Data-Driven ESG Analysis: Predicting ESG Scores with Reduced Feature Complexity

Berend van Nieuwland
Cornell Tech
bnv7@cornell.edu

Dennis Chen
Cornell Tech
dc2247@cornell.edu

Jet Semrick
Cornell Tech
js3788@cornell.edu

Abstract

Environmental, Social, and Governance (ESG) scores are vital for assessing corporate sustainability, yet divergence in calculation and interpretation hinder their reliability. This study develops predictive models for ESG scores using panel data from S&P 500 companies (2013–2024), emphasizing interpretable methods with reduced feature dependence. Among the models, the Fixed Effects approach excelled, achieving an R^2 of 0.8443 on the test set, leveraging lagged ESG scores to capture temporal and entity-specific patterns. By addressing limitations such as non-stationarity and potential feature leakage in alternative models, this study underscores the Fixed Effects model's robustness and reliability. These results highlight the robustness of ESG trajectories as predictors of future performance, offering actionable insights for data-driven, sustainable investment strategies.

1. Introduction

The growing importance of Environmental, Social, and Governance (ESG) scores in evaluating corporate sustainability has driven both investors and researchers to seek robust and interpretable models for their analysis. ESG scores are critical metrics used to assess a company's sustainability efforts, its exposure to reputational risks, and its alignment with long-term ethical and financial objectives. Despite their widespread adoption, challenges persist in their calculation and interpretation due to the inherent subjectivity of rating methodologies and divergence across scoring frameworks. This underscores the need for data-driven approaches that enhance the reliability and usability of ESG scores for decision-making.

In this project, we aim to address these challenges by developing predictive models that leverage panel data to forecast ESG scores over time while minimizing reliance on large feature spaces. By reducing the number of features required for prediction, we seek to provide a more straightforward and interpretable approach, avoiding the complexity and potential overfitting associated with excessive feature reliance. This simplification ensures that the model remains accessible and practical for stakeholders, enabling its application in real-world investment decisions.

Our study focuses on companies within the S&P 500, providing a consistent and high-quality dataset while maintaining relevance to the investment community. Through the application of clustering methods, ElasticNet regression, and panel data models such as Fixed Effects, we explore the predictive power of temporal and company-specific characteristics in understanding ESG dynamics.

By combining advanced statistical techniques with a structured dataset, this project seeks to bridge the gap between theoretical ESG modeling and practical investment strategies. Our goal is not only to provide accurate forecasts of ESG scores but also to ensure that the models remain efficient, interpretable, and actionable. This research contributes to the broader field of sustainable finance by offering insights into how ESG metrics can be integrated into decision-making processes to align portfolios with both ethical considerations and financial performance.

The dataset, provided by the London Stock Exchange (LSEG), covers S&P 500 companies and features 870 metrics, filtered for the companies from the last 12 years (2013-2024). The dataset contains information across three pillars: Environmental (e.g., Resource Use, Emissions, Innovation), Social (e.g., Workforce, Community, Product Responsibility), and Governance (e.g., Management, Shareholders, CSR Strategy) (LSEG., 2024). The variable of interest from the dataset is the 'ESG_Score' variable, denoting the ESG score for a company in that year.

To prepare the dataset for analysis, several preprocessing steps were implemented to improve structure, reliability, and usability. First, the raw dataset, initially in a stacked format, was transformed into a panel format where each row corresponds to a unique company-year combination indexed by 'isin' and year. Features from the 'fieldname' column became distinct columns, and values were aggregated using the first method to resolve any duplicates. After pivoting, column names were flattened and standardized to remove hierarchical indexing and ensure consistency.

To address missing data, a systematic approach was adopted. Features with more than 10% missing values were excluded to maintain the integrity of the dataset and avoid potential biases. For the remaining missing values, a combination of forward-fill and backward-fill imputation was employed within each company group, identified by `isin`. This two-step method ensured that gaps were effectively handled, where backward-fill came into play for the edge-cases where forward-fill was not possible.

