Metareview from Editor

Thank you for your submission to Data Mining and Knowledge Discovery.  I have received three reviews of your paper.  On the basis of these reviews I am unfortunately not able to accept your manuscript in its current form. In particular, while you managed to convince more or less two of the reviewers, reviewer 2 now become very negative. This is mainly due to the LR approach you now discuss again; it is good that you discuss it. However, the added discussion raised new issues raised by the Reviewer 2. Consequently, I am asking you to provide a more in-depth discussion of the relation between ELD and LR.

Thank you for this helpful summary of the main issues. We have restructured the paper to address this general concern, and added material to address a number of detailed comments. Our detail changes are described in-line below. Our high-level change is to present the paper as an instance of the **Exceptional Model Mining Framework** (EMM). This change does not affect the technical content, and is mainly in the abstract and introduction (Section 1). It also led to small changes in the discussion of empirical results (Section 8.2) and the conclusion. Advantages include the following.

1) Our work is now situated even more clearly in the context of the field. In particular, this clarifies that we are applying an existing but recent framework to a data type (network data) that is new for the framework.

2) Fit with journal. DAMI has published the major survey article on EMM (see Duivesteijn2016 et al. reference). But this survey article considers only propositional data. DAMI has also published the major survey article on outlier detection for multi-relational data (Akoglu2015 reference). Our paper combines these two recent directions for which the journal already is the place of record.

3) Framework not method. This change emphasizes that the main novelty is developing the EMM framework for multi-relational data, rather than proving that our new method (ELD) is better than any others. The message “EMM + SRL works well for multi-relational data” is a better summary of our results. Also, the reviewers accepted the strong performance of the two EMM methods (KLD and ELD), compared to other baselines that are not EMM; whereas one reviewer had doubts about the evidence comparing ELD and KLD. So emphasizing the merits of the EMM framework aligns with the reviewers’ reading of the empirical results.

While I do no agree with the reviewer that a full theoretical justification is required, also because the reviewer was more positive in the previous round, I do think that the issues raised should be addressed as much as possible. For instance, would the divergence of the conditional distribution indeed work?

We have aimed to address the reviewer’s issues, as described in our reply to the metareview and reviewer #2. For the conditional distribution question specifically, see our replies to reviewer #2 below.

How do we justify ELD given the rather strong theoretical foundations for LR?

We have added 2 pages of theoretical analysis (3 including proofs) in a new Section 6. The well-known result that KLD is approximated by Chi-square uses a Taylor series approximation. We use the same Taylor series analysis to clarify the key cancelling phenomenon: the first-order term for LR vanishes due to cancelling, whereas the first-order term for ELD is the well-established total variation distance (L1 norm). Relating ELD to a well-established distance concept should also allay worries about the theoretical foundation of this concept (naturally, there are interesting open theoretical questions for future work).

With such a in-depth discussion and probably some more empirical evidence I would be happy to overrule Reviewer 2.

We have added results about precision (complementing AUC), following previous outlier detection evaluation metrics by relational learning researchers such as Jiawei Han. (Section 8.2, Table 12) The precision results show the advantages of ELD more strongly.

So, if you revise the manuscript also paying careful attention to the other reviewers’ comments I would be pleased to receive the revision for another round of review.

Thank you. Detailed replies to the reviewers’ comments are below.

Please note that addressing one issue sometimes raises new issues that were not previously apparent, so there is no guarantee that the paper will be accepted even if you do address all issues raised.

Reviewer #1: The paper is definitely much improved.  
  
Structurally, it is is still odd to me that section 5 comes before section Section 5. Section refers to Table 5, which has computations for two scores functions that are only defined in Section 6.  I'd fix this.  
  
Minor points-per-game  
p2 of the object linked to it → objects  
p3 to baseline methods → to the baseline  
p3 is independent → is an  
p4 method(RIBL) → method (RIBL)  
p5 a similarity measures → a similarity measured  
p6 2 x Bayes network instead of Bayesian network as used in the rest of the paperwork  
  
Section 3.1 twices refers to Table 1 as a summary.   
P9: Fig 2 some of the probabilities seem to have extra )  
p9 where v\_i resp. v\_pa\_i... → this sentence is difficult to parse as written (also it should have a \noindent)  
  
p9 from context → from the context  
p9 of a match only → reads funny, please reword  
  
