"How Prolific Are You?": A Scientometric Analysis of Researchers

Rahul Yedida[1], Adel Aladwani[2], and Snehanshu Saha[1]

[1]Department of Computer Science & Engineering, PES University
[2]Department of Information Systems, Kuwait University

Abstract

Scholars in academia regularly write papers showing the results of their work. To gauge how relevant their work is, several measures are used, including citation count and the h-index. This paper describes the methods used to identify "prolific" authors, whose metrics are significantly higher than their peers, and provides a general discussion of the relationships between different metrics. We also discuss how this analysis is performed even with the high dimensionality of the dataset.

1 Introduction

Scientometrics is the study of measuring science and research. Several measures are currently in use to measure the output of researchers, and each of these individually may not be an indicator of the quality and repute of a scholar's work. Comparing researchers by several metrics at once is cumbersome to do manually, and it would be incorrect to use a function of several metrics to compare how "prolific" a researcher is. In this paper, we first identify linear relationships among the attributes. We assert that this is also useful in finding the "most important" features. Later, we use two algorithms to cluster the records and compare the outputs of both algorithms. Finally, we use these results to identify prolific authors.

The rest of the paper is structured as follows. Section 2 describes the data and shows correlations between the attributes. Section 3 discusses the methods we used to analyse this data.

2 Data

The data contains 618 records of scholars with 25 attributes. There are no missing values, and eight of the attributes are integers. For ease of analysis, we have labeled the attributes as *v1* through *v25*. A table explaining what each of these variables are is given in Table 1 of the Supplementary Materials section at the end of this paper.

2.1 Preliminary Analysis

For a preliminary analysis of the relationships between the attributes, we sought to discover monotonic relationships. This was done by computing pairwise Spearman's rank correlation coefficient. This revealed that there were only positive monotonic relationships. 42.08% of these coefficients were above 0.8, and 52% were above 0.7.

A logical next step, then, was to investigate the percentage of linear relationships among these. This was done by computing pairwise Pearson correlation coefficients. To visually understand pairs that had high linear correlations, we displayed this information in a matrix, maintaining only the upper half, and avoiding diagonal entries. This avoids counting self-correlations and counting the same pair twice. Further, when finding multiple linear regression equations, this finds only unique relations between the variables. Figure 1 shows this result. Asterisks indicate a Pearson R value of 0.8 or higher. The numbers at the right indicate how many other variables each variable is highly correlated with.

```
v01: - - - - - - * - - - - - - - - - - - - - - - -  1
v02:   - - * - * - - * * - - - * - * - * * - - - * -  9
v03:     - - - - - - - - - - - - - - - - - - - - - -  0
v04:       - - * - - * * * - * * * * * - * - - - - * *  12
v05:         - - - - - - - - - - - * - - - - - - - -  1
v06:           - - - * * - - * * - * * * * - - - - - *  9
v07:             - - - - - - - - - - - - - - - - - - -  0
v08:               - - - - - - - - - - - - - - - - - -  0
v09:                 - * * * * * * * * * - - - - - - -  9
v10:                   - * * * * * * * * * - - - - - -  9
v11:                     - - * * * * * * * - - - - - -  6
v12:                       - * - - * - * - - - - - - -  3
v13:                         - - * * * * * - - - - - -  5
v14:                           - * * * - - - - - - * -  4
v15:                             - * * * - - - - - - -  3
v16:                               - * * * - - - - * *  5
v17:                                 - * * - - - - * -  3
v18:                                   - - - - - - - -  0
v19:                                     - - - - - * *  2
v20:                                       - - - - - -  0
v21:                                         - - - - -  0
v22:                                           - * - -  1
v23:                                             - - -  0
v24:                                               - *  1
v25:                                                 -  0
```

Figure 1: Matrix showing pairs with Pearson R values 0.8 or higher

These correlations provide a starting point to find relationships between the attributes.

3 Methods

This section discusses in detail the methods used to analyse the data.

3.1 Regression Analysis

Our first analysis investigated the precise relationships between the attributes. We looked at linear relationships. Because v4 had the most number of correlations, we started with it.

Multiple regression analysis was performed as follows. For each attribute that the variable being analysed has a high correlation with, we ran an Ordinary Least Squares (OLS) regression model in two cases–with a constant term and without (that is, in a forward step-wise fashion).

In each case, for each variable, we check the p-value to ensure that all variables are statistically significant. Finally, we choose the model with the highest adjusted R^2 value, which avoids the pitfalls of "kitchen sink regression". We also make sure that there are no significant autocorrelations, using the Durbin-Watson test. [1] suggests that a conservative range for the acceptable values for this test is between 1 and 3. In some cases, we chose to pick a model that has a slightly lesser adjusted R^2 value in favor of using lesser variables. This results in a model that still has a high predictive power, yet is simpler. In addition, we also

check the residuals vs. fits plots and check that they are not very correlated.

We note here that this analysis may be useful for identifying "important" attributes in a dataset. If a model that predicts a dependent variable through the other attributes yields a high adjusted R^2 value, then we may safely remove this dependent variable, keeping only the statistically significant predictor variables from the regression analysis.

The full regression results can be found in the Supplementary Materials section at the end of this paper. Here, we present a summary of the results of the regression analysis that was performed.

3.1.1 Analysis on the full data

We first looked at v4, and did a step-wise regression analysis. *v6, v9, v11, v13, v14, v17, v19, v24,* and *v25* were statistically significant, and the adjusted R^2 was 0.972. For this model, the Durbin-Watson statistic was 2.06, which suggests very minor negative autocorrelations. The correlations between the residuals and the fitted values was -0.0047. Figure 2 shows the residuals vs. fits plot for *v4*.

