


[VLDB 2017](#)**43rd International Conference on Very Large Data Bases****Reviews For Paper**

Track Research -> March 2017
Paper ID 1160
Title Constellation Queries over Big Data


Masked Reviewer ID: Assigned_Reviewer_1**Review:**

Question	
Overall Rating	Weak Reject
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	<p>The paper introduces Constellation Queries, a concept to finding elements in large datasets that satisfies the geometrical characteristics specified by the query. It thoroughly describes the techniques that are used to mine the data for geometrical searches. The authors introduce the concept by giving two examples, one from astronomy and one from seismology. The focus in the rest of the paper is mainly on the astronomy case. The authors elaborate on an example from astronomy, where they search for Einstein-Cross objects in a subset of SDSS data. They present the algorithms applied, using quadtrees to reduce data complexity, and filter on candidates, The algorithms are elaborated deeper and explain the inclusion of matching attributes. The results and performance are presented and show performance tests. There is certainly a need for these kind of data-mining techniques in the Big Data databases and this work contributes with novel techniques to do so.</p> <p>I do have minor concerns that relate to the readability of the paper and therefore I recommend publication of this paper with minor improvements. The authors should consider those modifications to the corresponding sections. Below are my comments and I request the authors to address them.</p>
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	<p>S1 Need for presented techniques in application areas such as astronomy S2 Techniques thoroughly described and benchmarked S3 Practical usage demonstration</p>
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	<p>W1 Non-stringent use of terminology W2 Minor issues with domain science (Section 2) W3 Various minor presentation issues</p>
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution).	With some new ideas
Significance	Improvement over existing work

Technical Depth and Quality of Content	Syntactically complete but with limited contribution
Experiments	OK, but certain claims are not covered by the experiments
Presentation	Reasonable: improvements needed
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	<p>Throughout the paper, I got confused about the intermingled terms of additive factor, additive error (bound) whether they can be interpreted as an uncertainty on the distance. I suggest the authors to be more clear about this and mention explicitly its interpretation in the contexts.</p> <p>Below is the remaining of the review with detailed comments on the individual sections. </p> <p>Abstract</p> <ul style="list-style-type: none"> • The Einstein Cross looks like a cross by co-incident, the astronomical phenomenon of gravitational lensing should not be presented as if a cross is generated in general. <p>Section 1</p> <ul style="list-style-type: none"> • P2, left column, last paragraph: The authors state that three techniques are involved. However, in the next lines they sum four items. <p>1</p> <p>Section 2</p> <ul style="list-style-type: none"> • Third paragraph. Observation of Einstein Cross was not the confirmation of GR, but one of many. The authors should take some caution here, idem for the "indicate the presence of dark matter". It is rather a contribution in the research of dark matter. <p>Section 3</p> <ul style="list-style-type: none"> • Here is the introduction of ϵ. What is the relation of ϵ to the resolution of the SDSS? Later on in the text it becomes clear that the performance is sensitive to ϵ. I suggest the authors add a few lines here to its definition. • Third paragraph. It would be helpfull if the units of the distance are presented here Section 4.2.2 • In the matrix computations, $M(B_i, B_{i+1})$ has elements of value 1. It is not clear to me whether this is an integer/boolean value or if it represents a probability of 1 of a distance match. I suggest the authors spend a line on this. <p>Section 7.1.2</p> <ul style="list-style-type: none"> • First paragraph, the authors mention the seismic dataset, but in the next session it would be helpfull it was written explicitly where it was used. I suggest the authors show where the results are related to the seismic and astronomy datasets. <p>Section 7.4</p> <ul style="list-style-type: none"> • Third paragraph. The main element influencing the computational costs is ϵ. What is the relationship between the critical ϵ values and the survey resolution? The authors state that its value depends on the point where filtering out anchor nodes becomes efficient. Can the authors specify if there is or is not a relation to this value and survey resolution.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall Rating	Weak Reject
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	This paper introduces algorithms for finding sets of objects satisfying geometrical patterns (called constellation queries) on large datasets. While the problem has interesting applications and the paper presents some interesting approaches (such as using matrix multiplication), the paper needs some work to be ready for publication for several reasons: lack of clarity in the writing, lack of depth in experimental evaluation, including no comparison with related work.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	<ol style="list-style-type: none"> 1. The problem presented in this paper has very interesting applications (finding effects of gravitational lensing, salt domes, etc.). 2. A very interesting aspect is that the methods presented in this paper can find objects "by example" - the user can provide, for instance, the Einstein's cross, and the system would find similar patterns in the data. (Unfortunately, this strong point is not emphasized enough in the paper.) 3. Some of the techniques presented in this paper are interesting, such as using matrix multiplication 
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results	<ol style="list-style-type: none"> 1. The writing is sometimes unclear. This is especially true in the problem definition (Sect 3). The paper overall needs substantial improvement in the writing. See detailed evaluation. 2. The first technique presented in the paper (Reducing data complexity using quadrees) seems a simple adaptation of the typical algorithm for answering range queries in metric spaces. If this is the case, it should be stated, and this technique should perhaps not receive so much space in the paper (it currently has an entire page devoted to it). Otherwise, its novelty should be made clearer. Other parts

