Highlights of the First Twenty Years of Algorithms Research at ICSI

ICSI 20th Anniversary Celebration

October 17, 2008

General Goals

- Understand computational complexity, cryptography, power of randomization.
- Understand decision making under uncertainty: on-line algorithms, reinforcement learning.
- Create bridges between theory of computation and the natural sciences.
- Support other ICSI projects.

Networking Goals

- Improve reliability and speed of bulk data transfer over the Internet: Tornado codes, LT codes.
- Develop scalable content-addressable distributed networks: distributed hash tables.

Biology Goals

- Understand how genes and proteins interact to regulate cellular processes: analysis of protein interaction networks, global alignment of genomes.
- Discover the causal links between genetic variation and disease: analysis of genotypes, case-control association studies.

Historical Stages

- 1988 Algorithms group founded
- 1989-1998 Active visitor program
- 1997 Luby develops Tornado codes
- 1999 Creation of ACIRI
- 2001 Design of distributed hash tables
- 2003 Emphasis on functional genomics
- 2006 Statistical genetics research (Halperin)

How Much is it Worth to Know the Future ?

- Online algorithm: must make a sequence of decisions based on past events without knowledge of future.
- Examples: scheduling, query processing, optimization, investment, life itself.
- Competitive ratio: worst-case ratio between cost of optimal online policy and cost of optimal psychic policy.

Paging

- Slow memory contains N pages.
- Fast memory has space for k pages.
- Sequence of page requests arrives.
- Page fault: requested page not in memory.
- Goal: minimize number of page faults.
- Optimal psychic algorithm: replace the page whose next request is furthest in future.

Deterministic Paging Algorithms

- No algorithm can have competitive ratio less than k.
- Proof: adversary can always request page just evicted.
- Marking algorithm: Mark each requested page; when all pages in fast memory are marked, unmark them all; never evict a marked page. Achieve competitive ratio k.

Randomized Marking Algorithm

- Competitive ratio: is worst case ratio of *expected* cost of online algorithm to cost of psychic algorithm.
- H(k) = 1 +1/2 + 1/3 +…+1/k
- Randomized marking algorithm: evict a random unmarked page. Competitive ratio 2H(k) – 1.
- Best possible competitive ratio H(k*)***.**

Tornado and LT Codes

- Message: set of packets to be transmitted to one or many receivers.
- Erasure: failure of packet delivery.
- Goal: use minimal redundancy required redundancy to recover message despite erasures.
- Obtain very fast encoding and decoding algorithms.

Low-Density Parity Check Codes

- Use check packets to infer erased message packets.
- Code design: each check packet is XOR of several message packets.
- Simple decoding $C = X + Y + Z$

  $D = V + Z + T$

Solution process propagates through system of check equations.

..

Tornado Codes (Luby, Mitz., Spielman, Shokrollahi)

- Code viewed as random bipartite graph between message nodes and check nodes. Design of optimal degree distribution for message nodes to maximize probability of reconstructing lost message packets. Fast decoding algorithm.

LT Codes (Luby)

