A Proposal for Statistical Outlier Detection in Relational Structures

Fatemeh Riahi, Oliver Schulte, Qing li
{oschulte, sriahi, liqing}@sfu.ca
School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

Abstract
This paper extends unsupervised statistical outlier detection to the case of relational data. For nonrelational data, where each individual is characterized by a feature vector, a common approach starts with learning a generative statistical model for the population. The model assigns a likelihood measure for the feature vector that characterizes the individual; the lower the feature vector likelihood, the more anomalous the individual. A difference between relational and nonrelational data is that an individual is characterized not only by a list of attributes, but also by its links and by attributes of the individuals linked to it. We refer to a relational structure that specifies this information for a specific individual as the individual’s database. Our proposal is to use the likelihood assigned by a generative model to the individual’s database as the anomaly score for the individual; the lower the model likelihood, the more anomalous the individual. As a novel validation method, we compare the model likelihood with metrics of individual success. An empirical evaluation reveals a surprising finding in soccer and movie data: We observe in the data a strong correlation between the likelihood and success metrics.

Introduction
Outlier detection is an important data analysis task in many domains (Hodge and Austin 2004). Statistical approaches to unsupervised outlier detection are based on a generative model of the data. The generative model represents normal behavior. An individual entity is deemed an outlier if the model assigns sufficiently low likelihood to generating its features. We propose a new method for extending statistical outlier detection to the case of relational data. While in this paper we employ Bayes nets as our generative model class, the anomaly detection method can be used with any statistical-relational model (for example, Markov Logic Networks, Probabilistic Relational Models, Bayes Logic Programs etc.) Our approach opens up outlier detection as a promising new application area for statistical-relational models.

A generic Bayes net (BN) structure is learned with data for the entire population. The nodes in the BN represent features for links, of multiple types, and attributes of individuals, also of multiple types. A difference between relational and nonrelational data is that an individual is characterized not only by a list of attributes, but also by its links and by attributes of the individuals linked to it. We refer to the substructure comprising this information as the individual database. The model likelihood is computed by using the model parameter values that assign the highest possible likelihood to the individual database (Vuong 1989). The model likelihood defines an outlier score: the lower the model likelihood, the more anomalous the individual. A low model likelihood indicates that there are edges missing from the generic BN that represent an association or correlation that is present in the individual database.

As an alternative anomaly measure, we also consider a variant where a generic class-level BN model is compared with a specific BN model learned for the individual database. The model log-likelihood ratio (LR) is the log-ratio of the individual model likelihood to the generic model likelihood. A high model log-likelihood ratio indicates that the individual BN contains edges that are missing from the generic BN, but represent an association that is present in the individual database. Figure 1 shows a flow chart for computing the model likelihood and likelihood ratio. Note that the likelihood ratio is an inherently relational concept: if an entity is characterized by a flat feature vector, the individual database would contain just a single data point, which would not support learning a individual BN model.

Evaluation
We analyze two real-world data sets, from the UK Premier Soccer League and the Internet Movie Database (IMDB). Unsupervised anomaly detection methods are notoriously difficult to evaluate because of the lack of ground truth labellings of anomalous cases (Cansado and Soto 2008; Xu and Shelton 2010). In this paper we take a new approach, that compares likelihood-based anomaly metrics with other meaningful and independent metrics for ranking individuals. Our reference metrics are success rankings of individuals, e.g. League Standing of a team. Success rankings are one of the most interesting features to users, and therefore correlations to success rankings are of special interest. Strong correlations between likelihood metrics and meaningful independent success metrics provide evidence that the likelihood metric is meaningful as well. What we discover in the
data are surprisingly strong correlations between typicality and success.

**Contributions** The generative modelling approach to outlier detection is well established for nonrelational data. The novelty diagram in Figure 2 illustrates conceptual similarities and differences between that previous work and our new proposal. Our main contributions may be summarized as follows.

1. Two new statistical outlier metrics for relational data.
2. A new application of relational learning methods, for outlier detection.