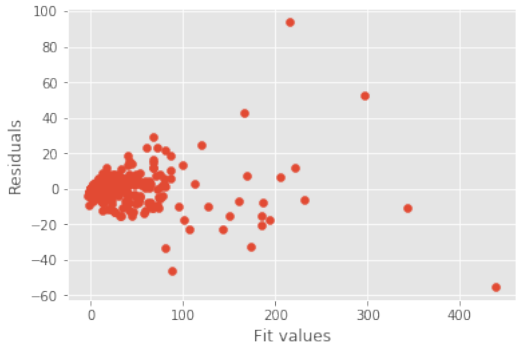


Figure 2: Residuals vs. fits plot for v4

Our next variable of interest was *v2*, having high Pearson R correlations with nine other variables. Only *v9* was not statistically significant. However, we also chose to discard *v24* because it added no predictive power to the model, and the adjusted R^2 for this model was 0.993, and the Durbin-Watson statistic was 1.853, suggesting only minor positive autocorrelations. All the p-values were less than 10^{-3}. The correlation between the residuals and the fits was 0.023. Figure 3 shows the residuals vs. fits plot for *v2*.

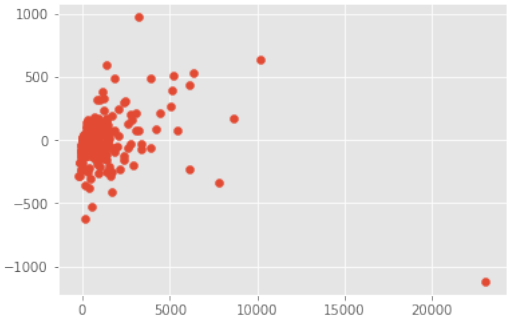


Figure 3: Residuals vs. fits plot for v2

We then looked at *v6*, also having high correlations with nine other variables. *v9* and *v14* were not statistically significant. We also discarded *v25* to obtain a simpler model with an adjusted R^2 value 0.001 less. This model had an adjusted R^2 of 0.913, and a Durbin-Watson statistic of 1.64. All p-values were below 10^{-3}.

Next, we looked at *v9*. Only *v13* was not statistically significant, but we also discarded *v10* and *v11*, which resulted in a model whose adjusted R^2 was 0.986 (a reduction of 0.002 from when the two vari-

ables were included). This model had a Durbin-Watson statistic of 1.99.

Finally, for *v10*, all the predictor variables were statistically significant, and we discarded *v19*, which resulted in a model with the same adjusted R^2. All the p-values were below 10^{-3}, except *v13*, whose p-value was 0.043 (95% C.I [0.07, 0.425]). The Durbin-Watson statistic was 1.861.

It should be noted that in all of the regression results of the above variables, the constant was not statistically significant, and the results reported are those for the models that did not include the constant.

3.1.2 Analysis on subset of the data

Our next step of analysis was to pick a subset of the variables, and redo the same analysis discussed above on this subset. Once again, it turns out that for all the regression equations obtained, the constant was not statistically significant.

The variables we chose were *v9* through *v21* and *v24*. Again, we created a matrix of Pearson R correlations, and picked out those that were higher than 0.8. Figure 4 shows this result.

Figure 4: Pearson R correlation matrix for subset of the data

Using this, we redid the same analysis as in the previous subsection. Again, we only discuss the results here, but the full list of experiments is in the Supplementary Materials section.

We first looked at v9. *v13* was not significant when all the variables it was correlated to were considered. However, when *v10* was not considered, it was statistically significant. We chose to sacrifice 0.2% adjusted R^2 in favor of a model that had two fewer predictor variables. The Durbin-Watson value was 1.99.

While predicting *v10*, dropping *v19* resulted in no loss of adjusted R^2. We further dropped *v18*, and dropped *v12* because it was not significant. While dropping *v18* caused a 0.3% drop in adjusted R^2, we were able to use two less variables. The Durbin-Watson statistic was 1.991.

We only looked at two other variables, because the adjusted R^2 for the others was below 0.95. For *v11*, dropping *v17* and *v18* resulted in a 0.2% drop in adjusted R^2, and we were able to use only four variables to get an adjusted R^2 of 0.971 and a Durbin-Watson of 2.081. Finally, for *v13*, all variables were statistically significant, and we discarded *v19* with no loss of adjusted R^2.

Perhaps more interesting than the above results was when we plotted the maximum adjusted R^2 (without discarding any variables) against the number of predictor variables used (on this subset of the data). Figure 5 shows this result, where a logarithm function fits quite well to the data. This meets our intuition of "diminishing returns"

as we increase the number of predictor variables. The equation obtained was

$$\text{Adj. } R^2 = 0.125465 \log n + 0.746654 \quad (1)$$

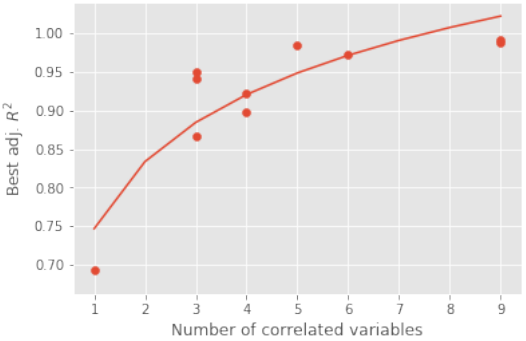


Figure 5: Adjusted R^2 vs. number of predictor variables

The sum of the squared residuals for this fit was 0.01456.