Ultimately, the steps included were: pivoting the dataset from long-format into a panel format, filtering features based on missing data thresholds, and imputing missing values. Testing the function confirmed that it produced a clean and analyzable dataset, structured as a time-series panel dataset indexed by isin and year. This preprocessing resulted in a dataset with 5833 entries, 99 features, corresponding to 483 companies over the last 12 years.

2. Related Work

Environmental, Social, and Governance (ESG) scores are well-studied metrics to assess the sustainability efforts of companies and their related operational risks. Previous work has primarily focused on calculating ESG scores for the purpose of making investment decisions. The primary issue faced by users is score divergence where ratings agencies and firms use different features and approaches to calculate scores.

Studies have divided divergence into three sources: scope, measurement, and weight (Berg et al. 2022). Scope divergence refers to situations where ratings are based on different sets of features. For example, one firm could evaluate lobbying activity whereas another firm might not which will create a divergence of the two scores.

Measurement divergence refers to a situation where rating agencies measure the same attribute using different indicators or methods. For example, a firm's labor practices could be evaluated on the basis of workforce turnover or by the number of labor-related court cases taken against the firm. Each method could be independently rationalized and lead to a significant divergence in scores.

Finally, weight divergence emerges when ratings agencies take different views on the relevance of certain features. In the above examples, one firm could argue labor practices are more relevant than lobbying whereas another firm could argue the opposite.

The conclusion of this study is that ESG scores are complex and challenging to consistently model due to the large number of available features. Specifically, climate risk, product safety, corporate governance, corruption, and environmental management typically see the most divergence in practice.

There are multiple potential solutions to this problem. First, firms could index multiple ESG ratings to normalize their own. Second, ESG features could be normalized between firms for specific companies. Third, researchers can construct more specific hypotheses around sub-categories of ESG performance, such as GHG emissions or labor practices. This avoids the problems of weight and scope divergence, but the risk of measurement divergence remains. Therefore, researchers should ideally work with raw data that can be independently verified.

Other studies focus on the usage of ESG scores rather than their calculation. Specifically, we can look at surveys of institutional investors to understand their use of ESG information for investment strategies (Amel-Zadeh et al. 2018).

The key result from this paper is that investors are more concerned by the reputational and financial risk related to ESG scores than other factors. Ultimately, firms care more about the financial implications of ESG compared to ethical implications.

ESG scores can help measure reputational risk by providing a measure of exposure to environmental controversy or regulatory change due to unsatisfactory practice. Firms use ESG data to alter their investment behavior or make alterations to existing investments to reduce risk.

ESG scores have been criticized for providing an inaccurate rating that is divorced from reality (Svanberg et al. 2022). By nature, the methodology of an ESG score is subjective and can lead to divergence, as explained above. This undermines the goal of having reliable sustainability metrics for companies and investors.

Suggestions to solve this problem revolve around taking a more data-centered and objective approach to creating a rating. Research shows that machine learning methods can be used to identify factors that are most correlated to environmental controversy, which allows firms to create models with less subjectivity.

In summary, the existing literature has demonstrated the ESG score is well studied and used frequently in the investment industry. However, challenges still exist around simplifying ESG calculations and providing less complex analysis for firms to use in their investment decisions. This gap highlights the importance of our study to develop strategies to accurately calculate ESG scores with a smaller set of features compared to existing calculation strategies.

3. Methods

This section will discuss the methodology employed in this exploration. First the clustering methodology will be discussed, followed by the details on predictive modeling.

3.1. Clustering Methodology

3.1.1 KMeans Model

The ESG dataset is multidimensional and has many unique features containing valuable information. In order to resolve this issue, we used principal component analysis (PCA) to reduce the dimensionality of the data and then trained a KMeans model to cluster the data. Clustering is a useful strategy because it allows us to group companies with similar ESG characteristics without perfecting the weight of each input.

To initialize the model, we started with principle component analysis and generated enough new components to capture 95 percent of variance in the dataset. Then, the elbow curve method was used to determine the optimal number of clusters for the dataset was three. Finally, these parameters were used to generate the clusters from the dataset.

As discussed in the related works section, firms often create diverging models for ESG scores. KMeans clustering provides a solution not relying on subjective weights or features. Rather we try to find patterns with correlated variables and categorize instead of predicting the exact ESG score for each company. In our model, we identified three clusters: low risk, moderate risk, and high risk. This model seeks to accurately categorize ESG risk, but is unable to give a precise prediction of the exact ESG risk score.

