

Research Statement

Tuğkan Batu

Technological advances of the last decades such as cheap computing power, high-volume data storage, and the omnipresence of networking made it possible to collect and store enormous amounts of data. Some examples of massive data sets include banking and sales transactions, stock-trading data, network-traffic and web-server logs, phone-call data, data collected by scientific experiments or sensory devices such as satellites, and genomic data. In particular, big department-store chains report to have accumulated transaction databases of sizes measured in tens of terabytes. The objective of utilizing and extracting information from data of this magnitude has lead researchers to explore what we can compute very efficiently, specifically, in sublinear time in the size of the input.

Various models of computation in which the time and space resources of the algorithm are sublinear have been proposed. The common theme and the main difficulty facing the algorithms in these models is the inability to read, or, at least, remember the entire input. Recent results have shown that one can solve with such sublinear constraints many interesting problems beyond those that straightforward statistical sampling techniques can solve. For instance, combinatorial optimization problems such as maximum cut and minimum spanning tree can be approximated in sublinear time. However, the study of computation with such severe constraints is a new and emerging field of study, and much remains to be explored.

I am interested in the design and analysis of algorithms, especially under sublinear time and space constraints. In particular, my research focus is on property testing, statistical analysis of data, string algorithms, and streaming data models. I am also interested in the complexity theoretic aspects of these models.

Sublinear-time algorithms for statistical data analysis

Many problems that arise in data mining, machine learning applications, and analysis of experimental data in scientific fields such as physics and biology require that we understand global statistical properties of data. Such global properties may include similarities, correlations, or the randomness of data. For this class of problems, one critical parameter that affects the efficiency of the algorithm is the size of the domain over which the data values are distributed. Standard statistical techniques perform poorly when the data is distributed over a large domain. For example, we can compare the age distributions in the population twenty years ago and today by inspecting a small number of samples from the respective distributions. On the other hand, a similar comparison of the distributions of the population on the postal codes twenty years ago and today would require many more samples.

Traditionally-used methods for such analysis include standard statistical tests such as χ^2 -test, normalized frequencies (what is often called plug-in estimates), and methods that rely on “learning” the distribution (e.g., through the use of Chernoff bounds). These methods require that the available data represents every “reasonably-likely” event sufficiently often to make reliable inferences. When the data takes values from a large domain, such a requirement can translate into a daunting linear (or, even superlinear) sample complexity in the domain size. In recent years, our algorithmic perspective has lead to efficient solutions to some problems that are studied traditionally in other fields.

SIMILARITY TESTS. Consider the following three questions, arguably central to reasoning about data sources: (i) Given data originating from a single source, is the data generated according to

a given specification? For example, given information on the flu cases from the last year, how much data do we need to discover whether the distribution this year is similar? (ii) When the data is emitted by two separate sources, do these sources generate data identically, or at least, similarly? For example, one might want to know if the destination addresses of IP packets going through two separate routers are similarly distributed? (iii) For data with multiple components, are certain components of the data independent? For instance, one might want to demonstrate that the cases of diseases and the postal codes of the patients are uncorrelated.

In two papers [5, 4], my colleagues and I have given procedures with sublinear sample complexity for the tasks mentioned above (e.g., the procedure for (ii) requires roughly $O(n^{2/3})$ samples for a domain of size n). Moreover, we have given information theoretic lower bounds that demonstrate near-optimality of our procedures.

ENTROPY ESTIMATION. One interesting attribute of a distribution is the amount of randomness or information content, which is formalized by Shannon's entropy function. The entropy of a source is also a natural measure of complexity, compressibility, and predictivity. Due to its various interpretations, estimating the entropy of a distribution from samples finds applications in many fields. For example, neurobiologists commonly use entropy estimation techniques to analyze the information transmitted by neural signals. Strong et al.¹ and Ma² use techniques with sublinear sample complexity to provide lower bounds on the entropy of a distribution from samples. In [1], we have presented an algorithm for estimating the entropy of a distribution from samples with relative error and sublinear sample complexity. In the same paper, we have also studied entropy estimation in several other models, corresponding to different modes of access to the data. In all these models, we have provided entropy-estimation procedures with sublinear sample complexity and information theoretic lower bounds for such procedures.

MONOTONE AND UNIMODAL DISTRIBUTIONS. One interesting direction of research is characterizing commonly-encountered classes of distributions that yield to efficient procedures for tasks such as above. Classical statistics literature contains tests on samples from the Gaussian distributions. Unimodal distributions (i.e., the distributions with a single peak such as the Gaussian distribution) in general embody many common distributions. Starting with [1], we have been exploring reasonable assumptions on the input distribution to yield more efficient procedures. In a recent work [6], we have shown that, in addition to the entropy estimation, similarity tests mentioned above can be performed with only polylogarithmic sample complexity for the class of unimodal distributions, which is almost an exponential gain over the general case.