3.2 Cluster Analysis

We next sought to cluster the points, as a first step towards identifying "prolific" authors. Because finding the number of clusters in such high dimensional data is difficult, we chose to use two different algorithms–DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[2] and mean-shift clustering [3] from Python's sklearn package [4]. We picked two subsets of the data to perform this analysis on–the first was the set of variables that were highly correlated with *v4* (denote this subset as S_1)–we chose these because it had an especially high number of correlated variables–and the second was the subset discussed in the previous subsection (which we shall denote as S_2. We also performed the clustering analysis with DBSCAN using PCA instead of manually picking features. The results of that are discussed in the Supplementary Materials.

Because DBSCAN requires an *eps* parameter, we chose to find a reasonable guess. We did this by setting a threshold that no more than roughly 10% of the points (we chose 60 points) should be in individual custers. We started with an initial *eps* guess of 0.5 and made increments of 0.1 till this condition was satisfied.

We did not scale the features before clustering, because the distributions of all the features were different (running a Kolgomorov-Smirnov test on every pair of features gave p-values < 0.05). To prove that feature scaling is an incorrect step empirically, we performed both z-standardization and feature scaling to [0, 100]. In both cases, we found the value of *eps* in the same way discussed above, but starting from 0.001 and going in increments of 0.001. In both cases, DBSCAN returned 60 individual clusters, and one other cluster containing all the points.

For S_1, the value of *eps* we arrived at was 45. Apart from 60 individual clusters, the algorithm identified three clusters of points. The results of this are summarized in Table 1.

Cluster	Number of points
-1	60
0	532
1	20
2	6

Table 1: DBSCAN clustering results on S_1

The cluster -1 represents individual

clusters. The rest of the clusters were arbitrarily numbered from zero. Next, we ran mean-shift clustering, which automatically finds clusters by estimating a *bandwidth* parameter. Mean-Shift clustering identified 12 clusters, out of which five were individual clusters. We summarize these results in Table 2.

Cluster	Number of points
-1	5
0	509
1	72
2	19
3	4
4	4
5	3
6	2

Table 2: Mean-Shift clustering results for S_1

We then checked whether the points in similar-sized clusters in each algorithm were the same. We briefly do this analysis for S_1, but conduct a more thorough investigation for S_2.

We first looked at cluster 0 of both algorithms. Of these, 507 points were common, indicating that the majority of the dataset was clustered into the same cluster by both algorithms. Interestingly, however, this seems to be the only major commonality between the outputs of these algorithms. For example, cluster 1 of DBSCAN and cluster 2 of Mean-Shift, which had a very similar number of points, had no common points at all. Finally, 21 of the points that were marked as individual clusters by DBSCAN were put into cluster 1 by Mean-Shift.

Next, we performed the above analysis for S_2. The value of *eps* we arrived at was 18.6, and DBSCAN identified three clusters apart from sixty individual clusters. A summary is provided in Table 3, again, with cluster -1 meaning individual clusters.

Cluster	Number of points
-1	60
0	541
1	3
2	14

Table 3: DBSCAN clustering results for S_2

Table 4 summarizes the results of mean-shift clustering, which identified seven "real" clusters and eight individual clusters.

Cluster	Number of points
-1	8
0	504
1	82
2	15
3	3
4	2
5	2
6	2

Table 4: Mean-Shift clustering results for S_2

From these two tables, we followed through to track each point in both clusters. The full results of this are given in the Supplementary Materials section, but a summary is as follows. DBSCAN marked 60 points as individual clusters, but this was due to our 10% threshold. On the other

hand, mean-shift only identified eight individual clusters. However, mean-shift identified four other clusters that had three points each or less, and all of these were marked as individual clusters by DBSCAN. More surprisingly, a cluster marked by mean-shift as having fifteen points was identified by DBSCAN as fifteen individual clusters. The rest of the 28 individual clusters identified by DBSCAN were a part of mean-shift's second-largest cluster of 82 points. The analysis so far may suggest that we should lower the threshold for DBSCAN, which would increase *eps* (because if fewer points have to be identified as individual clusters, the distance that the algorithm is willing to consider must increase). However, looking at the bigger clusters weakens this position. Mean-shift's largest cluster of 504 points was a proper subset of DBSCAN's largest cluster. A large part of mean-shift's second-largest cluster, 37 out of 82 points, was also a part of DBSCAN's single largest cluster. This seems to suggest that we should reduce *eps* to make the outputs of both algorithms more similar.

While the two conclusions drawn above may seem contradictory, they are two separate observations. Increasing *eps* would mean DBSCAN would find less clusters as single points. Decreasing *eps* would certainly increase the number of individual clusters, which we certainly do not want, but this would shift some points in the larger clusters around, which may be desirable.

3.3 Identifying Prolific Authors

Finally, we proceed to test our hypothesis–are the researchers in the individual clusters more "prolific" than the others? For this, we used the results of the DBSCAN clustering because they were more succinct. We used the Welch's t-test [5], which does not assume equal variance between the two samples, and computed the p-values for all pairs of clusters (see Table 3 for the list of clusters) at 5% and 10% level of significance. Table 5 shows the results of this, where C_1 and C_2 are the cluster numbers.

Clearly, looking at the first and third rows, there is a statistically significant difference between the means of the individual clusters and the first two largest clusters. In the third row, the 13/14 variables were statistically significant at 95% confidence level, the insignificant one being *v24* (star count). We also looked at the "goodness" of the clustering using the Silhouette score [6], which was 0.571. [7] notes that this value means that "a reasonable structure has been found".

C_1	C_2	**95% significant**	**90% significant**
-1	0	1.0	1.0
-1	1	0.857	0.857
-1	2	0.928	1.0
0	1	0.786	0.857
0	2	1.0	1.0
1	2	0.643	0.786

Table 5: Percentage of significant Welch's test p-values at 95% and 90% confidence levels

References

[1] A. Field, *Discovering statistics using SPSS*. Sage publications, 2009.

[2] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[3] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[6] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[7] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

Supplementary Materials

1 Data

The 25 variables that were used in this analysis are below.

