DBMS – PROJECTS

M.Tech. (CS), First Year, 2019–2020

Deadline: June 30, 2020

Total: 40 marks

SUBMISSION INSTRUCTIONS

STEP I:

1. Submit a solution sketch in a single file by the deadline. The solution must be self-explanatory.
2. The solution should include the sections (if applicable): Introduction, Related Work, Terminologies and Definitions, Theory, Methods, Results, Conclusion.
3. Include names and roll numbers of all of your group members (at most 3).
4. Naming convention for your submission file (assuming M is your project number): `projM` (.docx, .doc, .pdf, .tex, etc.).
5. To submit a solution file (say `projM.docx`), ensure that it is not password protected and mail to `malaybhattacharyya@isical.ac.in`.

STEP II:

1. Deliver a solution sketch through team presentation on (or immediately after) the deadline.
2. The presentation slides should contain the components (if applicable): Introduction, Related Work, Terminologies and Definitions, Theory, Methods, Results, Conclusion.

NOTE: The contribution must be novel and non-trivial.

Project 1: **[Conjuctive Queries]** The conjunctive queries (CQs) are one of the most fundamental classes of database queries that principally correspond to select-project-join and select-from-where queries in relational algebra and SQL, respectively [1]. There are several interesting problems related to semantic query optimization having a focus on transforming a query into an equivalent acyclic one [2]. Some of these problems are still open in tractable classes of CQs. To name a few, the problems related to the decidability status of the problems under keys/FDs and bounded hypertreewidth modulo equivalence [2]. You are required to contribute by finding novel solutions of any one of these problems.

[1] Pablo Barceló, Andreas Pieris, and Miguel Romero. "Semantic optimization in tractable classes of conjunctive queries," *ACM SIGMOD Record*, 46(2):5-17, 2017. (Link: `https://dl.acm.org/citation.cfm?id=3137588`)

[2] Pablo Barceló. "Semantic Optimization in Tractable Classes of CQs and CRPQs," In *ICDT Open Problems in Database Theory Session*, Venice, Italy, 2017. (Link: `https://databasetheory.org/sites/default/files/blog\_attach/barcelo-open-problems-semantic-acy_0.pdf`)

Project 2: [**Hidden Database**] Data enrichment helps to extend a local database with new attributes from external data sources. There are recent research results that highlight how one can progressively crawl hidden databases (the deep web) through a keyword-search API for enriching a local database in an effective way [1]. The major challenge in such problems is selecting the best query at each iteration.

There are many interesting open problems that can be studied in this domain. These are related to creating run-time samples such that the upfront cost can be amortized over time, supporting not only keyword-search interfaces but also other popular query interfaces such as form-based search and graph-browsing, and exploring approaches of crawling a hidden database for other purposes such as data cleaning and row population [1]. You are required to contribute by finding novel solutions of any one of these problems.

[1] Pei Wang, Ryan Shea, Jiannan Wang, and Eugene Wu. "Progressive Deep Web Crawling Through Keyword Queries For Data Enrichment." In Proceedings of the 2019 International Conference on Management of Data, pp. 229-246. ACM, 2019. (`Link:https://dl.acm.org/citation.cfm?id=3319899`)

Project 3: [**Distributed ML on RDBMS**] A number of popular systems, most notably Googles TensorFlow, have been implemented from the ground up to support Machine Learning (ML) tasks. Recent research in this direction has considered making subtle changes to a modern relational database management system (RDBMS) to make it suitable for distributed learning computations [1]. Several discrepancy have came up related to this attempt because this implementation was made on top of a research prototype. Developing high-latency Java/Hadoop system, reducing the said gap, is an attractive target for future work.

You are required to address the problem of dealing with distributed ML on a Java/Hadoop system. You can consider a NoSQL database management scheme too.

