Rustam Tagiew, Dmitry I. Ignatov, Andreas Hilbert, Radhakrishnan Delhibabu (Eds.)

EEML 2016 – The 3rd International Workshop on Experimental Economics and Machine Learning

Workshop co-located with the 13th International Conference on Concept Lattices and Their Applications (CLA 2016)

July 18, 2016, Moscow, Russia

Volume Editors

Rustam Tagiew
Polarez Engineering, Dresden, Germany

Dmitry I. Ignatov,
Department of Data Analysis and AI, Faculty of Computer Science
National Research University Higher School of Economics, Moscow, Russia

Andreas Hilbert
Business Intelligence Research, Faculty of Business and Economics
Technische Universität Dresden, Germany

Radhakrishnan Delhibabu,
Institute of Information Technology and Information Systems
Kazan Federal University, Russia

Printed by the National Research University Higher School of Economics.

The proceedings are also published online on the CEUR-Workshop web site, Vol. 1627, in a series with ISSN 1613-0073.

Preface

This volume contains the papers presented at the Third International Workshop on Experimental Economics and Machine Learning held on July 18, 2016 at the National Research University Higher School of Economics, Moscow.

This proceedings concentrates on an interdisciplinary approach to modelling human behavior incorporating data mining and expert knowledge from behavioral sciences. Data analysis results extracted from clean data of laboratory experiments are of advantage if compared with noisy industrial datasets from the web and other sources. In their turn, insights from behavioral sciences help data scientists. Behavior scientists see new inspirations to research from industrial data science. Market leaders in Big Data, as Microsoft, Facebook, and Google, have already realized the importance of Experimental Economics know-how for their business.

In Experimental Economics, although financial rewards restrict subjects preferences in experiments, the exclusive application of analytical game theory is not enough to explain the collected data. It calls for the development and evaluation of more sophisticated models. The more data is used for evaluation, the more statistical significance can be achieved. Since large amounts of behavioral data are required to scan for regularities, Machine Learning is the tool of choice for research in Experimental Economics. In some works, automated agents are needed to simulate and intervene in human interactions. This proceeding aims to create a forum, where researchers from both Data Analysis and Economics are brought together in order to achieve mutually-beneficial results.

This year the workshop has hosted nine regular papers and two research proposals on a variety of topics related to different aspects of human behavior in games, demography, economy crises, stock markets, etc. Each paper has been reviewed by two PC members at least; all these papers rely on different data analysis techniques and the presented results are supported by data.

The representatives of R&D department of Imhonet company, Vladimir Bobrikov and Elena Nenova, have presented a keynote talk concerning how to consistently value recommendations produced by recommender systems.

We would like to thank all the authors of submitted papers and the Program Committee members for their commitment. We are grateful to our invited speaker and our sponsors: National Research University Higher School of Economics (Moscow, Russia), Russian Foundation for Basic Research, and ExactPro. Finally, we would like to acknowledge the EasyChair system which helped us to manage the reviewing process.

July 18, 2016
Moscow

Rustam Tagiew
Dmitry I. Ignatov
Andreas Hilbert
Radhakrishnan Delhibabu

Organisation

Program Committee

Fadi Amroush	University of Granada, Spain
Danil Fedorovykh	National Research University Higher School of Economics, Moscow, Russia
Daniel Karabekyan	National Research University Higher School of Economics, Moscow, Russia
Alexander Karpov	National Research University Higher School of Economics, Moscow, Russia
Mehdi Kaytoue	LIRIS - INSA de Lyon, France
Michael Khachay	Krasovsky Institute of Mathematics and Mechanics of Ural Branch of RAS, Yekaterinburg, Russia
Natalia Konstantinova	University of Wolverhampton & First Utility, UK
Xenia Naidenova	Military Medical Academy, Saint Petersburg, Russia
Amedeo Napoli	INRIA – LORIA, Nancy, France
Sergey Nikolenko	Steklov Mathematical Institute & National Research University Higher School of Economics, Saint Petersburg, Russia
Henry I. Penikas	National Research University Higher School of Economics, Moscow
Heather Day Pfeiffer	Akamai Physics, Inc.
Artem Revenko	Semantic Web Company, Vienna, Austria
Peter Romov	Yandex Data Factory, Russia
Evgeniy Sokolov	Moscow State University & Yandex Data Factory, Russia

Additional Reviewers

Dmitry Dagaev	National Research University Higher School of Economics, Moscow

Table of Contents

Keynote Talk

Regular Papers

Project Proposals and Abstracts

What is a Fair Value of Your Recommendation List?

Vladimir Bobrikov[1], Elena Nenova[1], and Dmitry I. Ignatov[2]

[1] Imhonet
vcomzzz@gmail.com, enenova@imhonet.ru
https://imhonet.ru
[2] National Research University Higher School of Economics
Moscow, Russia
dignatov@hse.ru

Abstract. We propose a new quality metric for recommender systems. The main feature of our approach is the fact, that we take into account the set of requirements, which are important for business application of a recommender. Thus, we construct a general criterion, named "audience satisfaction", which thoroughly describe the result of interaction between users and recommendation service. During the criterion construction we had to deal with a number of common recommenders' problems: a) Most of users rate only a random part of the objects they consume and a part of the objects that were recommended to them; b) Attention of users is distributed very unevenly over the list of recommendations and it requires a special behavioral model; c) The value of the user's rate measures the level of his/her satisfaction, hence these values should be naturally incorporated in the criterion intrinsically; d) Different elements may often dramatically differ from each other by popularity (long tail – short head problem) and this effect prevents accurate measuring of user's satisfaction. The final metric takes into account all these issues, leaving opportunity to adjust the metric performance based on proper behavioral models and parameters of short head problem treatment.

Keywords: recommender systems, quality metric, explicit feedback, movie recommendations, AUC, cold start, recommendations for novices

1 Introduction

Every recommender system aims to solve a certain business problem. Successful recommendations can be assessed in terms of specific business results, such as the number of visitors, sales, CTR, etc. However, it is too difficult to measure the quality of recommendation algorithm in this way since it depends on a vast variety of conditions, where the recommendation algorithm itself can bring a small contribution.

Therefore it turns out that developers need to come up with a formal numerical criteria for recommendation algorithms in isolation from the business

goals. As a result, a lot of papers on recommendation systems are produced every year. However, the numerical metrics they apply are useful, but usually are overly abstract compared to the problem they solve.

The approach we suggest is based on the idea that every metric should be constructed for a specific business problem. In this paper, we will focus on a concrete example, a movie recommendation service on `www.imhonet.ru`. Although we have tested the proposed approach on movies, this case can be generalized and applied to any similar objects (domain) of recommendation.