Variable	Meaning
v1	Number of papers
v2	Total citations
v3	Production years
v4	Cites per year
v5	Cites per paper
v6	Cites per author
v7	PapersAuthor
v8	Authors per paper
v9	h-index
v10	g-index
v11	hc-index
v12	hl-index
v13	hl-norm
v14	AWCR
v15	AW index
v16	AWCR per author
v17	e-index
v18	hm-index
v19	CitesAuthorYear
v20	CitesAuthorYearArticle
v21	Annual hl
v22	h Coverage
v23	g Coverage
v24	Star count
v25	Adjusted star count

Table 1: List of attributes in the data

2 Regression Results

2.1 Analysis on the full data

The regression experiments for *v4* are shown in Table 2. The chosen results are in bold.

IVs	**Adj.** R^2	**Significant IVs**
v6	0.759	All
v6,v9	0.796	All except constant
v6,v9,v10 (no const)	0.868	All
v6,v9,v10	0.835	All
v6,v9,v10,v11 (no const)	0.874	All
v6,v9,v10,v11	0.846	All
v6,v9,v10,v11,v13 (no const)	0.893	All
v6,v9,v10,v11,v13	0.866	All except constant
v6,v9,v10,v11,v13,v14 (no const)	0.896	All
v6,v9,v10,v11,v13,v14,v15 (no const)	0.896	All except v15
v6,v9,v10,v11,v13,v14,v16 (no const)	0.893	All except v16
v6,v9,v10,v11,v13,v14,v17 (no const)	0.912	All
v6,v9,v10,v11,v13,v14,v17,v19 (no const)	0.953	All except v10
v6,v9,v11,v13,v14,v17,v19 (no const)	0.961	All
v6,v9,v11,v13,v14,v17,v19,v24 (no const)	0.961	All
v6,v9,v11,v13,v14,v17,v19,v25 (no const)	0.953	All
v6,v9,v11,v13,v14,v17,v19,v24,v25 (no const)	**0.972**	**All**

Table 2: Regression experiments for *v4*

The chosen regression experiment is shown in Figure 1.

OLS Regression Results

Dep. Variable:	v4	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	2429.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:16:28	Log-Likelihood:	-2143.2
No. Observations:	618	AIC:	4304.
Df Residuals:	609	BIC:	4344.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v6	-0.0065	0.002	-2.775	0.006	-0.011	-0.002
v9	2.0218	0.218	9.275	0.000	1.594	2.450
v11	-0.8051	0.219	-3.669	0.000	-1.236	-0.374
v13	-3.1005	0.282	-10.993	0.000	-3.654	-2.547
v14	0.0197	0.007	2.713	0.007	0.005	0.034
v17	0.7742	0.080	9.659	0.000	0.617	0.932
v19	1.8746	0.060	31.088	0.000	1.756	1.993
v24	11.5548	0.560	20.634	0.000	10.455	12.655
v25	-28.5114	1.771	-16.095	0.000	-31.990	-25.033

Omnibus:	470.481	Durbin-Watson:	2.060
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53677.047
Skew:	-2.556	Prob(JB):	0.00
Kurtosis:	48.370	Cond. No.	3.20e+03

Figure 1: Regression results for v4

The regression experiments on *v2* are shown in Table 3.

IVs	**Adj.** R^2	**Significant IVs**
v4	0.804	All
v4 (no const)	0.818	All
v4,v6	0.924	All
v4,v6 (no const)	0.93	All
v4,v6,v9	0.924	All except constant
v4,v6,v9 (no const)	0.93	All
v4,v6,v9,v10	0.925	All
v4,v6,v9,v10 (no const)	0.931	All except v9
v4,v6,v9,v10,v14	0.933	All
v4,v6,v9,v10,v14 (no const)	0.939	All
v4,v6,v9,v10,v14,v16	0.981	All except constant
v4,v6,v9,v10,v14,v16 (no const)	0.983	All
v4,v6,v9,v10,v14,v16,v18	0.981	All except constant and v18
v4,v6,v9,v10,v14,v16,v18 (no const)	0.983	All except v18
v4,v6,v9,v10,v14,v16,v18,v19	0.992	All
v4,v6,v9,v10,v14,v16,v18,v19 (no const)	**0.993**	**All except v9**
v4,v6,v9,v10,v14,v16,v18,v19,v24	0.992	All
v4,v6,v9,v10,v14,v16,v18,v19,v24 (no const)	0.993	All except v9

Table 3: Regression experiments for v2

The chosen regression experiment is shown in Figure 2.

OLS Regression Results

Dep. Variable:	v2	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	1.255e+04
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:16:31	Log-Likelihood:	-3849.7
No. Observations:	618	AIC:	7713.
Df Residuals:	611	BIC:	7744.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v4	17.9865	0.539	33.391	0.000	16.929	19.044
v6	2.9565	0.030	99.132	0.000	2.898	3.015
v10	-7.1974	1.155	-6.230	0.000	-9.466	-4.929
v14	2.8990	0.172	16.892	0.000	2.562	3.236
v16	-12.9553	0.677	-19.137	0.000	-14.285	-11.626
v18	15.0059	3.381	4.438	0.000	8.366	21.646
v19	-40.4789	1.372	-29.500	0.000	-43.174	-37.784

Omnibus:	183.865	Durbin-Watson:	1.853
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13139.347
Skew:	-0.267	Prob(JB):	0.00
Kurtosis:	25.583	Cond. No.	392.