3.2. Predictive Methodology

3.2.1 ElasticNet Regression

One of the models used in this exploration is ElasticNet. ElasticNet combines the L1 penalty from Lasso (promoting sparsity and feature selection) with the L2 penalty from Ridge (offering stability and controlled coefficient shrinkage). By adjusting the relative weighting of these penalties, ElasticNet effectively reduces the feature space while maintaining model robustness, ultimately providing a balanced and flexible approach to regularized regression.

To determine the optimal ElasticNet hyperparameters, we conducted a grid search over a predefined parameter grid, varying both the penalty strength ('alpha') and the ratio between L1 and L2 regularization ('l1_ratio'). Using cross-validation, with 5 folds, and a scoring metric based on the negative mean squared error, we identified the combination of parameters that minimized the prediction error on the training set. This approach ensured a thorough and unbiased selection of model settings.

Lastly, to ensure robust evaluation and mitigate temporal bias, a temporal train-test split was employed. The training set spans 10 years (2013–2021), while the test set encompasses the three subsequent years (2022–2024). This approach ensures that the model is trained on past data and validated on future data, simulating a real-world prediction scenario.

3.2.2 Augmented Dickey-Fuller Stationarity Tests

As part of determining which model to use for the time-series analysis, Augmented Dickey-Fuller (ADF) tests were performed to determine stationarity of the time-series. The ADF test is designed to detect the presence of a unit root in a autoregressive process. It does so by regressing the first-differenced series on its lagged level and additional lagged differences to account for higher-order autocorrelation. Formally, for a time-series y_t, the ADF test estimates the following regression:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \epsilon_t$$

where $\Delta y_t = y_t - y_{t-1}$, is the first-differenced time series, γ measures the presence of a unit root, and the β_i terms capture autocorrelation patterns. Under the null hypothesis, $\gamma = 0$ indicating a unit root and thus non-stationarity. Rejecting this null hypothesis suggests $\gamma < 0$, meaning the series is stationary. The time-series investigated was `ESG_Scores`, aggregated over the companies and grouped per year.

3.2.3 Fixed Effects Model

To leverage the panel structure of the dataset and relax the assumption of stationarity, this study employs a Fixed Effects model to predict ESG scores. The model is constructed with two lagged ESG scores—one with a 1-year lag and another with a 2-year lag—as the independent variables.

The general form of a Fixed Effects model is expressed as:

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it} \tag{1}$$

where Y_{it} is the ESG score (`ESG_Score`) of company i at time t, α_i is the entity-specific, time-invariant fixed effect for company i, β the regression coefficient, X_{it} the regressors, which are the lagged ESG scores, and u_{it} the idiosyncratic error term.

The inclusion of entity-specific fixed effects (α_i) allows the model to control for unobserved heterogeneity across companies, such as differences in governance structures, industry, or long-term ESG strategies. By focusing on within-entity variation, the Fixed Effects model effectively isolates the impact of lagged ESG scores on current ESG scores, removing biases associated with time-invariant entity-specific factors. While the model focuses on changes within each company over time, it also utilizes data from multiple companies to estimate consistent regression coefficients. This approach balances the ability to uncover broader patterns across companies with the precision of capturing firm-specific dynamics.

Lastly, a temporal split, just like in the ElasticNet methodology, was implemented for the training and test set split. In the case of the Fixed Effects model, the train set spanned seven years (2015-2021) instead of starting from 2013, due to the introduction of the lagged variables. The test set remained the horizon of three subsequent years (2022-2024).

3.2.4 Hausman Test

The Hausman Test is a statistical procedure used to evaluate whether a Fixed Effects (FE) model or a Random Effects (RE) model is more appropriate for a given panel dataset. The test specifically examines whether the unique entity-specific effects (α_i) are correlated with the regressors (X_{it}, as in an RE model the effects are assumed to be uncorrelated with the regressors, whilst in an FE model this is not a necessary assumption. Subsequently, the Hausman Test evaluates the hypothesis:

$$H_0 : Cov(\alpha_i, X_{it}) = 0$$

$$H_1 : Cov(\alpha_i, X_{it}) \neq 0$$

Thus, if we reject the null-hypothesis, the test suggests the entity-specific effects are correlated with the regressors, implying that a Fixed Effects model would be appropriate. Conversely, if we do not reject the null-hypothesis, an RE model might be more suitable.