that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	<p>that seem more novel, such as using matrix multiplications, are, instead, not presented clearly and exhaustively, and should receive more space. More in general, the presentation of the techniques is a bit messy (see detailed evaluation).</p> <p>3. The main techniques, such as using matrix multiplication, are not explained in a clear way.</p> <p>4. The experimental evaluation lacks in several aspects: It only evaluates running time performance, but how are the algorithms performing in terms of quality (precision and recall of the returned sets)? It is stated that the experiments are performed on both Einstein cross and seismic dome, but in the results it is not clear which one is which, and it seems that the results of only one query are reported; There is no experimental comparison with other techniques, such as package queries or Searchlight (the related work discussion about them is either incomplete or unclear).</p> <p>5. The related work section lacks clarity and completeness. In particular: Could one use subgraph matching if the graph had a distance value on each edge? What does it mean to "label tuples" to support local and global constraints in package queries? Searchlight (Kalinin et al.) seems related to this work in that it can find regions of the sky satisfying certain properties, but there is no discussion or comparison with it.</p>
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution).	With some new ideas
Significance	The paper is going to start a new line of research and products
Technical Depth and Quality of Content	Syntactically complete but with limited contribution
Experiments	OK, but certain claims are not covered by the experiments
Presentation	Reasonable: improvements needed
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	<p>Abstract:</p> <ol style="list-style-type: none"> 1. The very first sentence refers to "relative distances", which are only defined later in the paper. At that point, it is very unclear what the authors mean by relative distances. 2. It is stated that the problem grows exponentially not only in the size of the data but also of the pattern. I'm unsure whether this is the case. In fact, if the pattern is as large as the dataset ($k = D$), then there would only be one possible candidate, making the problem easy to solve. 3. The abstract first defines a "geometrical pattern" and then renames it "constellation query". If the authors prefer the latter name, I would suggest to use that term throughout. 4. I would suggest to avoid the last sentence "To the best of our knowledge..." as it adds uncertainty regarding the novelty of the work. <p>Introduction:</p> <ol style="list-style-type: none"> 1. The first sentence is misleading as for the purposes of the paper. What are these "new interpretations" of natural phenomena? For instance, is the Einstein cross a new interpretation of a natural phenomenon? But more importantly, is the problem solved by this paper that of finding new interpretations, or that of solving constellation queries? 2. The name "Bucket_NL" is not introduced. 3. Last paragraph, last sentence: "discusses" is repeated multiple times. <p>Sect 2:</p> <ol style="list-style-type: none"> 1. First paragraph: The paragraph is written as if it were describing a general methodology to collect sky data, whereas I believe this is, more specifically, the methodology used by SDSS. 2. "A constellation in the SDSS scenario would be defined by a sequence of objects from the catalog whose spatial distribution forms a shape." - Avoid uncertain definitions ("would"). Is a constellation a "sequence" or a "set" of objects? 3. Paragraph: "In a constellation query, an astronomer..." - The example in this paragraph implies a "query by example" paradigm. If this is the main application solved by the paper, it should be reported as such throughout. Note that this was not clear in the introduction.

Sect 3:

1. It is hard to understand the problem definition. This section needs to be rewritten in a clearer and more mathematically precise way.
2. Is a constellation query a “sequence” or a “set”? If it’s a sequence, is there an order in the sequence? How would this order be defined?
3. What is the domain Dom? Real numbers?
4. This sentence is hard to read: “the distances between the centroids of each pair of query elements that define the query shape and size with an additive allowable factor ϵ ”.
5. Is ϵ the same across different pairs?
6. This sentence is hard to read: “an element-wise function $f(e, q)$ that computes the similarity (e.g. in brightness at a certain wavelength) between elements e and q up to a threshold θ ”.
7. Is θ the same across different pairs?
8. In the second half of the section, $f(e, q)$ gets written as “ $f e$ ”.
9. Sections 2 and 3 could be combined (section 2 probably does not need a section of its own)

Sect 4:

1. Reading the general description of the two strategies (“Index the data...” and “After filtering...”) it is not clear what the difference between the two strategies is. It seems that they do the same job: filtering out candidates.
2. Fig 3 and 4: The distinction between centroids and data points is lost when seeing the paper in gray scale. Try using different shapes.
3. The technique “Reducing data complexity using quadtrees” seems a simple adaptation of the typical algorithm for answering range queries in metric spaces. If this is the case, it should be stated, and this technique should perhaps not receive so much space in the paper (it currently has an entire page devoted to it). Otherwise, its novelty should be made clearer.
4. In 4.1: The concept of a “bucket” appears for the first time at the end of the section, but it was not introduced before. The meaning of bucket B_i should be introduced beforehand - this is the output of the algorithm presented in the section.
5. In 4.2.2: The section (“... approaches”) title suggests more than one matrix multiplication approach, but the section only describes one such approach. Also, it is stated that this approach precedes bucket nested loops. Perhaps, it should be presented before that. It is also very hard to understand why matrix multiplication works. Can the authors provide a small (possibly visual) example?
6. The writing of 4.2.4 is poor, making it hard to understand. 4.2.5 should give more explanations about the algorithmic idea.