Let us shortly outline several relevant papers to our study. In [1] one of the most recent and consistent survey on evaluation of recommender systems can be found. Thus the authors discuss peculiarities of offline and online quality tests. They also review widely used quality metrics in the community (Precision, Recall, MAE, Customer ROC, Diversity, Utility, Serendipity, Trust, etc.) noting trade-off between these set of properties. Similar trade-off effects for top-n recommendations were noticed and studied earlier [2]: "algorithms optimized for minimizing RMSE do not necessarily perform as expected in terms of top-N recommendation task". In [3], importance of user-centric evaluation for recommender systems through a so called user experiment is stressed; in fact, this type of experiments suggests an interactive evaluation procedure that also extends conventional A/B tests. In [4], the authors proposed a new concept of unexpectedness as recommending to users those items that different from what they would expect from the system; their method is based on the notions of utility theory of economics and outperforms baselines on real datasets in terms of such important measures as coverage, aggregate diversity and dispersion, while avoiding accuracy losses. However, the first approach which is close to our proposed metric is based on the usage of ROC curves for evaluation of customer behaviour and can be found in [5]; here, the authors modified conventional ROC curves by fixing the size of recommendation list for each user. Later, two more relevant papers that facilitated our findings appeared: 1) [6] continues studies with incorporation of quality measures (in the original paper, serendipity) into AUC-based quality evaluation framework and 2) [7] combines precision evaluation with a rather simple behavioral model of user's interaction with the provided recommendation list. In the forthcoming sections, we extend and explain how these concrete ideas can be used for derivation of our user-centric evaluation measure to fulfill business needs of the company.

The paper is organized as follows. In Section 2, we describe the main measured operation, i.e. the interaction between our service that provides recommendations and its consumers. It is important to list all the cases of possible types of interaction that our service can meet. Based on that cases, in Section 3, we substantiate the use of a common recommender's *precision* as a starting point of our inference. Then, in Section 4 we show how a common *precision* could be transformed into a stronger discounted metric even with the help of rather simple behavioral model. Section 5 is devoted to users' rates values; it describes how two different merits of metric, namely, the ability to evaluate a ranked list and the ability to be sensitive to rate values, could be joined in one

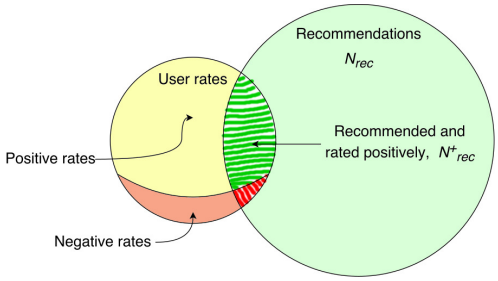


Fig. 1. Comparison of users' rates and recommended items.

term. In Section 6, we discuss how a *short head* problem could be treated. This specific recommender's problem makes it difficult to use those types of metrics that include sums of ratings. Section 7 summarizes the all previous considerations into the final expression for the metric and discusses several additional problems of its application. Section 8 demonstrates several illustrative cases of *Imhonet's* metric application from our daily practice. Section 9 concludes the paper and outlines future work.

2 Service product and users' feedback

The output of our service is a personalized list of recommended movies. The feedback comes directly from users in a form of movie rates. We shall start with the comparison of types of user's feedback, rates, and the lists of items that we recommend; they are shown in Figure 1. We need to consider the four situations.

Hit A movie has been recommended to a user, and it has received a positive rate from this user. This would normally mean that the recommendation was precise. We are interested in such cases, so let us call them "successes" and maximize their number.

Mishit A movie has been recommended but received a negative rate. We should avoid such situations. It is even possible that the avoidance of such cases is more important than maximizing the number of the cases of success. Unexpectedly, but we have learned from the experience that such cases can be ignored. The probability of the coincidence of negative signals with the elements from the list of recommendations, given by any proper recommender system, is too small to significantly affect the value of the metric. Therefore, it is not necessary to consider this case. This means that the metric is insensitive to negative rates.

Recommended but not rated If a movie has been recommended, but there has been no a positive signal, it seems that it does not mean anything. We do not know why it happened: a user has not seen the movie yet or has not rated it. As a result, it seems reasonable not to take into account these cases. The practice has shown that these cases constitute a majority. It happens due to two reasons. First, we always recommend a redundant number of movies, i.e.

more movies than a user could watch (N_{rec} is relatively large). Second, most of the users tend to not give rates for every single movie they have seen, as if we could access only a fraction of users' rates.

Rated but not recommended It is the opposite case, the movie has not been recommended, but it has received a positive signal; hence, the recommender has not used an opportunity to increase the amount of successful recommendations. As long as these cases exist, it is still possible for the recommender to improve its efficiency. If all positively rated movies have been recommended, it means that the recommendation system's accuracy is the highest possible and there is no room for improvement.

3 Precision

If instead of just a number of successes we use the value of *precision (p)*, i.e. we divide the number of successes N^+_{rec} by N_{rec}, there will be no a significant change: instead of the number of successes we will maximize the same value, only divided by a constant:

$$p = \frac{N^+_{rec}}{N_{rec}}. \tag{1}$$

However, as we will see later, this division provides an opportunity to make the metric sensitive to a very important aspect of our problem. (It allows us to make it discounted, in other words, – to take into account the order of the elements in the recommendation list.) Moreover, the value of p has a clear meaning, which can be described in a probabilistic language. Assume that a user consumes our product, namely, go through all the elements of the recommendation list. Then p shows the probability for him to find in this list a suitable element, i.e. the one that will satisfy him in the future. Denote *precision* for the user u as p_u:

$$p_u = \frac{N^+_{rec}(u)}{N_{rec}}. \tag{2}$$

Now we can generalize this formula for our entire audience (or a measured sample) of *Users*:

$$P_{N_{rec}} = \underset{u \in Users}{mean}(p_u) = \frac{1}{|Users|} \cdot \sum_{u \in Users} \frac{N^+_{rec}(u)}{N_{rec}} = \frac{N^+_{rec}}{N_{rec} \cdot |Users|}. \tag{3}$$

Every user looks through his own list and chooses what he/she needs, so $P_{N_{rec}}$ shows the average probability of success for all occasions. The value on the right side is the total number of successes in the whole sample.

4 Discounting

So far we have evaluated a list of elements as a whole, but we know that its head is more important than the tail – at least, if the list is not too short. The

metric, that takes this into account and, therefore, depends on the order of the elements in the list, is called a *discounted metric*.

The list's head is more important than the tail due to the uneven distribution of user's attention: people more frequently look at the first element, they less frequently look both at the first and the second elements of the list, etc. This means that a proper discounting requires a behavioral model and the data that can support and train this model.

