protocols, evaluate them experimentally through micro- and macro-benchmarks, and argue that the improved fault tolerance is cost-effective for a broad range of uses, opening up new avenues for practical, more reliable services.

A pervasive network tracing framework: Modern Internet systems often combine different applications (e.g., DNS, web, and database), span different administrative domains, and function in the context of network mechanisms like tunnels, VPNs, NATs, and overlays. Diagnosing these complex systems is a daunting challenge. Although many diagnostic tools exist, they are typically designed for a specific layer (e.g., traceroute) or application, and there is currently no tool for reconstructing a comprehensive view of service behavior. In [29] we propose X-Trace, a tracing framework that provides such a comprehensive view for systems that adopt it. We have implemented X-Trace in several protocols and software systems, and we discuss how it works in three deployed scenarios:

DNS resolution, a three-tiered photo-hosting website, and a service accessed through an overlay network.

Application placement in hosting platforms: Today's Web transactions involve a large variety of components that are unseen by the user. In particular, replicated application servers often do much of the heavy-lifting for large web services. These servers are increasingly hosted on shared hosted platforms. One particularly attractive hosting service model calls for physical servers to be dynamically allocated among multiple applications, with the active application (or applications, if sharing is allowed) dependent on the current workload. These servers therefore must be able to take applications in and out of service in a dynamic fashion. While this notion has been previously developed, the solutions essentially require the overall application churn to be low due to the heavy application startup costs. In this project we investigate techniques to make these application servers more agile by *(i)* running all applications simultaneously and suspending those not in use and *(ii)* using new operating system memory management techniques to reduce the cost of both paging a process out and back in when it is to be activated. We have implemented our solution and demonstrated its effectiveness [12].

2.6 Sensornets

Reliable bulk transport for multihop wireless networks: We have developed Flush [37], a reliable, single-flow transport protocol for sensornets. Flush provides end-to-end reliability, minimizes transfer time, is energy- and memory-efficient, and adapts robustly to changing network conditions. The protocol requires no special control packets to adjust its rate. Flush nodes propagate rate information against the direction of data flow by snooping on next hop traffic, which allows it to track the maximum achievable fixed rate over a wide range of path lengths. Flush is useful for many sensor network applications whose main requirement is to transmit all collected data to the edge, which include environmental monitoring, structural health monitoring, and protocol testing. Indeed, we collected the Flush performance data using Flush itself.

A declarative sensornet system: Sensor networks are notoriously difficult to program, given that they encompass the complexities of both distributed and embedded systems. To address this problem, we have developed a declarative sensor network platform, DSN: a declarative language, compiler and runtime suitable for programming a broad range of sensornet applications [45, 17]. Our approach is a natural fit for sensor networks by specifying several very different classes of traditional sensor network protocols, services and applications entirely declaratively these include tree and geographic routing, link estimation, data collection, event tracking, version coherency, and localization. We address a number of systems challenges that arise when building a generic compiler and runtime environment for the sensornet context; these include not only issues of limited resources, but also the management of asynchrony and requirements of predictable execution. Our results suggest that the declarative approach is well-suited to sensor networks, and that it

can significantly improve software productivity and quality while still producing efficient, resource-constrained code.

New sensornet designs for energy savings: Energy-efficiency has pervaded nearly every aspect of wireless sensor network research, and a range of system software techniques for low-power operation have emerged. Despite these advances, sensornet applications report short lifetimes and low data yields, much to the chagrin of their developers and users. In [21], we focus on a common class of sensor applications we call simple data collection, and explore how their lifetimes might be increased, and their data yields improved, by delaying transmissions as long as possible. We discovered that delay provides little to no practical benefits for polled and scheduled protocols. We identify radio wakeup latency and clock skew as the fundamental constraints limiting communications efficiency and highlight promising research efforts in these areas. We also show that artificially introducing delay raises new research challenges like quickly establishing routing gradients and neighbor tables, and we sketch some possible avenues to address these challenges.

Energy management archictecture: To provide a basis for sensornet energy management, we propose an architecture that allows sensornet application writers to treat energy as a fundamental design primitive [32]. Building systems that manage energy as a critical resource is not a new concept; research in a number of areas harnesses this idea. In fact, our architecture incorporates many of these concepts, including classifying energy as a first-class OS resource, prioritizing resource requests, accounting for fine-grained energy consumers, allocating resources based on dynamic constraints, and providing quality-ofservice (QoS) guarantees by using feedback. In addition, we adopt a three component decomposition that is common for architectures managing scarce shared resources: (1) a policy interface for user input, (2) a mechanism to monitor and control system and component resource usage, and (3) a management module for enforcing policy directives.

By employing and adapting concepts from traditional networking, and architecture literature, we envision an energy management architecture (EMA) that allows sensornet applications to accurately view and efficiently manage the energy budget of a node. The primary contributions of this work beyond existing architectural efforts are facilities for: (1) individual accountability for management of computational units relevant to sensornets, (2) priority of policy directives to enable graceful degradation and predictable operation in a dynamic environment, and (3) expression of network-level policy that can be used by individual nodes for local optimization and shared among nodes to adjust the behavior of the network as a whole. In its entirety, the EMA promotes prioritized enforcement of policy during runtime operation as well as enables improved system behavior visualization for debugging during application development.

A retrospective look at modular sensornet architectures: To deal with severe resource constraints, the first wave of sensornet programmers built tightly-integrated and monolithic system stacks. While the resulting systems were far more energy, memory, and bandwidth efficient than traditional systems, they had two unfortunate properties. First, they were extremely difficult to program, as they had to confront myriad new networking

and sensing challenges and doing so required very low-level control. Programming entire systems using low-level C code is a daunting task for experienced programmers, and all but impossible for the intended users of sensornets: general scientists. Second, the implementations provided few clearly defined internal interfaces, leading to limited code reuse and interoperability between applications from various programmers.

In order to deal with these challenges, Culler et al. proposed an overall sensornet architecture [19]. Such a modular architecture decomposes the system into a set of services, specifies a set of interfaces to these services, and can define its own protocols, including packet formats and communication exchanges. Although the Internet architecture provides inspiration, its applicability is limited in this context by the substantial differences between the Internet and sensornets. After two years of progress filled with successes and failures, we took a step back, and reflected on the evolution of our architectural outlook into its present day form, drawing from our design experiences. The results of this retrospective are captured in [44].

2.7 Game-Theoretic Approaches

Distributed algorithmic mechanism design: Most discussions of algorithmic mechanism design (AMD) presume the existence of a trusted center that implements the required economic mechanisms. This project, summarized in the book chapter [23], focuses on mechanism-design problems that are inherently distributed, i.e., those in which such a trusted center cannot be used. Such problems require that the AMD paradigm be generalized to distributed al- gorithmic mechanism design (DAMD).

Hidden actions in routing: In multi-hop networks, the actions taken by individual intermediate nodes are typically hidden from the communicating endpoints; all the endpoints can observe is whether or not the end-to-end transmission was successful. Therefore, in the absence of incentives to the contrary, rational (i.e., selfish) intermediate nodes may choose to forward packets at a low priority or simply not forward packets at all. Using a principal-agent model, we show (in [24]) how the hidden-action problem can be overcome through appropriate design of contracts, in both the direct (the endpoints contract with each individual router) and recursive (each router contracts with the next downstream router) cases. We further demonstrate that per-hop monitoring does not necessarily improve the utility of the principal or the social welfare in the system. In addition, we generalize existing mechanisms that deal with hidden-information to handle scenarios involving both hidden-information and hidden-action.

2.8 Internet Community Activities

Mark Allman chairs the IRTF's Internet Measurement Research Group (IMRG), co-chairs the IETF's TCP Maintenance and Minor Extensions (TCPM) and is a member of the IETF's Transport Area Directorate and General Area Review Team. Sally Floyd chairs the Transport Modeling Research Group (TMRG) of the Internet Research Task Force (IRTF), and is a member of the IETF Transport Area Directorate. Vern Paxson served as vice chair of ACM SIGCOMM and program co-chair of ACM SIGCOMM HotNets. He

is currently a member of the NSF Future Internet Design Planning Committee and the President's Council of Advisors on Science and Technology. Nick Weaver is one of the developers of the NSF-sponsored "Capture the Flag" security defense competition to be held during USENIX Security 2008. Scott Shenker serves on the GENI Science Council and was its first chair.

2.9 Outreach

Bro Workshop 2007: Jointly with the Lawrence Berkeley National Laboratory, we organized a three-day workshop focused on our "Bro" network intrusion detection system. The workshop, held at the San Diego Supercomputer Center in July 2007, attracted more than 30 participants from universities, government labs, and industry, including several from outside the USA. The workshop was aimed at computer security staff who want to learn more about the Bro scripting language and how to customize Bro based on each site's policy. Topics included a Bro scripting language tutorial, Bro customization, Bro log file analysis and active blocking using Bro. The workshop received very positive feedback, leading us to plan on a follow-up event in 2008.

NIDS Lectures and Labs at Aachen University, Germany: Two staffmembers spent several weeks as visiting researchers at RWTH Aachen, Germany. The visit included lectures for more than 40 students on the state of the art in network monitoring and intrusion detection, traffic analysis tools, and the Bro network intrusion detection system (NIDS). We held a version of the 2007 Bro Workshop in the form of a two-day student lab, in which the students solved progressively more difficult network monitoring tasks. Assignments ranged from simple tuning of Bro's default *alarm* and *notice* policies to a customizable and persistent database of services running on the monitored network's hosts. After two days, the students were able to implement the latter in less than 100 lines of Bro code.

References

[1] M. Allman (2007), "Personal Namespaces,", Proceedings of ACM SIGCOMM HotNets, November 2007.

[2] M. Allman, K. Avrachenkov, U. Ayesta, and J. Blanton (2007), "Early Retransmit for TCP and SCTP," Internet-Draft draft-allman-tcp-early-rexmt-05.txt, June 2007.

[3] M. Allman, K. Christensen, B. Nordman, and V. Paxson (2007) "Enabling an Energy-Efficient Future Internet Through Selectively Connected End Systems," *Proceedings of ACM SIGCOMM HOTNETS*, November 2007.

[4] M. Allman, C. Kreibich, V. Paxson, R. Sommer and N. Weaver (2007) "The Strengths of Weaker Identities: Opportunistic Personas," Proceedings USENIX Hot Security, August 2007.

[5] M. Allman, C. Kreibich, V. Paxson, R. Sommer, N. Weaver (2007), "Seeking Visibility Into Network Activity Across Time And Space," ACM Computer Communication Review, November 2007.

[6] M. Allman, L. Martin, M. Rabinovich, and K. Atchinson (2008) "On Community-Oriented Internet Measurement," Proceedings of the Passive and Active Measurement Conference, April 2008.

[7] M. Allman, V. Paxson, and W. Stevens (1999) "TCP Congestion Control," IETF RFC 2581, April 1999.

[8] M. Allman and V. Paxson (2007) "Issues and Etiquette Concerning Use of Shared Measurement Data," Proceedings ACM Internet Measurement Conference, October 2007.

[9] M. Allman, V. Paxson, and E. Blanton (2007) "TCP Congestion Control," IETF Internet Draft draft-ietf-tcpm-rfc2581bis-04.txt, December 2007.

[10] M. Allman and V. Paxson (2008) "A Reactive Measurement Framework," Proceedings of Passive and Active Measurement Conference, 2008.

[11] M. Allman, V. Paxson, and J. Terrell (2007) "A Brief History of Scanning," Proceedings ACM Internet Measurement Conference, October 2007.

[12] Z. Al-Qudah, H. Alzoubi, M. Allman, V. Liberatore, M. Rabinovich (2008) "Efficient Application Placement Mechanisms in a Hosting Platform," USENIX Annual Technical Conference, January 2008.

[13] D. Andersen, H. Balakrishnan, N. Feamster, T. Koponen, D. Moon, and S.Shenker (2007), "Holding the Internet Accountable", HotNets 07.

[14] K. Argyraki, P. Maniatis, O. Irzak, A. Subramanian, and S. Shenker (2007), "Loss and Delay Accountability for the Internet,", ICNP 2007.

[15] J. Blanton, E. Blanton, and M. Allman (2007), "Using Spurious Retransmissions to Adapt the Retransmission Timeout," Internet-Draft draft-allman-rto-backoff-05.txt, July 2007.

[16] M. Casado, M. Freedman, J. Pettit, N. McKeown, and S. Shenker (2007), "Ethane: Taking Control of the Enterprise,", SIGCOMM 2007.

[17] D. Chu, L. Popa, A. Tavakoli, J. Hellerstein, P. Levis, S.Shenker, and I. Stoica (2007), "The design and implementation of a declarative sensor network system,", SenSys '07.

[18] B.-G. Chun, P. Maniatis, S. Shenker, and J. Kubiatowicz (2007), "Attested Append-Only Memory: Making Adversaries Stick to their Word,", SOSP 2007.

[19] D. Culler, P. Dutta, C. T. Ee, R. Fonseca, J. Hui, P. Levis, and J. Zhao (2005), "Towards a Sensor Network Architecture: Lowering the Waistline,", HotOS, 2005.

[20] M. Demmer, K. Fall, T. Koponen, and S. Shenker (2007), "Towards a Modern Communications API,", HotNets 07.

[21] P. Dutta, D. Culler, and S. Shenker (2007), "Procrastination Might Lead to a Longer and More Useful Life,", HotNets 07.

[22] C.T. Ee, V. Ramachandran, B.-G. Chun, K. Lakshminarayanan, and S. Shenker (2007), "Resolving Inter-Domain Policy Disputes,", SIGCOMM 2007.

[23] J. Feigenbaum, M. Schapira, and S. Shenker, "Distributed Algorithmic Mechanism Design,", Algorithmic Game Theory, Cambridge University Press, 2007.

[24] M. Feldman, J. Chuang, I. Stoica, and S. Shenker (2007), "Hidden-Action in Network Routing,", IEEE Journal on Selected Areas in Communications, Vol. 25, No. 6, 2007.

[25] S. Floyd and M. Allman (2007), "Specifying New Congestion Control Algorithms," August 2007, RFC 5033, BCP 133.