[1] Dimitrije Jankov, Shangyu Luo, Binhang Yuan, Zhuhua Cai, Jia Zou, Chris Jermaine, and Zekai J. Gao. "Declarative recursive computation on an RDBMS: or, why you should use a database for distributed machine learning." Proceedings of the VLDB Endowment 12(7): 822-835, 2019. (`Link:http://www.vldb.org/pvldb/vol12/p822-jankov.pdf`)

Project 4: [**Subjective Databases**] In order to support experiential queries, a database system needs to model subjective data. Users should be able to pose queries that specify subjective experiences using their own words, in addition to conditions on the usual objective attributes. Based on this demand, Subjective Databases has emerged as a new alternative to satisfy user needs more effectively and accurately than alternative techniques through experiments with real data. OpineDB is the first one of this kind of databases [1].

Subjective databases introduce several new future research challenges. Subjective databases can designed to take into consideration a user profile to provide better search results in case the user chooses to share such a profile. Moreover, the system should be able to suggest queries to the user based on their profile and based on what may be unusual in the domain. More generally, the challenge is to model the user's expectations and point out the unexpected experiential aspects. Finally, one has to deal with the bias in review data. You are required to contribute by finding novel solutions of any one of these problems.

[1] Yuliang Li, Aaron Xixuan Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and

Wang-Chiew Tan. "Subjective Databases." Proceedings of the VLDB Endowment 12(11): 1330-1343, 2019. (Link: `http://www.vldb.org/pvldb/vol12/p1330-li.pdf`)

Project 5: [**Querying in Probabilistic Databases**] The probabilistic databases are motivated by the need to store large-scale uncertain data, and query it efficiently. There are promising advances in representing and querying large-scale automatically extracted data in probabilistic knowledge bases [1]. Determining the computational complexity of querying problems on probabilistic databases is a challenging domain of research [2]. There are some recent progresses in this direction [2, 3]. However, there are scopes for future research by extending the present results to other classes (such as equality generating dependencies) and, on the other hand, to obtain even more refined results (such as classification results) [3]. You are required to make novel theoretical contribution of any kind on this.

[1] Stefan Borgwardt, Ismail Ilkan Ceylan, and Thomas Lukasiewicz. "Recent advances in querying probabilistic knowledge bases," In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI)*, Stockholm, Sweden, 2018. (Link: https://www.ijcai.org/proceedings/2018/0765.pdf)

[2] Eric Gribkoff, Guy Van den Broeck, and Dan Suciu. "The most probable database problem," In *Proceedings of the SIGMOD Workshop on Big Uncertain Data (BUDA)*, Utah, USA, 2014. (Link: www.sigmod2014.org/buda/papers/p3.pdf)

[3] Ismail Ilkan Ceylan, Stefan Borgwardt, and Thomas Lukasiewicz. "Most probable explanations for probabilistic database queries," In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia, 2017. (Link: https://www.ijcai.org/proceedings/2017/0132.pdf)

Project 6: [**Probabilistic Query Evaluation**] Query evaluation on tuple-independent probabilistic databases (TIDs) are interesting problems. Conceptually, all tuples are independent probabilistic events in a tuple-independent relation within a TID. A tuple-independent relation is a relation $R(A_1, \ldots, A_m, P)$ in which the tuples are associated with marginal tuple probabilities $P \in [0, 1]$ [1]. There are some open problems related to probabilistic query evaluation on such databases and the dichotomies in their complexity [2]. You are required to attempt any one these problems and make a novel contribution.

[1] Ronald Fagin, Benny Kimelfeld, and Phokion G. Kolaitis. "Probabilistic data exchange," *Journal of the ACM*, 58(4):15, 2011. (Link: https://www.researchgate.net/publication/220430432_Probabilistic_Data_Exchange)

[2] Benny Kimmelfeld. "Dichotomies in the Complexity of Query Answering over Probabilistic Databases," In *ICDT Open Problems in Database Theory Session*, Venice, Italy, 2017. (Link: https://databasetheory.org/sites/default/files/blog_attach/open-problems-probabilistic-db_0.pdf)

Project 7: [**Data Visualization**] Visualization of data is considered to be an important area of Data Science. Data visualization symbolizes the efforts that help people in understanding the significance of data in a better way through representing it in a visual context. The patterns, trends and correlations that might go undetected in raw data can be exposed and

recognized effectively with data visualization methods [1]. There are plenty of data visualization approaches in the literature, starting from the basic ones like Venn diagrams, pie charts, boxplots, and scatter plots, and ranging upto the recent ones like alluvial diagrams and sunbursts [2]. You are required to recognize the limitations of the existing visualization approaches and suggest a novel data visualization method. Note that, development of your approach is suggested but deployment is not necessary.