Let us imagine an arbitrary user who looks through the elements of the list one by one, starting with the first element, then the second, the third... and then, at some point, stops. There is no need to identify a specific user, because sometimes the same person wants to see the list of 2 elements, and sometimes the list of 20. It might be useful to know the average probability of transition from one element to another, but we do not need such precise data. If there is a probability w_N that an arbitrary user goes through a list of N elements for any plausible N, then we can average the value of P_N according to the law of a total probability, where P_N is the average probability of success for the part of our audience, that went through the list of N elements. It can be described by the following definition:

$$AUC = \sum_{N_{rec}=1,\ldots,\infty} w_{N_{rec}} \cdot P_{N_{rec}} \tag{4}$$

In this definition N was replaced with N_{rec}. It turns out that in contrast to precision, AUC value estimates the average probability of success of personal recommendation lists in a real life environment, when the users' attention is unevenly distributed. Note that in order to derive the value of the AUC we used the dependence of precision $P_{N_{rec}}$ on the size of the recommendation list N_{rec}, which was considered as fixed earlier.

Let us note that the term AUC is also used to represent the *precision by recall integral*, which is sometimes used as a quality metric for classifiers [8]. The sum we calculated in Formula 4 is an analogue of this metric: different N_{rec} values simulate different Recall values.

4.1 An easy way to estimate w_N values

There is a simple evaluation model for w_N, which allows not to handle all of the transition probabilities, but provides a qualitative description of user's behavior. The only model parameter is the probability Q that an arbitrary user moves to the second page in the list, which is available through pagination web logs. Assume that each page contains m elements and users proceed to the next viewing element with the same probability p (or leave with the probability $(1-p)$). Then p can be easily obtained from the $Q = p^m$ ratio, assuming that the first element is viewed with a probability of 1. Then, the probability w_N that a user sees N elements and then stops can be easily calculated with the following equation:

$$w_n = p^{(N-1)} \cdot (1-p). \tag{5}$$

A similar approach, where transition probability p is set to be a constant was used in [7].

5 Rate value and satisfaction

So far we have only been using positive signals while ignoring the fact that we have their values. Clearly, it will be unreasonable to neglect this data. If the rate scale is tuned well, the rate value will be, on average, proportional to the user's satisfaction level. Taking into consideration the above, we can try to replace the counter of successes N^+_{rec} in Equation 4:

$$AUC = \sum_{N_{rec}=1,..\infty} w_{N_{rec}} \cdot P_{N_{rec}} = \sum_{N_{rec}=1,..\infty} w_{N_{rec}} \cdot \frac{N^+_{rec}}{N_{rec}|Users|} \tag{6}$$

with more informative sum of the positive rates:

$$AUC^r = \sum_{N_{rec}=1,..\infty} w_{N_{rec}} \cdot \frac{1}{N_{rec}|Users|} \cdot \sum_{r \in S_{N_{rec}}} (r-5), \tag{7}$$

where $S_{N_{rec}}$ is the set of successful rates, i.e. the positive rates which were counted in N^+_{rec}. On our 10-stars scale we consider six stars and more as a positive rate, so for the convenience we subtract 5 from all rates (implying that we sum up only the "satisfaction stars").

The role of the positive rates in the metric can also be represented in a different way:

$$AUC^r = \sum_{N_{rec}=1,..\infty} w_{N_{rec}} \cdot P_{N_{rec}} \cdot r^+_{mean}(S_{N_{rec}}), \tag{8}$$

where

$$r^+_{mean}(S_{N_{rec}}) = \frac{1}{N^+_{rec}} \cdot \sum_{r \in S_{N_{rec}}} (r-5). \tag{9}$$

We can think about the last term, that it is an average rate (positive and successful) among the audience that went through the list of N_{rec} elements. The product of the success probability and an average positive rate in case of success could be described as the total *satisfaction level*, that can be provided by the recommendation algorithm. In this way the metric, although losing a purely probabilistic interpretation, is now better suited for our purposes.

6 Long tail and short head

Amusingly, when it comes to movies and other media products, the popularity distribution of the elements is extremely uneven. The number of movies that are well-known and has been rated by a large amount of people is very small, it is a *short head*. The vast majority of movies stay unknown to the audience, it

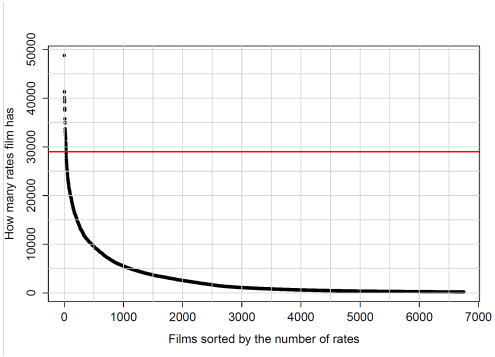


Fig. 2. Rates distribution on Imhonet.ru.

is a *long tail*. For example, the Imhonet rates distribution looks like the one in Figure 2.

The short-head rates distribution curve is so steep, that any rates summation will lead to the short-head movies domination, giving an 80% contribution to the metric. This causes its numerical instability: appearance or disappearance of a few short-head objects in the top of recommendation list can dramatically change the value of the metric.

Let us not forget that the goal of recommendation system is an effective personalization. It is primarily associated with the ability to select items from the long tail, because the elements of the short head are familiar to everyone, so there is no need to include them in the recommendation.

Therefore it seems reasonable to reduce a *short head* weight in the metric. In order to do it correctly, why do not make a start from the metric problem, which is the fact, that the numerical instability reduces sensitivity. We model the test cases for the metric to distinguish and try to achieve its maximum sensitivity.

As a starting point we take the situation when we know nothing about the users in our sample. In that case all personal lists of recommendations reflect an average movie rating and hence look exactly the same. In order to construct this rating we can use a simple probability, based on an average rating, that the movie will be liked. For example, it may be the probability, that an average movie score is higher than five points:

$$P(r > 5) = \frac{|The\ movie\ rates\ greater\ than\ 5|}{|All\ the\ movie\ rates|} \tag{10}$$

Fortunately, in addition to the rates we have a lot of different information about the users from the questionnaire: gender, age, preferences, interests, etc. For example, if we know the gender, we can build two different recommendation lists instead of a generalized one. It can be done using Bayes' formula:

$$P(r > 5|man) \propto P(man|r > 5) \cdot P(r > 5) \tag{11}$$

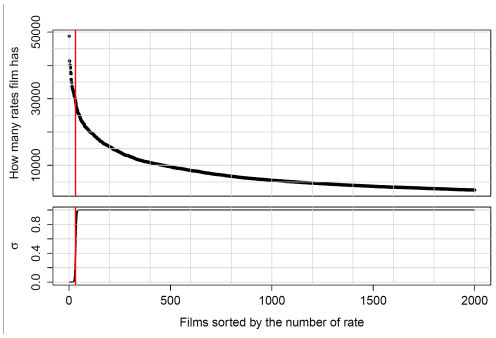


Fig. 3. The *short head* and sigmoid penalty function.

$$P(r > 5|woman) \propto P(woman|r > 5) \cdot P(r > 5) \tag{12}$$

Here, the probability for our starting point $P(r > 5)$ works a priori. Since two different recommendation lists are better than one, we can expect growth in the value of the metric.

