White Paper CIF21 DIBBs: STORM: Spatio-Temporal Online Reasoning and Management of Large Data, NSF Award 1443046

The increasing presence of smart phones and various sensing devices has led to humongous amounts of spatio-temporal data, and the imperative needs for rich data anlytics over such data. Many data from a measurement network and social media data sources are inherently spatial and temporal. As a result, numerous analytical tasks based on such data have a spatial and/or temporal extent.

The current practice is to develop customized data management software and the supporting hardware infrastructures (when necessary) by each group, institution, or organization, often on a project-by-project basis. Scientists also have to act as a "data" expert, or rely on the expertise from some extra data personnel, to design and develop a customized query engine. This engine translates inputs from a query interface or analytic task to specific query languages or programs (e.g., SQL queries as a simplest example) run against an underlying storage engine (e.g., a relation database). For each new data set or each new query type or analytic task on an existing data set, a new query interface and the underlying query program must be developed. Each new data type or schema and/or format update requires changes to the existing, customized query engines. Valuable efforts and manpower are invested in storing, processing, and querying large heterogeneous spatio-temporal data; time which scientists should have spent on their domain-specific investigation, research, and development.

Our objective is to design and develop the STORM system, *an automatic and broadly-applicable query and analytical engine for large, heterogeneous spatio-temporal data*. STORM offers *Spatio-Temporal Online Reasoning and Management at scale over any such data, and for single or multiple data sources*.

Even though various forms of spatial and spatio-temporal analytics have been extensively studied, the ever-increasing size of spatio-temporal data introduces new challenges. In particular, when the underlying data set is large, reporting all points that satisfy a query condition can be expensive, since there could be simply too many points that satisfy a query. The CPU cost of performing an analytical task or computing an aggregation using all these points adds additional overhead, and may not scale well with increasing number of points. Hence, waiting for the exact analytical or aggregation results may take a long time.

An important observation is that approximate results are often good enough, especially when approximation guarantees are provided. It is even more attractive if the quality of an approximation improves continuously over time until the exact result is obtained in the end. On big spatial and spatio-temporal data sets, waiting for exact results may take a while. The user faces a *dilemma*: *either* waits for the current query to complete *or* terminates the current query and issue the new query. And the number of possible combinations a user wants to investigate in order to find interesting patterns, even for a small region like NYC and first quarter, can be daunting.

The STORM system solves this dilemma. STORM uses spatio-temporal online reasoning and management to achieve online aggregation and analytics on large spatio-temporal data. STORM uses *spatial online sampling* to achieve its objectives. In particular, spatial online sampling continuously returns randomly sampled points from a user specified spatio-temporal query region, until user terminates the process or enough samples have been obtained to meet an accuracy requirement. An unbiased estimator, tailored towards a given analytical query, is built using the spatial online samples, and its approximation quality improves in an online fashion while more samples are being returned.

To make it easy for users and different applications to enjoy the benefit of spatio-temporal online analytics and aggregation, STORM also implements a data connector, so that it can easily import data in different formats and schemas, and enable spatio-temporal online analytics over such data without much efforts. Lastly, it features a number of built-in analytical and visualization modules so that a set of common analytical and visualization tasks can be executed without further engineering efforts once data have been imported. More complex and other analytical tasks can be built in a customized fashion.