[26] S. Floyd and M. Allman (2007), "Comments on the Usefulness of Simple Best-Effort Traffic," Internet-Draft draft-floyd-tsvwg-besteffort-01.txt, August 2007.

[27] S. Floyd, M. Allman, A. Jain, and P. Sarolahti (2007), "Quick-Start for TCP and IP," January 2007. RFC 4782.

[28] S. Floyd and E. Kohler (2007), "TCP Friendly Rate Control (TFRC): the Small-Packet (SP) Variant," April 2007. RFC 4828.

[29] R. Fonseca, G. Porter, R. Katz, S. Shenker, and I. Stoica (2007), "X-Trace: A Pervasive Network Tracing Framework", NSDI 07.

[30] J. Franklin, V. Paxson, A. Perrig, and S. Savage (2007), "An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants," Proceedings ACM Computer and Communication Security Conference, October 2007.

[31] J. Gonzalez, V. Paxson, and N. Weaver (2007) "Shunting: A Hardware/Software Architecture for Flexible, High-Performance Network Intrusion Prevention," Proceedings ACM Computer and Communication Security Conference, October 2007.

[32] X. Jiang, J. Taneja, J. Ortiz, A. Tavakoli, P. Dutta, J. Jeong, D. Culler, P. Levis, and S. Shenker(2007), "An Architecture for Energy Management in Wireless Sensor Networks,", International Workshop on Wireless Sensor Network Architecture, 2007.

[33] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan (2004) "Fast Portscan Detection Using Sequential Hypothesis Testing," IEEE Symposium on Security and Privacy, May 2004.

[34] L. Juan, C. Kreibich, C-H. Lin, and V. Paxson (2008) "A Tool for Offline and Live Testing of Evasion Resilience in Network Intrusion Detection Systems," Proceedings iCAST Conference, January 2008.

[35] J. Jung, R. Milito and V. Paxson (2007), "On the Adaptive Real-Time Detection of Fast-Propagating Network Worms," Proceedings Fourth GI International Conference on Detection of Intrusions & Malware, and Vulnerability Assessment, July 2007.

[36] J. Jung, R. Milito and V. Paxson (2008), "On the Adaptive Real-Time Detection of Fast-Propagating Network Worms," Journal on Computer Virology, 2008.

[37] S. Kim, R. Fonseca, P. Dutta, A. Tavakoli, D. Culler, P. Levis, S. Shenker, and I. Stoica (2007), "Flush: a reliable bulk transport protocol for multihop wireless networks," SenSys 2007.

[38] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K.H. Kim, S. Shenker, and I. Stoica (2007), "A Data-Oriented (and Beyond) Network Architecture," SIGCOMM 2007.

[39] K. Lakshminarayanan, M. Caesar, M. Rangan, T. Anderson, S. Shenker, and I. Stoica (2007), "Achieving Convergence-Free Routing using Failure-Carrying Packets," SIGCOMM 2007.

[40] D. Liu, M. Allman, S. Jin, and L. Wang (2007), "Congestion Control Without a Startup Phase," Protocols for Fast, Long Distance Networks (PFLDnet) Workshop, February 2007.

[41] V. Paxson, R. Sommer, and N. Weaver (2007), "An Architecture for Exploiting Multi-Core Processors to Parallelize Network Intrusion Prevention," Proceedings IEEE Sarnoff Symposium, May 2007.

[42] C. Reis, S. Gribble, T. Kohno, and N. Weaver (2008), "Detecting In-Flight Page Changes with Web Tripwires," Proceedings 5th USENIX Symposium on Networked Systems Design and Implementation, April 2008, to appear.

[43] P. Sarolahti, M. Allman, and S. Floyd (2007), "Determining an Appropriate Sending Rate Over an Underutilized Network Path,", Computer Networks Special Issue on Protocols for Fast, Long-Distance Networks, 51(7), May 2007.

[44] A. Tavakoli, P. Dutta, J. Jeong, S. Kim, J. Ortiz, D. Culler, P. Levis, and S. Shenker (2007), "A Modular Sensornet Architecture: Past, Present, and Future Directions,", International Workshop on Wireless Sensor Network Architecture, 2007.

[45] A. Tavakoli, D. Chu, J.Hellerstein, P. Levis, and S. Shenker (2007) "A Declarative Sensornet Architecture", International Workshop on Wireless Sensor Network Architecture, 2007.

[46] M. Vallentin, R. Sommer, J. Lee, C. Leres, V. Paxson, and Brian Tierney (2007) "The NIDS Cluster: Scalable, Stateful Network Intrusion Detection on Commodity Hardware," Proceedings Recent Advance in Intrusion Detection, September 2007.

[47] M. Vutukuru, H. Balakrishnan, and V. Paxson (2008), "Efficient and Robust TCP Stream Normalization," in submission.

[48] N. Weaver, V. Paxson, and J. Gonzalez (2007), "The Shunt: An FPGA-Based Accelerator for Network Intrusion Prevention," Proceedings FPGA, February 2007.

[49] N. Weaver and R. Sommer (2007), "Stress Testing Cluster Bro," Proceedings USENIX DETER Community Workshop on Cyber Security Experimentation and Test, August 2007.

3 Algorithms

The members of the Algorithms group during 2007 were Richard M. Karp (Group Leader), Eran Halperin(Research Scientist), John Moody (Research Scientist), Martin Gairing, Gad Kimmel, Bonnie Kirkpatrick, Jyri Kivinen, Henry Lin, Aron Rempel, Mat Saffel and Sriram Sankararamanan..

Much of the research in the Algorithms group continues to be focused on the development of computational tools that may help in studies of the genetics of complex diseases such as cancer, Alzheimer's disease and Parkinson's disease. In order to understand the genetic factors underlying complex diseases, disease association studies are performed. In these studies, a set of cases (individuals carrying the disease) and controls (healthy individuals) are genotyped, and the genetic variation across the two populations is compared. The information collected from each individual usually consists of a set of positions in the genome called single nucleotide polymorphisms (SNPs). During 2007 the number of large-scale datasets suitable for disease association studies has escalated, underscoring the need for algorithms to extract meaning from this type of data.

We have also continued our research in computer networking and remained active in core areas of computer science of longstanding interest to the group, such as combinatorial algorithms and reinforcement learning, and have become active in the emerging field of computational game theory.

Highlights Both of this years highlights are related to the central statistical issue of population effects in genetic association studies.

We have developed an efficient computational method, based on dynamic programming and importance sampling, for avoiding the confounding effects of population structure on measuring the significance of SNP-based association tests [10].

The genomes of individuals of mixed ancestry consist of a mosaic of short segments that have originated from the different ancestral populations. We have devised methods of inferring the structure of an individual's mosaic through an analysis of single-nucleotide polymorphisms [18, 19].

3.1 Genetic Association Studies

WHAP - Weighted Sum of HAPlotypes (E. Halperin, N. Zaitlen, H.M. Kang, E. Eskin)

In the 2006 Annual Report we described a method for improving the efficiency of association studies by using "indirect association," in which only a subset of SNPs are collected.They serve as proxies for the uncollected SNPs, taking advantage of the correlation structure between SNPs. This method has now been published [21], and we have built a web server (http://whap.cs.ucla.edu/) that allows people to use the methods through a web application.

Pedigree-Based Association Studies (B.Kirkpatrick, E. Halperin, R.M. Karp)

The ultimate goal of this project is to study associations between SNPs and phenotypes such as disease using data about the genotypes of individuals within pedigrees. In the first phase of the study our goal is to make probabilistic assignments of haplotypes to the

founders of the pedigrees, and determine an *a posteriori* distribution of haplotypes among the members of the pedigrees, based on the available genotyes and an *a priori* distribution of haplotypes. We devise methods applicable to two situations: one in which the available pedigrees are numerous, small and simple, and one in which large inbred pedigrees are available.

Population Association Test (G. Kimmel, M. Jordan, E. Halperin¡ R. Shamir, R.M. Karp)

A principal difficulty in drawing causal inferences from whole-genome case-control association studies is the confounding effect of population structure. Differences in allele frequencies between cases and controls may be due to systematic differences in ancestry rather than to association of genes with disease. This issue needs careful attention in forthcoming large-scale association studies, given the lack of knowledge regarding relevant ancestral history throughout the genome, and the need to aggregate over many individuals in order to achieve high levels of discriminatory power when many markers are being screened.

We have recently studied this problem and shown [10] how to perform a randomization test akin to a permutation test, that takes account of the population structure.We assume that a baseline estimate is available for the probability that each individual has the disease independent of their genotype, which we call the baseline probability vector. We then consider a null hypothesis in which each of the individuals has the disease independently according to the corresponding component of the baseline probability vector. Carried out naively, the computation underlying our method would be feasible in general only for small problems. As we show, however, the combination of importance sampling and dynamic programming renders this approach quite feasible for typical whole-genome studies. Our simulations show that our method achieves higher power and lower false positive rates than existing methods.

Estimating Local Ancestry in Admixed Populations (G. Kimmel, S. Sankararamanan, S. Sridhar, M. Jordan, E. Halperin).

The problem of inferring population sub-structure is especially challenging when admixed populations are involved. In these populations (e.g., African Americans and Latinos), two or more ancestral populations have been mixing for a relatively small number of generations, resulting in a new population in which the ancestry of every individual can be explained by different proportions of the original populations. Due to recombination events, the genomes of these individuals consist of a mosaic of short segments that have originated from the different ancestral populations.

In a recent study [18] we explored this problem, and suggested a new method to locally evaluate the ancestries of each individual. Our empirical results show that our method is significantly more accurate and more efficient than existing methods for inferring locus-specific ancestries, enabling it to handle large-scale data sets.

In [19] we presented a probabilistic model to represent the ancestral information of such admixed populations.Previous model-based approaches used hidden Markov models (HMMs) to model the problem, and the recombination events are modeled implicitly by

transition probabilities. In this work, we introduced a set of indicator variables that directly model the recombination events. These variables are used to determine the transition matrices of an HMM. The suggested model is an instance of a switching hidden Markov model, and it differs from a standard HMM in that it does not assert the same conditional independences as an HMM. The new model includes a number of biologically-motivated parameters, which can be inferred via a Markov chain Monte Carlo (MCMC) algorithm or an expectation-maximization (EM) algorithm, or some combination. Using this model, we further improved the accuracy of our previous method. Thus, we are able to retain the advantages of a model-based method while not sacrificing performance. In particular, the complete model-based approach allows the possibility of easily applying it in different scenarios, in which a different subset of the variables are unobserved. One example is a scenario where the allele frequencies are given for only one ancestral population. To the best of our knowledge, this is the first time that this realistic problem has been addressed.

Tag SNPs and Power (G. Kimmel, O. Davidovich, E. Halperin, R. Shamir)

Every association study is preceded by selection of the tag SNPs to be typed, whether explicitly by the researchers or previously, by chip producers. The success of the study critically depends on its power. We have recently shown [4] that using the prediction accuracy criterion instead of the commonly used correlation r^2 criterion for tag SNP selection, improves the power of association studies. The advantage, measured by the relative power, is quite substantial, reaching up to 11 percent for realistic scenarios.

Population Stratification and Pooled Genotypes (G. Kimmel, M. Jordan)

We have recently started a collaboration with colleagues in Celera. Our main project is to analyze preliminary data typed for thousands of individuals, in order to select tag SNPs for the next stage of the study. Obviously, one would like to choose the most promising SNPs to optimize the power. There is an additional challenge in this specific project: the data set contains pooled genotypes, i.e., several individuals typed together. Our first effort is to try to correct for population stratification. The approach is to use our recently published stratification correction method, and to adjust it to pools of genotypes.

Association Mapping via Phylogeny (G. Kimmel, R.M. Karp, M. Jordan, E. Halperin)

In many of the current association studies, researchers test for association separately SNP by SNP. There are new approaches that take into account the unobserved genealogy of the population. Briefly, the main idea is that the best information that can be obtained about association of a SNP is to know the full coalescent genealogy of the sample at that position. Although these methods have been shown to have higher power, they are computationally intensive, and cannot be applied to current whole genome data sets.

We have recently started to tackle this problem. The main idea is to exploit the special combinatorial properties of these trees (which hold a perfect phylogeny structure), to perform conditional, and hence more efficient, sampling. As we show in simulations, this approach gives significantly higher power and more accurate locus mapping than former methods.

Detecting Disease-Specific Dysregulated Pathways (I. Ulitzky, R.M. Karp, R. Shamir)

In [20] we present a method for identifying connected gene subnetworks significantly enriched for genes that are dysregulated in specimens of a disease. These subnetworks provide a signature of the disease potentially useful for diagnosis, pinpoint possible pathways affected by the disease, and suggest targets for drug intervention. Our method uses microarray gene expression profiles derived in clinical case-control studies to identify genes significantly dysregualated in disease specimens, combined with protein interaction data to identify connected sets of genes. Our core algorithm searches for minimal connected subnetworks in which the number of dysregulated genes in each diseased sample exceeds a given threshold. We have applied the method in a study of Huntington's disease caudate nucleus expression profiles and in a meta-analysis of breast cancer studies. In both cases the results were statistically significant and appeared to home in on compact pathways enriched with hallmarks of the diseases.

3.2 Discrete Structures, Networks and Algorithms

Routing in the Internet and in Sensornets Two investigations of network routing described in the 2006 Annual Report came to fruition in 2007 and led to publications. In [13] we describe a method for decomposing Internet-like graphs into small interlinked components, permitting the use of local, rather than global, routing tables.In [16] we give a continuum model of large, homogeneous sensornets, leading to a formulation of minimum-congestion routing as a problem in the calculus of variations that can be solved numerically.

Congestion Games (M. Gairing)

Congestion games [17] and variants thereof have long been used to model non-cooperative resource sharing among selfish players. Examples include traffic behavior in road or communication networks or competition among firms for production processes.

The *price of anarchy* [12] is a measure of the extent to which competition approximates the global objective; e.g., the minimum total travel time (latency) in the case of road networks. Typically, the price of anarchy is the worst-case ratio between the value of a global objective function (usually coined as *social cost*) in some state where no p[layer can unilaterally improve its situation, and that of some global optimum. The price of anarchy represents a rendezvous of Nash equilibrium [15], a concept fundamental to Game Theory, with approximation, an omnipresent concept in Theoretical Computer Science today.