It is more convenient to evaluate the relative increase:

$$\frac{AUC(man/woman) - AUC_0}{AUC_0} > 0. \tag{13}$$

The increase of the metric will take place every time we use any additional user information, essential for the users' preferences of movies. The more metric increase, the more it is sensitive to the information. Since we are not specifically interested in gender, in order to avoid over-fitting on this particular case, we will average AUC increase based on the variety of the criteria we use to segment the audience. In our experiment we used the users' answers to a 40 questions questionnaire.

Let us move on to an optimization problem. It can be described as searching for the metric with the best sensitivity. As you remember, the basic solution of the short-head/long-tail problem is to reduce the weight of the short-head elements. We denote the function responsible for the short-head elements penalties as σ. The σ-function must provide a maximum sensitivity:

$$\underset{\sigma}{argmax} \left(\underset{g \in G}{mean} \left(\frac{AUC_g(\sigma) - AUC_0(\sigma)}{AUC_0(\sigma)} \right) \right), \tag{14}$$

where G is the set of audience segmentations with relevant recommendations for each segment. In a simple experiment, which proved to be effective, we have used a step function σ (approximation of sigmoid) to null the *short head* elements weight as it is shown on the Figure 3. This means that the optimization problem 14 needs to be solved for a single parameter σ, which determines the position of the step in the list of movies, sorted by the number of their rates.

7 Final formula

Here is the final formula of the metric, that takes into account all the above reasoning:

$$AUC^r = \frac{1}{|Users|} \cdot \sum_{N_{rec}=1..z} \frac{w_{N_{rec}}}{N_{rec}} \cdot \sum_{r_{ui} \in S_{N_{rec}}} \sigma(i) \cdot (r_{ui} - 5), \qquad (15)$$

- N_{rec} is the length of recommendation list;
- z is the *precision* values summarizing limit. For long lists precision values are very small, as well as multiplier $w_{N_{rec}}$, so significantly large z value does not affect the metric;
- $|Users|$ is the number of users in the sample;
- $w_{N_{rec}}$ is the probability that a random user will look through a list of N_{rec} elements exactly;
- S_{N_rec} is the number of positive rates in the recommendation list of N_{rec} elements;
- $\sigma(i)$ is the penalty function value for an element i;
- r_{ui} is the rate of the movie i received from the user u.

8 Experiments

In this part we will discuss some practical examples of the metric application[3]. We have used a set of special machine learning models of *imhonet.ru* recommendation system. These models are not described here in greater details, since we only want to illustrate using the metric.

Cold start is one of the most crucial problems for recommendation system. There are two kinds of the cold start problem: *a new user* and *a new element*.

8.1 New users

Let us compare the quality of recommendations based on an arbitrary user rates along with the quality of recommendations based on the additional information about the user. The latter recommendations can be designed in order to solve the cold start problem by the following methods:

1. Finding out user's age and gender;
2. Giving a user few simple questions to answer;
3. Using user's rates given to non-movies elements.

The results of metric calculation for all these methods are presented in Figure 4. The black horizontal line at the bottom of the plot represents the quality of recommendation list in case we have no information about users and suggest all of them the same list of recommendations calculated by 10.

[3] All the datasets used in the experiments are available from the first author of this paper by e-mail request

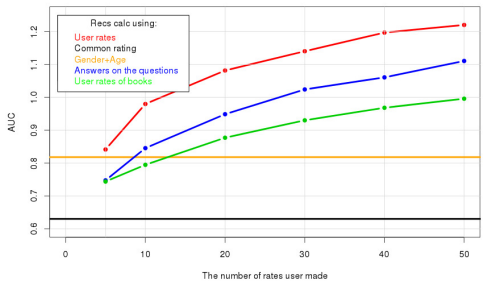


Fig. 4. Decision of new-users problem.

The topmost red curve shows the dependence of recommendations quality from the number of objects rated by user. The more objects user rates, the more precise his/her list of recommendations is.

The orange horizontal line near 0.8 *AUC* level shows the quality of recommendations in case we know only gender and age of the users. Information about user's gender and age makes the metric more precise as much as about 5 rates.

Now consider the blue curve under the red one. For a common user it is not always easy to put their preferences directly into rates, so we offer newcomers few simple questions, such as "Do you love anime?" or "Do you like action?", that are chosen from a few hundred questions list. Although questions are not as informative as rates (for example, 10 rates are equivalent to 25 answers), they are still useful, since for the majority of users it is easier to answer a question rather than to give numerical rate.

Let us explain the lowermost green curve. Sometimes we have to deal with the users who has already got a profile, based on the rates of the elements from non-movie domains. If there is a connection between preferences in different domains, we can try to use it. In our case, the users have already got profiles in fiction books section, and we are trying to use this information in order to recommend them some movies.

8.2 New items

It is important to be able to recommend new elements before they have received enough rates from users. Clearly, this can only be done on the basis of information about the movie itself. In our recommendation system movie properties that have a key influence on the recommendations are as follows: *genre*, *director*, *actors*, *screenwriter* These metadata make it possible to recommend a movie as if "experienced" users (not newcomers) have already given it about 27 ratings, which you can see in Figure 5.

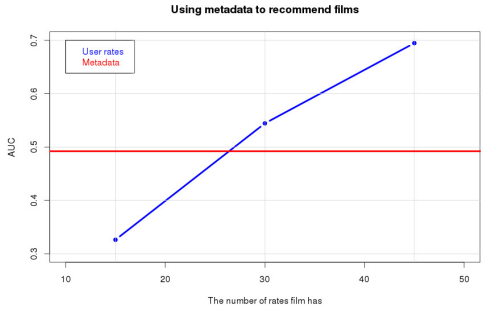


Fig. 5. Solution of the new-items problem.

A similar problem was described in [9]. Their *SVD*-based recommendation models for new elements were evaluated by RMSE. 10 user rates appeared to be more valuable than the metadata description.

9 Conclusion and future work

In this paper we have described the consistent inference of the metric for a recommendation system which is based on explicit users' rates and designed to provide a ranked list of recommended elements. We hope that the metric possesses a set of specific properties, such as: it is sensitive to the order of the items on the list or, more precisely, discounted in accordance with a simple behavioral model, when the user goes through the recommendations one by one, from the top to the bottom; it takes into account the value of positive ratings, so it can measure not only the concentration of successes, but also the amount of satisfaction; correctly handles short head/long tail problem — penalizes short head elements to optimize the sensitivity;

The main purpose of the metric inference is to develop an effective tool, that could be used for the recommendation algorithm optimization accompanied by the improvement of the business metrics. It means that in order to estimate the metric efficiency we will have to compare the target business metrics with the dynamic of the recommendation system metric. Although, as we have noticed in the introduction, this procedure is quite complicated and will be discussed further later.

