

# SHUHENG ZHOU

## ANNOTATED BIBLIOGRAPHY

### High-dimensional Statistical Inference

S. ZHOU, J. LAFFERTY AND L. WASSERMAN, Compressed Regression, in *Advances in Neural Information Processing Systems 20* (NIPS 2007). Full version submitted to IEEE Trans. on Information Theory.

Recent research has studied the role of sparsity in high dimensional regression and signal reconstruction, establishing theoretical limits for recovering sparse models from sparse data. This line of work shows that  $\ell_1$ -regularized least squares regression can accurately estimate a sparse linear model from  $n$  noisy examples in  $p$  dimensions, even if  $p$  is much larger than  $n$ . In this paper we study a variant of this problem where the original  $n$  input variables are compressed by a random linear transformation to  $m \ll n$  examples in  $p$  dimensions, and establish conditions under which a sparse linear model can be successfully recovered from the compressed data. A primary motivation for this compression procedure is to anonymize the data and preserve privacy by revealing little information about the original data. We characterize the number of random projections that are required for  $\ell_1$ -regularized compressed regression to identify the nonzero coefficients in the true model with probability approaching one, a property called “sparsistence.” In addition, we show that  $\ell_1$ -regularized compressed regression asymptotically predicts as well as an oracle linear model, a property called “persistence.” Finally, we characterize the privacy properties of the compression procedure in information-theoretic terms, establishing upper bounds on the mutual information between the compressed and uncompressed data that decay to zero.

S. ZHOU, J. LAFFERTY AND L. WASSERMAN, Time Varying Undirected Graphs, *Manuscript*, December, 2007

Undirected graphs are often used to describe high dimensional distributions. Under sparsity conditions, the graph  $G$  can be estimated using  $L_1$  penalization methods. However, all the current methods assume that the data are independent and identically distributed. If the distribution—and hence the graph—evolves over time, then the data are no longer identically distributed. In this paper, we show how to estimate the sequence of graphs for non-identically distributed data and establish theoretical results.

### Optimization Problems in Network Routing

S. RAO AND S. ZHOU, Edge Disjoint Paths in Moderately Connected Graphs, in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*. †<sup>1</sup>.

---

<sup>1</sup>Authors of papers marked with † are ordered alphabetically, following theoretical computer science convention.

We study the Edge Disjoint Paths (EDP) problem in undirected graphs: Given a graph  $G$  with  $n$  nodes and a set  $\mathcal{T}$  of pairs of terminals, connect as many terminal pairs as possible using paths that are mutually edge disjoint. This leads to a variety of classic NP-complete problems when the number of pairs in  $\mathcal{T}$  is a variable, for which approximability is not well understood. We show a poly-logarithmic approximation algorithm for the undirected EDP problem in general graphs with a moderate restriction on graph connectivity; we require the global minimum cut of  $G$  to be  $\Omega(\log^5 n)$ , where  $n$  denotes the number of vertices in  $G$ . Before our work, the only classes of undirected graphs for which constant or poly-logarithmic approximation algorithms were known were the classes of trees with parallel edges, expanders, grids and grid-like graphs, and most recently, even-degree planar graphs. These graphs either have special structure (e.g., they exclude minors) or there are large number of short disjoint paths. Our algorithm uses techniques that significantly extend the previous ideas, and apply to a large class of graphs; for example, our graphs can have high diameters and contain very large minors.

## Hierarchical Routing and Hierarchical Decompositions

A. GUPTA, B. MAGGS, AND S. ZHOU, On Hierarchical routing in Doubling Metrics, in *Carnegie Mellon University, Technical Report CMU-PDL-04-106*, December 2004. †

In a seminal work, Kleinrock and Kamoun (1977) described a hierarchical routing scheme based on an “optimal” hierarchical clustering model of nodes in the network. They further showed that for a class of large distributed networks, by following their routing scheme, it is possible to achieve a substantial routing table reduction with essentially no increase in the *average path length* over all source-destination pairs in the network. Essentially, the family of networks upon which it is possible to apply such an “optimal” hierarchical clustering scheme satisfies certain growth properties.