**Related Work on Outlier Detection**

To our knowledge, this is the first work that uses a Bayes net generative model for relational outlier detection. Our work applies two new advances in statistical-relational learning, a recent field that combines AI and machine learning (Getoor and Taskar 2007; Domingos and Lowd 2009). (1) We use a tractable definition of the likelihood function for a Bayes net given a (sub)database that generalizes the standard definition for the nonrelational case (Schulte and Khosravi 2012; Alsanie and Cussens 2012). (2) We apply the learn-and-join algorithm (LAJ), a state-of-the-art Bayes net structure learning method for relational data (Schulte and Khosravi 2012).

**Relational Outlier Detection.** Many approaches to relational outlier detection are based on discovering rules that represent the presence of anomalous associations for an individual or the absence of normal associations (Maervoet et al. 2012; Gupta et al. 2013). Such rules are informative, but they are not based on a statistical model and do not provide a single aggregate outlier measure for each individual.

The community detection approach uses the relational structure to identify communities (Gao et al. 2010; Sun et al. 2013). Given a (soft) assignment of individuals to communities, outliers can be defined similar to the nonrelational feature vector setting, where normal behavior is measured with respect to a community. Our model also aggregates information from individual features and links of different types, but does not require the identification of communities.

**Nonrelational Outlier Detection.** Bayesian networks have been successfully applied for unsupervised anomaly detection (Hill, Minsker, and Amir 2007; Xu and Shelton 2010; Babbar and Chawla 2010) with nonrelational i.i.d. data. For nonrelational data, previous evaluation methods include the following. (i) synthetic models with synthetic normal and abnormal data points that provide ground-truth labels for comparison. (ii) Real-world datasets where domain expertise was used to evaluate whether an outlier identification was correct. In future work we will extend our evaluation to include this methodology.

**Bayes Nets for Relational Data**

We adopt the Parametrized Bayes net (PBN) formalism (Poole 2003) that combines Bayes nets with logical syntax for expressing relational concepts.

**Functor Terms** A functor is a function or predicate symbol. Each functor has a set of values (constants) called the **domain** of the functor. The domain of a predicate is \( \{ T, F \} \). A predicate of arity at least two is a **relationship** functor. Relationship functors specify which individuals are linked.
Other functors represent features or attributes of an individual or a tuple of individuals (i.e., of a relationship). A population is a set of individuals. A term is of the form \( f(\sigma_1, \ldots, \sigma_k) \) where \( f \) is a functor and each \( \sigma_i \) is a first-order variable or a constant denoting an individual. In the context of Bayes nets, we refer to terms as \textbf{Parametrized Random Variables} (PRVs). The grounding concept represents moving from the population-level to the individual level. A grounding replaces each first-order variable in a term by a constant; the result is a ground term. A grounding may be applied simultaneously to a set of terms. Relational data can be represented in terms of data tables that show the value of ground terms.

**Example** The Opta dataset represents information about premier league data (MCFC Analytics 2012). The basic populations are teams, players, matches, with corresponding first-order variables \( T, P, M \). Table 1 specifies values for ground terms.

Table 1: Sample Population Data Table (Soccer). The first three column headers show first-order variables for different populations. The remaining columns represent features.

<table>
<thead>
<tr>
<th>MatchId</th>
<th>TeamId</th>
<th>PlayerId</th>
<th>First Goal</th>
<th>TimePlayed</th>
<th>ShotEff</th>
<th>Result</th>
<th>ShotEff(WA)</th>
<th>TeamFormation(WA)</th>
<th>Result(WA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>117</td>
<td>WA</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0.53</td>
<td>loss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>WA</td>
<td>0</td>
<td>0</td>
<td>85</td>
<td>0.57</td>
<td>loss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>WA</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0.55</td>
<td>win</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A novel aspect of our paper is that we learn generative models for specific individuals as well as for the entire population. The appropriate data table for a target individual is formed from the population data table by restricting the relevant first-order variable to the target individual. For example, the individual database for target individual Team \textit{Wigan Athletic}, forms a subtable of the data table of Table 1 that contains only rows where TeamId = WA; see Table 2.

**Bayesian Networks** A \textbf{Bayesian Network} (BN) is a directed acyclic graph (DAG) whose nodes comprise a set of random variables and conditional probability parameters (Pearl 1988). A \textbf{Parametrized Bayesian Net} (PBN) is a Bayesian network structure whose nodes are PRVs. For most of the paper we refer to PBNs simply as Bayesian networks, and to PRVs simply as random variables.

