Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

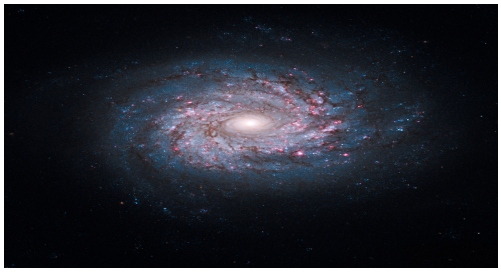
Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Introduction

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology

What to do About REALLY Big Data?



Spiral galaxy NGC 3982. Credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA)
Read more: http://www.universetoday.com/30168/galaxies/#ixzz2lgecrxox
SOURCE: Google

No Time

What can we hope to do without viewing most of the data?

Small World Phenomenon

- The social network is a graph:
  - "node" is a person
  - "edge" between people that know each other
- "6 degrees of separation"
  - are all pairs of people connected by path of distance at most 6?

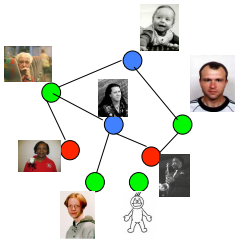


Photo & Image Source: Google

Really Big Data

- Impossible to access all of it?
- Accessible data is too enormous to be viewed by a single individual?
- Once accessed, data can change?

The Gold Standard

- linear time algorithms:
  - for inputs encoded by *n* bits/words, allow *cn* time steps (constant *c*)
- Inadequate…

What Can We Hope To Do Without Viewing Most of the Data?

- **Can't answer "for all" or "exactly" type statements:**
  - *exactly* how many individuals on earth are left-handed?
  - are *all* individuals connected by at most 6 degrees of separation?
- **Compromise?**
  - *approximately* how many individuals on earth are left-handed?
  - is there *a large group* of individuals connected by at most 6 degrees of separation?

What Types of Approximation?

- **Property Testing** – distinguish data that has a certain property from data that is *far* from having the property
- **Classical Approximation** – approximate correct output of computational problem

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Introduction

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Property Testing Algorithms

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

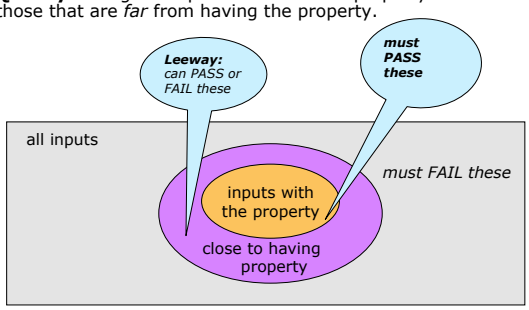
What Types of Approximation?

Property Testing

Distinguish data that *has* a certain property from data that is *far* from having the property.

Property Testing:

Quickly distinguish inputs that *have* the property from those that are *far* from having the property.



Why Property Testing?

- **Can often answer such questions much faster**
- **May be the natural question to ask**
  - *When some "noise" always present*
  - *When data constantly changing*
  - *Gives fast sanity check to rule out very "bad" inputs*
  - *Model selection problem in machine learning*

Which Properties?

- **Properties of any object, e.g.,**
  - *Functions, graphs, strings, matrices, codewords, …*
- **Examples:**
  - *clusterability, small diameter graph, increasing order, close to a codeword, linear or low degree polynomial function*
  - *Lots and lots more…*

Property Testing

- **Model must specify**
  - *representation of object and allowable queries*
  - *notion of close/far, e.g., number of bits/ words that need to be changed*

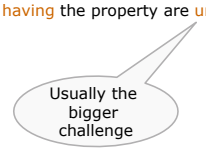
Examples

- Can test if the social network has 6 degrees of separation in CONSTANT TIME
- Can test if data is clusterable in CONSTANT TIME

[Parnas Ron][Alon Dar Parnas Ron]

Constructing a Property Tester:

- Find characterization of property that is
  - Efficiently testable
  - Robust
    - objects that have the property satisfy characterization,
    - and objects far from having the property are unlikely to PASS



Example: 6 Degrees of Separation

- **A "bad" testing characterization:**
  - For *every* node, *all* other nodes within distance 6.
- **Another bad one:**
  - For *most* nodes, *all* other nodes within distance 6.
- **Good characterization**: *[Parnas Ron]*
  - For *most* nodes, there are *many other* nodes within distance 6.