Figure 2: Regression results for v2

From v6 onward, the experiments were automated. The screenshots of the program results are shown instead. The regression experiments for *v6* are shown in Figure 3.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v6	v9	0.724	All
2	v6	v9 (no const)	0.69	All
3	v6	v9,v10	0.728	All
4	v6	v9,v10 (no const)	0.693	All
5	v6	v9,v10,v13	0.733	All
6	v6	v9,v10,v13 (no const)	0.692	v9,v10
7	v6	v9,v10,v13,v14	0.795	All
8	v6	v9,v10,v13,v14 (no const)	0.791	All
9	v6	v9,v10,v13,v14,v16	0.875	constant,v9,v13,v14,v16
10	v6	v9,v10,v13,v14,v16 (no const)	0.883	v9,v13,v14,v16
11	v6	v9,v10,v13,v14,v16,v17	0.875	constant,v9,v13,v14,v16
12	v6	v9,v10,v13,v14,v16,v17 (no const)	0.884	v9,v13,v14,v16
13	v6	v9,v10,v13,v14,v16,v17,v18	0.887	constant,v9,v13,v16,v17,v18
14	v6	v9,v10,v13,v14,v16,v17,v18 (no const)	0.893	v9,v13,v14,v16,v17,v18
15	v6	v9,v10,v13,v14,v16,v17,v18,v19	0.907	constant,v9,v13,v16,v18,v19
16	**v6**	**v9,v10,v13,v14,v16,v17,v18,v19 (no const)**	**0.913**	**v10,v13,v16,v17,v18,v19**
17	v6	v9,v10,v13,v14,v16,v17,v18,v19,v25	0.907	constant,v9,v13,v16,v18,v19,v25
18	v6	v9,v10,v13,v14,v16,v17,v18,v19,v25 (no const)	0.914	v10,v13,v16,v17,v18,v19,v25

Figure 3: Regression experiments for v6

The chosen regression results are shown in Figure 4.

OLS Regression Results

Dep. Variable:	v6	R-squared:	0.914
Model:	OLS	Adj. R-squared:	0.913
Method:	Least Squares	F-statistic:	1078.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	1.73e-321
Time:	19:16:32	Log-Likelihood:	-4003.3
No. Observations:	618	AIC:	8019.
Df Residuals:	612	BIC:	8045.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v10	18.4844	2.768	6.677	0.000	13.048	23.921
v13	-74.8686	7.615	-9.832	0.000	-89.823	-59.915
v16	5.4741	0.490	11.176	0.000	4.512	6.436
v17	-11.7171	3.150	-3.720	0.000	-17.903	-5.532
v18	58.0314	7.318	7.930	0.000	43.659	72.404
v19	13.1330	1.058	12.413	0.000	11.055	15.211

Omnibus:	613.817	Durbin-Watson:	1.640
Prob(Omnibus):	0.000	Jarque-Bera (JB):	84084.932
Skew:	3.994	Prob(JB):	0.00
Kurtosis:	59.583	Cond. No.	86.2

Figure 4: Regression results for v6

Figure 5 shows the regression experiments for *v9*.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v9	v10	0.94	All
2	v9	v10 (no const)	0.966	All
3	v9	v10,v11	0.945	v10,v11
4	v9	v10,v11 (no const)	0.969	All
5	v9	v10,v11,v12	0.954	All
6	v9	v10,v11,v12 (no const)	0.974	All
7	v9	v10,v11,v12,v13	0.956	constant,v10,v11,v13
8	v9	v10,v11,v12,v13 (no const)	0.975	v10,v11,v13
9	v9	v10,v11,v12,v13,v14	0.957	constant,v10,v11,v13
10	v9	v10,v11,v12,v13,v14 (no const)	0.976	v10,v11,v13,v14
11	v9	v10,v11,v12,v13,v14,v15	0.961	v10,v11,v13,v14,v15
12	v9	v10,v11,v12,v13,v14,v15 (no const)	0.978	v10,v11,v13,v14,v15
13	v9	v10,v11,v12,v13,v14,v15,v16	0.963	v10,v11,v12,v13,v14,v15,v16
14	v9	v10,v11,v12,v13,v14,v15,v16 (no const)	0.98	All
15	v9	v10,v11,v12,v13,v14,v15,v16,v17	0.974	v10,v11,v13,v14,v15,v17
16	**v9**	**v10,v11,v12,v13,v14,v15,v16,v17 (no const)**	**0.985**	**v10,v11,v13,v14,v15,v17**
17	v9	v10,v11,v12,v13,v14,v15,v16,v17,v18	0.98	v10,v11,v12,v14,v15,v16,v17,v18
18	v9	v10,v11,v12,v13,v14,v15,v16,v17,v18 (no const)	0.988	v10,v11,v12,v14,v15,v16,v17,v18

Figure 5: Regression experiments for v9

Figure 6 shows the chosen regression results.

OLS Regression Results

Dep. Variable:	v9	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	7123.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:16:32	Log-Likelihood:	-1040.4
No. Observations:	618	AIC:	2093.
Df Residuals:	612	BIC:	2119.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v10	0.5551	0.022	25.250	0.000	0.512	0.598
v11	0.3324	0.045	7.383	0.000	0.244	0.421
v13	0.4785	0.036	13.131	0.000	0.407	0.550
v14	0.0038	0.001	3.461	0.001	0.002	0.006
v15	-0.1440	0.056	-2.557	0.011	-0.255	-0.033
v17	-0.3342	0.019	-17.388	0.000	-0.372	-0.296

Omnibus:	111.690	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1976.264
Skew:	0.078	Prob(JB):	0.00
Kurtosis:	11.759	Cond. No.	169.