The relationships and features in an individual database define a set of nodes for Bayes net learning. Figure 3 shows a sample Bayes net for the entire Team population and another one for Wigan Athletic. We use the following notation.

- \( D_P \) is the database for the entire population; cf. Table 1.
- \( D_t \) is the restriction of the input database to the target individual; cf. Table 2.

![Figure 3](image-url)
The log-likelihood ratio $LR$ is the difference between the log-likelihood of the individual BN on the individual database and the log-likelihood of the generic BN on the individual database:

$$L(M_t, D_t) - L(M_P, D_t),$$

which we typically abbreviate as $LR$.

### Anomaly Score for Bayesian Networks

We now illustrate how the above recipe can be applied with Bayesian networks, which we utilize in this paper. For the model likelihood function $L(M, D)$, where $M$ denotes a Bayesian network, we employ the random selection pseudo-log-likelihood (Schulte 2011). This pseudo log-likelihood can be computed as follows for a given Bayesian network structure and database.

1. For each parent-child configuration, use the maximum-likelihood estimate of the conditional probability of the child given the parent.
2. Multiply the logarithm of this estimate by the instantiation proportion of the parent-child configuration. The instantiation proportion is the number of instantiating groundings, divided by the total number of possible instantiations.
3. Sum this product over all parent-child configurations and all nodes.

**Example.** In the BN structure of Figure 3, consider the parent-child configuration as shown in Figure 3(b). Suppose that the data show that team WA played 20 matches in formation 1 with team shot efficiency at level 2 and won 12 of these. Then the maximum likelihood estimate of winning given the parent values is $12/20 = 0.6$. The instantiation count is 10, so the instantiation proportion is $10/20 = 0.5$. This parent-child configuration contributes $\ln 0.5 \cdot 0.6 = -0.415$ to the aggregate pseudo-likelihood.

The combination of Bayesian network and the pseudo-likelihood has several attractive properties as an anomaly metric. (1) It can be computed in closed form given the sufficient database statistics (the instantiation counts). This is faster than computing conditional likelihoods for each individual ground term. (2) Instantiation counts are standardized to be proportions. This avoids giving exponentially more influence to features with more groundings (Schulte 2011).

### Bayes Net Learning

We need to learn a Bayes net structure for the entire population database as well as to each individual database. We apply the learn-and-join algorithm to obtain these Bayes net models. Any structure learning method can be used, the details of the method do not matter here. Table 5 shows the learning times for applying the learn-and-join algorithm to the whole database (generic BN) and the average for the individual databases (individual BN). The learning algorithms were executed with 64-bit Centos with 4GB RAM and an Intel Core i5-480 M processor. Learning on the soccer dataset is fast. Learning on the IMDB database is feasible at around 50 minutes. It takes the longest

---

**Table 3: Attribute Features.** $\mu =$ average, $\sum =$ sum. For relationships please see text.

<table>
<thead>
<tr>
<th>Individuals</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer-Player per Match</td>
<td><code>TimePlayed, Goals, SavesMade, ShotEff, PassEff, WinningGoal, FirstGoal, PositionID, TackleEff, DribbleEff, ShotsOnTarget</code></td>
</tr>
<tr>
<td>Soccer-Team per Match</td>
<td><code>Result, TeamFormation, \sum Goals, \mu ShotEff, \mu PassEff, \mu TackleEff, \mu DribbleEff</code></td>
</tr>
<tr>
<td>IMDB-Actor</td>
<td><code>Quality, Gender</code></td>
</tr>
<tr>
<td>IMDB-Director</td>
<td><code>Quality, avgRevenue</code></td>
</tr>
<tr>
<td>IMDB-Movie</td>
<td><code>year, isEnglish, Genre, Country, RunningTime, Rating by User</code></td>
</tr>
<tr>
<td>IMDB-User</td>
<td><code>Gender, Occupation</code></td>
</tr>
</tbody>
</table>

**Table 4: Summary Statistics for the IMDB and Soccer datasets.**

<table>
<thead>
<tr>
<th>Premier League Statistics</th>
<th>IMDB Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Teams</td>
<td>20</td>
</tr>
<tr>
<td>Number Players</td>
<td>484</td>
</tr>
<tr>
<td>Match Numbers</td>
<td>380</td>
</tr>
<tr>
<td>avg player per match</td>
<td>26.01</td>
</tr>
</tbody>
</table>
Empirical Distribution of Likelihood-Metrics

We show the distribution of the $-L$ and $LR$ values on the soccer and the IMDB datasets. We observe that subtracting the common trend in the population, tends to make individual deviations become more pronounced: The tail of the $LR$ distribution is longer than that of the $-L$ distribution. This is because the log-likelihood ratio cancels out areas of agreement between the generic and the individual BN models.

