

# 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 fea-

tures 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

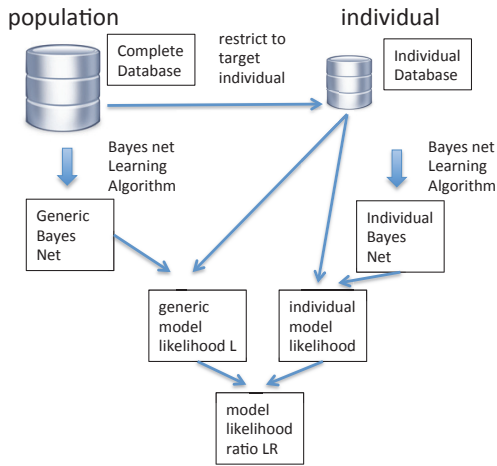


Figure 1: Computation of likelihood metrics for scoring anomalies. **Generic Model Likelihood  $L$** : Learn a generic Bayes net on the complete database. Score the generic Bayes net on the individual database. **Model Likelihood Ratio  $LR$** : Learn an individual Bayes net on the individual database. Compute the ratio of the generic BN/individual BN scores on the individual database.

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.

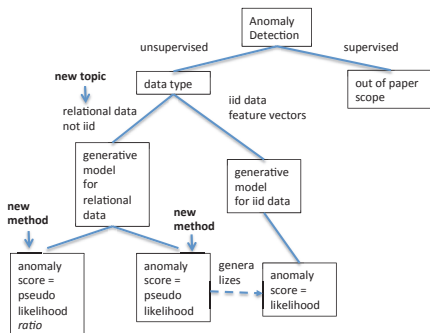


Figure 2: Novelty Diagram for conceptually positioning the new work in comparison with previous work.

## 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  $f(\sigma_1, \dots, \sigma_k)$  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 **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.

MatchId $M$	TeamId $T$	PlayerId $P$	First_goal( $P, M$ )	TimePlayed( $P, M$ )	ShotEff( $T, M$ )	result( $T, M$ )
117	WA	McCarthy	0	90	0.53	win
147	WA	McCarthy	0	90	0.35	loss
148	WA	McCarthy	0	85	0.57	loss
15	MC	Silva	1	90	0.59	win
175	MC	Silva	0	90	0.61	win
...	...	...	...	...	...	...

Table 2: Sample Individual Data Table, for team  $T = WA$ .

MatchId $M$	TeamId $T = WA$	PlayerId $P$	First_goal( $P, M$ )	TimePlayed( $P, M$ )	ShotEff( $WA, M$ )	result( $WA, M$ )
117	WA	McCarthy	0	90	0.53	win
147	WA	McCarthy	0	90	0.35	loss
148	WA	McCarthy	0	85	0.57	loss
...	WA	...	...	...	...	...

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 *WiganAthletic*, forms a subtable of the data table of Table 1 that contains only rows where TeamID = *WA*; see Table 2.

**Bayesian Networks** A **Bayesian Network (BN)** is a directed acyclic graph (DAG) whose nodes comprise a set of random variables and conditional probability parameters (Pearl 1988). A **Parametrized Bayes 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.

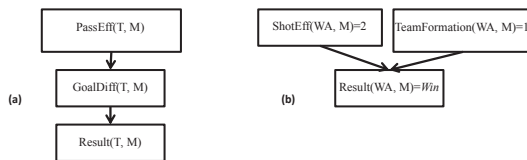


Figure 3: (a) Generic Bayes Net. (b) Individual Bayes net for Wigan Athletic (WA). The nodes in the individual BN are derived from the generic BN by setting  $T = WA$ .

- $\mathcal{D}_{\mathcal{P}}$  is the database for the entire population; cf. Table 1.
- $\mathcal{D}_t$  is the restriction of the input database to the target individual; cf. Table 2.
- $M_{\mathcal{P}}$  is a model (e.g., Bayesian network) learned with  $\mathcal{D}_{\mathcal{P}}$  as the input database; cf. Figure 3(a).
- $M_t$  is an individual model (e.g., Bayesian network) learned with  $\mathcal{D}_t$  as the input database; cf. Figure 3(b).

