# DAIS Qualifying Examination

*Fall 2014 (Sept 29, 2014)*

Department of Computer Science
University of Illinois at Urbana-Champaign

## Time Limit: 180 minutes

Please read the following instructions carefully before starting the test:

- The examination has 15 questions, including one basic concept question and 14 topic questions. You are required to answer the basic concept question and any **5** of the 14 topic questions of your choice. If you answer more than five topic questions, the committee will randomly select five to grade.
- The basic concept question has **18** subquestions. You are required to answer **9** of these 18 subquestions. If you answer more than 9 subquestions, the committee will randomly select 9 to grade.
- The 14 topic questions are distributed by areas as follows:
  - **Database Systems**: 3 questions
  - **Data Mining**: 3 questions
  - **Information Retrieval**: 3 questions
  - **Bioinformatics**: 5 questions
- Use a separate booklet for each question that you answer. That is, you will use a total of six booklets (1 for the basic concept question and the rest for the 5 chosen topic questions).
- To ensure the fairness of grading, all the exams will be kept anonymous. Your answer sheets should thus only bear the assigned ID but *no names*.
- Please make your answers clear and succinct; you will lose credit for verbose, convoluted, or confusing answers. *Succinctness does count!*
- The questions have been designed to avoid any ambiguity. However, in case you are not sure how to interpret a question, please assume one interpretation, write down your assumption, and solve the problem accordingly.
- On average, you will have 30 minutes for each of the six questions that you have chosen to answer. So plan your time accordingly, and keep in mind that some questions/subquestions may demand more writing than others. You may find it useful to periodically assess your time usage, and adjust your time allocation dynamically as needed to avoid running out of time on questions that you could have answered.

## Required Question (Basic Concepts): Problem 0

You are required to answer 9 out of the following 18 subquestions. Each subquestion is worth 1 point. Please focus on the major point and keep your answer concise; in general, an answer is expected to be no more than two sentences.

(1) (*Query Language*)
What is the most important property of a query language? Explain why.

(2) (*Parallel DBMS*)
How do you think parallel query processing in Parallel DBMS (such as Gamma) is different from Map-Reduce? Identify one major difference.

(3) (*Indexing*)
Indexing helps query processing. Other than speeding it up (making it more efficient), give one more major advantage that indexing enables for query processing.

**Note: Please start with a new answer sheet.**

Short answers: The answer of each of the following 3 questions is expected to be no more than two lines (i.e., no explanation nor reasoning is needed)

(4) (Data mining algorithms) It is interesting to mine homogeneous graphs (or networks) where all the nodes and edges are of one type. Name one algorithm for each of the following tasks:

1. mining frequent (sub)graph patterns
2. clustering subgrahs in a large graph, and
3. classifying (sub)graphs, given a training set of (sub)graphs.

(5) (Data mining or data cube concepts) Use one or two sentences to distinguish the following pairs of concepts or methods:

1. *PageRank* vs. *SimRank*, and
2. *support vector machines (SVM)* vs. *neural networks*.

(6) (Selection of data mining methods) Name or outline one data mining method that best fits each of the following tasks:

1. *identify patients that are best to be treated by a particular therapy*, and
2. *group bloggers by their political beliefs based on their blog contents*.

**Note: Please start with a new answer sheet.**

(7) (*IDF Weighting*) When we use the Okapi (BM25) retrieval function to score documents for a query that has only one term, the ranking of documents is not affected by IDF weighting. Use one to two sentences to briefly explain why.

(8) (*PageRank*) Use one to two sentences to briefly explain the main advantage of scoring a web page with PageRank over scoring a web page simply based on the raw count of the inlinks pointing to the page.

(9) (*Evaluation*) Let $R = (+, +, -, -)$ be the relevance status of four documents in a ranked list of retrieval results with "+" indicating a relevant document and "-" a non-relevant document. (Thus, the first two documents are relevant, while the last two are non-relevant). Suppose there are in total 10 relevant documents in the whole collection. Compute the precision, recall, and average precision. It's sufficient to give an expression; there is no need to reduce the expression to a value.

**Note: Please start with a new answer sheet.**

(10) (*Sequence alignment*) What does an affine gap penalty model mean in the context of pairwise sequence alignment ?