Section 4, uses subsubsection instead of subsections  
  
p13 Interpretability, which... → this is not a complete sentence.  I think "Interpretability. This..." is better  
  
p13 therefore an intuitively → and therefore is an intuitively  
  
p13 class and object distribution → distributions  
p13 F\_i,i=1,2 → F\_1 and F\_2 seems so much clearer  
  
p15 it can be shown → needs a \noindent in front  
p15 outier - outlier  
p19 in literature → in the literature  
p12 OutRank, recommended → OutRank as recommended  
  
p22 "have high performance to detect outliers" → reword  
p22 LR metrics → The LR metrics  
p22 in NHL dataset → in the NHL datasets  
p23 examine three top → examine the three top  
p25 Bayesian network Learning → Bayesian network learning  
p26 "we could found in" → that we could find in this dataset.  
P30 season.Assists → season. Assists  
Thank you for the comments, we have reorganized the sections. Now in section 4 and 5 we introduce the scores and the examples of score computations are at the end of section 5. We also fixed the typos and other inconsistencies that you kindly brought up.   
  
Reviewer #2: From my previous review:

Thank you for the extensive comments, from fundamental issues to typos.

"SUMMARIZE ANY CHANGES THAT MUST BE MADE FOR A REVISED VERSION OF THIS PAPER TO BE ACCEPTABLE FOR PUBLICATION.  
  
There must be a thorough comparison to likelihood-ratio tests (LR).  This comparison should include an explanation of when likelihood-ratio fails to detect outliers that ELD correctly detects, to help demonstrate the effective differences. A better theoretical explanation of ELD would also help, given the remaining weaknesses in the current explanation (discussed above)."

We think there is a high-level issue here about what we do and do not claim about LR. There are also detailed issues raised by your comments. Speaking first to the high-level issue, we have rephrased the introduction and some of the discussion to emphasize that the paper develops Exceptional Model Mining as a novel framework for relational outlier detection. EMM is a recent previously existing framework for subgroup discovery (see DAMI 2016 survey paper in our Duivesteijn2016 reference). It has not been applied to relational data before, so for relational data it is an entirely novel framework. EMM vs. the rest of the relational world: The message we wanted to get across that combining EMM with SRL provides a novel and successful approach to relational outlier detection. By making that explicit in the introduction we believe we have situated the paper even more clearly in the current research landscape for the reader.

Likelihood ratio vs. expected log distance. The main idea behind EMM is to turn model learning methods into exceptional mining methods by measuring how dissimilar (exceptional) an object model is from the population model (see Figures 1 and 6). They call such measures “quality measures”. It is understood that different quality metrics are possible, and the right one may depend on the application domain and data set. For example, even for simple linear regression models Duivesteijn et al. present several effective quality metrics. In that spirit, there is room for different quality metrics for log-linear relational models. Our idea was to propose and develop log-likelihood based quality metrics such as LR and ELD for such models. Taking these two metrics together as options for the EMM framework, we believe we have shown that EMM + SRL is a novel and effective approach for relational outlier detection. We did not mean to suggest that LR is a bad method that should not be used; when we introduce it, we refer to it explicitly as a good outlier detection method. In our experiments, it is close to ELD on the synthetic data as you note, and the best method on one real-world dataset. This is consistent with the general EMM view of developing and investigating different quality metrics as alternatives.

Now for detailed comments.

This revision adds LR to the experiments and discusses its relationship to ELD. However, it fails to provide a theoretical justification made to LR to obtain ELD. The stated motivations are avoiding the cancelation of differences and greater interpretability. There may be some interpretability advantage to ELD, although I think that LR could provide similar insights by looking at the divergence of each conditional distribution.

ELD and LR are equally interpretable, in the sense that the rankings can be explained by the same type of drill-down analysis presented in Section 8.3. We state this now explicitly in 8.3 and also in the introduction.

Avoiding cancelations, however, is not a strong justification for ELD, unless it can be shown that LR fails to detect clear outliers because of these cancelations. The toy examples demonstrate that LR obtains different numbers than ELD (as expected), but they do not show that LR ranks true outliers below normal instances.

LR is guaranteed to be positive for any example that is strictly different from a fixed normal distribution, so it works if one focuses on a single example as we do in our toy examples. One has to consider a set of instances as we do in our experiment. The point of the cancellation argument is that when one has a set of normal instances that deviate from the population distribution, and also a set of outliers that deviate from the population distribution, then the loss of information from cancelling differences with opposite signs makes it more difficult to separate the two. There is no theorem that guarantees this is always the case for every dataset, one can only check sample datasets to evaluate empirical trends, as we did in our evaluation section.

From the text: "Table 5 illustrates the undesirable cancelling effects in LR."  
We struck “undesirable”.