Figure 6: Regression results for v9

The regression experiments for *v10* are shown in Figure 7.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v10	v11	0.852	All
2	v10	v11 (no const)	0.911	All
3	v10	v11,v12	0.88	All
4	v10	v11,v12 (no const)	0.926	All
5	v10	v11,v12,v13	0.943	All
6	v10	v11,v12,v13 (no const)	0.965	All
7	v10	v11,v12,v13,v14	0.958	v11,v12,v13,v14
8	v10	v11,v12,v13,v14 (no const)	0.976	All
9	v10	v11,v12,v13,v14,v15	0.96	constant,v12,v13,v14,v15
10	v10	v11,v12,v13,v14,v15 (no const)	0.977	v12,v13,v14,v15
11	v10	v11,v12,v13,v14,v15,v16	0.961	constant,v12,v13,v14,v15,v16
12	v10	v11,v12,v13,v14,v15,v16 (no const)	0.977	v12,v13,v14,v15,v16
13	v10	v11,v12,v13,v14,v15,v16,v17	0.98	constant,v11,v13,v14,v15,v16,v17
14	v10	v11,v12,v13,v14,v15,v16,v17 (no const)	0.988	v11,v13,v14,v15,v16,v17
15	v10	v11,v12,v13,v14,v15,v16,v17,v18	0.987	All
16	**v10**	**v11,v12,v13,v14,v15,v16,v17,v18 (no const)**	**0.992**	**All**
17	v10	v11,v12,v13,v14,v15,v16,v17,v18,v19	0.987	constant,v11,v12,v14,v15,v16,v17,v18,v19
18	v10	v11,v12,v13,v14,v15,v16,v17,v18,v19 (no const)	0.992	v11,v12,v14,v15,v16,v17,v18,v19

Figure 7: Regression experiments for v10

Finally, the results of the regression for *v10* are shown in Figure 8.

OLS Regression Results

Dep. Variable:	v10	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.992
Method:	Least Squares	F-statistic:	1.020e+04
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:16:33	Log-Likelihood:	-1225.3
No. Observations:	618	AIC:	2467.
Df Residuals:	610	BIC:	2502.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v11	0.6603	0.061	10.896	0.000	0.541	0.779
v12	-0.5729	0.110	-5.207	0.000	-0.789	-0.357
v13	0.2161	0.106	2.030	0.043	0.007	0.425
v14	0.0354	0.002	15.377	0.000	0.031	0.040
v15	-0.7408	0.079	-9.420	0.000	-0.895	-0.586
v16	-0.0852	0.007	-12.879	0.000	-0.098	-0.072
v17	0.7858	0.022	35.053	0.000	0.742	0.830
v18	1.3468	0.075	17.896	0.000	1.199	1.495

Omnibus:	295.493	Durbin-Watson:	1.861
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3942.532
Skew:	-1.769	Prob(JB):	0.00
Kurtosis:	14.857	Cond. No.	258.

Figure 8: Regression results for v10

2.2 Analysis on subset of data

We now present the regression experiments and results on the chosen subset of the data. Please note that we renamed the variables for convenience, so *v1* through *v13* here are really *v9* through *v21*, and *v14* here is really *v24*.

The regression experiments for *v9* (shown as *v1*) are shown in Figure 9.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v1	v2	0.94	All
2	v1	v2 (no const)	0.966	All
3	v1	v2,v3	0.945	v2,v3
4	v1	v2,v3 (no const)	0.969	All
5	v1	v2,v3,v4	0.954	All
6	v1	v2,v3,v4 (no const)	0.974	All
7	v1	v2,v3,v4,v5	0.956	constant,v2,v3,v5
8	v1	v2,v3,v4,v5 (no const)	0.975	v2,v3,v5
9	v1	v2,v3,v4,v5,v6	0.957	constant,v2,v3,v5
10	v1	v2,v3,v4,v5,v6 (no const)	0.976	v2,v3,v5,v6
11	v1	v2,v3,v4,v5,v6,v7	0.961	v2,v3,v5,v6,v7
12	v1	v2,v3,v4,v5,v6,v7 (no const)	0.978	v2,v3,v5,v6,v7
13	v1	v2,v3,v4,v5,v6,v7,v8	0.963	v2,v3,v4,v5,v6,v7,v8
14	v1	v2,v3,v4,v5,v6,v7,v8 (no const)	0.98	All
15	v1	v2,v3,v4,v5,v6,v7,v8,v9	0.974	v2,v3,v5,v6,v7,v9
16	**v1**	**v2,v3,v4,v5,v6,v7,v8,v9 (no const)**	**0.985**	**v2,v3,v5,v6,v7,v9**
17	v1	v2,v3,v4,v5,v6,v7,v8,v9,v10	0.98	v2,v3,v4,v6,v7,v8,v9,v10
18	v1	v2,v3,v4,v5,v6,v7,v8,v9,v10 (no const)	0.988	v2,v3,v4,v6,v7,v8,v9,v10

Figure 9: Regression experiments for v9

The chosen regression model results are shown in Figure 10.

OLS Regression Results

Dep. Variable:	v1	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	7123.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:51:29	Log-Likelihood:	-1040.4
No. Observations:	618	AIC:	2093.
Df Residuals:	612	BIC:	2119.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v2	0.5551	0.022	25.250	0.000	0.512	0.598
v3	0.3324	0.045	7.383	0.000	0.244	0.421
v5	0.4785	0.036	13.131	0.000	0.407	0.550
v6	0.0038	0.001	3.461	0.001	0.002	0.006
v7	-0.1440	0.056	-2.557	0.011	-0.255	-0.033
v9	-0.3342	0.019	-17.388	0.000	-0.372	-0.296

Omnibus:	111.690	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1976.264
Skew:	0.078	Prob(JB):	0.00
Kurtosis:	11.759	Cond. No.	169.

