Analysis of the relationship in assessments of tutors' attractiveness using publicly available data

The evidence from the Higher School of Economics

A. Ternikov*

Higher School of Economics in Saint-Petersburg (HSE), Department of Economics
e-mail: ternikov.spb@mail.ru

ABSTRACT

Researchers of processes, which associated with the subjective evaluation of any social phenomenon, often face to problems such as data collection and analysis and revealing the dependence of objective data with the informal rules that have been established in a particular society. One way to avoid this kind of problem is to use publicly available sources of information and find on-line communities allied with them. This paper provides an example of tutors' rating of HSE and draws a parallel between the actual assessment according to the annual student voting and informal tutor estimation based on the attractiveness of his/her quotations published in the social network community at vk.com. Based on regression analysis, it is concluded that with sufficient accuracy of using the information obtained from informal communities, we can predetermine the actual result of evaluating the attractiveness of the tutor.

Key words. Social networks, open-source data, formal and informal rules

1. Introduction

According to many empirical researches it is indispensable to catch an effect between depending and depended variables. However, this is no a matter of fact that the obtaining of the result or estimation will be reliable and significant. The open-source data often provides a little bit limited information about some investigated phenomenon. One of the cases, that could help with data mining is to appeal to the other publicly available data source, which is connected to the subject of study. This could be, for instance, social network communities.

Accordingly to educational process, a lot of studies appreciate a role of social networks because today the majority of students and a big part of tutors appeal to internet communities (1; 2). Moreover, the importance of application of social network analysis related also to conducting of tutors' estimations in the context of their credibility in social network (3).

The Higher School of Economics is famous for its openness and transparency especially in the public access to the information about educational process and personnel structure. Despite this it is hard to get the eligible data about scoring distribution of annual tutors' rating. Moreover, the official site of the university provides a very compressed information offering binomial alternatives: whether or not tutor was recognized as the best according to the students' votes.

To overcome difficulties related to the lack of essential information it is needed to seek an appropriate community in social network system. The established hypothesis of this research is how effectively can be described the official open-source data using additional informal information from the social network in the case of ratings of tutors who work in Higher School of Economics.

2. Data

The data mining was held at three stages. First of all, it was chosen the informal source of students' estimations related to attractiveness of tutors in the context of their recognition as the best tutors. The choice fell on the community in the social network "*VKontakte*" named "*HSE tutors' quotes*" (http://vk.com/hseteachers). This community posts quotes of tutors which sent by followers of this community. Then these quotes are estimated under the attractiveness of finite post. At first sight, there are some limitations of this data connected to the relatively small list of tutors, whose quotes are published, and an opportunity to make an error in estimation due to the assessment of tutor's quotes but not so much their character. However, these problems can be smoothed by the representative sample size and similar uncertainties in official annual estimations.

The second stage related to mining and processing data from the social network. The information about amount of "*Like*"s of every post with the quote was collected.

The third stage is connected to the collection of data about tutors from official site of the university (http://www.hse.ru/org/persons). It takes the following parameters: the note about the status of the best tutor, campus in which tutor currently works, tutor's field of study, the possessiveness of academic degree, work experience in HSE and gender.

All available data was parsed automatically using "*R*" code and than manually corrected.

The total amount of posts is 1507 of 481 tutor. The percentage distribution by sex is 31.1% (women) to 68.9% (men). The biggest part of quotes accordingly to tutors came from Moscow (82%), the other campuses have less proportion (Saint-Petersburg — 11.2%, Nizhniy Novgorod — 5.6%, Perm' — 1.2%). Moreover, there are some the most mentioned subjects (fields of study) such as Economics (19.5%), Law (12.8%), Linguistics (12.2%) and Mathematics (10.4%). The distribution of "*Like*"s is left-side asymmetric but it can be reduced to normal using exclusion of outers and a huge amount of observations.

* Third-year student (135 group) of Higher School of Economics — Saint-Petersburg, Department of Economics

3. Methods

The aim of this research is to set up a connection between formal and informal data. To estimate and identify this coherency there is a need to use a regression analysis.

The specification of variables is presented in the table 1.

Table 1. Specification of variables.

Variable name	Description
Best*	Status of the best tutor
Like	Amount of "*Like*"s
Gender*	1 — male, 0 — female
Work	Work experience in HSE (year)
Degree*	Presence of academic degree
Moscow*	Campus in Moscow
SPB*	Campus in Saint-Petersburg
NN*	Campus in Nizhniy Novgorod
Perm*	Campus in Perm'
Subject	13 types of subjects**

* Binary variable

** Military studies, Orientalism, Computer Science, History, Mathematics, Management, Politology, Law, Psychology, Sociology, Linguistics, Philosophy, Economics

According to variables, there are no significantly strong correlations between independent variables (all of them less then 0.15 at the level of 95% of significance).