Singleton Congestion Games In [5] we study *singleton congestion games* where each player's strategy consists only of a single resource. A sample application for these modified games is load balancing. For such games, we prove a collection of new bounds on the price of anarchy for several interesting classes of congestion games.

Bayesian Malicious Congestion Games We introduce *Bayesian malicious congestion games* as a new model to study congestion games with malicious players. In such a game, each player is malicious with a certain probability - in which case her only goal is to disturb the other players as much as possible.

We show that it is NP-complete to decide whether a given game admits a Pure Bayesian Nash equilibrium.This result even holds when resource latency functions are linear and each player is malicious with some probability.

Moreover, we study the influence of malicious players on the price of anarchy. For certain subclasses of Bayesian malicious congestion games, we show that the presence of malicious players can be be either harmful or beneficial to the system. For both directions, we provide tight results on the maximum factor by which this can happen.

Streaming Algorithms for Selection and Approximate Sorting (R.M. Karp)

Companies such as Yahoo, Google and Microsoft maintain extremely large data repositories within which searches are frequently conducted. There is interest in developing such repositories in the public sector and applying them to massive data set problems of importance to the scientific community and society in general.

It is of interest to develop streaming algorithms for basic information processing tasks within such data repositories. We present such algorithms for selecting the keys of given ranks in a totally ordered set of n keys, and for a related problem of approximate sorting. in [8] we derive bounds on the storage and time requirements of these algorithms under the assumption that blocks of keys arrive in a random order, and show that these bounds are close to information-theoretic lower bounds for the problems.We assume random arrivals because these repositories support random access to disc blocks.

The *α-quantile* of a totally ordered set of n keys is the $\lfloor \alpha n \rfloor$th smallest element. We present near-optimal algorithms (simultaneously for time and storage), under the random arrivals assumption, for the following problems:

1. **Selection**: compute the α-quantile for a given α.
2. **Multiple selection**: compute α-quantiles for many given values of α.
3. **Parallel selection**, in which the input is divided into streams, each with its own buffer, and the different streams communicate by message passing.
4. **Approximate selection**: given α and ϵ, find a key whose rank differs from αn by at most ϵn.
5. **Approximate sorting**: Given a small positive constant ϵ, compute an ordering of the keys in which the rank assigned to each key agrees with its rank in the true ordering, within a relative error of ϵ.

Finally, as a byproduct of our analysis of approximate sorting, we give an elegant method for computing the expected number of comparisons for some classical randomized algorithms for selection and selection.

Bounds on the Probability of a Union (I. Adler, R.M. Karp, S. Ross)

Let $A_1, A_2, ..., A_n$ be events in a sample space. Given the probability of the intersection of each collection of up to k of these events, what can we say about the probability that at least r of the events occur? This question dates back to Boole in the 19th century, and it

is well known that the odd partial sums of the Inclusion-Exclusion formula provide upper bounds, and the even partial sums provide lower bounds. We give a combinatorial characterization of the error in these bounds, and use it to derive the strongest possible bounds of a certain form. The new bounds use more information than the classical "Bonferroni-type" inequalities, and are often sharper.

3.3 Direct Reinforcement Learning

John Moody's research on direct reinforcement learning focused on two application areas: portfolio management and adaptive videoconferencing. This effort included collaborations with an ICSI postdoc and three members of the ICSI Spanish visitors program.

In collaboration with Alberto Suarez and Matthew Saffell, we designed and implemented a recurrent reinforcement learning system for the dynamic management of financial portfolios. The system induces investment policies from asset price series and various financial and economic indicators. The search for investment policies is direct and avoids explicit modeling or forecasting of the asset returns series.

An important goal of the project was to analyze the effect of transaction costs on the type of policies that are learned. In the absence of transaction costs, the policies identified tend to exhibit switching behavior, heavily investing at a given time in the asset that is expected to perform best in the next period. As the transaction costs are increased, the current portfolio composition becomes a relevant input in the final decision on the subsequent reallocation of capital. In this manner, the investment policies learned become smoother and the portfolios selected allocate their investment more evenly across the different assets.

This behavior is in agreement with common financial views on the benefits of diversification and gradual rebalancing of investment portfolios. The quality of the strategies learned was assessed on real-world financial problems, including asset allocation across global stock markets.

Previous work with Juan Botia and Pedro Ruiz on use of direct reinforcement learning for adaptive multimedia resulted in completion of a technical report.

References

[1] J. A. Botiá, J. Moody and P. M. Ruiz (2007) "Recurrence and Cost within Stochastic Direct Reinforcement to Control Adaptive Videoconference Applications," Technical Report TR-01/07. Departamento de Ingenieria de la Información y las Comunicaciones. Universidad de Murcia. Murcia, Spain, Feb, 2007.

[2] C. Daskalakis, A.G. Dimakis, R.M. Karp, and M. J. Wainwright "Probabilistic analysis of linear programming decoding," To appear in IEEE Transactions in Information Theory; also in SODA 2007:385-394.

[3] C. Daskalakis, R.M. Karp, E. Mossel, S. Riesenfeld, and E. Verbin (2007) "Sorting and selection in posets," CoRR abs/0707.1532 (2007).

[4] O. Davidovich, G. Kimmel, E. Halperin, and R.Shamir (2008) "Increasing association power via prediction-based tag-SNP selection," Submitted for publication.

[5] M. Gairing and F. Schoppman (2007) "Total Latency in Singleton Congestion Games," Proc. of 3rd WINE. LNCS 4858, Springer Verlag (2007) 381-387.

[6] I. Gat-Viks, R.M. Karp, R. Sharan, and R. Shamir (2007) "Reconstructing chain functions in genetic networks," Siam J. Discrete Math. vol. 20 727-740(2007).

[7] R.M. Karp (2006) "George Dantzig's Impact on the Theory of Computation," To appear in Discrete Optimization.

[8] R.M. Karp (2007) "Streaming algorithms for selection and approximate sorting," FSTTCS 2007: 9-20.

[9] R.M. Karp, and R. Kleinberg "Noisy binary search and its applications." SODA 2007: 881-890.

[10] G. Kimmel, M.I. Jordan, E. Halperin, R. Shamir, and R.M. Karp (2007) "A randomization test for controlling population stratification in whole-genome association studies," Amer. J. Human Genetics (2007).

[11] B. Kirkpatrick, C.S. Armendariz, R.M. Karp, and E. Halperin (2007) "HAPLOPOOL: improving haplotype frequency estimation through DNA pools and phylogenetic modeling," Bioinformatics vol.23, no. 22 3048-3055 (2007).

[12] E. Koutsoupias and C.H. Papadimitriou (1999) "Worst-Case Equilibria," Proc. of 16th STACS, LNCS 1563, Springer-Verlag 404-413 (1999).

[13] H. Lin, C. Amanatidis, M. Sideri, R.M. Karp, and C. Papadimitriou (2008) "Linked decompositions, Internet routing, and the power of choice in Polya urns," To appear in SODA 2008.

[14] M. Narayanan and R.M. Karp (2007) "Comparing protein interaction networks via a graph Match-and-Split algorithm," J. Computational Biology, vol. 14, no. 7 892-907(2007).

[15] J.F. Nash (1951) "Non-Cooperative Games," Annals of Mathematics 54(2) 286-295 (1951).

[16] L. Popa, A. Rostami, R.M. Karp, C. Papadimitriou, and I. Stoica (2007) "Balancing the traffic load in wireless networks with curveball routing," Proceedings of Mobihoc, 2007.

[17] R.W. Rosenthal (1973) "A Class of Games Possessing Pure-Strategy Nash Equilibria," International Journal of Game Theory 2 65-67 (1973).

[18] S. Sankaramanan, S. Sridhar, G. Kimmel, and E. Halperin (2008) "Estimating local ancestry in admixed populations," To appear in American Journal of Human Genetics.

[19] S. Sankaramanan, G. Kimmel, E. Halperin, and M.I. Jordan "A switching HMM for inferring ancestries in admixed populations," Submitted for publication.

[20] I. Ulitzky, R.M. Karp, and R. Shamir (2008) "Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles," To appear in RECOMB 2008.

[21] N. Zaitlen, H.M. Kang, E. Eskin, and E. Halperin (2007) "Leveraging the HapMap Correlation Structure in Association Studies," American J. Human Genetics 80:683-691, 2007.

[22] A. Zemla, B. Geisbrecht, J. Smith, M. Lam, B. Kirkpatrick, M. Wagner, T. Slezak, and C. Zhou (2007) "ATRALCP-Structure alignment-based clustering of proteins," Nucleic Acids Research (2007).

4 Artificial Intelligence and its Applications

In 2007, the Artificial Intelligence group continued both its basic and applied thrusts in probabilistic modeling, language learning, semantic resource development, and neurobiology. In 2007, the applied role was expanded to incorporate the methodologies and algorithms developed by the group in applications in the areas of Semantic Extraction, Computational Biology, Natural Language Processing (NLP) for developing regions, Predictive Analysis, and Multilingual Semantic Resources.

The core scientific and technical work of the group is done within the three articulating efforts of the AI group. These are

1. The Neural Theory of Language (http://www.icsi.berkeley.edu/NTL) is a long-standing project investigating biologically plausible models of conceptual memory, language learning, and language use.

2. FrameNet (http://framenet.icsi.berkeley.edu) is an ongoing project led by Charles Fillmore that is building a semantically rich on-line lexicon based on the theory of Frame Semantics. The initial effort was an English Lexicon, but as described here, the effort has expanded to multiple languages.

3. Applications of the groups' research included the following efforts in 2007.

   (a) A multi-university project (the ICSI AI group, Prof. Chris Manning's group from Stanford University, Prof. Harabagiu's group from University of Texas, Dallas) on semantically based NLP funded by the DoD through the ARDA AQUAINT grant.

   (b) Industrial applications in semantic extraction and ontology mapping from technical documents (sponsored by Cisco systems) and in NLP based smart search (sponsored by Ask.com).

   (c) A project with Lawrence Livermore National Laboratories (LLNL) to build parallel realizations of semantic parsing on graphics hardware. This project involved cooperation between the ICSI NTL and FrameNet groups and the GAIA group at LLNL.

   (d) A new IARPA project on predictive analysis that was a result of a highly competitive selection process. The ICSI AI group was selected to provide the basic probabilistic modeling infrastructure for IARPA analysis applications.

   In all these cases, our main research goal is to use semantic tools and techniques developed by the the group to advance the automated analysis of information for a variety of tasks.

In addition, in 2007 the group continued work on hybrid state models of biological processes. Joseph Makin, a PhD student working with Srini Narayanan on the hybrid State model of the coagulation pathway completed his Qualifying exam and is expected to complete his dissertation in 2008. Our previous work resulted in the first comprehensive computational simulation of the mammalian coagulation pathway. In 2007, we combined

methods from nonlinear control theory with hybrid system modeling and verification to perform sensitivity analyses and design controllers for therapeutic interventions in the presence of specific clotting disorders.

In 2007, Srini Narayanan won a highly competitive Google faculty research grant. The grant is being used to start a new effort on investigating NLP techniques for providing multilingual health care information services to rural populations in developing countries. The ICSI AI group is partnering with a well known non-profit organization, Hesperian Press (http://www.hesperian.org), who books in over 80 languages on primary health care are being used by rural health workers in community lead efforts in over 100 countries around the world. This project is expected to be one of the major foci or the ICSI AI group in 2008.

Detailed accounts of progress in specific projects in 2007 follows.

4.1 The Neural Theory of Language

The NTL project of the AI group works in collaboration with other units on the UCB campus and elsewhere. It combines basic research in several disciplines with applications to natural language processing systems. Basic efforts include studies in the computational, linguistic, neurobiological and cognitive bases for language and thought and continues to yield a variety of theoretical and practical findings. In 2007, we have made significant progress on all of these aspects.

The group has developed a formal notation for Embodied Construction Grammar (ECG), which plays a crucial role in larger, simulation-based language understanding system. Jerome Feldman's book on the NTL project was published by MIT Press in June 2006. This has led to a number of seminars and public appearances related to our work and these are summarized on the book web site http://m2mbook.org/ . The ECG formalism is playing an increasingly important role in linguistics, both at Berkeley and beyond. Robert Porzel, a frequent visitor, has finished his doctoral thesis at U. Luebeck and Artjom Klein spent much of 2007 at ICSI working on his German Master's thesis.

A major new initiative is the production and release of a new ECG wiki, which will contain a tutorial and other pedagogical materials and will also serve as a coordination point for grammar development and analysis. The web site is: http://ecgweb.icsi.berkeley.edu/ECGweb/

Olya Gurevich completed her doctoral dissertation in linguistics and has taken a position at Powerset, Inc. Ellen Dodge passed her qualifying exams in Linguistics and has completed most of her dissertation. All of the CS students on the project have now passed their qualifying exams and are making good thesis progress.

One core NTL computational question is finding the best match of constructions to an utterance in linguistic and conceptual context. The general computational point is that our task of finding a best-fit analysis and approximate answers that are not always correct presents a more tractable domain than exact symbolic matching. More importantly, our integrated constructions are decidedly not context-free or purely syntactic.

John Bryant, a CS doctoral student, has completed a doctoral thesis on this topic and remains with the group as a post-doc. The resulting program pays an central role in the linguistics doctoral of Ellen Dodge, which is essentially complete and the CS doctoral work of Eva Mok. His system is also now in use at other labs internationally. This work in being

extended to include discourse and situational context in the semantic best fit computations. Bryant and Eva Mok showed how these ideas apply to languages like Chinese, that often omit words, in a paper given at the 2006 meeting of the Berkeley Linguistics Society. In addition 2007, Joseph Makin, Steve Sinha, and Leon Barrett completed their qualifying exams. Joseph and Steve are expected to complete their dissertations in 2008, and Leon Barrett is expected to complete in 2009.