Acknowledgments We are grateful to Prof. Peter Brusilovsky (University of Pittsburgh), Prof. Alexander Tuzhilin (NYU/Stern School), and Prof. Alexander Dolgin (HSE, Moscow, Russia) for discussions and helpful suggestions. We also would like to thank participants of clasess on recommender systems at NewProLab and HSE, Moscow, as well Sheikh Muhammad Sarwar (Dhaka University,

Banladesh). The third co-author was partially supported by Russian Foundation for Basic Research, grants no. 16-29-12982 and 16-01-00583.

References

1. Gunawardana, A., Shani, G.: Evaluating recommender systems. In: Recommender Systems Handbook. (2015) 265–308
2. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010. (2010) 39–46
3. Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Recommender Systems Handbook. (2015) 309–352
4. Adamopoulos, P., Tuzhilin, A.: On unexpectedness in recommender systems: Or how to better expect the unexpected. ACM Trans. Intell. Syst. Technol. **5**(4) (December 2014) 54:1–54:32
5. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: CROC: A new evaluation criterion for recommender systems. Electronic Commerce Research **5**(1) (2005) 51–74
6. Lu, Q., Chen, T., Zhang, W., Yang, D., Yu, Y.: Serendipitous personalized ranking for top-n recommendation. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2012, Macau, China, December 4-7, 2012. (2012) 258–265
7. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. Trans. Inform. Syst. 27 (December 2008)
8. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: ICML '06 Proceedings of the 23rd international conference on Machine learning. (2006) 233–240 ISBN:1-59593-383-2.
9. Pilászy, I., Tikk, D.: Recommending new movies: even a few ratings are more valuable than metadata. In: Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009. (2009) 93–100

Choice of the Group Increases Intra-Cooperation

Tatiana Babkina[1], Mikhail Myagkov[2], Evgeniya Lukinova[3], Anastasiya Peshkovskaya[4], Olga Menshikova[5], and Elliot T. Berkman[6]

[1] Skolkovo Institute of Science and Technology, Novaya, d.100, Karakorum Building, 4th floor, Skolkovo, 143025, Russian Federation, t.babkina@skoltech.ru,
[2] Department of Political Science, University of Oregon, 1585 E. 13th Avenue, Eugene, Oregon, 97403, United States
[3] New York University Shanghai, 1555 Century Ave, Pudong, Shanghai, China 200122
[4] Laboratory of Experimental Methods in Cognitive and Social Sciences, Tomsk State University, 36, Lenina Avenue, Tomsk, 634050, Russian Federation
[5] Russian Presidential Academy of National Economy and Public Administration (RANEPA), Prospect Vernadskogo, 82, Moscow, 119571, Russian Federation
[6] Department of Psychology and Center for Translational Neuroscience, University of Oregon, 1227 University of Oregon, Eugene, Oregon, 97403, United States

Abstract. This research investigates how variation in sociality, or the degree to which one feels belonging to a group, affects the propensity for participation in collective action. By bringing together rich models of social behavior from social psychology with decision modeling techniques from economics, these mechanisms can ultimately foster cooperation in human societies. While variation in the level of sociality surely exists across groups, little is known about whether and how it changes behavior in the context of various economic games. Specifically, we found some socialization task makes minimal group members behavior resemble that of an established group. Consistent with social identity theory, we discovered that inducing this type of minimal sociality among participants who were previously unfamiliar with each other increased social identity, and sustained cooperation rates in the newly formed groups to the point that they were comparable to those in the already established groups. Our results demonstrate that there are relatively simple ways for individuals in a group to agree about appropriate social behavior, delineate new shared norms and identities.

Keywords: collective action, group formation, cooperation

1 Introduction

Humans as "social animals" like to be around each other. More than mere liking, it has been proposed that social affiliation, specifically feeling a sense of belonging in a social group, is a basic human need [4]. Given the fundamental importance of social belongingness to survival, it follows that humans have

evolved specialized affective and cognitive tendencies related to group membership. This can be seen both in the extreme suffering caused by social isolation and loneliness and also in the feelings of affiliation and acts of commitment associated with group membership. The focus of our program of research is on the influence of group membership on cognitive processes and their attendant behavioral outcomes such as economic decisions.

Definition 1. *Sociality, even in a very minimal form, serves as a natural mechanism to promote sustainable cooperation among group members.*

Definition 2. *Sociality, or social utility, is defined here as an additional component of the subjective utility function that reflects the value of contributing to group outcomes and cohesion, and is derived at least in part from a sense of social identity.*

A recent study from our group found that socialization in the form of brief group-based social interactions prior to a Prisoners Dilemma or Ultimatum Game added social utility to choices that punished group members who did not cooperate (e.g., free-riders), but not to choices that punished the group as a whole [5]. A current research goal is to find an effective mechanism of sociality induction that allows members of a minimal group to mimic the economic behavior of those from the established group.

Our laboratory model of sociality [13] combines the classic "minimal group" paradigm from social psychology [20] with group-based manipulations that induce a sense of social connectedness in humans to allow us to measure the degree of utility conferred by sociality that otherwise has no economic utility. Testing this model would enable social psychology and economics to make a connection, such that the former can represent group membership in terms of a value calculation and the latter can inform its decision making models with empirical insights into the factors that guide human behavior in a social environment.

Traditional economic analysis generally makes the simplifying assumption that people are exclusively utility-maximizing and self-regarding. However, the breakdown point of economic models is in explaining behaviors that are altruistic, fair and trusting [11]. Such behaviors are inseparably linked with social context. Factors such as group membership, social identity and affiliation motives induce prosocial behaviors through additional utility that is rarely included in formal models of economic behavior. There are many ways of manipulating sociality for the purpose of testing its effect on economic decisions. To our knowledge, a formal typology of the various kinds of sociality is not currently available, even though such a typology would be quite useful for the present line of research and related efforts. A key contribution of this paper is to use various manipulations of sociality, to compare their effects on human choice behavior, and select an effective one that boosts sociality of strangers in a group to the level of behavior seen within established groups. Comparing these manipulations will have a broad impact on the field because researchers are in need of procedures that can reliably manipulate sociality. Individuals that identify with certain social groups

are often involved in power struggles in that they try to establish, change, or defend a power structure. Sociality in the minimal form is a way to make struggles and protests lead to pro-social outcomes.