4. Results

In this section, the modeling results will be discussed. First, the clustering analysis is explored. After the clustering analysis, the results of the predictive modeling will follow.

4.1. Clustering Analysis

The KMeans model produced three clusters correlating to various risk profiles of ESG scores. In general, a higher ESG score represents lower risk to investors and a low ESG score is higher risk. In figure 1 we can see the data did not produce well defined clusters leading to overlapping categorization and potential mis-classification. The results indicate that clustering is at best a moderate model for ESG risk, but will not produce perfect results.

The accuracy of the cluster classification is 77.03 percent. This was calculated by taking the cluster value and comparing to the actual categorization of each entry in the dataset.

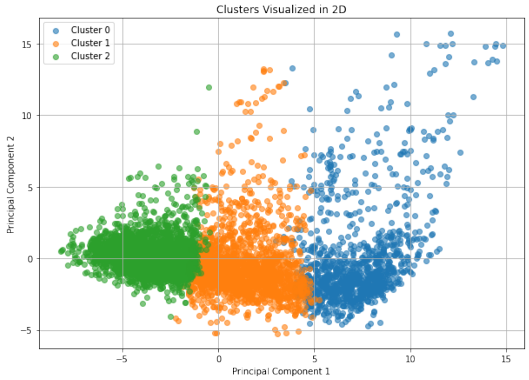

Figure 1. Visualization of Clusters

4.2. Predictive Analysis

4.2.1 ElasticNet Regression

We began our predictive modeling with an ElasticNet model, as it helps reduce the feature space and applies both L1 and L2 penalties for a balanced regularization approach.

Table 1. ElasticNet Regression Results

	RMSE	R^2	MAE
Training	0.0228	0.9853	0.0172
Testing	0.0202	0.9843	0.0155

The results are great... A little bit too great. This caused us to dive deeper into the validity of these results, performing residual analysis, investigating correlations and visualizing QQ-plots (these have been excluded from the report for brevity, and are not necessary for the actual prediction). The model passed all of these tests, perfectly. However, we also further investigated the scoring methodology of LSEG, where we uncovered that the ESG scores are calculated as a weighted average of the collected metrics. Given that ESG scores are calculated as a weighted average of other metrics, there's a real risk that features may inadvertently ”leak” target information and that the model is trying to learn the ESG-calculation formula instead of the underlying patterns, causing it to perform artificially well. This result, together with the aim to reduce dependency on a large feature space as much as possible, caused us to move towards investigating time-series models that require only a single (base) feature: lagged versions of the target variable.

4.2.2 Stationarity Test Results

Initially, we wanted to explore ARIMA(X) models as predictive methods for the time-series modeling. However, the stationarity tests performed through the ADF method suggested a key assumption for ARIMA(X) models was violated:

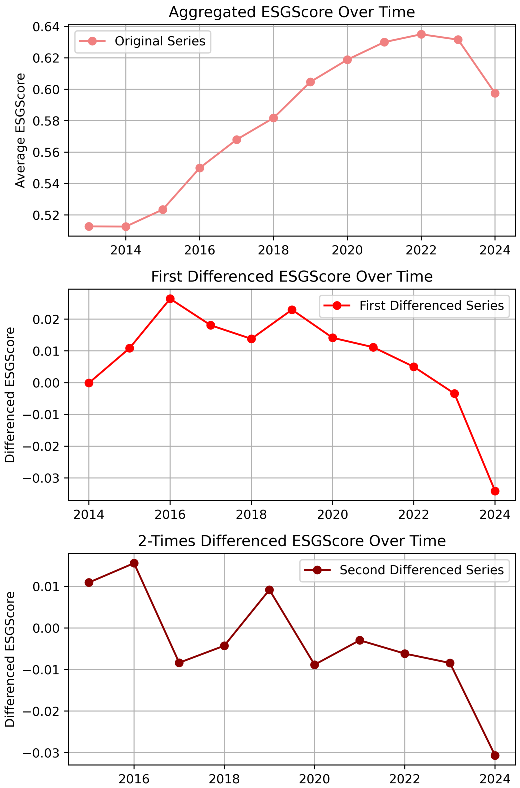

Figure 2. Visualization of ESG Score Time-Series

Table 2. ADF Stationarity Test Results

Series	**ADF Statistic**	**p-value**
Original	-0.8461	0.8052
1st Diff.	3.6350	1.0000
2nd Diff.	1.9664	0.9986

The results in Table 2 and Figure 2 suggest non-stationary data, even after first -and second-order differencing, and thus that ARIMA models are not suitable. This causes the need for exploring alternative methods. As we are in a panel structure setting, a logical next exploration are panel models, where we settled on a Fixed Effects model being able to best capture the problem at hand.

4.2.3 Fixed Effects Model Results

The Fixed Effects model proved to be the most reliable model. Its results are diplayed in Table 3.

The Fixed Effects model suggests that historical ESG performance of S&P500 companies is a strong predictor of their future ESG scores. The high R^2 values (over 0.82 on the training set and over 0.84 on the test set) indicate that a substantial portion of the variation in ESG scores can be explained by the model, underscoring the stability and predictive power of past ESG performance. The relatively low MAE and RMSE values on both the training and test sets further highlight the model's predictive accuracy. Moreover, the slightly lower errors on the test set suggest that the model not only captures the underlying relationship between historical and future ESG scores, but also generalizes well to out-of-sample data, reinforcing the model's robustness and reliability.

Table 3. FE Model Results and Performance Metrics

	RMSE	MAE	R^2
Train	0.0717	0.0573	0.8222
Test	0.0593	0.0464	0.8443

	Coeff.	Std. Err.
const	0.2206***	0.0079
ESGScore_lag1	0.6233***	0.0180
ESGScore_lag2	0.0392**	0.0177

The asterisks (*) indicate significance levels, with *** denoting significance at a 1% level and ** at a 5% level.

From an investor's perspective, these results highlight the importance of incorporating ESG trajectories into investment decision-making. Portfolios aiming to emphasize long-term sustainable growth can benefit from closely monitoring companies' historical ESG scores as these scores appear to be persistent and predictive indicators of future ESG performance.

Now to put our model to action, we decided to test it on NVIDIA, the largest chip-maker in the world, and the current powerhouse behind the AI boom. We re-trained our model on the entire dataset in order to make the best predictions possible, and forecasted the ESG scores over the next two years:

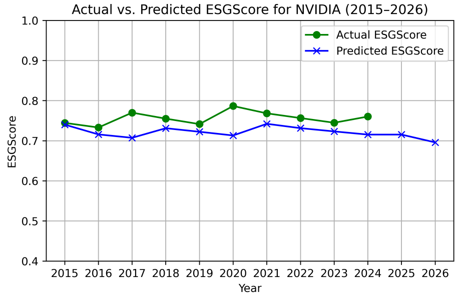

Figure 3. Visualization of NVIDIA's Predicted and Observed ESG Scores.

As shown in Figure [3], the predicted values track closely with the historical observations, indicating that our model captures the underlying trends effectively.

Looking ahead, the model suggests a slight decline in NVIDIA's ESG scores over the coming two years. One potential explanation is that the company's rapid expansion in production capacity and market reach—fueled by the AI revolution—may outpace corresponding improvements in its sustainability and governance measures. Additionally, evolving regulatory environments and market conditions, especially in key markets like the United States where Trump recently got re-elected, could influence corporate ESG efforts. From an investment standpoint, this potential ESG score decrease poses an interesting question: Will any marginal erosion in ESG performance outweigh NVIDIA's strong financial returns and continued leadership in high-growth technology segments? Ultimately, investors must balance these considerations when assessing the company's long-term risk and return profile.