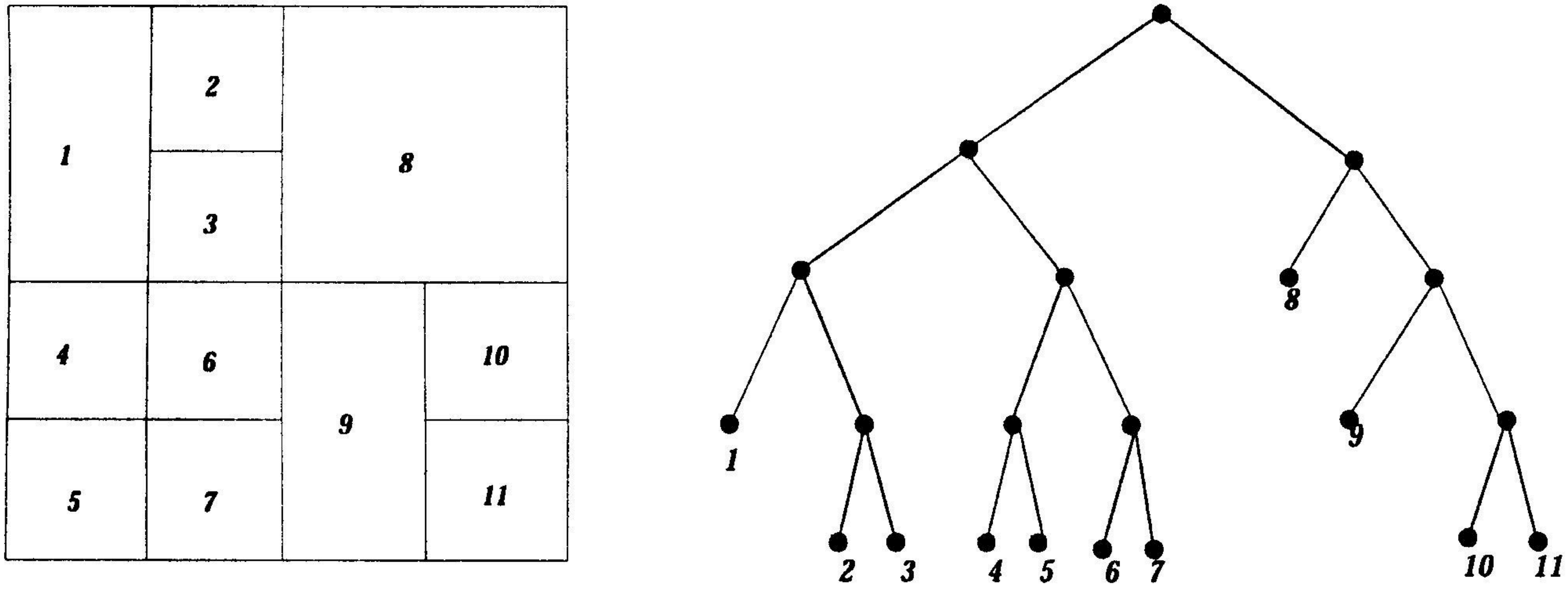
- Unlimited number of possible message nodes.
- Each message node is XOR of d random message packets, where probability distribution of d is chosen optimally.
- Digital fountain: doesn’t matter which message nodes are received provided their number is efficient.

Content-Addressable Network (CAN)

- A distributed, Internet-scale hash table.
- Scalable, fault-tolerant, self-organizing, robust, achieves low latency.
- Applications: peer-to-peer file sharing systems, large-scale storage management systems, distributed name server etc.

Design of CAN

- Virtual address space: d-dimensional box with opposite boundaries identified (as in torus).
- Entire space dynamically partitioned; each node owns a rectilinear zone and stores all records that hash to that zone.
- Overlay over Internet. Each node learns and maintains the IP addresses of nodes with neighboring zones.


2
1
3
8
4
6
10
9
5
7
11
8
1
9
2
3
4
5
6
7
10
11

Basic Operations

- Routing a message: pass it from neighbor to neighbor in direction of its hash address.
- Arrival of a new node: follow routing path to a random virtual address and split the zone containing it.
- Departure of a node: locally readjust zones using depth-first search.
- Failure recovery: each node monitors neighbors.

Refinements

- Long-range links (Kleinberg)
- Multi-dimensional coordinate spaces
- Multiple nodes sharing same zone
- Multiple coordinate spaces
- Topologically sensitive processor assignment
- Better routing metrics

Genomic Variation

- There are many differences between the genomes of any two individuals.
- We wish to study how genomic variation affects disease and other phenotypes.
- Much of this variation occurs at *polymorphic sites*, at which two different nucleotides commonly occur.

Association Studies

- Two populations: cases and controls.
- Data: each individual's pattern of genetic variations (SNPs).
- Problem: find the genetic variations indicative of disease and assess their statistical significance.
- Cases and controls may be drawn from different populations; a person's genome may be a mosaic of parts descended from different ancestral groups.

