

Opportunity Title: Multi-query search optimization: low cost set covers for large collections of regular expressions

Opportunity Reference Code: IC-17-19

Organization Office of the Director of National Intelligence (ODNI)

Reference Code IC-17-19

How to Apply **Create and release your Profile on Zintellect** – Postdoctoral applicants must create an account and complete a profile in the on-line application system. **Please note: your resume/CV may not exceed 2 pages.**

Complete your application – Enter the rest of the information required for the IC Postdoc Program Research Opportunity. The application itself contains detailed instructions for each one of these components: availability, citizenship, transcripts, dissertation abstract, publication and presentation plan, and information about your Research Advisor co-applicant.

Application Deadline 3/31/2017 6:00:00 PM Eastern Time Zone

Description **Research Topic Description, including Problem Statement:**

A regular expression (see Wikipedia), or REGEX, refers to a sequence of characters that defines a search pattern for use in a string searching algorithm. The complexity of regular expression matching was recently shown to depend on the expression's depth².

Suppose S is a large collection of text strings. In practice, elements of S will either be presented in a stream, or else will be accessible through Map Reduce operations.

There are n players, and player j creates a list $Q(j, 1), Q(j, 2), \dots, Q(j, r)$ of REGEX queries against S . Each collection T of strings in S has a value $V(T, j)$ to player j ; this function may be taken as submodular, or not. Moreover query $Q(j, k)$ has execution cost $C(j, k)$.

The players do not collude, and the set of strings returned by $Q(j, i)$ may have non-empty intersection with the set of strings returned by $Q(j', i')$ for j different to j' .

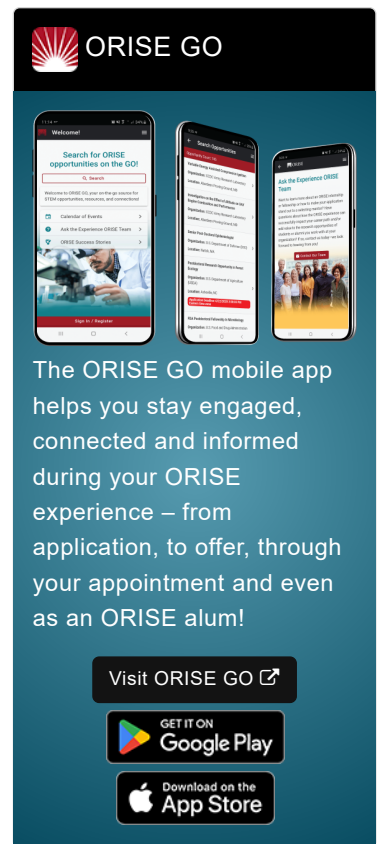
As a metaphor, introduce a broker, whose task is to inspect the union of all the queries $\{Q(j, k), 1 \leq j \leq n, k \geq 1\}$ and to devise a new set B of REGEX queries, of approximately least cost, which collectively returns all (or perhaps nearly all) of the union, over j and k , of the set of strings in S returned by $Q(j, k)$.

The broker's profit consists in the sum of fees paid by each player for the fulfillment of their queries, less the cost of executing his own set B of queries. His algorithm aims to maximize this profit. Profit is here a metaphor for cost savings achieved by multi-query search optimization.

Consider strings as vertices in a hypergraph, and a query as a hyperedge (i.e. the set of strings returned by the query), whose weight is determined by the query cost. Then a set of queries is a hyperedge-weighted hypergraph.

If the broker does not construct new queries, but merely selects a least costly subset of queries supplied by all the players, then the problem above is an instance of the NP-complete problem called set cover (see Wikipedia). This means: find a lowest weight subset of hyperedges whose union is all the strings sought by all players. There are approximation algorithms for set cover³.

The broker is not restricted to finding a set cover consisting of a subset of the queries supplied by



ORISE GO

The ORISE GO mobile app helps you stay engaged, connected and informed during your ORISE experience – from application, to offer, through your appointment and even as an ORISE alum!

Visit ORISE GO [↗](#)

GET IT ON
Google Play

Download on the
App Store

Opportunity Title: Multi-query search optimization: low cost set covers for large collections of regular expressions

Opportunity Reference Code: IC-17-19

the players. He may construct new queries, which may be less costly to execute in total. Hence this problem is more general than set cover.

² Arturs Backurs and Pitr Indyk, Which regular expressions are hard to match?, ArXiv:1511:07070, 26 September 2016

³ David P. Williamson, David B. Shmoys, The Design of Approximation Algorithms, Cambridge, 2011.

Example Approaches:

Proposals could consider some of the following approaches, or others not listed:

- Efficient algorithms for many non-colluding agents to query text.
- Efficient methods for searching large data sets by topic, rather than by keyword.
- Computational frameworks which allow a large number of queries by many users to be fulfilled with less computational load by a redacted set of queries.

Illustrations of recent work for finding set covers for massive data sets include:

- Barna Saha and Lise Getoor, On maximum coverage in the streaming model & application to multi-topic blog watch, Proc. SIAM Int'l Conf Data Mining (SDM), pages 697-708, 2009
- Ravi Kumar, Benjamin Mosely, Sergei Vassilvitskii, Andrea Vattani, Fast greedy algorithms in MapReduce and streaming, ACM Transactions on Parallel Computing, 2, 2015.
- Sarel Har-Peled, Piotr Indyk, Sepideh Mahabadi, Ali Vakilian, Towards tight bounds for the streaming set cover problem, PODS 2016.

Eligibility Requirements

- **Citizenship:** U.S. Citizen Only
- **Degree:** Doctoral Degree.
- **Discipline(s):**
 - **Business** ([11](#))
 - **Chemistry and Materials Sciences** ([12](#))
 - **Communications and Graphics Design** ([6](#))
 - **Computer, Information, and Data Sciences** ([16](#))
 - **Earth and Geosciences** ([21](#))
 - **Engineering** ([27](#))
 - **Environmental and Marine Sciences** ([14](#))
 - **Life Health and Medical Sciences** ([45](#))
 - **Mathematics and Statistics** ([10](#))
 - **Other Non-Science & Engineering** ([13](#))
 - **Physics** ([16](#))
 - **Science & Engineering-related** ([1](#))
 - **Social and Behavioral Sciences** ([28](#))