We show the construction of a novel hierarchical decomposition algorithm for any network  $G$ , whose shortest path distances  $d$  induce the doubling metric  $(X, d)$ , and specify how to route hierarchically given such decompositions. Note that a metric  $(X, d)$  has *doubling dimension*  $\dim(X)$  at most  $\alpha$  if every set of diameter  $D$  can be covered by  $2^\alpha$  sets of diameter  $D/2$ . A *doubling metric* is one whose doubling dimension  $\dim(X)$  is a constant. Following this routing scheme, the path length between every pair of nodes is at most  $(1 + \tau)$  times their shortest path length for any  $1 > \tau > 0$ , with only  $(\frac{\alpha}{\tau})^{O(\alpha)} \log \Delta \log \delta$  bits of routing information stored at each node, where  $\Delta$  and  $\delta$  denote the diameter and the maximum degree of  $G$  respectively.

Besides the description of the hierarchical decomposition and the routing scheme that specifies addressing, the routing tables and the forwarding algorithm, as appeared in the SODA05 paper, we also describe a modified version of the distributed Bellman-Ford algorithm that we use to build and perform dynamic routing table updates once nodes in the network have been assigned addresses given a decomposition, and characterize the hierarchical paths that network messages have to follow in the  $(1 + \tau)$ -stretch routing scheme.

T-H. CHAN, A. GUPTA, B. MAGGS AND S. ZHOU, On Hierarchical routing in Doubling Metrics, in *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*. † <sup>2</sup>

In addition to the result on routing as annotated in CMU technical report CMU-PDL-04-106, this paper also gives better constructions of sparse *spanners* for doubling metrics than those obtained from the routing tables above; for  $\tau > 0$ , we give algorithms to construct  $(1 + \tau)$ -stretch spanners for a metric  $(X, d)$  with maximum degree at most  $(2 + 1/\tau)^{O(\dim(X))}$ , matching the results of Das et al. for Euclidean metrics.

## Population Classification

SHUHENG ZHOU, Routing, Disjoint Paths and Classification. *Ph.D. Thesis*, Carnegie Mellon University, Technical Report CMU-PDL-06-109, August 2006. (This thesis contains results from three published papers: SODA05, ICALP06, and SODA07. In addition, it contains the following result by myself, whose proof is under preparation for submission and has not appeared in any other publication.)

We study the problem of partitioning a small sample of  $n$  individuals from a mixture, into  $k$  populations according to their population of origin based on their genotype data. Genotype contains a series of markers called SNPs (Single Nucleotide Polymorphisms), where each SNP has two variants (alleles). Population is described by a set of allele frequencies at each locus. SNPs have slightly different frequencies depending on which population they belong to. Given two populations, we use  $\gamma$  to denote the average  $\ell_2^2$  distance in allele frequencies for the chosen SNPs which measures the statistical divergence between them. We study the case assuming that SNPs are independently distributed across different loci, and an individual's genotype is generated by drawing alleles independently according to its population frequency distribution, following a parametric model as in Pritchard, Stephens and Donnelly (2000). Given a single bit from each SNP, this fits nicely with the statistical model of learning mixtures of  $k$  product distributions over Boolean cube  $\{0, 1\}^K$ , given a set of  $K$  SNPs.

This work demonstrates that, for a balanced input instance for  $k = 2$ , a certain graph-based optimization function returns the correct partition with high probability, where a weighted graph  $G$  is formed over  $n$  individuals, whose pairwise hamming distances between their corresponding bit vectors define the edge weights, so long as  $K = \Omega(\ln n/\gamma)$  and  $Kn = \tilde{\Omega}(\ln n/\gamma^2)$ . The function computes a maximum-weight balanced cut of  $G$ , where the weight of a cut is the sum of the weights across all edges in the cut. This result demonstrates a nice property in the high-dimensional feature space: one can trade off the number of features that are required with the size of the sample for certain tasks like clustering. Exploiting this tradeoff is of both theoretical and practical interests, given that whole genome sample are still scarce.

---

<sup>2</sup>This paper is a merge of two results: routing and spanners, for which S. Zhou and T-H. Chan worked independently, with the joint set of faculty. A full version on routing appears as the CMU tech report CMU-PDL-04-106.

K. CHAUDHURI, E. HALPERIN, S. RAO, AND S. ZHOU, Rigorous Analysis of Population Stratification with Limited Data, in *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms* (SODA 2007).<sup>†</sup>

Finding the genetic factors of complex diseases such as cancer, currently a major effort of the international community, will potentially lead to better treatment of these diseases. One of the major difficulties in these studies, is the fact that the genetic components of an individual not only depend on the disease, but also on its population of origin. Therefore, it is crucial to find methods that could reduce the population structure effects on these studies. This can be seen as a clustering problem, where the individuals are clustered according to their genotype data. The result addresses the clustering problem as described above using two bits from each SNP. In this paper, we present methods that can cluster individuals while attempting to optimize the number of features required. While we present some algorithms in this paper, we focus on showing that optimizing certain functions on the data yields the right clustering, under the random generative model.