(11) (*Gene prediction*) When an HMM is used for gene prediction, what do its states represent ?

(12) (*Phylogenetic tree*) I am given a phylogenetic tree T1. I multiplied every branch length by a factor of 2 to get a new tree T2. The probability of a nucleotide x changing to nucleotide y on any branch b is twice as much in tree T2 than in T1. True or false ?

(13) (*Phylogenetic tree reconstruction*) What is the maximum parsimony principle for evolutionary tree building?

(14) (*Gene expression analysis*) You have 10 samples of microarray data from mice with a specific disease and 10 samples from healthy mice. Name a statistical test that you might use to determine differentially expressed genes in the disease (versus healthy).

(15) (*Clustering of biological data*) Name a clustering method where you do not have to decide upon the number of clusters you want ahead of time (i.e., before running the clustering algorithm).

(16) (*Motif finding*) The advantage of using position weight matrices (PWMs) as a motif model is that you can calculate the probability of a site s having been sampled from a PWM W, i.e., the value of $\Pr(s|W)$. However, it is easy to see that a longer PWM will typically give smaller probability values, even to its best sites, than a shorter PWM gives to its best sites. How is this apparent bias addressed in scoring sites for matches to PWMs?

(17) (*DNA Sequencing*) Name a graph algorithm that can be used for assembling sequencing reads.

(18) (*Regulatory Genomics*) Give one reason why a transcription factor ChIP peak near a gene may not be indicative of that transcription factor regulating that gene.

# Database Systems: Problem 1

The paper "Incremental and Accuracy-Aware Personalized PageRank through Scheduled Approximation" addresses computing PPV efficiently– through incremental approximation.

**Part 1**

What does *accuracy-awareness* mean? Why is it important? (*2 points*)

**Part 2** The accuracy, in terms of L1 error, is given in the paper.

1) Give the derivation of this error. (*2 points*)

2) Identify the intuition behind the equation. (*2 points*)

**Part 3**

The technique requires selecting a set of *hubs*, for preprocessing to build precomputed PPV "segment" values.

The paper suggests a scheme considering the factors of *decaying power* and *overall popularity* of a node to select good hubs.

We ask you to disagree with this scheme. Give two *different* factors of considerations for hub selection. (*4 points*)

## Database Systems: Problem 2

The paper "Incremental and Accuracy-Aware Personalized PageRank through Scheduled Approximation" addresses computing PPV efficiently– through incremental approximation.

**Part 1**

What is the key insight of this paper for speeding up computation of Personalized PageRank? (*2 points*)

**Part 2**

How is "Personalized" PageRank different from the "generic" PageRank? Identify their similarities and differences. (*2 points*) Give each one an example application scenario. (*2 points*)

**Part 3**

PageRank is one type of graph ranking metric for measuring connection between nodes on a graph. There are other metrics proposed in the literature, such as SimRank or Hitting Time. You may think of more examples.

Do you think the general insight in this paper is applicable to other different metrics? Why? (*4 points*)

## Database Systems: Problem 3

The paper "Optimization for iterative queries on MapReduce" develops a query optimization approach for iterative queries in distributed environment.

**Part 1**

Given an example workload of "iterative querying". (*1 points*)

**Part 2**

Why is standard Map Reduce inappropriate for such iterative querying? Explain, using an example. (*2 points*)

What is the central idea in the paper for optimizing such querying? (*2 points*)

**Part 3**

The key concept of the paper lies in the identification and handling of *invariant views* and *variant views*.

1) Why are these two concepts important? (*2 points*)

2) How are they treated differently? Why? (*3 points*)

# Data Mining and Data Warehousing: Problem 1

Similar to a homogeneous network, one can define a *heterogeneous ego-net of degree k* of a node $p$ to be a subnetwork pulled out from a heterogeneous information network by selecting a node $p$ from the network and all of its local connections up to length $k$. For example, in a heterogeneous network formed by the DBLP data, an author Anne's ego network of degree 2 contains her papers (length 1, via metapath "$A - P$"), and their publication venues (length 2, via meta-path "$A - P - V$"), coauthors (via "$A - P - A$"), terms (via "$A - P - T$"), etc.

