CSoI Tier I seed grant proposal: Graph Inference based on Random Walks

CSoI Post-doc advisor: **Thomas Courtade**
courtade@stanford.edu (Stanford/Princeton University)

CSoI faculty advisor: **Jennifer Neville**
neville@cs.purdue.edu (Purdue University)

Student PI: **Victoria Kostina**
vkostina@princeton.edu (Princeton University, advisor: Sergio Verdú)

Student PI: **Suvidha Kancharla**
skanchar@purdue.edu (Purdue University, advisor: Jennifer Neville)

August 29, 2012

I. Problem statement

Estimating models of graph data from network samples is a task of fundamental importance for many problems today. For instance, the computation of page rank for search engines and the estimation of peer influence in social networks both rely on an estimate of the underlying network model. In many cases, the underlying graph is unknown (e.g., in the case of websites connected by directed hyperlinks), and must be estimated through a sampling mechanism. One of the most basic sampling methods is performed by *crawling* the graph in a random fashion. As a result, one obtains a random walk on the network (i.e., a Markov process whose statistics are determined by the underlying graphical structure).

This raises the following question: by performing k_n random walks in parallel, each of length n, how accurately can one infer the structure of the underlying graph? More generally, what is the tradeoff between the compression rate of the sampled data and the fidelity to which the graph can be reproduced? How is the compression rate influenced by the heterogeneity, sparsity, and/or clustering observed in real world social network graphs? From a data-compression standpoint, this problem is related to recent work on compression and estimation over large alphabets (cf. [1], [2]) and finite-blocklength compression schemes (cf. [3]). Furthermore, it is natural in this setting to allow probabilistic reproductions and measure the corresponding fidelity under Logarithmic Loss, thus relating the proposed problem to the recent work [4].

As this is a Tier I proposal, we do not expect to answer the above questions in their entirety during the award period. Rather, our goal is to bring together researchers with expertise in the relevant areas in order to formulate a precise research direction for which a Tier II proposal will be written. We anticipate producing a conference paper or technical report with preliminary findings.

A. Background

As mentioned above, lossy compression of sources with memory is one of the topics most related to our proposed direction of study. Known results in rate-distortion theory for sources with memory include the following. Coding theorems for ergodic discrete-alphabet sources with memory [5] show that the minimum asymptotically achievable rate is given by the rate-distortion function, which is expressed as a limit of a sequence of solutions of a certain convex optimization problem parameterized by blocklength n. Perhaps surprisingly, even in the simple case of a binary symmetric Markov source, an explicit expression for this limit is known only in the small distortion region. For higher distortions, upper and lower bounds allowing to compute the rate-distortion function in this case with desired accuracy have been recently shown in [6]. Gray [7] showed a lower bound to the rate-distortion function of finite-state finite-alphabet Markov sources with a balanced distortion measure and identified conditions under which it coincides with its corresponding upper bound. For variable-length lossy compression of sources with memory, Kontoyiannis [8] presented upper and lower bounds to the minimum achievable encoded length as a function of a given source realization and required fidelity of reproduction, which eventually hold with probability 1 for a sufficiently large blocklength n.

II. Proposed Activity

During the award period for this Tier I proposal, our main goals are to develop a precise formulation of a research problem (leveraging the interdisciplinary composition of the team) and to develop a Tier II proposal based on this formulation and corresponding preliminary results. The problem we have described in Section I is related to the "Knowledge Extraction" thrust of the center, as it directly tackles BigData problems such as page rank estimation and inference in social networks.

Successful formulation of a research problem and the subsequent development of preliminary results will require frequent interaction amongst the team members. Fortunately, Victoria will visit Stanford (as will her advisor, S. Verdú) from Sept. 2012 – Mar. 2013. This will enable Victoria and Tom to meet frequently in person. Likewise, Jen and Suvidha are both located at Purdue University and are also able to meet frequently in person. The Stanford and Purdue groups will meet in-person twice and monthly by virtual meetings. These virtual meetings will be held using the collaboration tools available through `soihub.org`.