We first show that when  $K > \Omega(\ln n/\gamma^2)$ , given two bits from each position, one can always classify for any size of  $n$ , for imbalanced cases with any number of mixtures, using essentially connected components algorithm on a certain weighted graph formed among individuals. The cleverness here is the construction of a diploid score at each locus, given any *pair of individuals*, which exploits the fact that every SNP in diploid organisms (like human beings) contains two bits that are drawn from the same distribution under no admixture model. In expectation, diploid scores are higher among pairs from different groups, than for pairs in the same group across all loci; Suppose we form a weighted graph  $G$  over the  $n$  individuals, whose pairwise distance as the sum of diploid scores at each locus defines the edge weights. Thus with high probability, weights on all edges must obey this ordering given a sufficient number of SNPs.

We next show a result similar to that in CMU Technical Report CMU-PDL-06-109 using a max-cut function computed over diploid scores, with the key difference in that this score requires two random bits at each locus.

The second approach explicitly estimates feature probabilities while clustering. Here, we show that the partition that maximizes the sum of the differences of the estimated probabilities yields the correct clusters when the number of features per individual is  $\Omega(\log N/\alpha^2)$ , where  $\alpha$  is the average absolute difference between the feature probabilities in the populations, and  $Kn = \tilde{\Omega}(\ln n/\alpha^4)$ . We also present polynomial time algorithms for the resulting max-cut and max margin problems which, for now, needs only slightly more data than stated above. Our methods can also be used to give a simple combinatorial algorithm for finding a bisection in a random graph that matches Boppana's convex programming approach and McSherry's spectral results.

A. BLUM, A. COJA-OGHLAN, A. FRIEZE AND S. ZHOU, Separating Populations with Wide Data: a Spectral Analysis, in *Proceedings of the 18th International Symposium on Algorithms and Computation* (ISAAC 2007).<sup>†</sup>

In this paper, we consider the problem of partitioning a small data sample drawn from a mixture of  $k$  product distributions. We are interested in the case that individual features are of low average quality  $\gamma$ , and we want to use as few of them as possible to correctly partition the sample. We analyze a spectral technique that is able to approximately optimize the total data size—the product of number of data points  $n$  and the number of features  $K$ —needed to correctly perform this partitioning as a function of  $1/\gamma$  for  $K > n$ . Our goal is motivated by an application in clustering individuals according to their population of origin using SNP data, when the divergence between any two of the populations is small.

## Other papers

S. ZHOU, G. GANGER AND P. STEENKISTE, Location-based Node IDs : Explicit Locality in DHTs, Carnegie Mellon University Technical Report CMU-CS-03-171, September 2003.

Current peer-to-peer systems based on DHTs struggle with routing locality and content locality because of random node ID assignment. To address these issues, we promote the use of location-based node IDs to encode physical topology and improve routing. This gives applications explicit knowledge about and control over data locality at a coarse-grain. Applications can place content in particular regions or route towards a close replica. Schemes to address the difficulties that ensue, particularly load imbalance, are discussed.

S. ZHOU, G. GANGER AND P. STEENKISTE, Balancing Locality and Randomness in DHTs, Carnegie Mellon University Technical Report CMU-CS-03-203, November 2003.

Embedding locations in DHT node IDs makes locality explicit and, thereby, enables engineering of the trade-off between careful placement and randomized load balancing. This paper discusses hierarchical, topology-exposed DHTs and their benefits for content locality, and administrative control and routing locality.

D. TAN, S. ZHOU, J. HO, J. MEHTA AND H. TANABE, Design and Evaluation of an Individually Simulated Mobility Model in Wireless Ad Hoc Networks, in *Proceedings of Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002)*.

Since there has been little focus on systematically creating user mobility and communication pattern traces, researchers in the mobile networking community often propose personal models with which to validate their routing algorithms. This approach has two problems: (1) invalid conclusions may be drawn from overly simplistic or unrealistic models, and (2) it is difficult to compare performance results of different algorithms due to the variety of models used. We describe a novel approach to generate user mobility patterns with an individually simulated behavioral model. In this model, the overall mobility pattern is the result of the interaction between the behaviors of individual nodes. We show that this general-purpose framework can reliably reproduce existing mobility patterns as well as create useful new ones. The model is validated with the statistical characteristics of representative scenarios and through the simulation with the Dynamic Source Routing(DSR) protocol.