1 Bibliography January 15, 2015

References

Abowd et al.: A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs raey

John M. Abowd, Lars Vilhuber, and William Block. “A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs”. In: *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer and Ilenia Tinnirello. Vol. 7556. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 216–225. ISBN: 978-3-642-33626-3. DOI: 10.1007/978-3-642-33627-0_17. URL: http://dx.doi.org/10.1007/978-3-642-33627-0_17.

Abstract: We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols. It is based on extensible tools and can be easily incorporated into existing instructional materials.

Lagoze et al.: Data Management of Confidential Data DBLP:journals/ijdc/LagozeBWAV13

Carl Lagoze, William C. Block, Jeremy Williams, John M. Abowd, et al. “Data Management of Confidential Data”. In: *International Journal of Digital Curation* 8.1 (2013). Presented at 8th International Digital Curation Conference 2013, Amsterdam. See also http://hdl.handle.net/1813/30924, pp. 265–278. DOI: 10.2218/ijdc.v8i1.259.

Abstract: Social science researchers increasingly make use of data that is confidential because it contains linkages to the identities of people, corporations, etc. The value of this data lies in the ability to join the identifiable entities with external data such as genome data, geospatial information, and the like. However, the confidentiality of this data is a barrier to its utility and curation, making it difficult to fulfill US federal data management mandates and interfering with basic scholarly practices such as validation and reuse of existing results. We describe the complexity of the relationships among data that span a public and private divide. We then describe our work on the CED2AR prototype, a first step in providing researchers with a tool that spans this divide and makes it possible for them to search, access, and cite that data.

Lagoze et al.: Encoding Provenance of Social Science Data: Integrating PROV with DDI LagozeEtAl2013

Carl Lagoze, William C. Block, Jeremy Williams, and Lars Vilhuber. “Encoding Provenance of Social Science Data: Integrating PROV with DDI”. In: *5th Annual European DDI User Conference*. 2013.

Abstract: Provenance is a key component of evaluating the integrity and reusability of data for scholarship. While recording and providing access provenance has always been important, it is even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. The PROV model, developed under the auspices of the W3C, is a foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We report on the results of our experimentation with integrating the PROV model into the DDI metadata for a complex, but characteristic, example social science data. We also present some preliminary thinking on how to visualize those graphs in the user interface.

Lagoze et al.: CEDfffdfffdAR: The Comprehensive Extensible Data Documentation and Access Repository LagozeJCDL2014

Carl Lagoze, Lars Vilhuber, et al. "CEDfffdfffdAR: The Comprehensive Extensible Data Documentation and Access Repository". In: *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014)*. Presented at the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014). ACM/IEEE. London, United Kingdom, Aug. 2014.

Abstract: Social science researchers increasingly make use of data that is confidential because it contains linkages to the identities of people, corporations, etc. The value of this data lies in the ability to join the identifiable entities with external data such as genome data, geospatial information, and the like. However, the confidentiality of this data is a barrier to its utility and curation, making it difficult to fulfill US federal data management mandates and interfering with basic scholarly practices such as validation and reuse of existing results. We describe the complexity of the relationships among data that span a public and private divide. We then describe our work on the CED2AR prototype, a first step in providing researchers with a tool that spans this divide and makes it possible for them to search, access, and cite that data.

Lagoze et al.: Encoding Provenance Metadata for Social Science Datasets LagozeEtAl2013b

Carl Lagoze, Jeremy Willliams, and Lars Vilhuber. "Encoding Provenance Metadata for Social Science Datasets". In: *Metadata and Semantics Research*. Ed. by Emmanouel Garoufallou and Jane Greenberg. Vol. 390. Communications in Computer and Information Science. Springer International Publishing, 2013, pp. 123–134. ISBN: 978-3-319-03436-2. DOI: 10.1007/978-3-319-03437-9_13. URL: http://dx.doi.org/10.1007/978-3-319-03437-9_13.

Abstract: Recording provenance is a key requirement for data-centric scholarship, allowing researchers to evaluate the integrity of source data sets and reproduce, and thereby, validate results. Provenance has become even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. Recent work by the W3C on the PROV model provides the foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We apply that model to complex, but characteristic, provenance examples of social science data, describe scenarios that make scholarly use of those provenance descriptions, and propose a manner for encoding this provenance metadata within the widely-used DDI metadata standard.

Li et al.: One Permutation Hashing NIPS2012˙1436

Ping Li, Art Owen, and Cun-Hui Zhang. “One Permutation Hashing”. In: *Advances in Neural Information Processing Systems 25.* Ed. by P. Bartlett et al. 2012, pp. 3122–3130. URL: http://papers.nips.cc/paper/4778-one-permutation-hashing.