Many More Properties Studied!

- Graphs, functions, point sets, strings, …
- Amazing characterizations of problems testable in graph and function testing models in constant time!

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Property Testing Algorithms

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Sublinear Time Approximation Algorithms

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

"Classical" Approximation

- **Output number close to value of the optimal solution (not enough time to construct a solution)**
- **Some examples:**
  - Minimum spanning tree,
  - vertex cover,
  - max cut,
  - positive linear program,
  - edit distance, …

What is Close?

- ***c*-multiplicative approximation**
  - output a number that is within a multiplicative factor of *c* of the best solution
    - e.g., *output < 2 x OPTIMAL*
- ***c*-additive approximation**
  - output a number that is additively within *c* of best solution
    - e.g., *output < OPTIMAL + c*

Example: Vertex Cover

- **Given graph G(V,E), a vertex cover (VC) C is a subset of V such that it "touches" every edge.**
- **What is minimum size of a vertex cover?**
  - NP-complete
  - Polynomial time multiplicative 2-approximation based on relationship of VC and maximal matching

"Classical" Approximation Examples

- **Can get CONSTANT TIME approximation for vertex cover on sparse graphs!**
  - Output y which is at most $2 \cdot OPT + \epsilon n$
- **How?**
- **Oracle reduction framework [Parnas Ron]**
  - Construct "oracle" that tells you if node u in 2-approximate vertex cover
  - Use oracle + standard sampling to estimate size of cover

But how do you implement the oracle?

Implementing The Oracle – Two Approaches:

- Sequentially simulate computations of a fast distributed algorithm [Parnas Ron]
- Figure out what a greedy maximal matching algorithm would do on u [Nguyen Onak]

Vertex Cover & Maximal Matching

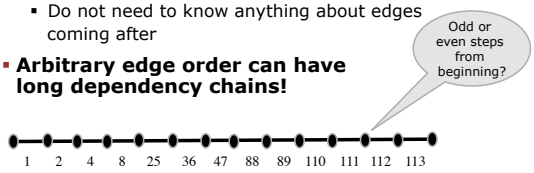
- **Maximal Matching:**
  - $M \subseteq E$ is a matching if no node is involved in more than one edge.
  - M is a maximal matching if adding any edge violates the matching property
- **Well known fact: Nodes in any maximal matching M give a pretty good Vertex Cover!**
  - That is, $|M| \leq VC \leq 2|M|$

Greedy Algorithm for Maximal Matching

- **Algorithm:**
  - Consider every edge (u,v) in arbitrary order:
    - If neither of u or v matched
      - Add (u,v) to M
  - Output M
- **Why is M maximal?**
  - If (u,v) not in M then either u or v already matched

Implementing the Oracle via Greedy

- **To decide if edge e in matching:**
  - Must know if adjacent edges that come before e in the ordering are in the matching
  - Do not need to know anything about edges coming after
- **Arbitrary edge order can have long dependency chains!**



Breaking Long Dependency Chains

- **Assign random ordering to edges**
  - Greedy works under any ordering
  - Important fact: random order has short dependency chains

[Nguyen Onak]

Better Complexity for VC

- Additional ideas yield query complexity nearly linear in average degree for general graphs [Yoshida Yamamoto Ito] [Onak Ron Rosen Rubinfeld]

Further Work

- More complicated arguments for sparse maximum matching, set cover, positive Linear Programming… [Parnas Ron + Kuhn Moscibroda Wattenhofer] [Nguyen Onak]
- Even better results for special classes of graphs [Hassidim Kelner Nguyen Onak] [Newman Sohler]
  - e.g., planar, hyperfinite

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Sublinear Time Approximation Algorithms

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Local Computation Algorithms

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

Big inputs

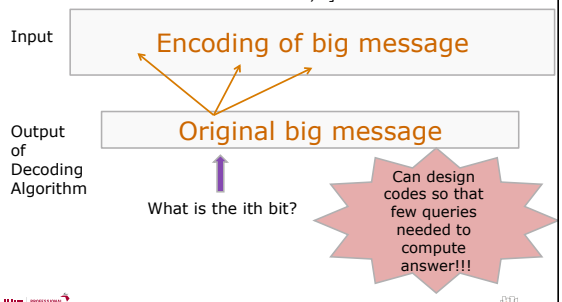
Big outputs

When We Don't Need to See All the Output…

do we need to see all the input?