[1] Cameron Chapman, "A Complete Overview of the Best Data Visualization Tools," *Toptal*, 2019. (Link: https://www.toptal.com/designers/data-visualization/data-visualization-tools)

[2] RAWGraphs. (Link: https://rawgraphs.io)

Project 8: **[Crowd Powered Databases]** Crowdsouring is a distributed approach of solving problems online by involving the crowd contributors, either as volunteers or in exchange of payments. In crowdsourcing databases, human operators are embedded into the database engine and they collaborate with the other conventional database operators to process the queries [1]. There are recent advances in developing various types of crowd-powered database functionalities. This also includes the query answering capabilities [2]. You are required to propose a novel crowd-powered approach to support any kind of database management operations.

[1] Sai Wu, Xiaoli Wang, Sheng Wang, Zhenjie Zhang, and Anthony KH Tung. "K-anonymity for crowdsourcing database," *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2207-2221, 2014.

[2] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. "CrowdDB: answering queries with crowdsourcing," In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD)*, pp. 61-72, 2011. (Link: http://nike.psu.edu/classes/ist501/2014-fall/ref/crowddb_sigmod2011.pdf)

Project 9: **[Consistent Query Answering in Dirty Databases]** Given a finite set of constraints and a Boolean query, a consistent query answering is a kind of decision problem that finds whether a query is true for every repair on an inconsistent database D, often referred to as a dirty database. There are several interesting problems related to evaluation of queries on databases, using the approach of consistent query answering and data repair [1]. However, the dichotomies/trichotomies of consistent query answering for broader classes of queries and constraints are still unknown. Contribute to address any one of these problems.

[1] Benny Kimelfeld and Paris Koutris. "Open Problems in Consistent Query Answering," In *ICDT Open Problems in Database Theory Session*, Venice, Italy, 2017. (Link: https://databasetheory.org/sites/default/files/blog_attach/open-problems-koutris-2_0.pdf)

Project 10: **[Probabilistic Unclean Databases]** The Probabilistic Unclean Databases (PUDs) comprise a framework for unclean data that follows a noisy channel approach to model how errors are added to the data. In a recent research proposal, two different types of statistical knowledge have been incorporated for a better modelling of unclean data [1]. The one first represents a belief of how intended (clean) data is generated, whereas the second one depicts a belief of how noise is introduced in the actual observed database. With this, the concept

and a novel framework of PUDs have been introduced. As a new research direction, it is possible to investigate the complexity of cleaning in more general configurations than the ones covered in the work by Sa et al. [1]. Moreover, in cases where probabilistic cleaning is computationally hard, it is of natural interest to find approximate repairs that have a probability (provably) close to the maximum. Another direction is the complexity of probabilistic query answering and approximation thereof, starting with the most basic constraints (e.g., primary keys) and queries (e.g., determine the marginal probability of a certain fact). Finally, an important direction is to devise learning algorithms for more general cases than those covered here. You are required to make a novel contribution in any one of the aforementioned directions.

[1] Christopher De Sa, Ihab Ilyas, Benny Kimelfeld, Christopher Re, and Theodoros Rekatsinas. "A Formal Framework for Probabilistic Unclean Databases," In *Proceedings of the 22nd International Conference on Extending Database Technology and 22nd International Conference on Database Theory (EDBT/ICDT)*, Lisbon, Portugal, 2019. (Link: http://pages.cs.wisc.edu/ thodrek/prob_unclean.pdf)

Project 11: **[Complex Event Processing]** Complex Event Processing (CEP) has emerged as a challenging field in the synergy of technologies that require processing and correlating distributed data sources in real-time. Unfortunately, there are no general techniques for evaluating CEP query languages with clear performance guarantees. A recent study presents a rigorous and efficient framework to CEP by proposing a formal language for specifying complex events, called CEL, that contains the main features used in the literature and has a denotational and compositional semantics [1]. A possible extension of this work is to study the evaluation of non-unary CEL formulas. Another relevant problem is to understand the expressive power of different fragments of CEL and the relationship between the different operators. Finally, one can have a focus on additional features of CEP languages like correlation, time windows, aggregation, consumption policies, etc. You are required to make a novel contribution in any one of the aforementioned directions.