Figure 4 presents six histograms, two for each type of individual, comparing the generic BN likelihood $-L$ to the $LR$. Density estimation curves smooth the empirical distribution. The density curves were produced by R’s default density function (the S3 Gaussian kernel estimator). If we think of the anomaly metric as measuring a distance from the generic individual (e.g., the generic team), we observe a pronounced central tendency for that distance. We may interpret this mean as an average degree of atypicality for an individual’s pattern of associations. There is a fairly large tail to the left of the mean that represents individuals who are more typical than most, and a smaller tail to the right that represents individuals who are definite outliers in their class. We next relate the likelihood and likelihood ratio metrics to metrics that identify successful individuals.

Likelihood Metrics vs. Success Metrics

The aim of this section is to compare the LR metric with other meaningful metrics for comparing individuals. Our reference metrics are success rankings of individuals, shown in Table 6. Success rankings are one of the most interesting features to users. Strong correlations between an anomaly metric and meaningful success metrics provide evidence that the anomaly metric is meaningful as well. We measure correlation strength by the standard correlation coefficient $\rho$. The coefficient ranges from -1 to 1, where 0 means no correlation and 1 or -1 indicate maximum strength (Fisher 1921).

Table 6: Success metrics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Success Metric</th>
<th>Min</th>
<th>Max</th>
<th>Standard Dev.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>Rating</td>
<td>1.5</td>
<td>9</td>
<td>1.14</td>
<td>6.3</td>
</tr>
<tr>
<td>Soccer-Player</td>
<td>TimePlayed</td>
<td>5.0</td>
<td>34.2</td>
<td>10.15.89</td>
<td>1484.0</td>
</tr>
<tr>
<td>Soccer-Team</td>
<td>Standing</td>
<td>1.0</td>
<td>20</td>
<td>5.91</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Our main observation is that typicality correlates with success. These correlations are remarkable in at least two respects. (1) The strength of the correlation is surprisingly high: magnitudes range from 0.3 to 0.6 where 1 is the maximum strength. (2) The trend holds across two different domains, different types of individuals, and different success metrics. The correlation between typicality and success becomes most apparent when we divide the population into different subgroups or strata.

Team Standing

The reference metric is Standing in the final Premier League table. The most successful team has Standing = 1 and the least successful team has Standing = 20. Table 7 shows the correlation between the model likelihood metrics and Team Standing. For teams the number of parameters in the individual BN correlates with Standing as well so we include it. If we distinguish Top Teams (Standing ≤ 10) from Bottom Teams (Standing > 10), a strong positive correlation emerges between $-L$ resp. $LR$ and Standing, meaning that unusual teams tend to have worse (i.e., higher) standing.

The number of parameters correlates positively with Standing for Top Teams and Bottom Teams. We interpret this result as saying that teams with complex correlation patterns tend to perform worse. The plane of best fit is shown in the 3-D scatterplots of Figure 5.

Player Time Played

The reference metric is TimePlayed, the total time that a player played over all matches in the season. This metric was shown to correlate strongly with other success metrics, such as salary, on MLS soccer data (T.Swartz 2013). Table 8 shows the correlations between the likelihood metrics and TimePlayed. For model likelihood (magnitude) $-L$, there is a strong positive correlation with TimePlayed. The correlation is even greater for subpopula-
Figure 4: Generic BN model log-likelihood $-L$ (magnitude) vs. individual BN model log-likelihood ratio. The x-axis shows the generic model log-likelihood $-L$ resp. the log-likelihood ratio $LR$. Higher x-values indicate more anomalous individuals. The y-axis shows densities derived from the counts of x-values divided by total number of individuals. The plots include density estimation curves. $N =$ number of individuals.