## Datasets

In this paper, data tables are prepared from Opta data (MCFC Analytics 2012) and IMDB (Internet Movie Database). Our datasets are available on-line (Khosravi et al. 2014). Table 3 lists the populations and features. Table 4 shows summary statistics for the datasets.

**Soccer Data** The Opta data were released by Manchester City. It lists all the ball actions within each game by each player, for the 2011-2012 season. For each player in a match, our data set contains eleven player features. For each team in a match, there are five features computed as player feature aggregates, as well as the team formation and the result (win, tie, loss). There are two relationship random variables,  $Appears\_Player(P, M)$ ,  $Appears\_Team(T, M)$ . We store the data in a relational database, with a table for each base population and a table for each relationship.

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

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

Table 4: Summary Statistics for the IMDB and Soccer data sets.

	Premier League Statistics	IMDB Statistics
Number Teams	20	Number Movies 3060
Number Players	484	Number Directors 220
Number Matches	380	Number Actors 98690
avg player per match	26.01	avg actor per movie 36.42

**IMDB Data** The Internet Movie Database (IMDB) is an on-line database of information related to films, television programs, and video games. The IMDB website offers a dataset containing information on cast, crew, titles,

technical details and biographies into a set of compressed text files. We preprocessed the data (Peralta 2007) to obtain a database with seven tables: one for each population and one for the three relationships  $Rated(User, Movie)$ ,  $Directs(Director, Movie)$ , and  $ActsIn(Actor, Movie)$ .

## Likelihood-Based Anomaly Metrics

In principle our approach can be used with any generative model class, such as Bayesian Networks, Markov Logic Networks, Probabilistic Relational Models, Bayes Logic Programs etc. All that is required is that a generative model define a **relational model log-likelihood function**

$$L(M, \mathcal{D})$$

that assigns a real number to a structure database  $\mathcal{D}$  given a model  $M$ . A common approach for defining the likelihood function to use a template semantics (knowledge-based model construction) for assigning a likelihood to a relational dataset (Getoor and Taskar 2007). To emphasize the general applicability, we state our definitions abstractly, then discuss how to implement them using Bayesian networks in the next section.

A generative model assigns a likelihood to a single data point. Low likelihood of a data point indicates an anomaly. Defining a likelihood for relational data is more complicated, however, because an individual is characterized not only by a feature vector, but by an individual database that lists the individual’s links and the attributes of linked entities. Our proposal in this paper is to use as an anomaly score the model likelihood assigned to the individual database. The individual database  $\mathcal{D}_t$  is defined by instantiating (grounding) the value of a population variable with the target individual, and extracting from the entire population database the facts that satisfy this instantiation; cf. Table 2. Every statistical-relational model defines a log-likelihood function  $L(M, \mathcal{D})$  for any structured database, so once the individual database  $\mathcal{D}_t$  is constructed, we can apply the log-likelihood function to obtain an anomaly score.

We examine two definitions of an anomaly metric that are based on the relational model likelihood function  $L(M, \mathcal{D})$ . Both metrics are defined so that *a higher value indicates a more anomalous individual*. The **generic likelihood metric** is the magnitude of the log-likelihood of a model learned for the general population, evaluated on an individual database (see Figure 1). Using the notation above, this quantity is denoted by

$$-L(M_{\mathcal{P}}, \mathcal{D}_t), \quad (1)$$

which we typically abbreviate as  $-L$ .

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:

$$L(M_t, \mathcal{D}_t) - L(M_{\mathcal{P}}, \mathcal{D}_t), \quad (2)$$

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  $L(M, \mathcal{D})$ , 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.