Genotypes and Haplotypes

- The human genome contains two copies of each chromosome, one received from each parent.
- A person's *genotype* specifies the pair of bases at each (polymorphic) site, but does not specify which base occurs on which chromosome.
- The *haplotypes* give the sequences of individual chromosomes.

Genotypes and Haplotypes

- Haplotypes: A T G G C
  A G C G T
- Genotype: A G/T C/G G C/T

- Binary notation :

  Haplotypes 0 1 1 1 0
  0 0 0 1 1

  Genotype 0 2 2 1 2

Determining Haplotypes

- Essential for understanding genetic variation and inheritance of complex diseases.
- *Haplotype mapping project* seeks to determine common haplotypes in several human subpopulations.
- It is cheaper experimentally to determine a person's genotype than to determine his/her haplotypes.
- *Haplotype phasing problem*: given the genotypes of a set of individuals, determine their haplotypes.

Perfect Phylogeny Assumption

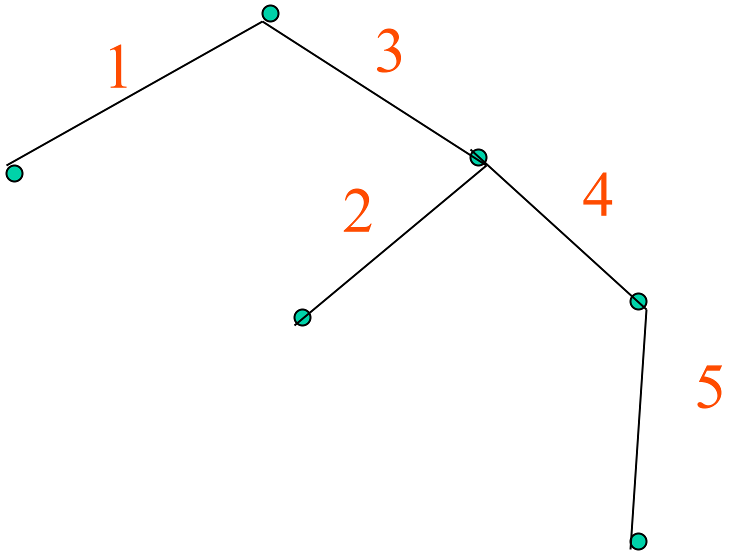
- Genome is divided into *blocks*. In each block only a small number of haplotypes occur.
- Assumption (Gusfield) : haplotypes observed within a block have evolved according to a *perfect phylogeny*, in which at most one mutation event has occurred at any site.
- In a perfect phylogeny, at most three of the four joint states of any two SNPs may occur.

Perfect Phylogeny on *m* Sites and *n* Haplotypes

- An unrooted tree with n vertices, each corresponding to a haplotype.
- Each haplotype is an element of $\{0,1\}^m$.
- Each site occurs as a label on exactly one edge (labels correspond to mutations).
- Each edge is labeled by at least one site.
- The haplotypes at the end points of an edge differ at precisely the sites labeling the edge.

A Perfect Phylogeny

- Haplotypes: 00000, 10000, 00100, 01100, 00110, 00111



Determining Haplotypes from Genotypes

- PHASE (Halperin and Eskin) A powerful and widely used program using perfect phylogeny assumption plus Occam’s razor (principle of parsimony).
- Core algorithm: fast test of whether a set of genotypes is compatible with a perfect phylogeny.
- Current research: determine haplotypes from genotypes within pedigrees.

Computational Processes in the Sciences

- Regulation of cellular processes
- Mechanisms of learning
- Immune system response
- Collective behavior of animal communities
- Molecular self-assembly
- Strategic behavior of companies
- Evolution of Web-based social networks

The Computational Lens

- In many sciences, the natural processes being studied are computational in nature.
- Complex systems such as the Web are more like natural processes than mathematically specified systems.
- Viewing natural or engineered systems through the lens of their computational requirements or capabilities provides new insights.
- Computer science is not only a science of the artificial.