Table 8: Players: Correlations of TimePlayed, with player-parameters, player $LR$ ratio and player log-likelihood $-L$.

<table>
<thead>
<tr>
<th>Genre</th>
<th>$p(LR, Rating)$</th>
<th>$p(L, TimePlayed)$</th>
<th>$p(-L, TimePlayed)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Genres</td>
<td>0.28</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Drama</td>
<td>0.28</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Action</td>
<td>0.42</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Sci-Fi</td>
<td>0.35</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Adventure</td>
<td>0.34</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Film-Noir</td>
<td>0.49</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Movies: Correlations of Average Rating with movie $LR$ ratio and movie log-likelihood $-L$, by Genre. Correlations with #Parameters are weak (not shown).

Movie Ratings The reference metric is the average user rating of the movie, called Rating. Table 9 shows the correlations between the likelihood metrics and Rating. For model likelihood (magnitude) $-L$, there is a positive correlation with Rating. This trend becomes stronger for strata defined by genres; similarly for the $LR$ metric. The same movie may fall into different genres so the strata are not disjoint. Thus unusual movies tend to receive higher ratings.

In summary, we observe strong correlations between the likelihood metrics and the success metrics. Among players and movies, less typical individuals tend to be more successful, whereas among teams, more typical individuals tend to be more successful. Thus the direction of the correlation appears to be domain-specific. The fact that success can be predicted from our proposed anomaly metrics is an interesting and potentially useful feature of these metrics. Our argument is that a correlation with a meaningful quantity provided evidence that our proposed metrics are meaningful as well.

Conclusion

We presented a new approach for applying Bayes nets to relational outlier detection. The key idea is to learn a graphical model to represent population-level associations, and compare how well the generic model fits individual-level associations. A variant learns another graphical model to represent individual-level associations, and compare how well each model fits the relational data that characterize the individual, using the model likelihood ratio.

We applied efficient state-of-the-art relational learning methods to construct one generic Bayes net for a given domain and many Bayes nets for individuals (e.g., 3060 models for IMDB movies). The likelihood ratio highlights individual deviations more than the simple model likelihood. The data show a surprisingly strong correlation between our proposed anomaly metrics and success metrics. We believe that closer investigation of the correlation between success and typicality is a fruitful topic for future work.

Instead of estimating parameters for the generic BN from the individual database, an option is to estimate parameters from the population database. The likelihood score then reflects not only individual deviations in qualitative model structure, but also in event frequencies. The difficulty is that individual frequency deviations conflate group membership features with causal connections. For instance, relatively few soccer players are goalies, so a BN with frequency estimates derived from the general population is biased towards viewing goalies as anomalous. Combining individual deviations in terms of qualitative structure, as discussed in this paper, and quantitative frequency into a single meaningful metric is a challenging and fruitful topic for future work.

Overall, our new model likelihood metrics for statistical-relational outlier detection showed promising results on the Soccer and IMDB dataset. They provide a new approach for applying AI techniques to a challenging and practically important topic.

References


Babbar, S., and Chawla, S. 2010. On bayesian network


Fisher, R. A. 1921. On the probable error of a coefficient of correlation deduced from a small sample. \emph{Metron} 1:3–32.


Gupta, M.; Gao, J.; Yan, X.; Cam, H.; and Han, J. 2013. On detecting association-based clique outliers in heterogeneous information networks. In \emph{ASONAM}. ACM.

Hill, D. J.; Minsker, B. S.; and Amir, E. 2007. Real-time bayesian anomaly detection for environmental sensor data. In \emph{In 32nd Conference of IAHR, the International Association of Hydraulic Engineering & Research}.


Internet Movie Database. Internet movie database. URL = http://www.imdb.com/.


Pearl, J. 1988. \emph{Probabilistic Reasoning in Intelligent Systems}. Morgan Kaufmann.


Sun, Y.; Han, J.; Zhao, P.; Yin, Z.; Cheng, H.; and Wu, T. 2013. Community distribution outlier detection in heterogeneous information networks. In \emph{ECML/PKDD (1)}, volume 8188 of \emph{Lecture Notes in Computer Science}, 557–573. Springer.

T.Swartz, Adriano Arce, M. P. 2013. Assessing value of the draft positions in major league soccer’s superdraft. \emph{The Sport Journal}.