Figure 10: Regression results for v9

The regression experiments for *v10* are shown in Figure 11.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v2	v3	0.852	All
2	v2	v3 (no const)	0.911	All
3	v2	v3,v4	0.88	All
4	v2	v3,v4 (no const)	0.926	All
5	v2	v3,v4,v5	0.943	All
6	v2	v3,v4,v5 (no const)	0.965	All
7	v2	v3,v4,v5,v6	0.958	v3,v4,v5,v6
8	v2	v3,v4,v5,v6 (no const)	0.976	All
9	v2	v3,v4,v5,v6,v7	0.96	constant,v4,v5,v6,v7
10	v2	v3,v4,v5,v6,v7 (no const)	0.977	v4,v5,v6,v7
11	v2	v3,v4,v5,v6,v7,v8	0.961	constant,v4,v5,v6,v7,v8
12	v2	v3,v4,v5,v6,v7,v8 (no const)	0.977	v4,v5,v6,v7,v8
13	v2	v3,v4,v5,v6,v7,v8,v9	0.98	constant,v3,v5,v6,v7,v8,v9
14	**v2**	**v3,v4,v5,v6,v7,v8,v9 (no const)**	**0.988**	**v3,v5,v6,v7,v8,v9**
15	v2	v3,v4,v5,v6,v7,v8,v9,v10	0.987	All
16	v2	v3,v4,v5,v6,v7,v8,v9,v10 (no const)	0.992	All
17	v2	v3,v4,v5,v6,v7,v8,v9,v10,v11	0.987	constant,v3,v4,v6,v7,v8,v9,v10,v11
18	v2	v3,v4,v5,v6,v7,v8,v9,v10,v11 (no const)	0.992	v3,v4,v6,v7,v8,v9,v10,v11

Figure 11: Regression experiments for v10

The chosen model for *v10* is shown in Figure 12.

OLS Regression Results

Dep. Variable:	v2	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.989
Method:	Least Squares	F-statistic:	8902.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:52:54	Log-Likelihood:	-1355.9
No. Observations:	618	AIC:	2724.
Df Residuals:	612	BIC:	2750.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v3	0.5665	0.072	7.841	0.000	0.425	0.708
v5	1.3477	0.056	23.986	0.000	1.237	1.458
v6	0.0438	0.002	18.660	0.000	0.039	0.048
v7	-0.6372	0.094	-6.785	0.000	-0.822	-0.453
v8	-0.0772	0.007	-11.541	0.000	-0.090	-0.064
v9	0.6080	0.023	26.854	0.000	0.564	0.652

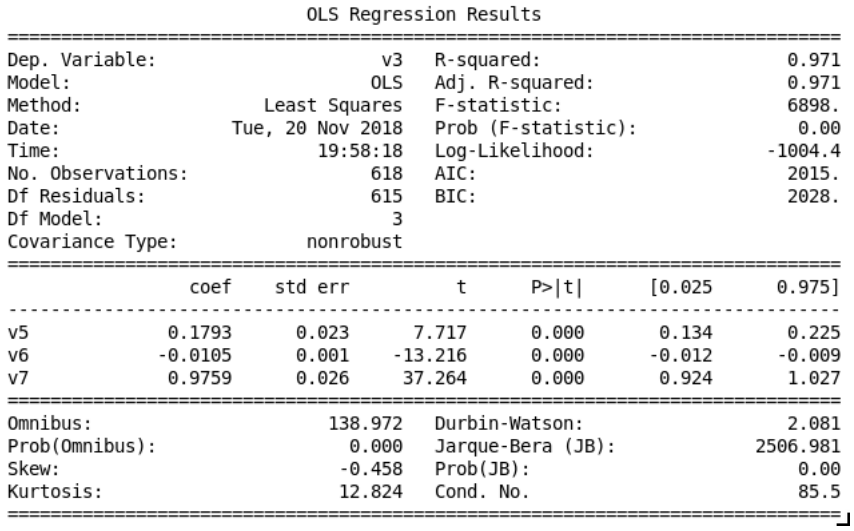
Omnibus:	264.178	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3118.984
Skew:	-1.563	Prob(JB):	0.00
Kurtosis:	13.553	Cond. No.	171.

Figure 12: Regression results for v10

Next, we show Figure 13 and 14, which show the experiments and results for *v11*.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v3	v5	0.809	All
2	v3	v5 (no const)	0.902	All
3	v3	v5,v6	0.832	All
4	v3	v5,v6 (no const)	0.905	All
5	v3	v5,v6,v7	0.936	All
6	v3	v5,v6,v7 (no const)	0.97	All
7	v3	v5,v6,v7,v8	0.936	constant,v5,v6,v7
8	**v3**	**v5,v6,v7,v8 (no const)**	**0.97**	**v5,v6,v7**
9	v3	v5,v6,v7,v8,v9	0.938	constant,v5,v6,v7,v9
10	v3	v5,v6,v7,v8,v9 (no const)	0.971	v5,v6,v7,v9
11	v3	v5,v6,v7,v8,v9,v10	0.939	constant,v5,v6,v7,v9,v10
12	v3	v5,v6,v7,v8,v9,v10 (no const)	0.972	v5,v6,v7,v9,v10

Figure 13: Regression experiments for v11



OLS Regression Results

Dep. Variable:	v3	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.971
Method:	Least Squares	F-statistic:	6898.
Date:	Tue, 20 Nov 2018	Prob (F-statistic):	0.00
Time:	19:58:18	Log-Likelihood:	-1004.4
No. Observations:	618	AIC:	2015.
Df Residuals:	615	BIC:	2028.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v5	0.1793	0.023	7.717	0.000	0.134	0.225
v6	-0.0105	0.001	-13.216	0.000	-0.012	-0.009
v7	0.9759	0.026	37.264	0.000	0.924	1.027

Omnibus:	138.972	Durbin-Watson:	2.081
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2506.981
Skew:	-0.458	Prob(JB):	0.00
Kurtosis:	12.824	Cond. No.	85.5

Figure 14: Regression results for v11

Finally(!), we show the same set of figures for *v13*.