Abstract: While minwise hashing is promising for large-scale learning in massive binary data, the preprocessing cost is prohibitive as it requires applying (e.g.,) k=500 permutations on the data. The testing time is also expensive if a new data point (e.g., a new document or a new image) has not been processed. In this paper, we develop a simple **one permutation hashing** scheme to address this important issue. While it is true that the preprocessing step can be parallelized, it comes at the cost of additional hardware and implementation. Also, reducing k permutations to just one would be much more **energy-efficient**, which might be an important perspective as minwise hashing is commonly deployed in the search industry. While the theoretical probability analysis is interesting, our experiments on similarity estimation and SVM & logistic regression also confirm the theoretical results.

File: http://papers.nips.cc/paper/4778-one-permutation-hashing.pdf.

Li et al.: b-Bit Minwise Hashing in Practice PingShrivastava2013

Ping Li, Anshumali Shrivastava, and Arnd Christian König. “b-Bit Minwise Hashing in Practice”. In: *Internetware 2013.* Oct. 2013. URL: http://www.nudt.edu.cn/internetware2013/.

Abstract: Minwise hashing is a standard technique in the context of search for approximating set similarities. The recent work [26, 32] demonstrated a potential use of b-bit minwise hashing [23, 24] for efficient search and learning on massive, high-dimensional, binary data (which are typical for many applications in Web search and text mining). In this paper, we focus on a number of critical issues which must be addressed before one can apply b-bit minwise hashing to the volumes of data often used industrial applications. Minwise hashing requires an expensive preprocessing step that computes k (e.g., 500) minimal values after applying the corresponding permutations for each data vector. We developed a parallelization scheme using GPUs and observed that the preprocessing time can be reduced by a factor of 20 80 and becomes substantially smaller than the data loading time. Reducing the preprocessing time is highly beneficial in practice, e.g., for duplicate Web page detection (where minwise hashing is a major step in the crawling pipeline) or for increasing the testing speed of online classifiers. Another critical issue is that for very large data sets it becomes impossible to store a (fully) random permutation matrix, due to its space requirements. Our paper is the first study to demonstrate that b-bit minwise hashing implemented using simple hash functions, e.g., the 2-universal (2U) and 4-universal (4U) hash families, can produce very similar learning results as using fully random permutations. Experiments on datasets of up to 200GB are presented.

File: http://ecommons.library.cornell.edu/bitstream/1813/37986/2/a13-li.pdf.

Li et al.: GPU-based minwise hashing: GPU-based minwise hashing LiSK12

Ping Li, Anshumali Shrivastava, and Arnd Christian König. "GPU-based minwise hashing: GPU-based minwise hashing". In: *Proceedings of the 21st World Wide Web Conference (WWW 2012) (Companion Volume)*. 2012, pp. 565–566. DOI: 10.1145/2187980.2188129. URL: http://doi.acm.org/10.1145/2187980.2188129.

Abstract: Minwise hashing is a standard technique for efficient set similarity estimation in the context of search. The recent work of b-bit minwise hashing provided a substantial improvement by storing only the lowest b bits of each hashed value. Both minwise hashing and b-bit minwise hashing require an expensive preprocessing step for applying k (e.g., k=500) permutations on the entire data in order to compute k minimal values as the hashed data. In this paper, we developed a parallelization scheme using GPUs, which reduced the processing time by a factor of 20-80. Reducing the preprocessing time is highly beneficial in practice, for example, for duplicate web page detection (where minwise hashing is a major step in the crawling pipeline) or for increasing the testing speed of online classifiers (when the test data are not preprocessed).

Li et al.: Entropy Estimations Using Correlated Symmetric Stable Random Projections NIPS2012˙1456

Ping Li and Cun-Hui Zhang. "Entropy Estimations Using Correlated Symmetric Stable Random Projections". In: *Advances in Neural Information Processing Systems 25*. Ed. by P. Bartlett et al. 2012, pp. 3185–3193. URL: http://papers.nips.cc/paper/4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.

Abstract: Methods for efficiently estimating the Shannon entropy of data streams have important applications in learning, data mining, and network anomaly detections (e.g., the DDoS attacks). For nonnegative data streams, the method of Compressed Counting (CC) based on maximally-skewed stable random projections can provide accurate estimates of the Shannon entropy using small storage. However, CC is no longer applicable when entries of data streams can be below zero, which is a common scenario when comparing two streams. In this paper, we propose an algorithm for entropy estimation in general data streams which allow negative entries. In our method, the Shannon entropy is approximated by the finite difference of two correlated frequency moments estimated from correlated samples of symmetric stable random variables. Our experiments confirm that this method is able to substantially better approximate the Shannon entropy compared to the prior state-of-the-art.

File: http://papers.nips.cc/paper/4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.pdf.