Just because some of the terms in the sum are negative does NOT mean that the method is wrong. Entropy also includes positive and negative terms in its summation; however, taking the absolute value of each term in the entropy calculation would destroy its theoretical properties without providing any clear benefit. This seems to be the difference between LR and ELD as well. LR has a much stronger theoretical foundation than ELD; therefore, ELD requires strong empirical evidence to justify its use.

We have introduced section 6 Theoretical Analysis and Comparison to add theoretical understanding if not justification. It depends on what one means by “theoretical foundation”. The Neyman-Pearson lemma states that LR is uniformly most powerful for a simple hypothesis test with fixed alternatives. Our setting involves a composite test. In that case, no theoretical results are known about the relative type I/II errors of different divergences, not even for standard divergences like \Chi^2, Hellinger, total variation that have been used for decades. If there were a result that says “metric A is always better than metric B” than Metric B would cease to be used. Of the established metrics, tehre is no single metric or outlier method guaranteed to outperform the others.

What we were able to do is apply the standard Taylor series analysis for both LR and ELD. For LR, it is used to prove the celebrated result that LR can be approximated by Chi^2. For ELD, we obtained an approximation by total variation distance (L1 norm). This is a standard well-established metric on distributions, with good theoretical properties, so perhaps this provides some reassurance that we can expect ELD to have satisfactory theoretical properties too. The Taylor series also gives a clear view of the cancelling phenomenon: the first-order term for LR cancels out to 0, and for ELD it is exactly total variation. This clarifies where and how the cancelling occurs.

In Table 11 (real data), ELD outperforms LR in 4 out of 5 cases. This is somewhat interesting, but not strong enough support given the lack of theoretical justification or simple illuminating examples.

We added precision@k type of results that show the failure to indicate large deviations more clearly than AUC. . Again, we think that both LR and ELD are good ways to apply the EMM framework to relational data; LR is tried and true and ELD is novel and promising, given our evaluation. We have rewritten accordingly the introduction, discussion of empirical results, and conclusion.

OTHER MINOR COMMENTS:  
  
1. Regarding pseudo-likelihood and likelihood: It appears that pseudo-likelihood is being used as defined by Schulte (2011). However, the most common definition of pseudo-likelihood is from Besag (1975), and this pseudo-likelihood has already been used for learning probabilistic relational models (e.g., MLNs -- (Richardson & Domingos, 2006)). Besag's definition is also the one found on Wikipedia.  
  
Please rename this function or at least clarify the difference from the standard, widely-known and widely-used pseudo-likelihood function.

We changed to the term “normalized log-likelihood” due to Xiang and Neville (they call it “normalized” because counts are standardized to frequencies).   
  
2. "Khot et al. introduced a non-parametric relational one-class classification based on first-order trees. They proposed a tree-based distance metric to discover new relational features and to differentiate relational examples [23]."  
  
Please explain why the work by Khot et al. isn't a suitable baseline. I imagine it could be justified, but it's worth including a sentence to make the reasoning explicit.

Khot et al emphasize that their method is not suitable for outlier detection. Their reasoning is discussed in their Related Work section, we gave a brief summary of it in our discussion in our Related Work Section 2.

3. "log-likelihood distance (ELD)" -- you should explain why the initials "ELD" are an appropriate abbreviation for a log-likelihood distance, because it's not obvious.

we meant expected log-likelihood distance. We made sure the abbreviation was explained before the acronym was used, including E = expectation.

4. In section 4, the subsections are listed as subsubsections (e.g., 4.0.1 instead of 4.1).

Thank you fixed.

5. Section 6.1: The definition of LR\_i is missing parentheses around the difference of logs.  
The formula is fixed. Thanks

6. "The case studies illustrate that our outlier score is easy to interpret, because the Bayesian network provides a platform that makes the detected outliers very easy to interpret."  This is circular reasoning.  
Right, thank you. Changed to “because the Bayesian network provides a local decomposition of log-likelihood differences.”

7. "statistical-relational learning, a recent field that combines AI and machine learning" -- I don't think this is a very good description of SRL, since it could apply just as well to Alpha-Go (which also combines AI and machine learning and is much more recent).  
Right. We were missing a reference to relational data, changed this thank you.

8. I suggest defining "i.i.d." before its first use.  
good point. Changed in abstract and introduction.

9. Typos: "the the likelihood ratio" --> "the likelihood ratio"; "decomposing the log ration" --> "decomposing the log ratio"

Thanks, we fixed them in the text.  
  
  
  
Reviewer #3: Table 4 is not mentioned in text. Moreover, it seems to be out of place. It should probably appear in Sec 3.3, after the PBN figure has been introduced.

Thank you so much for your detailed comments. We call out Table 4 now, but kept it in section 3.1 where database frequencies are defined.