III. Expected Outcomes

Since this is a Tier I proposal, we do not expect to answer the questions posed in Section I in their entirety during the award period. Rather, our goal is to bring together researchers with expertise in the relevant areas in order to formulate a precise research direction for which a Tier II proposal will be written. Specific outcomes include:

- **Short Term Outcomes**
  1) Thorough understanding of related work: Currently, each member of the proposed team is familiar with his/her specific area. However, one of the short term objectives is to combine this knowledge and also perform an in-depth literature survey to understand how the proposed problem fits into the context of existing work.
  2) Multi-institution collaboration via `soihub.org` tools: In order to facilitate multi-institution collaboration and track progress, we will use the collaboration tools available through `soihub.org`. Not only is this a required outcome for successfully funded projects, it will also facilitate project documentation and tracking.
  3) Precise problem formulation: As mentioned previously, one of the primary goals of this project is to exploit the interdisciplinary nature of the team in order to develop a clearly defined research problem related to the Knowledge Extraction thrust of the center. Specifically, the research problem should capture the essence of a graph inference problem in a way that is relevant to the computer science, information theory, and machine learning communities.

- **Long Term Outcomes**
  1) Tier II proposal: The ultimate goal of the work conducted during the award period for this Tier I proposal is to put together a set of well-formulated research problems which can be realistically solved. These problems will form the basis of a Tier II proposal which will be submitted next year.
  2) Conference paper or technical report: During the award period of this Tier I proposal, we anticipate that preliminary results will be established as we review the literature and work toward a precise formulation of the problem at hand. In particular, we hope to obtain results for compression of Markov sources in the finite blocklength regime, and also for graph inference/compression under logarithmic loss. Both of these are interesting problems in their own right, and constitute two of the basic building blocks for the more general graph inference problem we propose to study. If it is not possible to produce a conference paper, we will compile our findings into a technical report to support the Tier II proposal.

IV. Proposed Work Statement

Jen Neville and Tom Courtade will supervise the project from the Purdue and Stanford sides, respectively. Based on each member's background, the roles each person will play in the project are quite natural: Jen and Suvidha will lend their expertise in machine learning and social network analysis, while Victoria and Tom will bring their knowledge of lossy compression to the table.

As mentioned previously, Victoria and Tom will meet frequently in-person at Stanford, while Jen and Suvidha will frequently meet at Purdue to discuss progress (e.g., on a weekly basis). The two groups will meet virtually on a monthly basis, and will meet in-person twice using project funds. Additionally, members of the team will meet during the December, 2012 NSF site visit at Purdue University.

Virtual meetings will take advantage of the collaboration tools through `soihub.org`. This will facilitate communication, track progress, and the saved documentation will be used in the six-month progress and annual reports.

References

[1] A. Orlitsky and N. Santhanam, "Performance of universal codes over infinite alphabets," in Data Compression Conference, 2003. Proceedings. DCC 2003, march 2003, pp. 402 – 410.

[2] A. Dhulipala and A. Orlitsky, "Universal compression of markov and related sources over arbitrary alphabets," Information Theory, IEEE Transactions on , vol. 52, no. 9, pp. 4182 –4190, sept. 2006.

[3] V. Kostina and S. Verd ú, "Fixed-length lossy compression in the finite blocklength regime," IEEE Transactions on Information Theory, vol. 58, no. 6, pp. 3309–3338, June 2012.

[4] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," in Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, july 2012, pp. 761 –765.

[5] R. Gallager, Information theory and reliable communication John Wiley & Sons, Inc. New York, 1968.

[6] S. Jalali and T. Weissman, "New bounds on the rate-distortion function of a binary markov source," in IEEE International Symposium on Information Theory June 2007, pp. 571 –575.

[7] R. Gray, "Rate distortion functions for finite-state finite-alphabet markov sources," IEEE Transactions on Information Theory, vol. 17, no. 2, pp. 127 – 134, mar 1971.

[8] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," IEEE Transactions on Information Theory, vol. 46, no. 1, pp. 136–152, Jan. 2000.

