

You are free to use any resources you like (e.g., papers, publicly available code, homework solutions found on the Web) in completing this homework, subject to two conditions: (1) All such resources used must be clearly and prominently acknowledged. (2) Everything you write should be original (in your own words) and you should be able to explain everything from scratch. Each question below is worth 20 points.

1. Provide a concrete example that illustrates the benefits of optimizing a set of queries together instead of individually [4, page 428]. Specify the SQL queries and indicate the query plans generated under separate and combined optimization (with justifications). Depict the query plans in usual tree notation. Quantify the difference in expected performance using suitable assumptions and base-table sizes.
2. Provide a concrete example of a client-server query processing scenario in which the best strategy requires shipping client-cached data back to the server [4, page 441]. Justify your answer by calculating the expected cost of alternatives for the query plan.
3. Identify the three most significant sources of errors in the study on P2P workloads [3]. Identify the three most significant results in the paper. Indicate if and how these results are affected by the errors.
4. A naive method for dynamic replication is the repeated application of the square-root replication strategy [2, Section 4]. How does this method compare with the ADR algorithm [4, page 455]? Discuss the costs of the methods as well as the quality of the resulting solutions.
5. Consider the relational representation of XML as a generic graph [1, Section 2.6]. Characterize, as generally as you can, queries that require recursion in this representation but do not require recursion in a more specific representation, such as the one based on tuple-generating elements [1, Section 2.3].

Submission Submit your entire homework electronically as a single PDF or plain text file. Please check that the PDF is portable. At the very least,

the `gv` program on the CSIC cluster must display it properly. Name your file using the scheme `LastnameIJ-hw1-NNNN.pdf`, where `NNNN` is a 4-digit integer of your choice, and upload it by anonymous FTP to the `/incoming/chaw` directory on `ftp.cs.umd.edu`.

References

- [1] S. S. Chawathe. *Semistructured Data in Relational Databases*, chapter 3. Practical Handbook of Internet Computing. CRC Press, 2004. To appear.
- [2] E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM)*, pages 177–190, Pittsburgh, Pennsylvania, Aug. 2002.
- [3] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 314–329, Pittsburgh, Pennsylvania, Aug. 2003.
- [4] D. Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.