Nancy Chang and Eva Mok have continued developing representations and algorithms useful for an embodied approach to language acquisition and use. Eva Mok has extended the paradigm to include more complex constructions and to cover Asian languages, which behave differently. Chang and Mok published three papers recently on the representation and its use in modeling language acquisition [11, 12, 13]

Leon Barrett (with Feldman and MacDermed) completed a paper and system for a new solution to the variable binding problem. This has now been accepted by Neural Computation, the leading journal in the field. He also published two papers on language processing with collaborators on UCB campus [16, 17, 1].

Imaging (fMRI) results of work conducted in collaboration with Rich Ivry's group at UCB and Lisa-Aziz Zadeh at USC (Lisa was an ICSI PostDoc before joining the USC faculty) consistent with the NTL simulation hypothesis was accepted for publication in the journal Social Neuroscience in 2007 [4]. The 2006 annual report contains a detailed description of the experiments performed.

Ben Bergen (U. Hawaii) continues to cooperate with the group after completing a UCB linguistics thesis using a statistical, corpus-based approach in combination with psycholinguistic experimentation, to explore probabilistic relations between phonology on the one hand and syntax, semantics, and social knowledge on the other. Bergen and Feldman have completed a major invited article showing how NTL helps explain the ancient mystery of how people can learn new concepts [2]. Bergen and Narayanan collaborated with Cognitive Scientist Teenie Matlock (at the University of California, Merced) testing the NTL hypothesis that language understanding involves embodied simulation. The collaboration resulted in a paper in the Cognitive Science Journal in 2007 [6].

In 2007 there was a very significant increase in the use of the group's results in UCB courses and in linguistics research. Collaboration with the FrameNet project has been broadened and deepened with positive results for both efforts, some of which are described in this report. J. Feldman ran an interdisciplinary class in Spring 2007 and several of the research efforts from that class are being incorporated into the project. George Lakoff is incorporating much of the new NTL book into his undergraduate course and it is being widely used elsewhere. In Fall 2007, Feldman, Lakoff, and Eve Sweetser ran a graduate seminar on ECG and NTL. Again, this has added significantly to the research group's efforts. In addition to courses directly related to the group's activities, Srini Narayanan also taught the UC Berkeley undergraduate AI course (CS188) in Spring 2007.

There were also several invited talks and seminars presented by NTL members. Jerry Feldman presented the NTL project at the Computer Science Colloquium at the University of Southern California, and Srini Narayanan gave an invited seminar at the Artificial Intelligence Colloquium at the University of Texas, Austin.

4.2 Smart Search

In 2007, the ICSI AI group entered into a multi-year collaboration with Ask.com to research the use of machine learning and probabilistic modeling techniques to improve the Ask smart search results. This came as a follow-up to a small 2006 pilot project that explored the use a combination of Natural Language Processing (NLP) and machine learning techniques, many of which are already developed and in use at ICSI for general purpose NLP processing, to enhance people search and the presentation of results through the automatic extraction of key information on the person.

In 2007, the ICSI group used machine learning techniques to induce profiles from social network pages. Specifically, the inputs comprised of an unfilled registration form (that you need to fill to register), and the profile page (in html). The output was the filled-in registration form induced from the profile (what the person would have entered to result in the profile page for the specific social network site). The data (including training data) was provided by ASK for a variety of social network sites. The training data was typically around 30 examples, and the testing data was several orders of magnitude higher. The ICSI group extracted features from the html including content and meta-information (such as html tags) and used a variety of classification techniques (including SVM, Decision trees, Bayesian models) to compute the profile information. We evaluated our results against a gold-standard (built manually) for the different sites. We compared item by item the profiles extracted with the gold standard using a variety of sampling techniques and with specific guidance on extreme and important cases (provided by the sponsor).

Our results were very encouraging ($> 93\%$ accuracy) and the Ask team felt they could extend the general approach to other sorts of wrapper induction problems within their company. We have given them the training and testing code for the classifier and they are extending and evaluating the use of the technique for other problems. In 2008, John Bryant has joined the ICSI-ASK collaboration team as a PostDoctoral researcher. We are currently working on a set of problems related to smart search and query disambiguation which has the potential to make a qualitative difference in web search in cases where there are structured answer sources.

4.3 Model based Semantic Extraction

In 2007, we completed the semantic extraction pipeline, Montie, and handed over the tool to Cisco for their internal use. Our goal was to fine the the best semantic mapping from the Cisco command documents to the Cisco information model of network services and routers, Chameleon. In 2007, we extended our 2006 work both in terms of accuracy of extraction (using deep linguistic features) and modularity of the toolkit which we handed over to the Cisco network management group for internal use within Cisco. Specifically, in 2007 we

- Worked on extraction of syntactic and semantic constituents from the source (command reference) documents. Due to the novelty of the text domain (it is very different than general English), we emphasized a lot on the preprocessing of the text to get the best out the Part of Speech (POS) tagging and syntactic parsing. Once we were

able to pre-process the information, we parsed the Cisco source documents using an open source off-the-shelf parser (Collins parser). We also designed a component for computing semantic similarity at the sentence and document level using a variety of linguistic resources such as WordNet and (wherever possible) FrameNet.

- Implemented a modular software architecture and pipeline for delivery to Cisco. In 2006, the extraction pipeline, Montie, was monolithic in that a single invocation would do everything (converting input documents, extracting information, doing linguistic processing, mapping, populating the database). This made integration of the work at Cisco with the work at ICSI difficult. In 2007, Montie was modularized, and datastructures that were transient are now persisted to files. These structures, in xml form, can be manipulated by different tools. So Montie-derived features are accessible to other tools, and these tools can add and manipulate features independently of Montie. The abstraction of a FeatureBase has been introduced. This stores the features associated with an individual element. The mapping process now operates on these, and can be used whether the FeatureBase is built from Cisco or ICSI based models. As a consequence of this decoupling of the different components, it is now possible to store the component scores, then try many different methods of aggregation and compare them.

The extraction pipeline and the Montie software was handed over to the sponsors at the completion of the project in 2007.

4.4 Semantic Role Labeling with Parallel Hardware

In 2007, the ICSI AI group and the GAIA group at Lawrence Livermore National Laboratory (LLNL) completed a pilot project where we jointly designed and implemented a real-time system for automatic assignment of FrameNet semantic roles. This system uses a Support Vector Machine for classification and parts of it are implemented on graphics hardware (GPUs).

The team at ICSI included Collin Baker and Jisup Yung from the FrameNet project, and EECS graduate student, Joseph Makin. The project was supervised by Srini Narayanan. The ICSI team worked closely with researcher Heidi Kuzma from the LLNL team led by Sheila Vaidya. ICSI was responsible for the design of the semantic role assignment system, and for working with Heidi Kuzma on the implementation of the system on the GPU hardware. Specifically, the ICSI team

1. Identified the basic task (semantic role assignment) and produced data in the right format.

2. Identified the optimum set of features for SVM training,

3. Worked with LLNL to implement the kernel similarity computations on the GPU architecture.

4. Evaluated and Documented the system.

Our current results indicate upto an order of magnitude speedup using the GPU implementation compared to the most optimized SVM based systems running on CPUs. These results are for the testing phase where the training (including feature engineering, and the Quadratic Programming (QP) solver) was implemented outside the graphics architecture on CPUs. Given the pilot system testing results, we concluded that investigating the optimal implementation of these aspects in the GPU architecture can result in significant speedup the training phase as well as the testing phase. Accomplishing this next level of speedup using GPUs would move us toward deep semantic parsing with streams and other real-time information sources.

4.5 Probabilistic Models for Analysis

In 2007, the ICSI AQUINAS group collaborated with the University of Texas and others on Phase III of the AQUAINT program. We used previous research by our group within the AQUAINT program (ICSI has been involved in all three phases of the AQUAINT Question Answering project funded by DoD through ARDA and DTO). to build a QA system capable of answering questions about domain events and actions of interest to intelligence analysts. CS Graduate Student Steve Sinha used the previously designed schemata of complex actions and event structure to populate ontologies with instantiations of these schemata, and translated the entries into a form suitable for effective inference based on Coordinated Probabilistic Relational Models (CPRM) which were designed at ICSI (see 2004 and 2006 report). CPRM are designed to model, simulate, and reason about the parallel, sequential, and coordinated evolution of events in dynamic and uncertain environments.

The models built were integrated into a multi-university team team demonstration on interactive Question Answering. A full end-to-end QA system with complex scenario models based on ICSI modeling software was presented at the AQUAINT III PI meeting in September 2007 to various scientists and intelligence analysts. Our system modeled a complex information analysis scenario of utility to the Intelligence community and our demonstrations showcased the ability of the ICSI modeling framework to outperform the state of the art baseline system on causal and event related questions for the selected intelligence scenarios [7, 5].

The feedback we received was uniformly positive and led to ICSI winning a highly competitive award for a new five year IARPA project on Predictive Analysis. ICSI's role in this new project is to build models of realistic threat analysis and other intelligence related situations in order to asses the likelihood of specific events or the possibility of preventing specific outcomes. This new project is expected to be one of the major foci of the ICSI group in 2008.

4.6 Hybrid System Models of Human Blood Clotting

The process of human blood clotting involves a complex interaction of continuous-time/continuous-state processes and discrete-event/discrete-state phenomena, where the former comprise the various chemical rate equations (which can be written as differential equations) and

the latter comprise both threshold-limited behaviors and qualitative facets of the coagulation cascade. We model this process as a hybrid dynamical system, consisting of both discrete and continuous dynamics. Previous blood clotting models used only continuous dynamics and perforce addressed only portions of the coagulation cascade. The model was implemented as a hybrid Petri net, a graphical modeling language that extends ordinary Petri nets to cover continuous quantities and continuous-time flows.

In 2006, we reported on the the first known computational simulation of the complete clotting system built by graduate student Joseph (working with Srini Narayanan). In 2007, we used tools from nonlinear control theory (such as feedback linearization, output tracking, and model predictive control) and discrete event analysis (such as model checking, reachability, and verification) to compute the impact of specific therapeutic interventions to hyper and hypo coagulation disorders. Our model is able to combine multiple sources of knowledge both quantitative and qualitative in simulating disease stages and make quantitative predictions of the effectiveness of therapeutic interventions including dosage levels and frequencies [3]. Results of the work in 2007 have resulted in two new journal submissions in 2008. The model was also presented at the BioEECS faculty seminar at UC Berkeley.

The model has attracted considerable interest from the medical community and we now have now in our team clinical experts from UCSF (Dr. Mark Shuman, Dr. Tracy Manchiello, and Dr. Patrick Fogarty) examining the model for accuracy and robustness and testing the model predictions on known quantitative data on disease pathologies. We have also been contacted by different Pharmaceutical companies to use the model as part of their development toolkit. We are currently testing and refining the model, including using the model to predict clinical data on various tests (such as PTtime and APTtime). Ultimately, we hope to be able to package and make available the model as a routine diagnostic aid for hematologists.

4.7 SemEval 2007: Frame Semantic Structure Extraction Task

A major activity of the FrameNet team[1] was organizing and running a task as part of the Semeval Workshop at the ACL 2007 annual meeting, held in Prague, Czech Republic. Semeval is a successor to the Senseval workshops, and the name change reflects a change in concept from simple word sense disambiguation to representing more of the semantics of a text.

The "Frame Semantic Structure Extraction" task was designed to showcase advances in automatic semantic role labeling (ASRL) and to demonstrate that a great deal of the semantic structure of a text can be derived from a FrameNet-style annotation. It was also intended to remedy the weaknesses of the FN-related task in Senseval-2 (at ACL 2004), in which the data was all fully released before the task, so that the "answers" were known in advance, and there was no good way to evaluate whether a particular ASRL system was overfitting the data. Also, the targets and frames were pre-marked, so the first step

[1]Primarily Collin Baker, Michael Ellsworth and Jisup Hong at ICSI, with assistance from Prof. Katrin Erk of U Texas, Austin

in frame-semantic ASRL, recognizing what frame is evoked by a particular word, was not tested.

In the 2007 task, he FrameNet team annotated the gold-standard texts very thoroughly, creating lexical units, frames, and frame elements as needed, but the annotation was not released until the evaluation was over. So the 2007 task was much more complex, as it included both frame recognition and semantic role recognition. Also, (to simplify the scoring) frame elements were not considered correct unless the boundaries matched the gold standard exactly, so that parsing errors such as incorrect PP attachment could lower the score. Furthermore, unlike WSD tasks, in which the senses are drawn from a predefined set, the FN task included the possibility that words in the texts not previously covered by FN would be covered in newly-defined frames, which the ASRL systems **could not** have seen in training. Since the gold standard included lexical units that had not been seen in training but were added to existing frame, full credit was given if the participant's data postulated an LU in that frame, but partial credit was given if the participant's data placed the LU in a different but related frame. Likewise, new frames were created for these texts that were related through inheritance, etc. to existing frames; partial credit was given if the new LUs were located in a related frame (since it was impossible for the participants to guess the name of the newly created frame). This is rather daunting, but it also represents a realistic situation, one in which the system of analysis is developing along with the software for applying it.

It may be that the task was **too** daunting, for only six or seven teams downloaded the data, and only three built working systems and turned in the data for evaluation. The system that performed the best achieved an F-1 score of roughly .60-.75 (across the three test texts) on the frame recognition portion and .44-.55 on the overall task, including labeling semantic roles. Considering the difficulty of the task, we consider that to be quite good. We are especially pleased that the team that built the best system, Richard Johansson and Pierre Nugues of Lund University in Sweden, have made their system freely available.

Semantic Role Structure vs. "flat" Labeling There were two types of scoring for the SemEval task; the first one one simply measured the overlap of the participants' labels with the gold standard labels. Consider the sentence

(1) This geography is important in understanding Dublin.

In the frame-semantic analysis of this sentence, there are two predicators which FN has analyzed: *important* and *understanding*, as well as one which we have not yet analyzed, *geography*. In addition, *Dublin* is recognized by the NER system as a location. In the gold standard annotation, we have the annotation shown in (2) for the **Importance** frame, evoked by the target *important*, and the annotation shown in (3) for the **Grasp** frame, evoked by *understanding*.