**Part 1**

(a) What are the similarities and differences of such an ego-net in comparison with an ego-net in a homogeneous network? (*2 points*)

(b) Discuss how to define *similar authors* effectively by exploring the concept of *heterogeneous ego-net*, and outline an efficient mechanism to implement it in large heterogeneous networks. (*2 points*)

(c) Discuss how to define *similar author-venue pairs* effectively by exploring the concept of *heterogeneous ego-net*, and outline an efficient mechanism to implement it in large heterogeneous networks. (*2 points*)

**Part 2**

(a) Based on this concept, experts may provide labels for some venues and ask the system to classify other venue into a specific field (e.g., AI). Outline an efficient classification mechanism to do it. (*2 points*)

(b) Based on this concept, one likes to identify anomalous authors, based on its heterogeneous ego-net up to degree 2. Outline an efficient mechanism to do it. (*2 points*)

# Data Mining and Data Warehousing: Problem 2

People nowadays move around with smart-phones, generating tremendous amount of mobility data.

**Part 1**

(a) Mobility sequential patterns can be generated, such as "many go to offices in the morning and then go to local restaurants for lunch, and then go back homes in late afternoon." Outline an efficient method that mines such patterns efficiently. (*2.5 points*)

(b) Suppose only a small portion of people keep their smart-phones on, and thus data so collected is sparse. Discuss how your algorithm should be revised in such a situation. (*2.5 points*)

**Part 2**

(a) Suppose multi-level and multi-dimensional information is associated with users and locations. One may like to find non-redundant and interesting patterns in multi-dimensional space flexibly, such as finding students in Department $A$ will likely to go to a Japanese restaurant for lunch on Fridays. Outline an efficient method to find such patterns. (*2.5 points*)

(b) Suppose text messages can be associated with smartphones and their users. Outline an efficient method that may make credible friendship recommendation to smart-phone users based on their interests and their mobility patterns. (*2.5 points*)

# Data Mining and Data Warehousing: Problem 3

## Part 1

(a) Tweets often discuss events reported in the news. To what extent that the contents in tweets and that in news complement each other in many cases? (*1 point*)

(b) What are the challenges on mining news alone, and mining tweets along, respectively? (*1.5 points*)

(c) Outline a mechanism that news and tweets can be mined together and their integrated mining may lead to more effective mining of both. (*2.5 points*)

## Part 2

(a) Many tweets are geo-coded (i.e., their geo-locations are known). Suppose a tweet contains user-id, time, location, hashtag, and other messages. Design an effective method that can distinguish hashtags which address issues that are *local*, vs. *nation-wide*, vs. *international*, based on *online streams* of tweet and news feeds. (*2.5 points*)

(b) Not every piece of news or tweets is truthworthy. Design a mechanism that may use both sources to identify what is likely to be the truth in news and tweets. (*2.5 points*)

# Information Retrieval: Problem 1. Economic Models of Search

[Luo et al. 14] refers to the following paper:

Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: dual-agent stochastic game in session search, SIGIR'2014.

[Azzopardi 14] refersto the following paper:

Leif Azzopardi. 2014. Modelling interaction with economic models of search, SIGIR'2014.

**Part 1**

a What are the two agents that the dual-agent stochastic game framework proposed in [Luo et al. 14] attempts to model? (*1 point*)

b Why is a traditional retrieval model such as BM25 inadequate to optimize session search? (*2 points*)

**Part 2**

The work [Azzopardi 14] proposed the following new cost function for modeling interactive IR:

$$c(Q, V, S, A) = c_q.Q + c_v.V.Q + c_s.S.Q + c_a.A.Q$$

a Compared with the existing work on a similar cost function, what components in the formula above are novel? What does each variable in these new components mean? (*2 points*)

b The author of [Azzopardi 14] derived an analytical solution for optimal A and optimal Q by optimizing an objective function with a constraint. What is the objective function optimized? What is the constraint about? (*2 points*)

**Part 3**

Both [Azzopardi 14] and [Luo et al. 14] proposed models for interactive information retrieval, yet they address different research questions. What is the most important difference in the research questions addressed by these two papers? The difference in their research questions led to the difference in how they perform evaluation. Using no more than 4 sentences to briefly explain what each paper attempts to prove through experiments. (*3 points*)

## Information Retrieval: Problem 2. Text Mining

[McAuley & Leskovec 13] refers to the following paper:

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text, RecSys'2013.