2 Economic Decisions in a Group Context

Prosocial behaviors are innate to humans because our survival individually and as a species depends on collective action and achievement of common goals [1, 6]. Since [8] researchers argued that distinctive features of human sociality resulted from selection among individuals who live in groups. Voluntary social integration of people into groups occur for the reasons of need for affiliation (inclusion) [4], need for power dependence (control), need for intimacy (affection) [18], need for achievement [14], proximity in distance and in social self (attachment), and making sense of the world (principle of "social proof" to validate your own existing beliefs).

Groups, in turn, are complex adaptive systems, they are "entities that emerge from the purposive, interdependent actions of individuals" [15]. In fact, observation of naturally occurring groups in public places reveal that dyads are common and few groups contain more than five or six people [7, 16], this justifies our decision to divide participants in groups of six people. In small self-organized groups norms emerge that guide coordinated action, including cooperation [2, 4]. Previous evidence suggests that endogenously formed groups, unlike exogenously formed alliances, lead to the creation of social ties and trust that favorably affect cooperation in economic games such as the Prisoners Dilemma (PD; [10, 21]. Moreover, autonomy in group membership can transform competition of individual outcomes to be a competition of generosity and other prosocial traits. It has been proposed that choosing a group in animal and human societies can reflect your willingness to follow the leaders of a group; in this case, group members will endorse the norms of the leader and the leader will influence the members to achieve efficacy in a task performance. We compared the economic behavior of players under variations of sociality in an experimental setting. Specifically, our central comparison is between the cooperative behavior of people in groups formed around minimal social characteristics and people in already-established groups that were based on relatively longstanding and salient identity features.

3 Current Study

This paper reports on the results of a series of laboratory experiments that systematically varied the type of sociality induction and measured economic decisions as participants interacted anonymously with others in their in-group or in an out-group. In particular, experiments with different social group composition were considered: participants in the Assigned condition (number of subjects: N=108) were randomly assigned to a group to socialize and play with; participants in the Established condition (number of subjects: N=60) were invited as

a part of established group (group formed long before the experiment) to socialize and play with; participants in the Choice condition (number of subjects: N=108) were randomly assigned to choose a group with whom they socialized and played.
In our previous study [5], we coupled a classic "minimal group" paradigm from social psychology with group interaction-based manipulations to induce a sense of social connectedness in our participants. This experiment allowed us to measure the degree of utility conferred by sociality that otherwise has no economic utility. We found that socialization, or group interactions immediately prior to the focal economic games, created and sustained fairness during the economic games even though this social interaction was logically irrelevant to the games. In this study, we changed the socialization pattern and altered a second dimension of sociality, i.e. the manner in which the group was assembled. By making only one change in our socialization task we kept sociality in a relatively minimal form: strangers are allowed to interact for 15 mins and form groups endogenously. Our desire is to find a group-based manipulation that causes economic decisions made under minimal sociality resemble those of established and real life cultural groups under strong sociality.

4 Materials and Methods

The study procedures involving human participants were approved by the Skolkovo Institute of Science and Technology (Skoltech) Human Subjects Committee. Written informed consents were obtained from participants. Experimental data are readily available on Dropbox.[7]
Subjects (total number of subjects: $N = 276$, 174 males) for this set of experiments were recruited from the students at the Moscow Institute of Physics and Technology (MIPT). The MIPT Experimental Economics laboratory was used to carry out all experiments. Each experiment (total number of experiments: 23) consisted of a different set of 12 students, pre-selected before the experiment to be unfamiliar with one another (except for the Established condition). We acknowledge that some of the pre-selected students might know each other, however, we tried to avoid it by selecting distant year students from different departments, distinct science orientation within one department (linked to a particular research institution) and by checking their Vkontakte.ru (biggest Russian social network) profile and friend circles. In the Established condition we purposefully invited groups of 6 people that know each other and share an affiliation of some sport club or a hobby. In order to carry out experiments in this condition an advertisement recruiting participants in Vkontakte specifically requested established groups of 6 formed around hobby or sport to sign up. In particular, one of the group members was asked to enter 6 names of participants including himself and the meaningful condition around which the group was formed.
Each experiment lasted a bit longer than 1 hour and was divided into 3 consec-

[7] https://dl.dropboxusercontent.com/u/7646503/ChoosingYourTeammatesdata.zip

utive phases (Fig.1) that occurred in a fixed order.

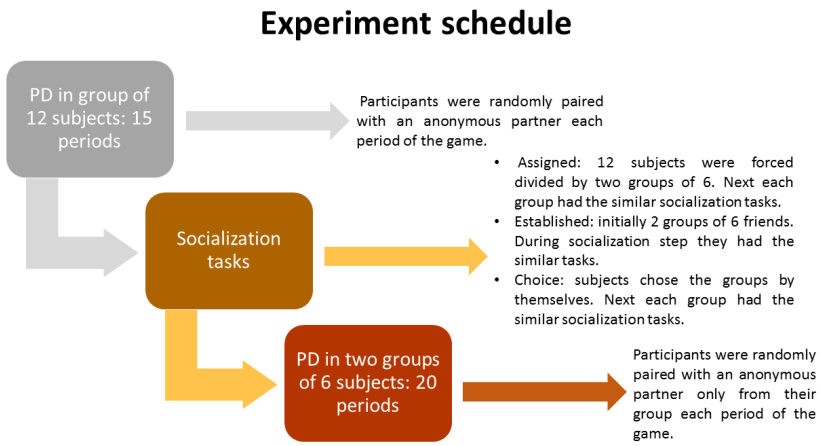


Fig. 1. Phases of the experiment depending on types of the groups.

1. Anonymous Game phase, where participants played the two-person one-shot Prisoners Dilemma (PD) (the parametres of the PD game are shown in Fig.2). Participants were randomly paired with an anonymous partner each period of the game (economists refer to this as one-shot games) and alternated roles on subsequent trials between column chooser and row chooser for the PD for the dynamic of the experiment. This game phase is repeated for 15 periods, though this number was not known to participants at the start of the phase. Each period participants were given information only about their profit for that period.