(2) [Factor This geography] [Cop is] IMPORTANT [Undertaking in understanding Dublin]. [Interested_party INI]

(3) This geography is important in UNDERSTANDING [Phenomenon Dublin]. [Cognizer CNI]

The second type of scoring took advantage of a new way of representing the semantics

of the sentence, as a tree of semantic frames, which can be built up from overlapping the FN annotations and limited parse information. Since the role fillers are dependents (broadly speaking) of the predicators, the full FrameNet annotation of a sentence is roughly equivalent to a dependency parse, in which some of the arcs are labeled with role names; and a dependency graph can be derived algorithmically from FrameNet annotation; an early version of this was proposed by [19]

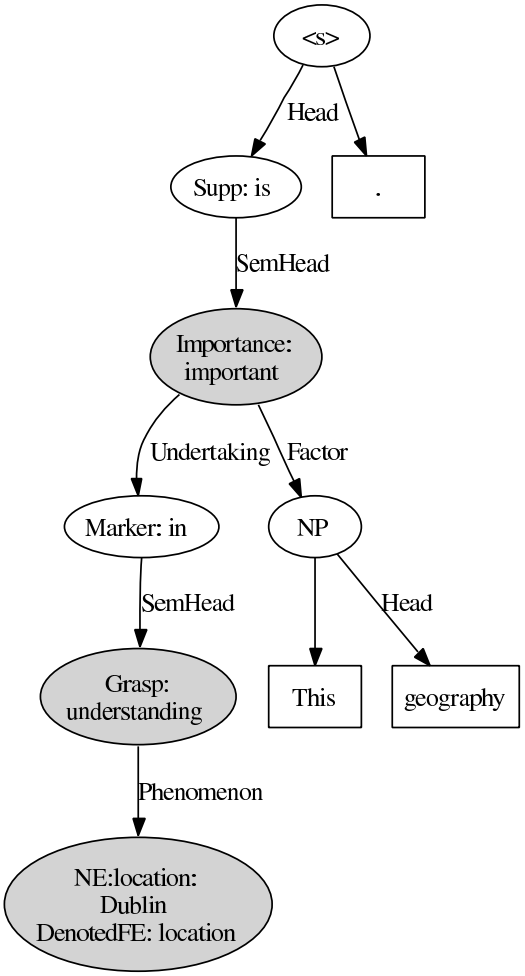

Figure 2: Semantic dependency graph

Fig. 2 on page 47 shows the semantic dependency graph derived from sentence (1). It shows that the top frame in this sentence is evoked by the word *important*, although the syntactic head is the copula *is* (here given the more general label "Support"). The labels on the arcs are either the names of frame elements or indications of which of the daughter nodes are semantic heads, which is important in some versions of the evaluation. The labels on nodes are either frame names (also colored gray), syntactic phrases types (e.g. NP), or the names of certain other syntactic "connectors", in this case, Marker and Support.

The FrameNet team created new scoring programs for both the flat label and semantic

dependency representations and made them available through the Semeval website. There were also scoring functions available with and without named entity recognition.

4.8 Refactoring of FrameNet codebase

A major accomplishment of our programmer/analyst, Jisup Hong, in 2007 has been a thorough refactoring of the existing Java code for the FrameNet desktop software, completed in early July. This was a necessary preparation for further development. We are now able to build several sections of the software independently, and to run test servers unconnected with the main applications server for development and testing.

At the same time, Mr. Hong updated the versions of Java, MySQL, and JBOSS/Enterprise Java Beans that FrameNet was using. This had previously been impossible due to bugs in the code which were fixed in the process of refactoring.

4.9 Collaboration with Adam Kilgarriff on Word Sketch Engine

In 2007 we continued to use the Word Sketch Engine (WSE) over the web as part of our vanguarding process, under license from Lexical Computing, of Brighton, U.K., headed by our consultant, Dr. Adam Kilgarriff. In December, 2006, we had also received some source code and data from Pavel Rychly of University of Brno, Czech Republic, who is lead programmer for the WSE. But after working with these materials, we realized that they constituted an earlier version of the Word Sketch Engine (WSE), without some of the features that we were using on their web server.

Dr. Kilgarriff came to Berkeley for consultation with FrameNet May 2-7, 2007. We discussed how we plan to use the Word Sketch Engine, and he suggested that we could design our vanguarding process to connect with their servers, without having a local copy. That might have been a faster way to get things running, but we were opposed to the idea that our day-to-day work would be dependent on a remote system over which we had no control. We made it clear that we needed the complete, latest source code together with documentation and e-mail support in order to integrate it into our vanguarding process. Dr. Kilgarriff agreed to provide these, and also gave us a detailed explanation of how the WSE computes its results, including how the **gramrel** tables are set up to define the grammatical relations used in the WSE tableaux in terms of sequences of part-of-speech tags. We intend to continue to use the POS sequences for the present, and to implement the use of full parses after the integration into the FN desktop is accomplished.

June 23-30, 2007, PI Baker attended the ACL in the Czech Republic and talked further with Dr. Kilgarriff and, on June 27th, met for the first time with Pavel Rychly. He promised to send us the full source code of the WSE package, as soon as a new version was complete. We received source code and binaries on August 22nd, and after some e-mail back and forth with Mr. Rychy, were able to get it running locally at ICSI. We ran into a further problem regarding the maximum size of corpus which can be encoded, which was resolved based on further advice from Mr. Rychly. As of early November, 2007, we have a local copy of the Word Sketch Engine running on a large corpus (North American Newswire from the LDC) and producing what seem to be correct results. We hope to complete an initial version of the integration into the FN Desktop software by the end of the year.

4.10 Development of the FrameNet Database

Accuracy	*accurate.a, on.prep, off.prep, inaccurate.a, exact.a, precise.a...*
Appointing	*accredit.v, appoint.v, accredited.a, designate.v, finger.v, tap.v, name.v...*
Diversity	*homogeneous.a, diversity.n, medley.n, uniform.a, assortment.n, mixed bag.n ...*
Earnings and losses	*earn.v, earnings.n, profit.n, loss.n, revenue.n, net.n, make.v, net.v, lose.v ...*
Fame	*famous.a, fame.n, stature.n, legendary.a, well known.a...*
Having or lacking access	*access.n, blocked.a, accessible.a*
Historic event	*historic.a, make history.v*
Interrupt process	*uninterrupted.a, interrupt.v, interruption.n*
Locale by event	*site.n, scene.n, venue.n, theater ((of war)).n, field.n, battlefield.n*
Nuclear process	*fissile.a, decay.v, fuse.v, fusion.n, fission.n, thermonuclear.a, radioactive.a...*
Possibilities	*possibility.n, way.n, choice.n, opportunity.n, alternative.n, option.n, future.n*
Relational political locales	*capital.n, county seat.n, see.n*
Relational quantity	*about.prep, a good.a, over.prep, almost.prep, approximately.adv, in neighborhood (of).prep, precisely.adv, at least.adv, at most.adv...*
Temporal collocation	*in.prep, on.prep, at.prep, now.adv, when.adv, modern.a, nowadays.n, no longer.adv, within.prep...*
Temporary stay	*stay.v, stay.n, lodge.v, room.v, guest.v, overnight.v, sleep over.v, stay over.v*
Version sequence	*preliminary.a, initial.a, draft.n, final.a, rough.a, working.a*
Visiting	*visitor.n, visit.v, visit.n, revisit.v, call.n*

Table 1: Examples of New Frames (and Lexical Units) Created in 2007

The FrameNet database, at the beginning of 2007, contained 828 semantic frames, covering 10,562 lexical units. 717 lexical units have been added as of Nov.13, 2007, including 269 in 59 new frames and 448 in existing frames. The recent hiring of an undergraduate annotator under an REU supplement (NSF 0737953) should speed up the vanguarding and annotation process. Examples of new lexical units defined in 2007 are shown in Tables 1 and 2.

Development of Image Schematic Frames A recent research direction has been an investigation of the frames needed to represent the image schemas described in current versions of Cognitive Linguistics. Properly representing image schemas is an essential step to being able to make the correct inferences from text. For example, in the Wikipedia sentence "...most indigenous peoples of the Americas descended from people who probably migrated from Siberia across the Bering Strait, anywhere between 9,000 and 50,000 years ago," image schemas provide much of the information needed to make the right inferences:

Age	*fresh.a*
Arriving	*influx.n*
Assessing	*reappraise.v, appraise.v, appraisal.n, reappraisal.n*
Attention	*closely.adv, close.a, keep a eye (on).a*
Contingency	*factor.n, variable.n*
Experiencer subj	*happily.adv, feverish.a, feverishly.adv*
Frequency	*never.adv, often.adv, once in a while.adv, rare.a, regularly.adv...*
Leadership	*high priest.n, diplomat.n, congressman.n, senate.n, vice president.n, mogul.n...*
Locale by use	*institute.n, pub.n, museum.n, square.n, college.n, canal.n, green.n, headquarters.n...*
Importance	*fundamental.a, main.a, serious.a, seriously.adv*
Imprisonment	*incarcerate.v, incarceration.n*
Indigenous origin	*stranger.n, non-native.a, local.n*
Indigenous origin	*stranger.n, non-native.a, local.n*
Law	*regulation.n, regime.n, policy.n*
Locative relation	*inside.prep, near.prep, between.prep, above-ground.a, opposite.prep, elsewhere.adv, here.adv, surrounding.a, distant.a, north.prep, ubiquitous.a, offshore.a...*
Luck	*luckily.adv, fortunately.adv, fortune.n*
Medical conditions	*plague.n, malnourishment.n, malnutrition.n*
Moving in place	*earthquake.n, temblor.n, quake.n*
Organization	*agency.n, union.n, brotherhood.n*
Political locales	*multinational.a, city-state.n, district.n*
Prevarication	*dissemble.v, dissembler.n, deceptive.a*
Quantity	*any.a, many.a, all.a, numerous.a, a bit.a, no.a, multiple.a, dose.n*
Relative time	*recent.a, last.a, past.a, on time.adv, punctually.adv, punctuality.n, next.a*
Stage of progress	*modern.a, cutting-edge.a, generation.n, state-of-the-art.a*
Subordinates and Superiors	*senior.a, junior.a*
Text	*list.n, manuscript.n, works.n, literature.n, history.n, brochure.n, material.n*
Undergo change	*metamorphosis.n, changeable.a, shift.n, shift.v, transition.n*
Weapon	*strategic.a, strategic nuclear weapon.n, ICBM.n*

Table 2: Examples of New Lexical Units in Existing FramesCreated in 2007

In addition to the literal image schema of Siberia and the Americas with the Bering Strait between them, and a physical movement of people from one to the other, there are two metaphorical image schemas, one for the location of this event in time between 9,000 and 50,000 years ago, and the other viewing the change from one generation to the next as a movement downward, a *decent*. The words *descend, from, across*, and *between* all evoke image schemas which combine to create most of the meaning of the sentence, and allow us to infer many facts which are not explicit: that the ancestors of the peoples of the Americas were in Siberia before 50,000 years ago, that they completed this movement by 9,000 years ago, that Siberia, the Bering Strait, and the Americas are connected linearly in that sequence, etc.

We are planning to add roughly 100 frames in this area, which will cover many common prepositions and adjectives. We expect that doing so will require some realignment of existing frames, but will also greatly improve the FN semantic representation for much of everyday language.

4.11 Other activities

Portuguese FrameNet Beginning with a visit to FrameNet in late March, 2007, by Margarida Salomão, Rector of the Universidade Federal de Juiz de Fora, in Minas Gerais, Brazil, plans have been underway for a Portuguese FrameNet. A later visit by Reginaldo Arcuri, President of the Brazilian Agency for Industrial Development led to a more general agreement for scientific and technical cooperation between ICSI as a whole and this agency; a formal Memorandum of Understanding was signed October 29th. The first project will be the collaboration with FrameNet; Margarida Salomão, will be PI of FrameNet Brazil, and is planning to start the project and ask for funding immediately. Spanish FrameNet will cooperate with Berkeley FrameNet to exchange information about our work on Spanish and Brazilian Portuguese.

Powerset In September, Michael Kaisser, a PhD student at the University of Edinburgh, Scotland, supervised by Prof Bonnie Webber spoke at ICSI on "Question Answering based on Semantic Roles", discussing both FrameNet and PropBank as possible resources for question answering. Mr. Kaisser was then doing an internship at Powerset, a San Francisco startup trying to create a search engine that uses LFG parsing on both queries and potential answers. John B. Lowe, former director of FrameNet and now an official at Powerset accompanied Mr. Kaisser, and invited the FrameNet team to visit the Powerset offices in San Francisco. Prof. Fillmore and Dr. Baker visited Powerset on Sept. 20th and spoke to a small but very attentive audience of Powerset staffers about how FrameNet might help search. The possibility of cooperation between FrameNet and Powerset is now being discussed.

Presentations Charles Fillmore: Presentation of FrameNet project to Cognitive Science 110 class, UC Berkeley, Prof. J. Feldman, instructor. February, 2007

Kyoko Ohara: "A Corpus-Based Account of Fictive Motion Sentences in Japanese

FrameNet". Presentation at 10th International Pragmatics Conference, Göteborg, Sweden. July, 2007.

Carlos Subirats gave a "Workshop on Spanish FrameNet" at the University of Juiz de Fora (Minas Gerais, Brazil). September 11-14, 2007.

References

[1] L. Barrett, J. Feldman, and L. Mac Dermed (2008) "A (Somewhat) New Solution to the Variable Binding Problem," Neural Computation, 2008

[2] B. Bergen and J. Feldman (2008) "It's the Body, Stupid: Concept Learning according to Cognitive Science," Elsevier Handbook of Embodied Cognitive Science, 2008 .

[3] J. Makin and S. Narayanan (2005) "A Hybrid System Model of Human Blood Clotting," http://www.icsi.berkeley.edu/ snarayan/clot.pdf (also ICSI TR-08-002)

[4] L. Aziz-Zadeh, C. Fiebach, S. Narayanan, J. Feldman, E. Dodge, and R. Ivry (2008) "Modulation of the FFA and PPA by language related to faces and places," Social Neuroscience, 2008.