[1] Alejandro Grez, Cristian Riveros, and Martin Ugarte. "A Formal Framework for Complex Event Processing," In *Proceedings of the 22nd International Conference on Extending Database Technology and 22nd International Conference on Database Theory (EDBT/ICDT)*, Lisbon, Portugal, 2019. (Link: http://drops.dagstuhl.de/opus/volltexte/2019/10307/pdf/LIPIcs-ICDT-2019-5.pdf)

Project 12: **[Citizen Science]** Citizen science is a new approach of conducting scientific research by involving the general public. For example, Zooniverse is a citizen science platform that is powered by common people [1]. The Zooniverse enables everyone to take part in real cutting edge research in many fields across the sciences, humanities, and more. Planet Hunters is a citizen science project under Zooniverse that tries to discover new planets by employing common people as volunteers [2]. The volunteers observe how the brightness of a star changes over time and report the same.

You are required to design a citizen science platform (say CrowS: Crowd-driven Science) and suitably connect it to a database (one that suits the best here) so that it enables people-powered research. You are free to choose a scientific domain (say species discovery) for designing the platform. The logins for requesters and responders must be password protected

and different.

[1] Zooniverse. (Link: https://www.zooniverse.org)

[2] First Validated PHT Planet!. (Link: https://blog.planethunters.org)

Project 13: **[Fair Price Platform]** The prices of everyday commodities is so dynamic and populated (unethically by some people) that they vary place to place even within the same locality. Suppose you wish to provide the details of prices of commodities in your nearby market places through an application. There are two parties in this application - the information provider and the information explorer.

Design and implement such a mobile App that would locate your place and show the details of nearby market prices product by product. You have to use the SQLite database to be accessed by the App.

[1] SQLite. (Link: https://www.sqlite.org/index.html)

Project 14: **[Model for NoSQL Databases]** Unlike relational databases, which are typically driven by the structure of available data, NoSQL data modeling often starts from the application-specific queries [1]. As with NoSQL approaches, we cannot model relations within the data, therefore, sometimes it is hard to visualize the database as structured collection of data.

Propose a novel data model for NoSQL databases for overcoming the said limitations. The model should be created with appropriate characterizations.

[1] NOSQL DATA MODELING TECHNIQUES. (Link: https://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques)

Project 15: **[Correctness of SQL Queries]** We often test the correctness of SQL queries by executing the query in question on some test database instance and compare its result with that of the correct query [1]. The problem of finding small counterexamples for different classes of queries, including those involving negation and aggregation, is in general known to be NP-hard. There are recent algorithms to address such problems [1]. Building user-friendly tools to learn and debug database queries is an interesting research direction. In particular, building a similar tool with the full functionality of SQL queries is a challenging open problem. Suggest a model for creating such a platform.

[1] Zhengjie Miao, Sudeepa Roy, and Jun Yang. "Explaining Wrong Queries Using Small Examples." In Proceedings of the 2019 International Conference on Management of Data, pp. 503-520. ACM, 2019. (Link: https://dl.acm.org/citation.cfm?id=3319866)

Project 16: **[Video Database Management]** A video database management system provides integrated support for spatio-temporal and semantic queries for videos. Some recent research in this domain put forward some benchmark that evaluates the performance of these systems. Visual Road is one such benchmark that comes with a data generator and a suite of queries over cameras positioned within a simulated metropolitan environment [1]. This contribution is too simple and does not fuse tiles, track objects across tiles, and supports increasingly complex procedurally-generated tiles.

You are required to model an advanced version of Visual Road with a new perspective and enhanced functionalities.