4.2.4 Hausman Test Results

Finally, as a sanity and validity check, the Hausman test was perfomed. The result from the test is displayed in the following table:

Table 4. Hausman Test Results

Metric	Value
Test Statistic	1382.1851
P-Value	0.000

From Table 4 we see that the test statistic is statistically significant at a <1% level, confirming that the Fixed Effects model is the correct choice in the context of this exploration. This result aligns with expectations, as ESG scores likely encapsulate company-specific strategies and behaviors that are not fully captured by observable variables. Such factors, including corporate culture or sustainability initiatives, are plausibly correlated with the lagged ESG scores, necessitating the use of the Fixed Effects model to produce consistent and unbiased estimates. ESG scores are influenced by both company-level characteristics (e.g., size, industry, management) and temporal changes (e.g., yearly ESG trends). These company-level effects are likely correlated with the predictors (e.g., lagged ESG scores, or other company-specific variables), which makes Fixed Effects more appropriate. ESG scores are not random in our setting; they reflect company-specific factors that remain constant over time (fixed effects) and are likely correlated with predictors such as lagged scores.

5. Discussion

While this study demonstrates the efficacy of panel data models, particularly the Fixed Effects model, in predicting ESG scores, several limitations and avenues for future research warrant discussion. The dataset used in this study originates from a single provider (LSEG) and focuses on companies in the S&P 500. While this ensures consistency in data collection and calculation methodologies, it limits the generalizability of the findings. Exploring other datasets, such as those from MSCI, Sustainalytics, or Bloomberg, could provide a more comprehensive view. Combining features from one dataset with target variables from another, for example, could mitigate the risk of target variable leakage, potentially uncovering meaningful relationships.

Additionally, The focus on S&P 500 companies inherently restricts the analysis to large, publicly traded corporations in the United States. These companies often exhibit specific characteristics, such as substantial resources for sustainability initiatives, which may not be representative of smaller firms or companies in other regions. Expanding the dataset to include firms from emerging markets, mid-cap indices, or private companies could enhance the applicability of the model across diverse corporate landscapes. Potentially, this could generate a dataset large enough for the implementation of deep learning models to capture more complex relationships between ESG metrics over time.

Pertaining specifically to the clustering part of this research: clustering could be a valuable tool for categorizing ESG risk, however in this implementation there were no promising results that could be useful for making investment decisions. In the future, further study could be done to use clustering models other than KMeans or to attempt to separate data by industry to try to have more defined clusters in the dataset. Currently, the dataset is too complex and overlapping for the KMeans model to categorize data with a high degree of accuracy.

By addressing the outlined limitations and exploring these future directions, researchers can enhance the predictive accuracy, robustness, and applicability of ESG scoring models, ultimately contributing to more data-driven sustainability assessments.

To conclude, this study lays the foundation for predicting ESG scores using panel data models, demonstrating their potential in leveraging temporal dependencies and controlling for entity-specific effects. Reliable ESG score forecasts are a valuable asset in aligning portfolios with sustainability objectives while managing risks associated with poor ESG performance. By integrating these forecasts into investment strategies, such as portfolio optimization and scenario analysis, investors can make informed decisions that balance financial returns with ethical considerations.

References

[1] Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, 74(3), 87–103. https://doi.org/10.2469/faj.v74.n3.2

[2] Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

[3] Svanberg, J., Ardeshiri, T., Samsten, I., Öhman, P., Rana, T., & Danielson, M. (2022). Prediction of environmental controversies and development of a corporate environmental performance rating methodology. *Journal of Cleaner Production*, 344, 130979. https://doi.org/10.1016/j.jclepro.2022.130979

[4] Carlei, V., Durbano, F., Rossetti, C., & Spazian, A. (2024). Can machine learning explain alpha generated by ESG factors? *Computational Economics*. Springer US. https://link.springer.com/article/10.1007/s10614-024-10602-8

[5] LSEG. (2024). ESG scores. Retrieved from https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores. Accessed 19 Dec. 2024.

[6] Chatterji, A., Durand, R., Levine, D. I., & Touboul, S. (2014). Do ratings of firms converge? Implications for managers, investors, and strategy researchers. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2524861.

[7] Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. (2022). Tackling climate change with machine learning. *ACM Computing Surveys*. Retrieved from https://dl.acm.org/doi/10.1145/3485128#sec-8-1.