4. Results

The process of finding of dependency between the status of the best tutor and the amount of "*Like*"s at informal internet community related to defining of variable's specification. The choice of Method of Least Squares is specified because of the approximation for big data volume. Moreover, the estimations, which are got from this analysis will be faithful due to the existence of the limit in "*Like*"s. Consequently, the binary variable can be estimated correctly.

On the first stage, it was built a model 1, where coefficients were estimated along the whole sample:

$$\text{Best} = \alpha_1 \cdot \text{Like} + \alpha_2 \cdot \text{Gender} + \alpha_3 \cdot \text{Work} + \alpha_4 \cdot \text{Degree} + \\ + \alpha_5 \cdot \text{Moscow} + \alpha_6 \cdot \text{SPB} + \alpha_7 \cdot \text{NN} + \sum_{i=2}^{13} (\alpha_{6+i} \cdot \text{Subject}_i). \quad (1)$$

Results of estimation of this model is presented in table 2. The obtaining of the strong result shows statistically significant coefficient front the variable of amount of "*Like*"s. It shows that the probability to recognize tutor as the best according to official students' voting increases at 0.03% due to adding a one "*Like*" ceteris paribus.

This result accepted the stated hypothesis about dependency between formal and informal data. However, there is a need to confirm and strengthen this estimation. The construction of the second model starts with changing the specification of independent variable of amount of "*Like*"s. Conceivably, if we reduce the sample to the amount of unique tutors, it is indispensable to aggregate the amount of "*Like*"s. At first glance, the quotes of tutors are distributed in different way: some of them have a few quotes with the high rate but the other have many quotes with low estimations. Consequently, it is correct to use the sum of "*Like*"s (variable "*Like_sum*") as a measure of aggregation:

$$\text{Best} = \beta_1 \cdot \text{Like\_sum} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot \text{Work} + \beta_4 \cdot \text{Degree} + \\ + \beta_5 \cdot \text{Moscow} + \beta_6 \cdot \text{SPB} + \beta_7 \cdot \text{NN} + \sum_{i=2}^{13} (\beta_{6+i} \cdot \text{Subject}_i). \quad (2)$$

The result of this estimation is presented in table 2. There is no significant differences between new (0.01%) and previous estimation of the coefficient front the variable of "*Like*"s' amount. This fact also confirms the previous hypothesis.

The other variables (almost all) in both models are statistically significant at the level of 95% of confidence. They have more strong influence on the dependent variable ceteris paribus but the purpose of this work is to confirm a significance of the "*Like*"'s coefficient.

Table 2. Regression models of the status of the best tutor.

	(1) Best	(2) Best
Like	0.000346***	
Like_sum		0.0000933***
Gender	0.187***	0.141**
Work	0.0272***	0.0265***
Degree	0.0686**	0.0908
Moscow	0.439***	0.422***
SPB	0.524***	0.495***
NN	-0.0977	0.00681
Subject== Orientalism	0.490***	0.343
Subject== Computer Science	0.237*	0.166
Subject== History	0.384***	0.477**
Subject== Mathematics	0.453***	0.434**
Subject== Management	0.603***	0.432**
Subject== Politology	0.554***	0.367*
Subject== Law	0.594***	0.472***
Subject== Psychology	0.630***	0.519**
Subject== Sociology	0.387***	0.306*
Subject== Linguistics	0.423***	0.384**
Subject== Philosophy	0.736***	0.457**
Subject== Economics	0.546***	0.370**
Constant	-0.786***	-0.683***
Observations	1507	481
Adjusted R^2	0.216	0.167
AIC	1783.3	629.5
BIC	1889.7	713.0

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Conclusions

1. The regression analysis shows us the dependency between formal and informal data, borrowed from different data sources. Furthermore, it is necessary to notice that the estimations taken from social internet community is so eligible to be taken into account as an official sources of information.
2. In many cases of any research there is a need to find and analyse the space of social networks because it can bring an

unexpected significant relation to the subject of a proper research. These sources should not be underestimated because of those latent inner explanatory power.

3. This paper provides a significant result of intercommunication between official and informal open-source data in the case of estimation of tutors' attractiveness in the one university. Consequently, the most crucial fact is to reckon informal exogenous factors (with endogenous nature) in any analysis in the sphere of economic and social interaction.

References

[1] Hew, K. F. (2011). Students' and teachers' use of Facebook. Computers in Human Behavior, 27(2), 662–676.

[2] Madge, C., Meek, J., Wellens, J., & Hooley, T. (2009). Facebook, social integration and informal learning at university: 'It is more for socialising and talking to friends about work than for actually doing work'. Learning, Media and Technology, 34(2), 141–155.

[3] Mazer, J. P., Murphy, R. E., & Simonds, C. J. (2009). The effects of teacher self-disclosure via Facebook on teacher credibility. Learning, Media and Technology, 34(2), 175–183.