[9] T. A. Courtade, "Two Problems in Multiterminal Information Theory," Ph.D. dissertation, University of California, Los Angeles, 2012.

[10] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," Information Theory, 2007. ISIT 2007. IEEE International Symposium on, June 2007, pp. 566 –570.

[11] N. Cesa-Bianchi and G. Lugosi, Prediction, Learning, and Games New York, NY, USA: Cambridge University Press, 2006.

[12] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in Data Compression Conference (DCC), Snowbird, UT, Mar. 2011, pp. 53–62.

V. Budget and Justification

Budget section removed

VI. Research statements

A. Thomas Courtade

Much of Tom's recent work has focused on lossy compression when the reproduction fidelity is measured under logarithmic loss. In this setting, the decompressor produces beliefs (i.e., soft decisions), rather than deterministic decision values. The logarithmic loss function is a method of measuring the quality of these beliefs with respect to the true realization of the data. Surprisingly, many longstanding open problems become tractable when studied in the setting where distortion is measured under logarithmic loss. These include the CEO problem, the multiterminal source coding problem, and the interactive lossy source coding problem, among others [9]. The study of source coding under logarithmic loss is not only attractive from a tractability perspective, it also has an axiomatic justification [10]. Indeed, the logarithmic loss function is a canonical penalty function in the fields of prediction, learning, and game theory [11].

Since our problem considers graph inference based on sampled data, it is natural to ask how well the graphical structure can be inferred from the sampled data. In our setting, we would like to measure the quality of our inference under logarithmic loss as a starting point.

B. Victoria Kostina

Part of Victoria's dissertation, which recently appeared in IEEE Transactions on Information Theory [3], is an in-depth treatment of the fundamental limits of lossy data compression in the non-asymptotic regime. While the core results of that paper, namely, new tight achievability and converse bounds to the minimum achievable source coding rate as a function of blocklength and tolerable distortion, allow for memory, analysis and numerical computation of those bounds has been performed only in the most basic setting of lossy compression of a stationary memoryless source.

Since a random walk on a given graph is essentially a Markov source, finite blocklength results for sources with memory would play a crucial role in the proposed project. More generally, as most real-world sources have memory, analysis of the finite blocklength bounds in [3] for sources with memory, such as Markov sources, constitutes a very interesting research direction in itself. As observed in [3], [12], in the stationary memoryless case the minimum finite blocklength coding rate admits an exceptionally simple approximation involving only two terms, the rate-distortion function, which describes the minimum rate achievable if the blocklength is allowed to grow indefinitely, and a new source characteristic termed the rate-dispersion function, which quantifies the overhead over the rate-distortion function incurred by the finite blocklength constraint. Does the minimum achievable finite blocklength coding rate of Markov sources possess a similarly simple representation?

C. Jennifer Neville

Jennifer Neville is an assistant professor at Purdue University with a joint appointment in the Departments of Computer Science and Statistics. Neville's main research interests lie in the areas of data mining, machine learning, and statistical network analysis, with a particular focus on statistical analysis tools and methodology for knowledge discovery in complex network domains (e.g., social networks, social media, communications networks, distributed systems). In these domains, the *relationships* are an important source of information that identifies potential statistical dependencies among instances (e.g., articulated friendships in online social networks often identify people with correlated attributes).

However, a critical challenge for developing accurate and efficient analysis methods for large, partially-observable social network datasets is to understand the complex interaction between local model properties and the global network structure, and the impact of these interactions on algorithm performance (e.g., learning, inference, and evaluation). Neville's current work, and the focus of her 2012 NSF CAREER award, is thus focused on exploiting methods from network modeling, learning theory, and statistics to better characterize the performance of statistical models and algorithms based on the graph and data structure in the underlying networks.

D. Suvidha Kancharla

Suvidha Kancharla is a new Masters student at Purdue University. She received a B.E. in Computer Science and an M.S. in Mathematics from BITS Pilani in 2009 and then worked at Oracle in Bangalore India for the last three years. She is particularly interested in the application of graph theory for social network analysis and web analytics—thus her interests and background are well aligned with the goals of this project.