Pg 10, line 32: How are database frequencies different from data table frequencies? Aren't data tables also databases.

A relational database contains typically several tables. In the special case where the database contains only one table, the database frequency as defined by us (actually Halpern and Bacchus) coincides with the standard data table frequency. From a logical point of view, the case with only one table corresponds to unary predicates with one argument only (see <https://oschulte.github.io/srl-tutorial-slides/ch1-data.pptx> for a tutorial explanation).

Page 11, line 26-28: It is not clear how are the probabilities in the equations computed. Can you please add the symbolic names (e.g P(shotEff = high...)) to equations.

We expanded equations like 0.1 \times \ln(0.44) with the symbolic terms associated with these numbers (see Section 3.4). We hope this is what you meant. It is a good idea in any case.

I appreciate the re-ordering but now the descriptions and images are completely out of order. Table 5 refers to FD and ELD that have not yet been introduced. Also we don't know what the "high" and "low" correlation cases mean. Section 7.1 describing the synthetic dataset should be introduced in Section 5. It is a good running example to compare the issues with various scoring methods. Table 5 should follow the description of FD and ELD (maybe in section 6.2).

Oh dear. we have reorganized the sections. Now in section 4 and 5 we introduce the scores and the examples of score computations are at the end of section 5. We have expanded the caption of figure 7 with a brief description of the high and low correlations, in addition to the description in the text.

Table 7 is not mentioned anywhere in the text.  
Thanks for noticing this, we have fixed this issue in section 7.2

Page 22, line 30: Can you elaborate on how the center was computed for discrete features and why that results in the same distances for each individual.

According to the RIBL work, the distance between two sets is the maximum distance between data points in that set. For example if we want to calculate the distance between these two sets {(cheese, butter), (cat, dog)} {(cheese, bread), (cat, dog)}, we need to calculate the distance between this data point (cheese, butter) and (cheese, bread) as the other members in the sets are equal so distance is zero. In case of our synthetic data, the normal population is {(1,1), (0,0), (1,0), (0,1), (1,1), (1,0) ….} and since we have only two binary features, the data points in the outlier object set would be the one that we have seen normal case. So there won’t be any instance different and the distance would be zero.

Section 8.3: Am I correct in assuming that the model gave a higher outlier score to some object in the normal class than most of the actual outliers ? Even with few outlier feature values, I would at least have expected most of the outlier class objects to score higher than the normal class object.  
You are right, we call those i- class outlier. An example would be a striker which is exceptional than other strikers in the normal class. This happens when the injected anomalies are from a class with similar features (e.g. midfielders vs. strikers). In-class outliers reflect the lack of ground truth for real-world data which is what makes evaluating outlier detection methods difficult: an exceptional object from the normal class may be statistically more different from the class average than members from another class of objects. This is another reason why being able to explain the outlier scores is so important.

Section 9: I would argue that an additional limitation would be the requirement that this approach needs sufficient data for an object to be able to reliably learn the object model.  
While we agree with your point we believe this is the limitation of most outlier methods and assumed it is known and did not list it as a limitation specific to our approach. It is also (indirectly) addressed by our first limitation of not providing a binary decision. For example, if the decision is made using a statistical test, an object with too few data will lack sample size to be judged significantly different from the normal population.

Page 26, line 13: Which table contains these coefficients ranging from 0.45 and 0.82 ?  
Table 16, now explicitly called out, thank you.

Scatter plots font size is too small and seems that there is enough space to use a larger font size.

We have increased the font size in those plots thank you.

Page 30, line 25: Seems a stretch to say that it can "identify future stars" given that the algorithm identifies them based on their successful performance. I would argue that these players are already stars, albeit relatively under-paid.  
Agreed. We now refer to under-paid players.

Section 10.3: Comparison to other metrics should be moved up before the sections that delve into the detailed analysis of the proposed method.

The difficulty is that we introduce the concepts mentioned in Section 10.3 and Table 20 gradually in Section 10 first (e.g. NHL success metrics). In this way Table 20 is meaningful to the reader when it is shown. As a compromise, we briefly summarize at the beginning of Section 10 that ELD has the strongest success correlation except for log-likelihood. This order is our best shot, but we wouldn’t insist on it.   
  
Minor issue:  
Pg6, Line 43: The domain of predicate -> The range of predicate  
Fig 2. P(Action) = T) -> P(Action) = T  
properly relational ->  truly relational ?  
Page 20, line 18: Table 7.3 -> Table 8  
Page 22, line 45: ration -> ratio  
Table 14: Bottom Team -> Teams

Thank you so much for pointing out the flaws and typos. We have fixed them in the text.