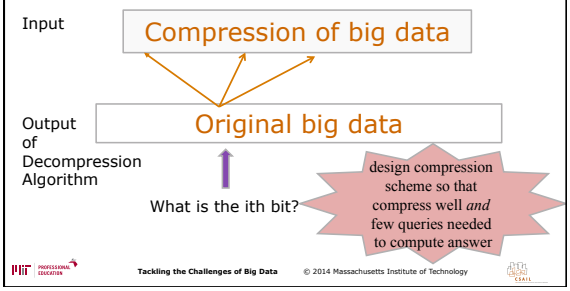
Locally (list-)Decodable Codes

[Sudan-Trevisan-Vadhan, Katz-Trevisan,…]

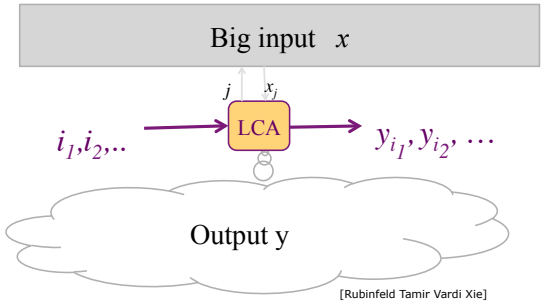


Local Decompression Algorithms

[Muthukrishnan Strauss Zheng][Chandar Shah Wornell][Sadakane Grossi] [Gonzalez Navarro] [Ferragina Venturini] [Kreft Navarro] [Billie Landau Raman Sadakane Satti Weimann] [Dutta Levi Ron Rubinfeld]



Local Computation Algorithms (LCAs)



[Rubinfeld Tamir Vardi Xie]

What More Can be Done in This Model?

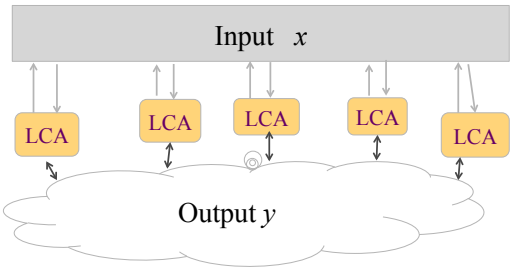
- **Optimization problems** [Rubinfeld Tamir Vardi Xie] [Alon Rubinfeld Vardi Xie]
  - Maximal independent set, certain constraint satisfaction problems,…
- **Graph sparsification** [Campagna Guo Rubinfeld] [Levi Ron Rubinfeld]

Even more:

- **Local algorithms for**
  - Ranking webpages
  - Graph partitioning

  [Andersen Borgs Chayes Hopcroft Mirrokni Teng] [Andersen Chung Lang] [Spielman Teng] [Borgs Brautbar Chayes Lucier]
- **Property preserving data reconstruction** [Ailon Chazelle Comandur Liu] [Comandur Saks] [Jha Raskhodnikova][Campagna Guo R.] [Awasthi Jha Molinaro Raskhodnikova] …

LCAs and the Cloud



Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Local Computation Algorithms

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I
Big Distributions

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

No Samples

What if data only accessible via random samples?

Distributions

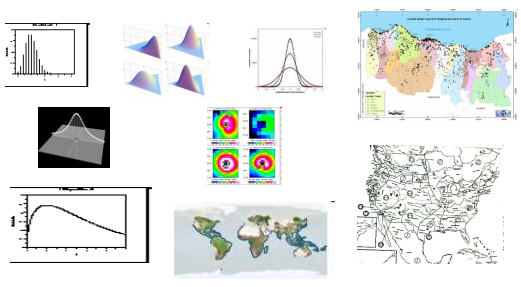


Photo & Image Source: Google

Play The Lottery?

Image Source: Google

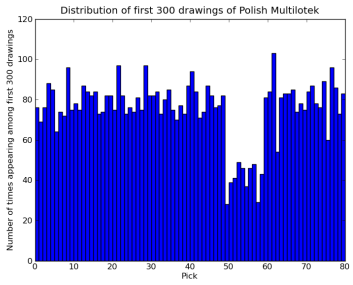
Is The Lottery Unfair?