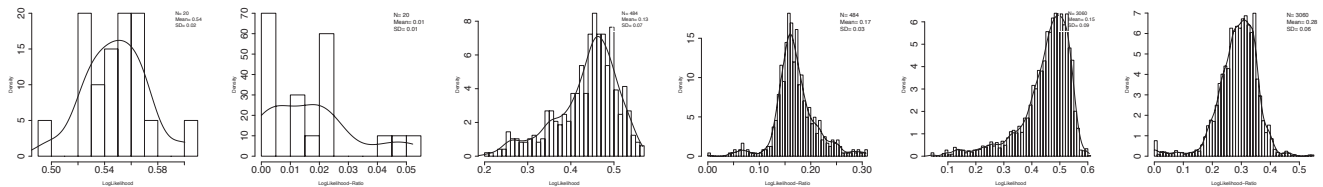
*Example.* In the BN structure of Figure 3, consider the parent-child configuration as shown in Figure 3(b). Suppose that the data show that team  $WA$  played 20 matches in formation 1 with team shot efficiency at level 2 and won 12 of these. Then the maximum likelihood estimate of winning given the 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 because of the large number of individuals (over 98,000 Actors) and fairly complex relationships. The subdatabase for an individual movie is small, so individual learning is fast, and it is possible to learn 3,060 individual models.

## 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



(a) Soccer: BN log-likelihood for each team.  $N = 20$  (b) Soccer: Log-likelihood ratio for each team.  $N = 20$  (c) Soccer: BN log-likelihood for each player.  $N = 484$ . (d) Soccer: Log-likelihood ratio for each player.  $N = 484$ . (e) IMDB: BN log-likelihood for each movie.  $N = 3,060$ . (f) IMDB: Log-likelihood ratio for each movie.  $N = 3,060$ .

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.

Database	Generic BN	Individual BN
Soccer-Team	10 min	0.33 min
Soccer-Player	10 min	0.25 min
IMDB-Movie	50 min	1.00 min

Table 5: Bayes Net structure learning time for different datasets.

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).

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 max-

Dataset	Success Metric	Min	Max	Standard Dev.	Mean
IMDB	Rating	1.8	9	1.14	6.3
Soccer-Player	TimePlayed	5.0	3420	1015.69	1484.0
Soccer-Team	Standing	1.0	20	5.91	10.5

Table 6: Success metrics.

	$\rho(\#Parameters, Standing)$	$\rho(LR, Standing)$	$\rho(-L, Standing)$
All Teams	-0.10	0.19	0.52
Top Teams	0.44	0.62	0.45
Bottom Teams	0.33	0.41	0.38

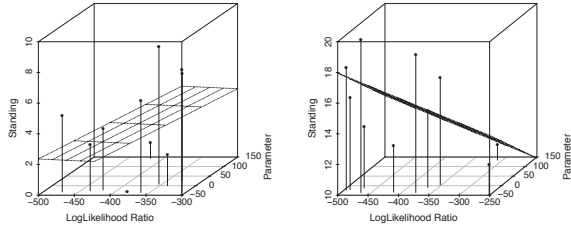
Table 7: Teams: Correlations of team-standing with team-#parameters, team  $LR$  ratio and team log-likelihood  $-L$

imum 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 ( $Standing \leq 10$ ) 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 scatter plots 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



(a) For bottom half teams. (b) For top half teams. The dependent variable is 10-standing.

Figure 5: Teams: Regression plane of Standing on the  $LR$  ratio and #parameters.