Expt. No.	DV	IVs	Adj. R^2	Significant IVs
1	v5	v7	0.809	v7
2	v5	v7 (no const)	0.901	All
3	v5	v7,v8	0.842	All
4	v5	v7,v8 (no const)	0.912	All
5	v5	v7,v8,v9	0.865	All
6	v5	v7,v8,v9 (no const)	0.923	All
7	v5	v7,v8,v9,v10	0.971	constant,v8,v9,v10
8	**v5**	**v7,v8,v9,v10 (no const)**	**0.984**	**All**
9	v5	v7,v8,v9,v10,v11	0.971	constant,v8,v9,v10
10	v5	v7,v8,v9,v10,v11 (no const)	0.984	v8,v9,v10

Figure 15: Regression experiments for v13

OLS Regression Results

Dep. Variable:	v5	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.985
Method:	Least Squares	F-statistic:	1.005e+04
Date:	Thu, 22 Nov 2018	Prob (F-statistic):	0.00
Time:	23:51:33	Log-Likelihood:	-769.75
No. Observations:	618	AIC:	1547.
Df Residuals:	614	BIC:	1565.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
v7	0.0419	0.020	2.136	0.033	0.003	0.080
v8	-0.0153	0.002	-8.471	0.000	-0.019	-0.012
v9	0.1492	0.008	18.628	0.000	0.133	0.165
v10	0.7589	0.015	49.960	0.000	0.729	0.789

Omnibus:	199.818	Durbin-Watson:	2.106
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4996.027
Skew:	-0.839	Prob(JB):	0.00
Kurtosis:	16.828	Cond. No.	29.5

Figure 16: Regression results for v13

3 Cluster Analysis

We now present the full description of where each point is in each clustering algorithm. The explanation in the paper supplements this well enough to clearly understand the segregation of points by both algorithms.

3.1 PCA

We performed a PCA before using DBSCAN. In all cases, we used the same method discussed in the paper to compute the value of *eps*. We did this analysis for one through twelve dimensions. For number of dimensions greater than six, the performance stagnates at 0.571. We argue however, that the new dimensions obtained by PCA are less interpretable than simply using the original features, and the maximum difference between the performance of DBSCAN using our hand-picked features (0.571) and using PCA (0.644, n=2) was only 0.073. Figure 17 shows the results of using PCA to various numbers of dimensions.

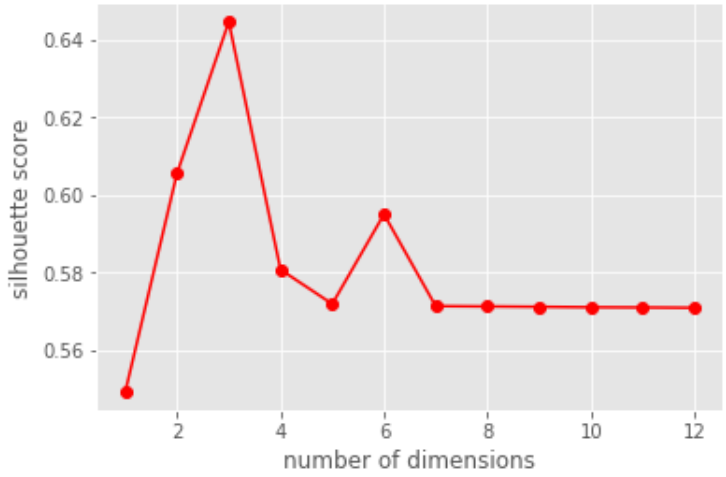


Figure 17: Results of using PCA and then clustering with DBSCAN

3.2 Analysis on S_1

Figure 18 shows the analysis for S_1. The Silhouette score for DBSCAN here was 0.635.

We make the following observations. The largest clusters for both the algorithms are very similar **(row 9)**. The two points in the MS cluster that were not in DBSCAN's largest cluster were in its second largest cluster **(row 2)**. All the clusters of the mean-shift algorithm that were of size 4 or less were marked as individual clusters by DBSCAN **(rows 5-8 and row 1)**. More surprisingly, the cluster of 19 points identified by mean-shift was marked as 19 individual clusters by DBSCAN **(row 4)**. Finally, a considerable part of mean-shift's second-largest cluster was also identified as individual points by DBSCAN **(row 3)**.

Index	DBSCAN cluster#	MS cluster#	DBSCAN cluster count	MS cluster count	Common
1	-1	-1	60	5	5
2	-1	0	60	509	2
3	-1	1	60	72	21
4	-1	2	60	19	19
5	-1	3	60	4	4
6	-1	4	60	4	4
7	-1	5	60	3	3
8	-1	6	60	2	2
9	0	0	532	509	507
10	0	1	532	72	25
11	1	1	20	72	20
12	2	1	6	72	6

Figure 18: Clustering by DBSCAN and Mean-Shift algorithms on S_1

3.3 Analysis on S_2

Figure 19 shows the analysis for S_2. As noted in the paper, the Silhouette score for DBSCAN for this subset was 0.571.

Index	DBSCAN cluster#	MS cluster#	DBSCAN cluster count	MS cluster count	Common
1	-1	-1	60	8	8
2	-1	1	60	82	28
3	-1	2	60	15	15
4	-1	3	60	3	3
5	-1	4	60	2	2
6	-1	5	60	2	2
7	-1	6	60	2	2
8	0	0	541	504	504
9	0	1	541	82	37
10	1	1	3	82	3
11	2	1	14	82	14

Figure 19: Clustering by DBSCAN and Mean-Shift algorithms on S_2