- From Hitlotto.com: Lottery experts agree, past number histories can be the key to predicting future winners.

True Story!

- **Polish lottery Multilotek**
  - Choose "uniformly" at random distinct 20 numbers out of 1 to 80.
  - Initial machine biased
    - e.g., probability of 50-59 too small
- **Past results: http://serwis.lotto.pl:8080/archiwum/wyniki_wszystkie.php?id_gra=2**

Image: Lotto, Multi Multi, Mini Lotto, Twój Szczęśliwy Numerek ...www.wynikilotto.mortin.pl
Image Source: Google



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)

New Jersey Pick 3, 4 Lottery

- **New Jersey Pick k (=3,4) Lottery.**
  - Pick k digits in order.
  - 10^k possible values.
  - Assume lottery draws iid
- **Data:**
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test gives 42% confidence
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.
    - fewer results than possible values
    - χ^2-test gives no confidence

Distributions on BIG domains

- **Given samples of a distribution, need to know, e.g.,**
  - entropy
  - number of distinct elements
  - "shape" (monotone, bimodal,…)
  - closeness to uniform, Gaussian, Zipfian…
- **No assumptions on shape of distribution**
  - i.e., smoothness, monotonicity, Normal distribution,…

Distributions on BIG domains, ctd.

- **Considered in statistics, information theory, machine learning, databases, algorithms, physics, biology,…**

Key Question

- **How many samples do you need in terms of domain size?**
  - Do you need to estimate the probabilities of each domain item?
  - Can sample complexity be sublinear in size of the domain?

Rules out standard statistical techniques, learning distribution

Our Aim:

Algorithms with ***sublinear*** sample complexity

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I
Big Distributions

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I:
Similarity of Big Distributions?

Ronitt Rubinfeld
Professor
Massachusetts Institute of Technology

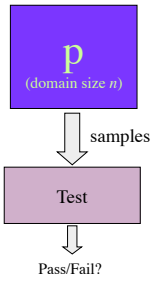
Our Aim:

Algorithms for understanding distributions with ***sublinear*** sample complexity

Similarities of Distributions

- **Are two distributions *p* and *q close* or *far*?**
  - *p* is given via samples
  - *q* is either
    - known to the tester (e.g. uniform)
    - given via samples

Is p uniform?



Theorem: ([Goldreich Ron] [Batu Fortnow Rubinfeld Smith White] [Paninski]) Sample complexity of distinguishing

$p=U$

from $||p-U||1>$ ☒ is ☒$(n\uparrow 1/2)$

Upper Bound for L_2 Distance

- L_2 distance:

$||p-U||_2^2 = \Sigma_{i\in[1..n]} (p_i - 1/n)^2$

$= \Sigma_i p_i^2 - 2\Sigma p_i/n + \Sigma 1/n^2$

$= \Sigma p_i^2 - 1/n$

- Estimate collision probability to estimate L_2 distance from uniform

[Goldreich Ron]

Testing Uniformity

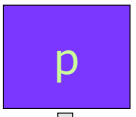
- **Upper bound: Estimate collision probability and use known relation between L1 and L2 norms**
  - Issues:
    - Collision probability of uniform is 1/n
    - Use O(sqrt(n)) samples via recycling
  - Comment: [Paninski] uses different estimator
- **Easy lower bound: $\Omega(n^{1/2})$ samples needed**
  - Can improve to $\Omega(n^{1/2}/\varepsilon^2)$ [P]

[Goldreich Ron, Batu Fortnow Rubinfeld Smith White]

Back to The Lottery…

plenty of samples!

Is p Uniform?



samples

Test

Pass/Fail?