	$\rho(\text{Parameter}, \text{TimePlayed})$	$\rho(LR, \text{TimePlayed})$	$\rho(-L, \text{TimePlayed})$
All Players	0.00	0.34	0.20
Goalkeepers	-0.37	0.58	0.60
Defenders	0.09	0.26	0.54
Midfielders	-0.20	0.42	0.63
Strikers	-0.15	0.17	0.60

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

likelihood metrics and TimePlayed. For model likelihood (magnitude)  $-L$ , there is a strong positive correlation with TimePlayed. The correlation is even greater for subpopulations defined by the player positions Goalkeepers, Defenders, and Midfielders. For the log-likelihood ratio  $LR$ , there is a substantive correlation as well. *Thus unusual players tend to play more minutes.*

**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 pre-

Genre	$\rho(LR, \text{Rating})$	$\rho(L, \text{Rating})$
All Genres	0.28	0.13
Drama	0.28	0.29
Action	0.42	0.29
Sci-Fi	0.35	0.32
Adventure	0.34	0.37
Film-Noir	0.49	0.5

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).

dicted 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

- Alsanie, W., and Cussens, J. 2012. Learning recursive prism programs with observed outcomes. In *Proceedings of the ICML-2012 Workshop on Statistical-Relational Learning*.
- Babbar, S., and Chawla, S. 2010. On bayesian network and outlier detection. In Kumar, P. S.; Parthasarathy, S.; and Godbole, S., eds., *COMAD*, 125. Allied Publishers.
- Cansado, A., and Soto, A. 2008. Unsupervised anomaly detection in large databases using bayesian networks. *Appl. Artif. Intell.* 22(4):309–330.
- Domingos, P., and Lowd, D. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan and Claypool Publishers.
- Fisher, R. A. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1:3–32.

- Gao, J.; Liang, F.; Fan, W.; Wang, C.; Sun, Y.; and Han, J. 2010. On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10*, 813–822. New York, NY, USA: ACM.
- Getoor, L., and Taskar, B. 2007. *Introduction to statistical relational learning*. MIT Press.
- Gupta, M.; Gao, J.; Yan, X.; Cam, H.; and Han, J. 2013. On detecting association-based clique outliers in heterogeneous information networks. In *ASONAM*. ACM.
- Hill, D. J.; Minsker, B. S.; and Amir, E. 2007. Real-time bayesian anomaly detection for environmental sensor data. In *In 32nd Conference of IAHR, the International Association of Hydraulic Engineering & Research*.
- Hodge, V., and Austin, J. 2004. A survey of outlier detection methodologies. *Artif. Intell. Rev.* 22(2):85–126.
- Internet Movie Database. Internet movie database. URL = <http://www.imdb.com/>.
- Khosravi, H.; Man, T.; Hu, J.; Gao, E.; and Schulte, O. 2014. Learn and join algorithm code. URL = <http://www.cs.sfu.ca/~oschulte/jbn/>.
- Maervoet, J.; Vens, C.; Vanden Berghe, G.; Blockeel, H.; and De Causmaecker, P. 2012. Outlier detection in relational data: A case study in geographical information systems. *Expert Syst. Appl.* 39:4718–4728.
- MCFC Analytics. 2012. The premier league dataset. URL = <http://www.mcfc.co.uk/Home/The%20Club/MCFC%20Analytics>.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Peralta, V. 2007. Extraction and Integration of MovieLens and IMDb Data. Technical report, APDM project, Laboratoire PRISM, Université de Versailles, Versailles, France.
- Poole, D. 2003. First-order probabilistic inference. In *IJCAI*, 985–991.
- Schulte, O., and Khosravi, H. 2012. Learning graphical models for relational data via lattice search. *Machine Learning* 88(3):331–368.
- Schulte, O. 2011. A tractable pseudo-likelihood function for Bayes nets applied to relational data. In *SIAM SDM*, 462–473.
- Sun, Y.; Han, J.; Zhao, P.; Yin, Z.; Cheng, H.; and Wu, T. 2013. Community distribution outlier detection in heterogeneous information networks. In *ECML/PKDD (1)*, volume 8188 of *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. *The Sport Journal*.
- Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2):307–333.
- Xu, J., and Shelton, C. R. 2010. Intrusion detection using continuous time bayesian networks. *J. Artif. Intell. Res. (JAIR)* 39:745–774.