FUTURE DIRECTIONS. A multitude of statistical properties are yet to be explored with this fresh perspective. My research studied and resolved some questions relating to one common similarity measure, namely, statistical variation distance. However, many other similarity measures between distributions are proposed, and different measures suit certain contexts or applications. For example, in information theoretic or learning theory settings, relative entropy (i.e., KL-divergence) is often the similarity measure of choice. Extending our studies to other similarity measures would benefit these applications.

Another example of an interesting statistical property is the mutual information between random variables. Mutual information is related to the correlations of random variables, and it is a quantity estimation of which would be a valuable data analysis tool. Our techniques for entropy estimation may be used for estimating the mutual information as well as other information theoretic quantities of interest such as Rényi entropy. In general, characterizing properties of dis-

¹S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80:197–200, 1998.

²S. K. Ma. *J. Stat. Physics.* 26, 221, 1981.

tributions that admit sublinear-time tests is an intriguing direction of research.

In addition to devising new procedures for other statistical properties, I am also interested in investigating how these procedures perform on real data. We have provided proven guarantees on the sample complexity and the error probability of our tests. However, it would be of great value, both practically and theoretically, to observe whether any stronger guarantees seem to hold when real data is used.

Efficient algorithms for string similarity computations

Many of the fundamental string problems are based on string similarity. String similarity problems arise in disguise in many fields such as databases, informational retrieval, and genome analysis: for example, sequence comparisons or similar document searches. Perhaps the most common measure of string similarity is the edit distance and its variants. Computing or approximating the edit distance between two strings along with the related string matching problems are of particular interest. However, when the input data is as large as a genomic sequence, standard solutions can be prohibitively expensive. Two papers of mine address this issue and give very efficient approximation schemes with tradeoffs for different ranges of running time: the first one is an algorithm that runs in linear-time (or more if further precision is desired) and the second one is a sublinear algorithm with a weaker approximation guarantee.

LINEAR-TIME EDIT DISTANCE APPROXIMATION. The main result in my most recent publication [3] was an approximation algorithm to the edit distance between two input strings of length n . We have presented an algorithm that can approximate the edit distance within a factor of nearly $n^{1/3}$ in almost linear time. We have also provided a tradeoff between the approximation quality and the running time, where we can get an n^ϵ approximation for an arbitrary constant $\epsilon > 0$ while increasing the running time. This is the best known algorithm for edit distance approximation. The main tool used in this algorithm is an embedding of strings into shorter strings over a larger alphabet that approximately (and almost optimally) preserves the edit distance.

SUBLINEAR-TIME EDIT DISTANCE APPROXIMATION. Previously in [2], we have presented a sublinear-time approximation algorithm for the edit distance. Our work has provided a procedure that, after reading only sublinear size portions of two strings, can distinguish pairs of strings that can be derived from each other with a small number of edit operations from pairs of strings that require considerably more edit operations. This procedure can be viewed as providing a weak approximation to the edit distance between two strings.

A classification task based on edit-distance over a database of strings can employ our procedure to reduce the workload in a couple of ways. When there are many pairs of strings to be compared, only a sublinear-time algorithm might be feasible. Even when exact computations or stronger approximation guarantees are needed, our procedure can be used to partition the database into groups of similar strings. Then, exact computations can be performed only on similar pairs. This general framework is also one of the common motivations for studying sublinear-time algorithms.

FUTURE DIRECTIONS. Improving the approximation guarantees of the algorithms mentioned above, hence, their practical value, is an immediate goal to be pursued. Certain variants of the edit distance such as the edit distance with affine gaps are considered to be better similarity metrics in certain contexts. Algorithms for approximating these metrics as well as other sequence analysis problems can immensely help fields such as computational biology, where efficient algorithms are particularly needed.

References

- [1] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.
- [2] Tuğkan Batu, Funda Ergün, Joe Kilian, Avner Magen, Sofya Raskhodnikova, Ronitt Rubinfeld, and Rahul Sami. A sublinear algorithm for weakly approximating edit distance. In *Proceedings of 35th Symposium on Theory of Computing*, 2003.
- [3] Tuğkan Batu, Funda Ergun, and Cenk Sahinalp. Oblivious string embeddings and edit distance approximations. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms*, 2006.
- [4] Tuğkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.
- [5] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of 41th IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.
- [6] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of 36th ACM Symposium on Theory of Computing*, pages 381–390, 2004.