2. Socialization (10-15 minutes). Experiments were randomly assigned to their sociality manipulation condition. Critically, participants were socialized with either assigned others (i.e., chosen by the experimenter) or chosen others (i.e., chosen by the participants). For the Established condition participants were pre-selected to be familiar with each other in a meaningful social group of 6.
Assigned: All 12 participants completed the Sociality induction, an icebreaker called "Snowball": First participant said his/her name and an adjective that started from the same letter, second subject repeated first subjects name and adjective and said his/her own name and adjective, and so on till the last participant said all names and adjectives in order. Then, in a reverse order each participant shared his/her life facts, including major, year, and hobby. Participants formed a circle during this task. Finally, participants were randomly divided into two groups of 6 and each group completed the task of identifying

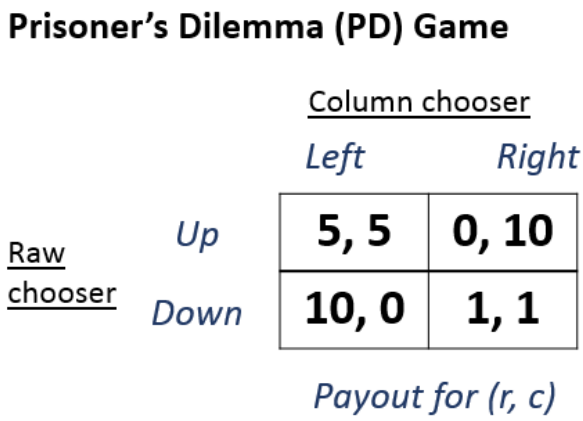


Fig. 2. Economic games to be played against anonymous partners from the socialized and non-socialized groups. Participants will be row and column chooser in the Prisoners Dilemma (left), and the offerer and responder in the Ultimatum Game (right).

five characteristics that everyone in their group shares, and then selected one of those characteristics as their groups name. The group then provided to the experimenter a list with the characteristics written down and the groups name circled.

Established: Participants in this condition completed a sociality "reminder" whereby they were divided into established groups that matched pre-selected meaningful social groups (e.g., members of aerobics team, dorm neighbors, football team fans, etc.) of 6 people, named their group, and in 3 mins shared the group name with the experimenter.

Choice: All 12 subjects participated in the same Sociality induction from Assigned condition, an icebreaker in which the participants in a sequence told their names and adjectives that started from the same letter and in a reverse order shared their life facts. After that participants were asked to raise their hands if they volunteer to be a leader. Participants did not know from instructions what responsibilities of a leader will entail. The first two participants with hands raised automatically become group leaders (rules of becoming a leader and rules of group formation are given in the Supporting Information). Players that were not leaders were asked to decide which leader they want to join on a piece of paper. The participants get to choose a group, which easily satisfies the minimal group requirement [19] and social identity theory [20]. Finally, as in Assigned condition each group selected their groups name based on identified

common characteristics and passed it to the experimenter.

3. Socialized Game phase, where participants played the PD with a random human partner from their socialized group of 6. Their anonymous partner changed each round of the game, making it a one-shot game. The participants switched roles on alternating trials: column chooser and row chooser for PD. There was total of 20 periods in this game phase. Number of periods was not known to participants. Each period participants were given information only about their profit for that period.

Our main hypothesis is that participants will behave more cooperatively, and do so in a more sustained way, in the Choice condition compared to the Assigned condition. This is due to the group-based social factors suggested by evolutionary and social psychologists (e.g., autonomy) that alter the expected patterns of economic behavior based solely on a motive to maximize ones own immediate utility. The Established condition was included as a solid test of the strength of this manipulation. This condition allowed us to compare cooperation rates among strangers who underwent a sociality induction in the laboratory (Assigned and Choice conditions) to those of people in established groups who underwent a similar procedure in the field.

5 Results

Result 1. Various types of socialization activate sociality and facilitate collective action.

Across all variations of sociality a cooperative equilibrium was created. Fig. 3 portrays the average mean cooperation rate across three variations of sociality. In each of the sociality manipulations, displayed on the x-axis, there was an increase in the cooperation rate from the Anonymous Game phase (the first stage of each experiment), when subjects played anonymously with other people who were strangers, to the Socialized Game phase (the third stage of each experiment), when the same subjects played anonymously with people who were part of the subjects socialized group. There was a significant elevation of group cooperation rates from the game phase before socialization to the game phase after socialization (Anonymous $M = 0.21, SD = 0.06$, Socialized $M = 0.53, SD = 0.12; t(20) = -5.49, p < .05$; Table 1). Each data point is the rate of cooperation (proportion of cooperative choices) averaged across all subjects and all periods within Anonymous and Socialized games in one experiment.

Result 2. Stability of cooperation in the newly formed groups with socialization is comparable to that in the established groups.

Fig. 4 shows the sustainability of cooperation across sociality manipulations in Socialized game phase. The closer the blue bar height (mean cooperation for the first 5 periods) is to the red bar height (mean cooperation for the last 5 periods), the more sustainability there is. It is apparent that there is an increase

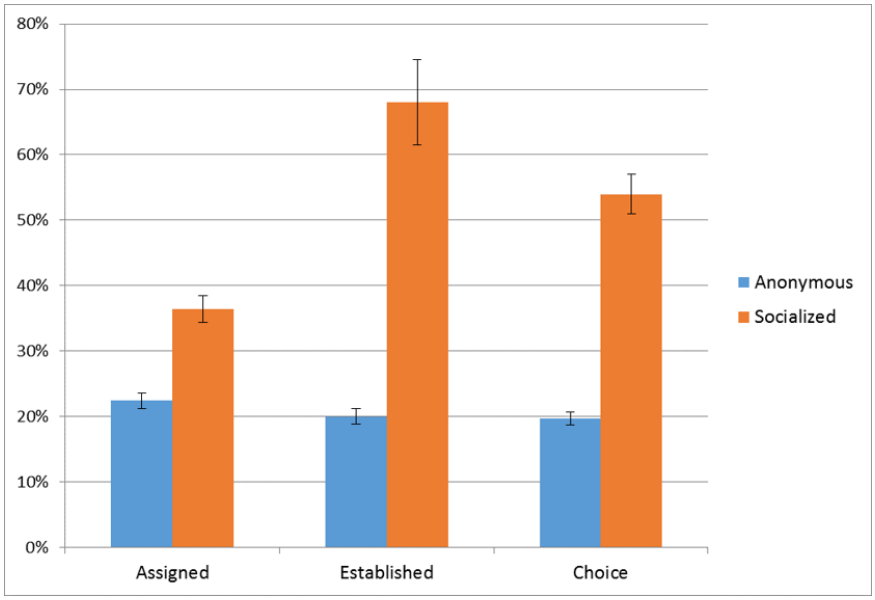


Fig. 3. Cooperation level of socialized group members, Socialized Game, (red) and non-socialized group members, Anonymous Game, (blue) across three types of sociality manipulations. Each bar represents the mean of the cooperation rate for all experiments of a certain condition. Error bars are SE.

from early to late cooperation in the Established group, there is only some decline in the Choice group, and there is a stark discrepancy between the two time intervals in the Assigned group. The average cooperation rates in the Socialized game phase significantly declined in the Assigned group (proportion of cooperative choices for the first 5 periods $M = .46$, $SD = 0.41$; last 5 periods $M = 0.26$, $SD = 0.35$; $t(47) = 3.37$, $p < .01$; see Table 1), whereas there is no significant decline for Choice socialization (average cooperation for the first 5 periods $M = .59$, $SD = 0.42$; last 5 periods $M = 0.53$, $SD = 0.41$; $t(59) = 1.53$, $p = 0.13$; see Table 1). Taking into consideration quartile of experiment duration (4 on 5 periods), there are significant main effects of condition as well as a

Table 1. Cooperation rates

Types	Assigned	Established	Choice
Anonymous Game	22%	20%	20%
Socialized Game	36%	69%	54%
Socialized Game (first 5 periods)	46%	64%	59%
Socialized Game (last 5 periods)	26%	72%	53%

condition-by-time interaction on cooperation, but not of time period, (two-way factorial ANOVA; D.V.: proportion of cooperative choices averaged across all subjects within one experiment; condition: $F(2, 59) = 107.13$, $p < .01$; quartile: $F(3, 59) = 1.81$, $p = 0.16$; condition on time: $F(6, 59) = 3.21$, $p < .01$).