[5] J. Scheffczyk, C. Baker, and S. Narayanan (2008) "Ontology-based Reasoning about Lexical Resources," Ontology and Lexical Resources in Natural Language Processing. Cambridge University Press, 2008.

[6] B. Bergen, T. S. Lindsay, T. Matlock, and S. Narayanan (2007) "Spatial and linguistic aspects of visual imagery in sentence processing," Cognitive Science 31 (2007), 733-764, 2007.

[7] S. Narayanan, K. Sievers, and S. Maiorano (2007) "OCCAM: Ontology-Based Computational Contextual Analysis and Modeling," Lecture Notes in Computer Science, Modeling and Using Context, Volume 4635:356-368, Springer Berlin/Heidelberg 2007.

[8] C. Baker, M. Ellsworth and K. Erk (2007) "SemEval-2007 Task 19: Frame Semantic Structure Extraction," Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007):99-104, 2007.

[9] B. Löeker-Rodman (2007) "Beyond Syntactic Valence: FrameNet Markup of Example Sentences in a Slovenian-German Online Dictionary," Proceedings of Slovko 2007, Fourth International Seminar: NLP, Computational Lexicography and Terminology, Bratislava, Slovakia, 25-27 October 2007,

[10] M. Ellsworth and A. Janin (2007) "Mutaphrase: Paraphrasing with FrameNet," Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing: 143-150, 2007.

[11] N. Chang and E. Mok (2006) "Putting Context in Constructions," The Fourth International Conference on Construction Grammar. Tokyo, Japan, 2006.

[12] N. Chang and E. Mok (2006) "A Structured Context Model for Grammar Learning," International Joint Conference on Neural Networks. Vancouver, BC, 2006

[13] E. Mok and N. Chang (2006) "Contextual Bootstrapping for Grammar Learning," 28th Annual Conference of the Cognitive Science Society. Vancouver, BC. 2006

[14] E. Mok and J. Bryant (2006) "A Best-Fit Approach to Productive Omission of Arguments," The 32nd Annual Meeting of the Berkeley Linguistics Society. Berkeley, CA, 2006.

[15] J. Feldman (2006) "From Molecule to Metaphor: A neural Theory of Language," Cambridge, MA. MIT Press, 2006.

[16] S. Petrov, L. Barrett, R. Thibeaux, and Dan Klein (2006) "Learning Accurate, Compact, and Interpretable Tree Annotation," International Conference of the Association of Computational Linguistics (ACL), 2006.

[17] S. Petrov, L. Barrett, and D. Klein (2006) "Non-Local Modeling with a Mixture of PCFGs," International Conference of Computational Natural Language Learning (CONLL), 2006.

[18] S. Atkins, M. Rundell, and H. Sato (2003) "The contribution of FrameNet to practical lexicography," International Journal of Lexicography, vol. 16, issue 3:333-357, September 2003. Editor T. Fontenelle.

[19] C. Fillmore, J. Ruppenhofer, and C. Baker (2004) "FrameNet and Representing the Link between Semantic and Syntactic Relations," Computational Linguistics and Beyond, Language and Linguistics Monographs Series B: 19-62, Institute of Linguistics, Academia Sinica Press, 2004.

[20] C. Baker, C. Fillmore and B. Cronin (2003) "The structure of the FrameNet database," International Journal of Lexicography, vol. 16, issue 3:281-296, September 2003. Editor T. Fontenelle.

[21] H. Boas (2005) "Semantic Frames as Interlingual representations for Multilingual Lexical Databases," International Journal of Lexicography, 18.4.445-478. 2005

[22] R. Cook, P. Kay, and T. Regier (2005) "The World Color Survey Database: History and Use," Cohen, Henri and Claire Lefebvre (eds.) Handbook of Categorisation in the Cognitive Sciences. Elsevier, 2005.

[23] J. Feldman and S. Narayanan (2004) "Embodied Meaning in a Neural Theory of Language," Brain and Language 89 (2004), 385-392, Elsevier Press, 2004.

[24] C. Fillmore (1976) "Frame semantics and the nature of language," In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, volume 280, 20–32, 1976.

[25] C. Fillmore, C. Baker, and H. Sato (2004) "Framenet as a "net"," Proceedings of LREC, volume 4, 1091–1094, Lisbon. ELRA, 2004.

[26] J. Hobbs and S. Narayanan (2003) "Spatial Representation and Reasoning," In Encyclopedia of Cognitive Science, Nature Publishing Group, MacMillan, London, 2003.

[27] P. Kay (2005) "Color Categories are Not Arbitrary," Cross Cultural Research(2005) 39, 72-8, 2005

[28] S. Narayanan (1997) "KARMA: Knowledge-Based Active Representations For Metaphor and Aspect," Ph.D. Dissertation, Computer Science Division, University of California, Berkeley, 1997.

[29] S. Narayanan and S. Harabagiu (2004) "Question Answering based on Semantic Structures," Proceedings of the International Conference on Computational Linguistics (COLING 2004), Geneva, August 2004.

[30] T. Regier (1996) "The Human Semantic Potential," MIT press, Cambridge, MA, 1996.

[31] S. Sinha and S. Narayanan (2005) "Model-Based Answer Selection," Textual Inference in Question Answering, AAAI 2005.

5 Speech Processing

2007's Speech efforts were headed by research staff members Dilek Hakkani-Tür, Adam Janin, Nikki Mirghafori, Nelson Morgan, Elizabeth Shriberg (ICSI and SRI), Andreas Stolcke (ICSI and SRI), and Chuck Wooters. Our work also continued to be bolstered by external collaborators such as Dan Ellis of Columbia University. Additionally, our researchers collaborated heavily with colleagues working with Hynek Hermansky and Hervé Bourlard of IDIAP, and Mari Ostendorf of the University of Washington. Other domestic and international colleagues have also played a critical role in our progress. Independent consultant George Doddington worked with the group to help formulate research directions for speaker recognition. As always, major contributions were also made by our team of students, research associates, postdoctoral Fellows, and international visitors. (see http://www.icsi.berkeley.edu/groups/speech/members.html for a current list of group members, collaborators, and alumni).

The sections below describe a number of the year's major activities in speech processing. The list of topics described is by no means exhaustive, but it should provide a useful overview of the major speech activities for the year.

5.1 Speech Recognition

In 2007 our speech recognition activity was split between work for GALE (the large DARPA project for which we were part of an SRI-led team), work on improving our meeting speech system (under support from the Swiss research network IM2 and the EU integrated project AMIDA), and more fundamental efforts in cortically-based features (funded by a combination of US Federal and European/Swiss support).

ASR for GALE In 2007 we conducted a number of experiments in the generation of discriminantly trained features, and then used our best results for the two DARPA evaluations that we were part of in the SRI-led team. In one set of experiments, we incorporated MLP outputs trained to generate articulatory feature categories; in the long run, we found these to give poor results, but our current view is that this was due to the choice of input variables and the fact that in the interest of time we used cross-language training. More promising were some results that we observed using features generated from MLPs that are "corrected" (during HMM training) to adjust for errors in the MLP phonetic classification. Additionally, we found that simple subsampling techniques permitted us to reduce MLP training time by almost an order of magnitude with little degradation in the ultimate word error rate. Finally, we worked with SRI and the University of Washington to provide the software capability to generate our current best features so that they can operate fairly independently under the pressure of DARPA evaluations.

ASR and related processing for meetings In 2007, we continued working on speech processing for meetings. Tuning of speech vs. non-speech priors in the ASR segmenter was found to dramatically reduce deletion rates, leading to a 1.7% reduction in word error rate on the NIST conference room evaluation set (which includes AMI meetings). Also,

improved speech activity detection for the Individual Headset Microphone (IHM) condition by using cross-channel delays in addition to log-energy differences.

Cortically-inspired features for speech recognition This research aims to develop, evaluate, and incorporate multi-stream spectro-temporal features for robust speech recognition. In prior work [18] [24], researchers have used 2-D Gabor filters to extract spectro-temporal features for speech recognition and speech discrimination. However, these studies have involved only single streams of task-optimized features or very large multi-dimensional representations of spectro-temporal responses. Therefore, there is a need to explore the use of multiple streams of spectro-temporal features, which may preserve the organizational map of spectro-temporal response fields and allow efficient utilization of larger feature sizes, in speech recognition.

Starting in the fall of 2007, multiple streams of spectro-temporal features, extracted using 2D Gabor filters, have been tested and tuned in digit-recognition experiments performed on the Numbers95 Corpus. The most-recent performance of the hybrid recognition system, using four streams of 506 to 529 non-task optimized spectro-temporal features each, has yielded a 5.3% word-error rate. The baselines for this experiment are the performances of the same hybrid system using a single stream of 27 perceptual-linear-prediction (PLP) features (word-error rate of 5.1task-optimized spectro-temporal features employed by Kleinschmidt [18] in his digit recognition experiments (word-error rate of 5.4the four streams of spectro-temporal features are combined with the stream of the PLP features, the performance of the hybrid system yields a word-error rate of 4.2%, a 0.9% decrease from that achieved using a single PLP-feature stream.

These initial findings suggest that multiple streams of non-task-optimized spectro-temporal features may be as robust as a single stream of task-optimized features when used in a hybrid recognition system. Furthermore, multi-stream spectro-temporal features may be used along with the more-conventional PLP features to obtain improved recognition performance. More recent results (which will be expanded upon in our 2008 report) show much more dramatic results in noise for which the system has not been trained, in which roughly 1/3 of the errors are removed.

5.2 Speaker Diarization

During the past few years, speaker diarization has largely improved in accuracy. The most successful approaches, among them the ICSI Speaker Diarization Engine, are based on agglomerative clustering. This technique, however, exhibits an inherent computational cost which makes diarization algorithms often several times slower than real-time. One of the major goals in speaker diarization research in 2007 therefore was to improve the performance of agglomerative clustering approaches in general and the ICSI Speaker Diarization engine in particular.

One of the major results was a fast-match framework to speed up the agglomerative clustering-based speaker diarization. The basic idea is to adopt a computationally cheap method to reduce the hypothesis space of the more expensive and accurate model selection via Bayesian Information Criterion (BIC). Three strategies based on high-level features

(including prosodic features), pitch-correlogram and unscented-transform based approximation of KL-divergence are used independently as a fast-match approach to select the most likely clusters to merge. The experiments were performed using the existing ICSI speaker diarization system. All three fast-match strategies speed up the diarization system without degrading the diarization performance. The best result is achieved using KL-divergence fast-match strategy, which performs only 14 % of total BIC comparisons needed in the baseline system, speeds up the system by 41 % without affecting Speaker Diarization Error Rate (DER). The result is a robust and rapid speaker diarization system.

In addition, the development of a very fast logarithm implementation, which we called ICSILog, gave us roughly another factor of two speedup. Given an IEEE 754 floating point number, the main idea is to use a quantized version of the mantissa as a pointer into a lookup table. The amount of quantization of the mantissa determines the table size and therefore the accuracy.

As part of the VACE project, the diarization system was integrated with other features (video activity, prosodics, turn-taking patterns, etc.) to classify the most dominant speaker in a meeting. For meetings in which human agreement on dominance was high, accuracy of automatic detection of dominance was also high.

However, rapid speaker diarization does not necessarily mean online diarization, which is inherently a much harder problem. For many interactive downstream multimedia/multimodal applications, online speaker diarization is desirable and more practical. Another part of the work performed in 2007 was exploring speaker anchor modeling, as an alternative to come up with speaker discriminant features. The basic idea of speaker anchor features is to construct a generic speaker space consisting of a set of speaker anchors and represent testing utterances indirectly with respective to the anchor speakers. Our initial experiments on the Timit corpus show that speaker anchor features are promising speaker discriminant features. Since this technique does not require the audio file to be batch processed, like agglomerative clustering, it might be interesting for online speaker diarization.

Overlap Detection In the past year we began work developing an overlapped speech detection system with the initial objective of improving meeting diarization. We investigated various features, with a focus on high-precision performance for use in the detector, and examined performance results on a subset of the AMI Meeting Corpus. In running these experiments we observed that the most reduction to diarization error rate (DER) came from using MFCCs, RMS energy, and diarization posteriors entropy. For the high-quality signal case of a single mixed-headset channel signal, we observed a relative improvement of about 7.4% diarization error rate (DER) over the baseline diarization system, while for the more challenging case of the single farfield channel signal the relative improvement was 3.6%. A paper describing this work was submitted and accepted to ICASSP08 [1].

The work revealed that the system is extremely sensitive to training/testing condition mismatches. To address this we have investigated the use of specific features such as the modulation spectrogram (MSG) features developed by Brian Kingsbury as well as the use of feature normalization techniques such as gaussianization with moderate success.

Another area explored for overlap detection was the use of a Tandem system using multilayer perceptron (MLP) outputs as features for the HMM segmentation system. Using

only the baseline cepstral features and deltas as inputs to the MLP, the overlap detection performance was increased from a precision of 0.54 and recall of 0.16 to a precision of 0.69 and a recall of 0.18. In addition, the diarization error rate relative improvement was 4.48significant improvement over previously obtained results.

5.3 Sentence Segmentation

Sentence segmentation from speech is part of a process that aims at enriching the unstructured stream of words output by standard speech recognizers. Its role is to find the sentence units in this stream of words. It is of particular importance for speech related applications, as most of the further processing steps, such as parsing, machine translation, information extraction, assume the presence of sentence boundaries. In 2007, we continued working on multi-lingual (English, Arabic, and Mandarin) sentence segmentation in the framework of the DARPA GALE project. We focused on semi-supervised learning methods for sentence segmentation [8], analysis of prosodic features for different languages [7] and genre [4, 3] and the effect of sentence segmentation on further processing such as machine translation [23] and information extraction [6].

