Reviewer Replies:

Thank you for the comments which have helped us to substantially improve the content and organization of the paper. We have reorganized the content so that our new ELD method appears immediately after the necessary definitions have been introduced. The likelihood-based comparison methods are then introduced following a lesion structure, where we explain how they remove different components of the ELD score.

\*I find it odd that Figure 2 is used for both a BN and PBN, I guess I'd use something that highlights the difference between them.

That’s a great suggestion, thank you. We’re building this up now much more slowly: first, we show a non-relational intuitive BN (about movie genres; Figure 2). Then we show how the standard propositional BN syntax is a special case of the more elaborate first-order syntax, by a simple expanstion of the first non-relational BN to match the PBN syntax. Then we present an example of the full PBN syntax, for which at this point the reader should be prepared.

The theoretical justification for ELD is still limited. There are two differences from LR: the terms in the summation are decomposed into marginal terms and mutual information terms, and absolute values are used in place of a direct sum. The reason given for this is that, in outlier detection, we do not want differences to cancel out -- the average difference between (0,0) and (-1,1) (over all dimensions) is zero, even though they're different. This example is somewhat misleading, since the KL divergence (or log of the likelihood ratio) is non-negative, and it is only zero when the distributions are identical. In other words, even when some differences are positive and some are negative, they will never fully cancel out. For each configuration of a variable's parents, the KL divergence between the global and object (conditional) distributions will never be negative. Thus, if terms are grouped by child variable and parent configuration, each group of terms will have a

positive sum.

Thank you for this observation. We have taken explicit note of the fact that the KLD or LR is guaranteed to be positive. We have also removed the example of the vectors (0,0) vs. (-1,1) which is perhaps misleading in the context of comparing probability distributions. (See section with new number 4.0.4).

The explanation of pseudo-likelihood vs. log-likelihood for the relational models is unclear. (2) appears to be the definition of log-likelihood divided by the number of objects in the dataset. However, this is referred to instead as pseudo-likelihood, which would make sense if the parent-child graph is cyclic, as in a dependency network. For the rest of the paper, I assumed that all of the distances were either based on a relational BN (in which case the log-likelihood is easy to compute) or a relational BN with cycles (in which case the summations are related but not equivalent to log-likelihood).

We read this point as asking whether the pseudo-likelihood score (2) is normalized as the term “likelihood” would imply. Our paper here follows previous work, especially reference [40]. This reference discusse the conditions under which the relational log-likelihood score is normalized; acycility is indeed one of them. As discussed in [40], the score (2) can be used for object-relational data whenever a normalized likelihood score would be used for non-relational object data. This is what we are doing in our paper. We have added an explicit pointer to the normalization issue and a footnote with some brief explanation. We are diffident about adding a footnote as we agree that the question of normalization is an important, interesting and profound point. However, it is not a focus for our paper, which is about scores for outlier detection so we hope a footnote will provide information for interested readers without interrupting the flow of our main argument.

With regards to finding correlations between success and anomaly, I would have expected really unsuccessful players and teams to also be marked as anomalous. Why doesn't the approach mark really bad teams/players as anomalous too ? I would rather see the correlation between the ELD score and |success measure - mean success measure|.

We discussed this issue above Section 11.1. Our main statement was “Our interpretation of the correlation between ELD and success, rather than failure, is that our domains featured skilled individuals, such that the average is quite successful already.”. We have rewritten this part so that it is clearer that it is discussing the point about not singling out unusual players. We also note this as an issue for future research, for example by applying outlier detection to datasets that feature a range of skills, rather than professional performance.