**Part 1**

a From the perspective of computation, what is the input and what is the intended output of the algorithm proposed in [McAuley & Leskovec 13]? (*2 points*)

b The proposed model in [McAuley & Leskovec 13] is a combination of two existing models. What are these two existing models? (*1 point*)

**Part 2**

a The objective function to be optimized in the proposed HFT model in [McAuley & LesKovec 13] can be decomposed into two parts, corresponding to the two existing models that the new model extends. What does each part of the objective function attempt to optimize (i.e., what is maximized or minimized)? Use no more than 3 sentences or sketch an illustrative formula to explain how these two parts in the objective function are connected (i.e., what variables in these two parts introduce dependency between the two parts). (*2 points*)

b Using no more than 3 sentences to briefly explain how the HFT model can be used to discover product categories. (*1 point*)

**Part 3**

In [McAuley & LesKovec 13], the rating prediction task is evaluated using Mean Squared Error (MSE) as the main measure. One might argue that for recommending products, MSE does not necessarily reflect the utility of a recommendation result from a user's perspective. Suppose we have 3 candidate products, A, B, and C, and their gold standard ratings are 5.0, 4.0, and 3.0, respectively, and we have two recommender systems, X and Y. Suppose we use X and Y each to recommend one of the 3 products to a user. Give an example of recommendation results where system X has achieved a lower MSE, but generated a worse recommendation from a user's perspective, than system Y. Can you suggest an alternative measure that might be better than MSE in such a case? (*4 points*)

## Information Retrieval: Problem 3. Product Search

[Li et al. 11] refers to the following paper:

Beibei Li, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. Towards a theory model for product search, WWW'2011.

**Part 1**

a According to [Li et al. 11], what is the main difference between product search and regular web document search? (*1 point*)

b In [Li et al. 11], the utility of product is defined as

$$U_p(X) = \sum_{k=1}^{K} \beta^k . x^k + \xi$$

What does each of $\beta^k$, $x^k$, and $\xi$ refer to? Give an example of $x^k$. (*2 points*)

**Part 2**

a The authors of [Li et al. 11] proposed two models of utility surplus, i.e., Logit Model and BLP Model. What is the key technical difference (i.e., difference in the formulas of the two models) between the two models? What is the difference between the two models in the required training data for estimating the model parameters? (*3 points*)

b In [Li et al. 11], the authors used the negative sign of the coefficient on "review_value rating" obtained using their model to conclude that hotels that receive a "high value" rating have lower demand. Considering that there are many other variables used in the model and some of them may be correlated, do you have any concern about this conclusion? (*1 point*)

**Part 3**

Even though the title of the paper [Li et al. 11] contains "search", the proposed model has not considered a user's query. Yet, one might argue that a user can potentially provide a keyword query or a structured query in some form. Propose a way to extend the model and product ranking algorithm proposed in [Li et al. 11] so as to also incorporate either a keyword query or a structure query in some form. Try to use formulas to illustrate your idea if you can.

(*3 points*)

## Bioinformatics: Problem 1. Stormo and Fields 1998

(a) The authors consider the binding constant of a site $X_i$ for a transcription factor T:

$$K_{eq}(X_i) = \frac{[T \cdot X_i]}{[T][X_i]}$$

They then define a normalized version of this binding constant, denoted by $K_s(X_i)$. Upon this normalization, the probability that a TF molecule binds to $X_i$, given that it has the possibility of binding anywhere in the genome, is given by what function of $K_s(X_i)$? (*1 points*)

(b) Does a gel shift assay allow direct measurement of the binding constant $K_{eq}(X_i)$? If not, does it allow measurement of some quantity (formula) related to $K_{eq}(X_i)$? (*2 points*)

(c) How are the quantities $\Delta G_s$ (binding energy) and $K_s$ (binding constant) related to each other? (*1 points*)