Sect 5:

1. As stated in the first sentence, this section presents the main algorithm to answer pure constellation queries, but it refers back to the previous section, which contained the actual algorithms. I suggest to present things in a different order: Algorithm 3 first, then Algorithm 4, Algorithm 1 and 2.
2. Also, Sect 4 and 5 could be merged. Perhaps, the composition algorithms could be in a different section, after the main algorithm is presented.
3. The names “pure” and “general” are somewhat misleading, as “general” actually refers to results satisfying relaxed scale factors.
4. In Algorithm 3: In line 1, “ qt ” is not yet defined (it is defined in line 2). In line 4, what is “ l ”?

Sect 6:

1. The writing of this section is very dense, and appears also rushed. There are no space boundaries between the various theoretical results (lemmas and proofs).
2. What is the intuition behind the choice of “scalebasic”? And what is the intuition behind the algorithm and its proof, more in general?

Experiments:

1. It is stated that “the techniques succeeded in spotting the right structures among billions of candidates”. I don’t see results that prove this claim. How do you define what the “right” structures are? How do the algorithms perform in terms of recall (finding all patterns) and precision (avoiding wrong patterns)?
2. In all plots, there is no x-axis and y-axis labels.
3. Suggestion: Use more descriptive captions (include what you are measuring (running time?) and what the results show).
4. The error bars are so low that you can perhaps avoid showing them.
5. The title of 7.3 should replace “effectiveness” with “efficiency” or, more generally, “performance”, as effectiveness is mostly used to denote precision/recall assessments.
6. Are the results of Table 1 generalizable to other queries? My guess would be that the thresholds shown in the table are actually query dependent.

7. A more general issue with the experiments is that it seems that they only show the results of one query, although 7.1 suggests that more than one dataset and query are used. In the plots, however, it is not clear which dataset/query they refer to.

8. The results on general queries are only described in text, but a plot similar to Fig 12 would be more convincing.

9. Finally, there is no experimental comparison with other techniques, such as package queries or Searchlight. Regarding package queries, can the distances be encoded as linear constraints?

Related work:

1. Could one use subgraph matching if the graph had a distance value on each edge?
2. What does it mean to "label tuples" to support local and global constraints in package queries? The explanation about why package queries would be impractical to solve constellation queries is unsatisfying.
3. Searchlight (Kalinin et al.) seems related to this work in that it can find regions of the sky satisfying certain properties, but there is no discussion or comparison with it.
4. Multiple references such as [[25],[18],[27]] should appear as [25,18,27].

Conclusion:

1. In the last paragraph, please expand on the future work opportunities.


References:

1. [6] looks weird. What is "00000"? Same problem in [3].
2. Extra comma in [28]

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall Rating	Reject
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	This paper presents algorithms based on quadtrees that evaluate spatial pattern queries, called constellation queries in the paper. The paper ignores previous work on spatial joins. The presented techniques are based on evaluating star patterns using index nested loops. The comparison to previous work is inadequate and the contribution is questionable.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	S1. The authors present some applications of constellation queries.
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	W1. The presentation is sub-standard. The problem definition is obscure and so is the description of the methods. W2. The authors do not implement or test state-of-the-art spatial join algorithms for solving their problem, although spatial joins have been studied extensively for at least 2 decades. W3. The proposed method is too slow as shown in the experiments.
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium	Novelty unclear

ratings for delta papers and papers on well known topics but still with some valuable contribution).	
Significance	No impact
Technical Depth and Quality of Content	Questionable work
Experiments	Obscure, not really sure what is going on and what the experiments show
Presentation	Sub-standard: would require heavy rewrite
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	<p>C1. The MM approach is unclear. I do not understand why you use matrix multiplication to implement a spatial join. </p> <p>C2. The authors ignore all previous work on spatial joins. For example, the problem presented here is not new. Spatial pattern queries have been studied for example in [Dimitris Papadias, Nikos Mamoulis, Vasilis Delis: Algorithms for Querying by Spatial Structure. VLDB 1998: 546-557]. Spatial intersection join algorithms have been used in evaluating multiway join queries in [Nikos Mamoulis, Dimitris Papadias: Integration of Spatial Join Algorithms for Processing Multiple Inputs. SIGMOD Conference 1999: 1-12]. Your constellation queries involve distances not intersections however it is straightforward to extend intersection joins algorithms to apply for distances. Your algorithms seem to be based on (indexed) nested loops, which is one of the worst choices for the problem.</p> <p>C3. Figure 7 shows that the proposed approach is very expensive (>10 sec even for less than 100K objects). Clearly, state-of-the-art spatial join algorithms are not used here.</p> <p>C4. It is unclear what the y-axis of Figures 10 and 11 measures. In general, you should put measures on all axes.</p>