Theorem: Sample complexity of distinguishing

$p=U$

from $||p-U||_1>\varepsilon$ is $\theta(n^{1/2})$

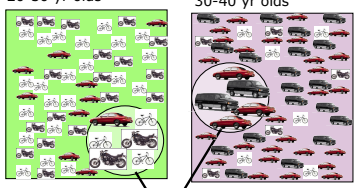
[Goldreich Ron][Batu Fortnow Rubinfeld Smith White] [Paninski]

Nearly same complexity to test if p is any known distribution
[Batu Fischer Fortnow Kumar Rubinfeld White]

Testing Closeness of Two Distributions

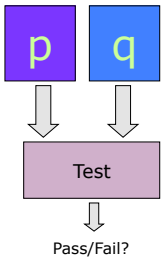
Transactions of 20-30 yr olds

Transactions of 30-40 yr olds



trend change?

Testing Closeness



Pass/Fail?

Theorem: Sample complexity of distinguishing

$p=q$

from $||p-q||_1 >\epsilon$

is $\theta(n^{2/3})$

[BFRSW] [P. Valiant][Chan Diakonikolas Valiant Valiant]

Why So Different?

- Collision statistics are all that matter
- Collisions on "heavy" elements can hide collision statistics of rest of the domain

Additively Estimate Distance?

New goal: Output $||p-q||_1 \pm \epsilon$

much harder, but still sublinear!

need $\theta(n/\log n)$ samples [G. Valiant P. Valiant]

Collisions Tell All

- **Algorithms:**
  - Algorithms use collisions to determine "wrong" behavior
    - E.g., too many collisions implies far from uniform [GR,BFSRW]
  - Use Linear Programming to determine if there is a distribution with the right collision probabilities and the right property [G. Valiant P. Valiant]

Collisions Tell All

- **Lower bounds:**
  - For symmetric properties, collision statistics are only relevant information [BFRSW] (see also [Orlitsky Santhanam Zhang] [Orlitsky Santhanam Viswanthan Zhang])
  - Need new analysis tools since not independent
    - Central limit theorem for generalized multinomial distributions [G. Valiant P. Valiant]

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I
Big Distributions

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I
Big Distributions – Final Words

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology

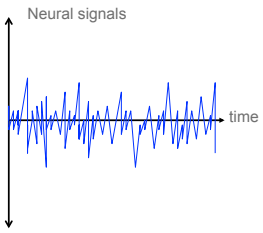
Information Theoretic Quantities

Entropy

Support size

Information in Neural Spike Trails

[Strong, Koberle, de Ruyter van Steveninck, Bialek '98]



- Each application of stimuli gives sample of signal (spike trail)
- Entropy of (discretized) signal indicates which neurons respond to stimuli

Compressibility of Data

GPS Tracking Reduces Illegal Garbage Dumping | GPS Tracking Blogtrackingtheworld.com

Image source: Google

Can We Get Multiplicative Approximations?

- **In general, no….**
- **What if entropy is at least some constant?**
  - Can γ-multiplicatively approximate the entropy with $\tilde{O}(n^{1/\gamma 2})$ samples (when entropy $>2g/\epsilon$) [Batu Dasgupta Rubinfeld Kumar]
  - requires $\Omega(n^{1/\gamma 2})$ [Valiant]
  - better bounds when support size is small [Brautbar Samorodnitsky]
  - Similar bounds for estimating support size [Raskhodikova Ron Rubinfeld Smith] [Raskhodnikova Ron Shpilka Smith]

Additive Approximations for Entropy and Support Size

- need $\theta(n/\log n)$ samples [Raskhodnikova, Ron, Shpilka, Smith] [Valiant Valiant]

More Properties:

- Independence and Limited Independence [Batu Fischer Fortnow Kumar Rubinfeld White] [Alon Andoni Kaufman Matulef R. Xie] [Haviv Langberg]
- K-histogram distributions [Levi Indyk Rubinfeld]
- K-modal distributions [Daskalakis Diakonikolas Servedio]
- Poisson Binomial Distributions [Daskalakis Diakonikolas Servedio]
- Monotonicity over general posets [Batu Kumar Rubinfeld] [Bhattacharyya Fischer Rubinfeld P. Valiant]
- Properties of multiple distributions [Levi Ron Rubinfeld]

Conclusion

- **For many problems, we need a lot less time and samples than one might think!**
- **Many cool ideas and techniques have been developed**
- **Lots more to do!**

Tackling the Challenges of Big Data

Big Data Analytics

Fast Algorithms I

Big Distributions – Final Words

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics
Fast Algorithms I

Ronitt Rubinfeld

Professor

Massachusetts Institute of Technology