(d) *Fill in the blank at the end of this sentence*: Given a collection of binding sites and assuming these to be strong binding sites, one can show that the "weight matrix" given by the formula $W(b,j) = \log_2 \frac{f(b,j)}{p(b)}$ (where $f(b,j)$ is the frequency of base b at position j of the sampled sites and $p(b)$ is the frequency of base b in the genome) maximizes ... (*1 points*)

(e) The authors propose that given the $\Delta G_s$ values of a set of sites (e.g., single nucleotide mutants of the consensus site), one can predict the $\Delta G_s$ value for any site. What is the key assumption about binding energies that goes into this claim? (*2 points*)

(f) If the assumption alluded to in part (d) above is not true, what changes might you make to model binding site strengths? (*3 points*)

## Bioinformatics: Problem 2. Degner et al. 2012

(a) What is an eQTL? (*2 points*)

(b) What is a dsQTL? (*2 points*)

(c) What is a possible mechanism to explain dsQTLs? (*2 points*)

(d) What relationship do the authors note between dsQTLs and QTLs in this paper? (*2 points*)

(e) What relationship do the authors note between dsQTLs and transcription factor binding? (*2 points*)

## Bioinformatics: Problem 3. Gutenkunst et al. 2007

(a) The authors measure the change in model behavior with varying parameters $\theta$ by the $\chi^2$ score:

$$\chi^2(\theta) = \frac{1}{2N_c N_s} \sum_{s,c} \frac{1}{T_c} \int_0^{T_c} [\frac{y_{s,c}(\theta, t) - y_{s,c}(\theta^*, t)}{\sigma_s}]^2 dt$$

Here, an average is taken over all molecular entities $s$ and all conditions $c$. Explain the quantity whose average is taken, i.e., the term inside the summation. (*3 points*)

(b) The authors calculate the Hessian matrix of the $\chi^2$ score:

$$H^{\chi^2}_{j,k} = \frac{d^2\chi^2}{d\log\theta_j d\log\theta_k}$$

and compute the eigenvalues of this Hessian matrix. These eigenvalues are what they use to quantify the "sloppiness" of the model around the optimum $\theta^*$. Briefly explain their rationale for doing so. That is, why are they looking at the Hessian matrix? What do the eigenvalues of the matrix reflect intuitively? (*4 points*)

(c) The authors argue that even if parameters of a model are poorly determined, i.e., there is great uncertainty in the values of some of the parameters, this may not necessarily be a problem for the modeler. Why is this so? (*3 points*)

## Bioinformatics: Problem 4. Guan et al. 2008

(a) What is the classification problem addressed in this paper? (What is the output label to be predicted, and what data forms the input vectors?) (*2 points*)

(b) The baseline classifier utilized in this work is an SVM with bootstrap aggregation (bagging). What does 'bagging' mean? (*1 points*)

(c) Briefly describe the main idea behind "Bayesian hierarchical combination of SVM classifiers" on the GO hierarchy. You do not need to describe the actual inference algorithm, which was described in an earlier paper. Just describe how the Bayesian network is set up and what are the observables and hidden nodes in the network. (*4 points*)

(d) One of the techniques explored in this work was a naive Bayes classifier to combine multiple classifiers' outputs. What was the difference among these classifiers whose outputs were combined? (*2 points*)

(e) Which of the four methods tested in this work (baseline SVM, Hierarchy:MarkovBlanket, Hierarchy:Tree, naive Bayes) worked best most of the time? (*1 points*)

## Bioinformatics: Problem 5. Nourmohammad and Lassig 2011

(a) The first analysis that the authors show in this paper is that of computing an autocorrelation function on cis-regulatory modules. The autocorrelation function is defined by them as $\Delta(r) = c(r) - c_0$. Explain this formula. (That is, what are $r, c(r)$ and $c_0$?) (*1 points*)

(b) Provide two different sources suggested by the authors for the autocorrelation signal seen by them using the above formula. (They actually list more than two sources.) (*2 points*)

(c) How do the authors define the "similarity information" of a pair of sites? (*3 points*)

(d) What is the maximization problem they solve for a given cis-regulatory module, using dynamic programming, in their attempt to study the extent of local duplications in these modules? (*2 points*)

(e) What is the one message you will take home from this paper about evolution of binding sites in Drosophila cis-regulatory modules? (*2 points*)