Typically, statistical methods are used for sentence segmentation. However, they require significant amounts of labeled data, preparation of which is time-consuming, labor-intensive, and expensive. We investigated the application of semi-supervised learning algorithms on the sentence boundary classification problem by using lexical and prosodic information. The aim is to find an effective semi-supervised machine learning strategy when only small sets of sentence boundary-labeled data are available. We especially focused on two semi-supervised learning approaches, namely, self-training and co-training. We also compared different example selection strategies for co-training, namely, agreement and disagreement. Furthermore, we proposed another method, which is a combination of self-training and co-training. The experimental results obtained on the ICSI Meeting (MRDA) Corpus show that co-training and self-combined methods outperform self-training. Sentence segmentation is very appropriate for multi-view learning since the data sets can be represented by two disjoint and redundantly sufficient feature sets, namely, using lexical and prosodic information. Performance of the lexical and prosodic models was improved by 26% and 11% relative, respectively, when only a small set of manually labeled examples was used. When both information sources were combined, the semi-supervised learning methods resulted in 7% relative reduction in the segmentation error rate over the baseline.

Previous work on sentence segmentation combined lexical and prosodic features. This usually results in significant gain in terms of performance but can also impose significant computational challenges because of the large size of feature sets. Little is understood about which features most benefit performance, particularly for speech data from multiple languages and different speaking styles. We compared sentence segmentation for speech from broadcast news versus natural multi-party meetings, using identical lexical and prosodic feature sets across genres. Similarly, we compared sentence segmentation for English, Arabic and Mandarin with identical features. Results based on boosting and forward selection for this task showed that (1) features sets can be reduced with little or no loss in performance, and (2) the contribution of different feature types differs significantly by genre and language. We concluded that more efficient approaches to sentence segmen-

tation and similar tasks can be achieved, especially if genre and language differences are taken into account.

We also analyzed the effect of punctuating speech on machine translation and information extraction and found that careful optimization of the sentence segmentation parameters directly for the following task improves the performance of these tasks in comparison to independent optimization for segmentation quality of the predicted sentence boundaries.

We analyzed sentence segmentation performance as a function of many feature types and combination methods, as well as manual versus automatic transcription of broadcast news speech and meetings. Results showed that: (1) overall, features for broadcast news transfer well to meetings; (2) pitch and energy features perform similarly across corpora, whereas other features (duration, pause, turn-based, and lexical) show differences; and (3) the effect of speech recognition errors is remarkably stable over features types and corpora, with the exception of lexical features for meetings.

5.4 Information Distillation

Information distillation aims to extract the most useful pieces of information related to a given query from massive, possibly multilingual, audio and textual document sources. For example, given a set of multilingual audio and text sources, the purpose of distillation in this case could be to extract the biography of a person, or list arrests from a given organization during a specific time period with explanation. The participants are given a set of query templates with a variable portion. The goal of a distillation system is to output an ordered segments called *snippets* that can be considered as an answer to the query. A snippet can range from a fragment of a sentence to a paragraph. Below is an example query (in which the location and date range are variables) with some related snippets:

Query: *Describe attacks in [the Gaza Strip] giving location (as specific as possible), date, and number of dead and injured. provide information since [28 sept 2000].*

Snippets:

- *attack against a school bus filled with Israeli children*
- *The militant Islamic Jihad claimed responsibility*
- *Car hit by helicopter missile fire in Gaza Strip*

In 2007, while working on the DARPA GALE project, we focused on extending the previous year's classification based approach by using syntactic and semantic graphs (i.e., charts) [20], use of source and target language information extraction annotations for document retrieval [11, 9], and use of unsupervised methods, inspired from blind feedback in information retrieval, for information distillation [16].

A critical component in a distillation engine is detecting sentences to be extracted from each relevant document. For each sentence to assess, we employ statistical classifiers to decide if the sentence is "on template", and whether it accounts for query slots: free-text descriptions of names, organizations, topic, events, etc. that templates are centered around. We extract the features used for classification features from *charts*, compilations of elements from various annotation levels, such as word transcriptions, syntactic and semantic parses, information extraction annotations and others. In our experiments we showed

that integrating higher levels of information significantly improves algorithm's performance by about 30% relative compared to using only lexical information.

Another critical component for information distillation is document retrieval, which aims to find the documents that contain relevant snippers for a query. This year, we incorporated multi-lingual information extraction annotations to augment document retrieval for distillation. This work was motivated by the fact that some of the distillation queries can be associated with annotation elements introduced for the NIST Automatic Content Extraction (ACE) task. We experimentally showed that using the ACE events to constrain the document set returned by an information retrieval engine significantly improves the precision at various recall rates for various query templates.

We also proposed an iterative unsupervised sentence extraction method to answer open-ended natural language queries about an event, such as *Describe the facts about EVENT*, with possible definitions of *EVENT* such as *Looting of Iraqi museums after U.S. invasion*. The approach consists of finding the subset of sentences that are very likely to be relevant or irrelevant for the query from candidate documents, and iteratively training a classification model using these examples. Our results indicate that performance of the system may be improved by around 30% relative in terms of F-measure, by using the proposed method.

5.5 Social Network Analysis

Social Network Analysis (SNA) aims to study the relationships (ties) among social entities (nodes), such as communications and interactions between members of a group, nations, or organizations, as well as patterns and implications of these relationships [31]. SNA is widely studied and used in the social and behavioral sciences, economics, marketing and industrial engineering to understand the social networks for various reasons.

Previous work on social network analysis mainly focused on examining interaction patterns, and statistical co-occurrences, not considering their content. The analysis graphs are created based on who interacted with who and when. In the case of spoken interactions, speaker diarization methods have been employed to determine who talk when [30]. The SNA research which considers the content has focused on analyzing text via using information extraction (IE) methods [15]. In 2007, we initiated work on social network analysis from meetings and used content information as well as SNA based metrics, such as *centrality*, and have shown that for speaker role detection, one can benefit from using both sources of information.

5.6 Paraphrasing

In 2007, we developed a preliminary version of Mutaphrase, a system that generates paraphrases of input sentences. The algorithm generates a large number of paraphrases with a wide range of syntactic and semantic distances from the input. For example, given the input "I like eating cheese", the system outputs the syntactically distant "Eating cheese is liked by me", the semantically distant "I fear sipping juice", and thousands of other sentences. The wide range of generated paraphrases makes the algorithm ideal for a range of statistical machine learning problems (such language modeling), as well as other semantics-

dependent tasks such as query, language generation, summarization, etc. On a preliminary language modeling task, using mutaphrases reduced perplexity by up to 12%.

5.7 Sound Analysis in Real Environments: Binaural Cues, Speech Models, and Nonspeech Audio Events

A key problem is to identify the speech information available in noisy, real-world recordings made by microphones that may not be close to speakers. Two complementary approaches to this problem are (a) using spatial information derived from multiple microphones to differentiate energy coming from a particular source (speaker), and (b) exploiting prior knowledge of the behavior of speech signals to separate plausible speech information from other interference. On the spatial side, we have been developing a full probabilistic model of the way that speech signals are distorted by reverberant reflections, and how they interact in mixtures. Optimal inference allows us to identify just the reliable information in time and frequency, to associate the information with each source, as well as extracting the spatial information characterizing the position of each source. In particular, we have investigated several different assumptions about the way spatial position affects level difference in binaural recordings, and revealed a tradeoff between the resolution of the modeled spatial effect and the severity of the conditions (level of reverb, number of interfering sources) [22].

Given partial or mixed speech signals, we can estimate the full signal, or the underlying linguistic message, by using models trained from uncorrupted speech. This is of course the basis of conventional speech recognizers, but by using models that have a much finer spectral resolution than normally employed in ASR, and by adapting the recognizer to accept combinations of states, we can use these models to separate speech from interference. The success of this separation depends on how closely the model matches the particular voice being separated; since it is impractical to assume speaker-dependent models for all speakers of interest, we have used the "Eigenvoice" speaker subspace technique to build parametric, high-resolution, speaker-dependent models, and used these for separating single-channel speech mixtures [32]. We are currently looking at combining these two threads, to infer speech signals from reverberant mixtures by jointly estimating the spatial information and the speaker characteristics, then inferring the most likely original speech sequence.

In the realm of non-speech sounds, one common difficulty lies in defining an appropriate vocabulary of relevant sound events. Rather than attempting to enumerate these from scratch, we have developed a scheme for learning these directly from the audio by looking for short events that recur in very long duration recordings. We use a fingerprinting scheme based on time and frequency distances between spectral peaks to be robust to variations in channel and background, provided the structure of the sound event is strongly consistent (such as cellphone ringtones). The efficient comparisons between compact fingerprints allow us to do full searches comparing every frame in recordings of many hours with every other frame. The assumption is that identifying the audio structures that recur most frequently will both define those components as forming a whole, and end up building models for only the recurrent events that are actually present in the target recording [26].

5.8 Spoken Language Systems - SmartWeb

As part of the development of the SmartWeb 1.0 system, ICSI added new English parser rules for triggering of recommendations, and tested and revised other rules so that existing functionality continued to work properly in the SmartWeb 1.0 system. We also performed integration testing.

We used the SmartWeb English speech recognition system to investigate eliminating processing delay in vocal tract length normalization (VTLN) by using a VTLN warp factor determined from previous utterances. Our results were preliminary but promising. Our partners at Sympalog were intrigued and they plan to follow up with their own experiments.

We experimented with a novel "swing filter" approach to noise removal for speech recognition. The swing filter approach attempts to avoid the difficult problem of modeling a changing noise spectrum by exploiting differences in amplitude between the speech and the noise instead of modeling the noise spectrum. Our initial results were disappointing, perhaps because the swing filter had too harsh an effect on the speech. In the future, we would like to try using a less aggressive swing filter combined with other noise reduction approaches.

5.9 Speaker Recognition: Modeling Idiosyncrasies in Speaking Behavior

Word-conditioned HMM Supervectors for Speaker Recognition We improved upon the current Hidden Markov Model (HMM) techniques for speaker recognition by using the means of Gaussian mixture components of keyword HMM states in a support vector machine (SVM) classifier [19]. We achieved an 11% improvement over the traditional keyword HMM approach on SRE06 for the 8 conversation task, using the original set of keywords. Using an expanded set of keywords, we achieved a 4.3% EER standalone on SRE06, and a 2.6% EER in combination with a word-conditioned phone N-grams system, a GMM-based system, and the traditional keyword HMM system on SRE05+06. The latter result improves on our previous best. Our approach was inspired by Campbell et al's [2], which used the Gaussian mixture means from a GMM-based system in an SVM classifier. Unlike their approach, however, we used time-dependent acoustic feature information and applied keyword constraining.

Performance Analysis Although a great deal of progress has been made in the task of automatic speaker recognition, there are still many challenges that remain. A relatively limited amount of work has been done to find or characterize speakers who may be inherently hard to recognize. To that end, we plan to analyze automatic speaker recognition systems, with a focus on identifying what speaker characteristics make the systems perform better or worse for different speakers. By considering a range of intrinsic speaker qualities, including physical attributes, prosodic characteristics, and accents or dialects, we aim to find shared attributes among the speakers who cause the fewest or the most recognition errors.

As a starting point for the analysis, we consider a cepstral Gaussian mixture model (GMM) system, with an experiment on data from Switchboard-1. Switchboard-1 is an

appropriate corpus to begin with, for several reasons: it includes less channel variation than more recent corpora (making it easier to isolate intrinsic factors); there are existing detailed human transcriptions of the speech data; and, useful information about the speakers is available, including age, educational level, and dialect area. A model is trained for each of the almost 5000 conversations, and each model is tested against all conversations, in order to obtain a complete picture of the system performance for these speakers. Initial analysis of score distributions indicates that it is possible to identify both good- and bad-performing speakers. After such speakers have been determined, we will look for correlations with a number of attributes in order to find commonalities that can account for the differences in recognition performance. A second component of the analysis will determine the value of different speech-units in performing a speaker verification task. We will consider a range of speech-units, such as phones, diphones, syllables, words, or phrases, in order to find the most useful information for discriminating between speakers.

Furthermore, we plan to look at other types of speaker recognition systems to determine whether speakers behave in the same way for different appraoches. Additional extensions include exploring the effects of using more than one conversation to train the target speaker models, as well as analyzing data from recent NIST Speaker Recognition Evaluation corpora. Ultimately, the goal of such analyses will be to find methods for improving the performance of speaker recognition systems.

MLP-based methods We have begun using neural networks for speaker recognition based on the approach described by Niesen and Pfister in [25]. The inputs to the network are short-duration acoustic features from a test speaker and a target speaker. Output is binary – 1.0 if the test speaker is the target speaker, and 0.0 otherwise. The current system evaluates all possible combinations of frames from the two speakers, and is therefore quite slow. We are investigating methods to speed up the system by evaluating only on phonetically similar frames. Using standard ASR features and a very simple combination scheme, an equal error rate of 24% has already been achieved. Improvements using features more appropriate to speaker recognition and better combination methods are underway.

Automatic Segmentation of Laughter Audio communication contains a wealth of information in addition to spoken words. Specifically, laughter provides cues regarding the emotional state of the speaker, topic changes in a conversation, and the speaker's identity. Currently, our goal is to develop an automatic speaker recognition system which relies on features from laughter segments.

Since most speaker recognition datasets do not consistently transcribe laughter, we needed to first build an automatic laughter segmenter. Previously, we used multilayer perceptrons (MLPs) trained with short-term features (including Mel-cepstral coefficients, pitch, and energy) to compute the probability that each frame was laughter. This system had an 8% equal error rate (EER). While the EER was quite promising we found that within a laughter segment, the output probability varied more than desired causing the system to classify sequential frames as both laughter and non-laughter. In order to improve laughter segmentation, we utilized our results from the MLPs to implement the hybrid HMM/ANN (hidden Markov model/artificial neural network) system, which used

the probability that each frame was laughter calculated by the MLP (or ANN) to calculate the emission probabilities of the HMM and finally used Viterbi decoding to parse the input into laughter and non-laughter segments. This approach achieved a 4.7% false alarm rate and an 11.0% miss rate. In order to compare this system with the MLP only system we chose a point where the false alarm rates were similar. When the MLP only system had a false alarm rate of 4.3%, the miss rate was 18.3%.

The hybrid HMM/ANN system outperformed the MLP only system since the miss rate was significantly lower when the false alarm rates were approximately the same. Furthermore, the hybrid system parsed the data into segments as opposed to classifying each 10 ms frame as laughter or non-laughter. In the future we plan on comparing these results with the tandem system which uses the output of the MLP as the input features of the HMM.