[1] Brandon Haynes, Amrita Mazumdar, Magdalena Balazinska, Luis Ceze, and Alvin Cheung. "Visual Road: A Video Data Management Benchmark." In Proceedings of the 2019 International Conference on Management of Data, pp. 972-987. ACM, 2019. (Link: `https://dl.acm.org/citation.cfm?id=3324955`)

Project 17: **[Stable Rankings in Databases]** Decision making through weighted rankings is a challenging task. It is ideally targeted that the ranked order would not change with nominal changes in these weights. This property is known as stability of the ranking. A recent paper has developed a framework that can be used to assess the stability of a provided ranking and to obtain a stable ranking within an "acceptable" range of weight values [1]. The existing definition of stability considers a pair of rankings to be different if they differ in one pair of items. An alternative is to allow minor changes in the ranking. Moreover, a weight vector is a single point in a stable region, which can be extended to characterize the boundaries of the stable region for some applications.

You are required to make novel contributions in one of the aforementioned directions of research.

[1] Asudeh, Abolfazl, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich. "On obtaining stable rankings." Proceedings of the VLDB Endowment 12, no. 3 (2018): 237-250. (Link: `http://www.vldb.org/pvldb/vol12/p237-asudeh.pdf`)

Project 18: **[Growth Analysis of COVID-19]** As the statewise daily counts and other details of COVID-19 affected people are available [1], it is interesting to assume that a distributed database can be designed to keep this kind of data. Given such a distributed data, how can we model the state-by-state growth of COVID-19 affected people based on demographic details like population density, border sharing of states, migration of workers, etc.

You are required to develop a model and fit the data on daywise count of COVID-19 infects in different states of USA into it. The data is to be stored in a graph database.

[1] https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_India

Project 19: **[Sequence Analysis of SARS-CoV-2]** The DNA or RNA sequence of a species is characterized by a string of four nucleotides. The sequences of SARS-CoV-2 (causing COVID-19) are available from multiple countries collected over a timeline of a couple of months [1]. Based on the k-mer distribution of these sequences, can we classify them based on location or collection timeline.

You are required to develop a classifier that can serve the said purpose.

[1] https://www.ncbi.nlm.nih.gov/labs/virus/vssi

Project 20: **[Visualization of COVID-19]** You are required to visualize any kind of COVID-19 data available online in a novel way such that it reflects new things. Your observation should be novel. You can also design a website to reflect your results.

Project 21: [**Open COVID-19 Challenge**] You can come up with any kind of new model or novel exploratory analysis to understand COVID-19 better and report something new. You are open to use your own idea. The only requirement is that the data is to be represented in NoSQL format.

Project 22: [**COVID-19 Crowdsourced Data**] The crowdsourcing based tracking of COVID-19 affected people is being done in India [1]. Given this data, can you mine any new information.

You are required to state a novel hypothesis and validate that on the data.

[1] https://t.me/covid19indiaops

Project 23: [**Re-creating COVID-19 Analysis**] You can perform any one of these analyses in an Indian context. Alternatively, use the data and explore something novel.

(a) http://www.healthdata.org/data-visualization/covid-19-us-state-state-projections

(b) https://www.genomedetective.com/app/typingtool/cov

(c) https://www.washingtonpost.com/graphics/2020/world/corona-simulator

(d) https://www.go-fair.org/implementation-networks/overview/vodan

(e) https://github.com/nytimes/covid-19-data

(f) https://github.com/ieee8023/covid-chestxray-dataset

(g) https://healthweather.us/?mode=Atypical

Take some ideas from this link:
https://drive.google.com/file/d/1vDcb6HeS-hufNgqH0dDhIEGjuJpnnkzT/view.

You can even create a repository of COVID-19 details in the form of a website.

Project 24: [**DREAM COVID-19 Challenge**] The DREAM challenge on COVID-19 provides clinical data and asks the question "Of patients who have at least one clinical encounter/visit at UW Medicine and who were tested for COVID-19, can we predict who is positive?" [1]. Machine learning approaches are required to be applied for taking part into this challenge. Participate in this challenge and compete with others by making a good contribution. You have to team with the participant named "malayb", who is already registered in this challenge.

[1] https://www.synapse.org/#!Synapse:syn21849255/wiki/601865