5.10 Language Recognition

The ICSI 2007 language recognition system constitutes a variant of the classic PPRLM (parallel phone recognition followed by language modeling) approach which is commonly used for this task. The basic idea of PPRLM is to model the phonotactic characteristics of the languages in the test by means of a statistical language model.

The novel aspect of our approach is that we used a combination of a hidden Markov model (HMM) open loop phone recognizer (the English open-loop DECIPHER recognizer developed by SRI) and multiple in-house frame-by-frame multilayer perceptron (MLP) phone classifiers. Our version of DECIPHER uses gender-dependent 3-state hidden Markov models for open-loop phone recognition. The Markov models were trained using MFCC features of order 13 plus first and second order derivatives with overall dimensionality of 39, on the Switchboard I and II corpora. The MLP phone classifiers were built for English, Arabic, and Mandarin Chinese languages. The inputs comprised PLP features plus first and second derivatives, with a fast GMM-based estimate for vocal tract length normalization, over a local context of 9 consecutive frames. A feed-forward network structure fully connects the inputs to a large hidden layer, which is connected to output units corresponding to the phones of a language. For each frame of a test utterance, a phone label is determined as the network's output unit with the maximal activation. For English, gender-specific MLPs with 20800 hidden units were trained on 2000 hours of 8kHz conversational telephone speech, classifying 46 phones; gender was detected with GMM likelihoods. For Arabic, a gender-independent MLP with 10000 hidden units was trained on 465 hours of 16kHz broadcast news, classifying 36 phones. For Mandarin, a gender-independent MLP with 15000 hidden units was trained on 870 hours of 16kHz broadcast news, classifying 71 phones; to aid discrimination of tonal vowels, the PLP inputs were augmented with pitch-based features. For the latter two, the 8kHz samples are up-sampled to 16kHz prior to the recognition.

Taking into account a recent enhancement of the PPRLM approach on the backend , we used the n-gram counts of phones as features to train support vector machines (SVM) instead of building an actual maximum-likehood language model. Besides a more sophisticated decision function, this variant is characterized by supporting a combination of different n-grams, say bigrams and trigrams. It also supports an immediate combination

of multiple frontends on the feature level. The relative n-gram counts were first rank-normalized in order to obtain comparable ranges for all features and to map the n-gram frequencies to a uniform distribution.

The number of components of the feature vectors was reduced by only choosing 30% of the trigrams This lead to a total number of 136591 features. For each language, a one-against-all SVM with a second order polynomial kernel was trained, using the training examples of that particular language as positive examples and the training examples all other languages as negative examples. The bias which results from a larger number of negative examples was compensated by choosing an appropriate parameter cost-factor by which training errors on positive examples outweigh errors on negative ones. T-norm score normalization was applied to the scores. With T-norm, scores for a test utterances are generated against the impostor models in order to estimate the impostor score distribution.

The system was evaluated on the 2007 NIST language recognition evaluation. It obtained an average cost of 0.078 for the general language recognition task (14 languages) in the 30 seconds condition. This cost value roughly corresponds to an equal error rate (EER) of 7.8%. Although the best performing system achieved an average cost of 0.01 (1% EER), ICSI was among the best first time participants. Replacing the sub-optimal decision threshold, an average cost of 0.06 (6% EER) could be obtained in post-evaluation experiments.

5.11 Speech Coding

In partnership with colleagues at IDIAP and Qualcomm, we have been working on the development and evaluation of a new speech/audio coding method based on Frequency Domain Linear Prediction (FDLP). FDLP uses an auto-regressive model to approximate Hilbert envelopes in frequency sub-bands for relatively long temporal segments. Although the basic technique achieves good quality of the reconstructed signal, there is a need for improving the coding effi- ciency. In last year's work, our Swiss colleagues developed a novel method for the application of temporal masking to reduce the bit-rate in a FDLP based codec. ICSI'S focus in this work was putting together structured listening tests following the MUSHRA approach in order to compare the evolving coding system to more established standards, and also to begin to look at entropy coding methods to reduced the bit rate further. As of the end of 2007, we were able to demonstrate essentially identical performance with commercial approaches, and we have a number of comparatively standard tricks to still apply to reduce the bit rate below what was achieved with the temporal masking approach.

5.12 Technology and Infrastructure for Emerging Regions (TIER)

Primary activities in 2007 were the completion of beta versions of three open source software applications designed for use in multilingual, low-literacy environments. All three contain an audio component, support multilingual usage, and include an editor component for easy modification and customization. The software applications are:

- Open Spell - Educational software that targets reading and writing skills in a native

language or dialect. Based loosely on the popular game Speak and Spell, this game is easily modifiable and supports most writing systems used in the world today.

- Open Sesame - an interactive Spoken Dialog System that allows an operator to create multimedia touchscreen and speech driven applications for educational and training purposes. Open Sesame includes two example applications. The first is a demonstration in the Tamil language of agricultural techniques for growing banana crops. The second, also in Tamil, is a demonstration of drip irrigation methods.

- Open Survey - audio enabled touch screen survey and survey editor co-designed with Blanca Gordo and implemented by Richard Carlson, which is now currently in use by California Prevention and Education Project for AIDS awareness and outreach in Oakland.

The associated tasks this year include:

- Secured funding from Elsevier and AT&T Foundation. I also unsuccessfully applied for funding from International Development Research Centre, Open Society Institute, Deshpande Foundation, Ford Foundation, and Google.

- Open Sesame completed GUI editor which allows easy creation and modification of all audio/visual aspects of the tool. Editor includes a single word recognizer editor which integrates dictionary files from DictionaryMaker, an open software tool for creating pronunciation dictionaries in new languages. The recognizer editor allows you to select training data, label it, select a dictionary and simple grammar, and thereby generate a recognizer. Ported the recognizer from HTK to ATK which is supposedly better for live input. Enabled video output effectively creating the framework for a navigable video library.

- Deployment in summer of Open Sesame to MSSRF headquarters in Chennai, India. Udhaykumar Nallasamy met with content coordinator for the NGO who made several requests for the tool, which were integrated after the meeting.

- Open Spell finalized GUI editor which allows easy creation of keyboards, words, and game commands in any language with a writing system. Tested the editor on 8 native speakers of languages other than English, including languages with syllabic writing systems (Japanese, Hindi, Bengali, Tamil) and a language that write from right to left (Farsi). Bug list was kept during these test runs and those bugs were fixed. Posted software on Source Forge and requested a UI review from a colleague.

- Published a conference paper with Ozgur Cetin and Udhaykumar Nallasamy at AI in ICT for Development Workshop of the 20th International Joint Conference on Artificial Intelligence and a journal article with Udhaykumar Nallasamy in the Information Technologies and International Development journal. Unsuccessfully submitted a paper with same co-authors at ICASSP.

- Reviewed an article for Cognitive Science journal.

Other work:

- Organized several kick off meetings for the Hesperian Digital Library Project.
- Built a wiki for the Hesperian Digital Library Project.
- Researched parallel efforts and technologies in online digital book viewing.
- helped recruit and interview an experimental coordinator for continuation of the LDC mixer collection.

References

[1] K.A. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland (2008) "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," Proceedings of IEEE ICASSP Las Vegas, NV, April 2008.

[2] W.D. Campbell, D.E. Sturim, D.A.Reynolds (2006) "Support vector machines using GMM Supervectors for Speaker Verification," IEEE Signal Processing Letters, Vol. 13, pp. 308-3 2006.

[3] S. Cuendet, D. Hakkani-Tür, E. Shriberg, J. Fung, and B. Favre (2007) "Cross-Genre Feature Comparisons for Spoken Sentence Segmentation," IEEE International Conference on Semantic Computing Irvine, California, 2007.

[4] S. Cuendet, E. Shriberg, B. Favre, J. Fung, and Dilek Hakkani-Tür (2007) "An Analysis of Sentence Segmentation Features for Broadcast News, Broadcast Conversations and Meetings," Proceedings of SIGIR Workshop on Searching Conversational Spontaneous Speech Amsterdam, The Netherlands, 2007.

[5] A. Faria and N. Morgan "Corrected Tandem Features for Acoustic Model Training," Accepted for IEEE ICASSP Las Vegas, NV, April 2008.

[6] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tür, and M. Ostendorf (2008) "Punctuating Speech for Information Extraction," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Las Vegas, Nevada, 2008.

[7] J. Fung, D. Hakkani-Tür, M. Magimai-Doss, E. Shriberg, S. Cuendet, and N. Mirghafori (2007) "Prosodic Features and Feature Selection for Multi-lingual Sentence Segmentation," Proceedings of International Conference on Spoken Language Processing (Interspeech) Antwerp, Belgium, 2007.

[8] U. Guz, S. Cuendet, D. Hakkani-Tür, and G. Tur (2007) "Co-Training using Prosodic and Lexical Information for Sentence Segmentation," Proceedings of International Conference on Spoken Language Processing (Interspeech) Antwerp, Belgium, 2007.

[9] D. Hakkani-Tür, H. Ji, and R. Grishman (2007) "Using Information Extraction to Improve Cross-Lingual Document Retrieval," Proceedings of Recent Advances in Natural Language Processing (RANLP) Workshop on Multilingual Information Extraction and Summarization Bulgaria, 2007.

[10] D. Hakkani-Tür and G. Tur (2007) "Statistical Sentence Extraction for Information Distillation," Proceedings of ICASSP Honolulu, Hawaii, 2007.

[11] D. Hakkani-Tür, G. Tur, and M. Levit (2007) "Exploiting Information Extraction Annotations for Document Retrieval in Distillation Tasks," Proceedings of International Conference on Spoken Language Processing (Interspeech) Antwerp, Belgium, 2007.

[12] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez (2008) "Estimating the Dominant Person in Multi-Party Conversations Using Speaker Diarization Strategies," Accepted for IEEE ICASSP Las Vegas, NV, April 2008.

[13] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez (2007) "Using audio and video features to classify the most dominant person in meetings," Proceedings of ACM Multimedia 2007, pp. 835-838 Augsburg, Germany, September 2007.

[14] M-Y. Hwang, G. Peng, W. Wang, A. Faria, A. Heidel, and M. Ostendorf (2007) "Building a Highly Accurate Mandarin Speech Recognizer," IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 07) Kyoto, Japan, 2007

[15] H. Jing, N. Kambhatla, and S. Roukos (2007) "Extracting Social Networks and Bibliographical Facts from Conversational Speech Transcripts," Proceedings of 45th Annual Meeting of Association for Computational Linguistics (ACL) Prague, Czech Republic, 2007.

[16] K. Kamangar, D. Hakkani-Tür, G. Tur, and M. Levit (2008) "Punctuating Speech for Information Extraction," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Las Vegas, Nevada, 2008.

[17] K. Kamangar, D. Hakkani-Tur, G. Tur, and M. Levit (2008) "An Iterative Unsupervised Learning Method for Information Distillation," Accepted for IEEE ICASSP Las Vegas, NV, April 2008.

[18] M. Kleinschmidt "Localized spectro-temporal features for automatic speech recognition," Proc. Eurospeech 2003.

[19] H. Lei and N. Mirghafori (2007) "Word-conditioned phone N-grams for speaker recognition," Proceedings of ICASSP Honolulu, Hawaii, 2007.

[20] M. Levit, D. Hakkani-Tür, and Gokhan Tur (2007) "Integrating Several Annotation Layers for Statistical Information Distillation," Proceedings of IEEE 10th biannual workshop on Automatic Speech Recognition and Understanding Kyoto, Japan, 2007.

[21] M. Magimai-Doss, D. Hakkani-Tür, O. Cetin, E. Shriberg, J. Fung, and N. Mirghafori (2007) "Entropy based classifier combination for sentence segmentation," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2007.

[22] M. Mandel and D. Ellis (2007) "EM localization and separation using interaural level and phase cues," Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio WASPAA-07, Mohonk NY, pp. 275-278, October 2007. http://www.ee.columbia.edu/ dpwe/pubs/MandE07-ild.pdf

[23] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney (2007) "Improving Speech Translation with Automatic Boundary Prediction," Proceedings of International Conference on Spoken Language Processing (Interspeech) Antwerp, Belgium, 2007.

[24] N. Mesgarani, M. Slaney, and S. Shamma (2004) "Speech discrimination based on multiscale spectro-temporal features," Proc. ICASSP May, 2004.

[25] U. Niesen and B. Pfister (2004) "Speaker verification by means of ANNs.," Proceedings of ESANN '04, Bruges (Belgium), pages 145-150 April 2004.

[26] J. Ogle and D. Ellis (2007) "Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings," Proc. ICASSP-07 Hawai'i, pp.I-233-236. (4pp) http://www.ee.columbia.edu/ dpwe/pubs/OgleE07-pershash.pdf

[27] M. Plauché, O. Cetin, and N. Uhdaykumar (2007) "How to build a Spoken Dialog System with Limited (or no) Resources," ICT for Development Workshop of the Twentieth International Joint Conference on Artificial Intelligence Hyderabad, India, January, 2007.

[28] L. Stoll, J. Frankel, and N. Mirghafori (2007) "Speaker Recognition Via Nonlinear Discriminant Features," Proc. NOLISP Paris, May 2007.

[29] S. Stoyanchev, G. Tur, D. Hakkani-Tur (2008) "Name-aware Speech Recognition for Interactive Question Answering," Accepted for IEEE ICASSP Las Vegas, NV, April 2008.

[30] A. Vinciarelli, F. Fernandez, and S. Favre (2007) "Semantic Segmentation of Radio News Using Social Network Analysis and Duration Distribution Modeling," Proceedings of IEEE International Conference on Multimedia and Expo (ICME) Beijing, China, 2007.

[31] S. Wasserman and K. Faust (1994) "Social Network Analysis," Cambridge University Press 1994.

[32] R. Weiss and D. Ellis (2007) "Monaural speech separation using source-adapted models," Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio WASPAA-07, pp. 114-117 Mohonk, NY, October 2007. http://www.ee.columbia.edu/ dpwe/pubs/WeissE07-spkrs.pdf