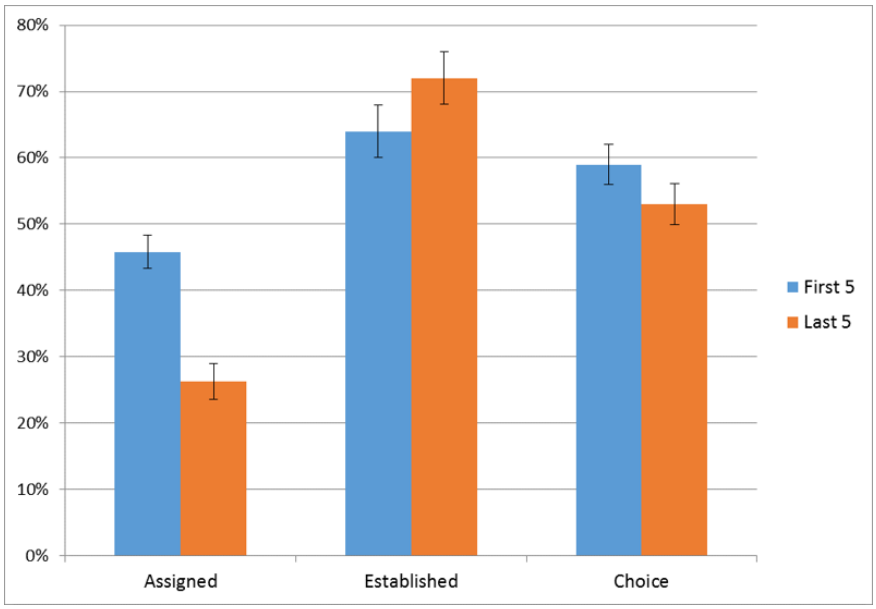


Fig. 4. Cooperation level during first five periods of the Socialized Game Phase (blue) and the last five periods of the Socialized Game Phase (red) across three types of sociality manipulations. Each bar represents the mean rate of cooperation for 5 periods for all experiments of a certain condition. Error bars are SE.

6 Discussion

A persistent and curious feature of human behavior in social environments is "prosocial" actions such as fairness, cooperation, and the provision of collective good even at the cost of individual gain. Thus, identifying the mechanisms of collective action in different circumstances is a major puzzle for social science. This paper integrates theories and decision models from economics, social psychology, and neighboring fields, makes them comprehensible for a broad general science audience, and provides insight into the mechanisms of human sociality in the context of during economic choice.

In this paper, we tested predictions from social psychological theory that suggested ways to increase cooperation. In doing so, we identified a promising way to

increase a sense of sociality in a group of relative strangers. We compared different types of sociality manipulations and concluded that each of them facilitated collective action through increasing cooperation among individuals. Through one of the minimal sociality manipulations, i.e. voluntary group choice, group members, who were strangers before experiment, were able to achieve and sustain cooperation comparable to that of the meaningful social groups.

Our research team started with the idea that individual will economically value the outcomes of others in a group to a greater extent when that individual strongly identifies with the group. This idea has been extensively developed by psychologists and suggest that prosocial behaviors are driven not by miscalculations of utility but instead by social factors (which are not mutually exclusive with each other) such as the salience of social identity, the presence of in-group favoritism and group norms, and evolutionary adaptations to foster group success. Economists have only recently begun to incorporate these insights from psychology into their models and studies of decision-making and choice, but at this point several groups are converging on the parsimonious explanation that aspects of the social world impart subjective utility to individual choices that favor group outcomes. Of course, the links between minimal group, social identity, and expected utility theories on the one hand, and the recent developments in the life sciences and neuro disciplines on the other are still tenuous at best. In this paper we have obtained strong evidence to support the idea that "sociality," or an individuals sense of belonging and connectedness to a group, holds positive subjective utility and thereby can influence economic decisions when they take place in a social context. We conclude that choosing ones group and interacting with your group for a small period of time achieves the level of in-group favoritism of already established groups and believe that our results will encourage other researchers to explore this potentially very profitable domain.

Acknowledgements. We thank Rinat Yaminov for writing the programming code for experiments, Aleksander Chaban for technical help in conducting experiments at MIPT, and Ivan Menshikov for useful suggestions in data analysis.

References

1. Aronson, E.: The social animal. WH Freeman and Co. New York (1995)
2. Arrow, H., Bennett, R., Crosson, S., Orbell, J.: Social poker: A paradigm for studying the formation of self-organized groups. Institute of Cognitive and Decision Sciences, University of Oregon (1999)
3. Bagozzi, R., Dholakia, U.: Intentional social action in virtual communities. J INTERACT MARK 16(2): 2-21. (2002)
4. Baumeister, R., Leary, M.: The need to belong: desire for interpersonal attachments as a fundamental human motivation. PSYCHOL BULL 117(3): 497529. doi:10.1037/0033-2909.117.3.497. (1995)
5. Berkman, E., Lukinova, E., Menshikov, I., Myagkov, M.: Sociality as a Natural Mechanism of Public Goods Provision. PLOS ONE 10(3): e0119685. (2015)

6. Bowlby, J.: Attachment and loss. Vol. 1. Attachment. Hogarth, London (1969)
7. Caporael, L., Baron, R.: Groups as the minds natural environment. In Simpson JA, Kenrick DT (eds) Evolutionary social psychology. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 317-344. (1997)
8. Darwin, C.: The descent of man and selection in relation to sex. John Murray, 2nd edition, London (1874)
9. Goette, L., Huffman, D., Meier, S.: The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. AM ECON REV: 212-216. (2006)
10. Goette, L., Huffman, D., Meier, S.: The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. AM ECON 4(1): 101-115.(2012)
11. Hoffman, M., Yoeli, E., Nowak, M.: Cooperate without looking: Why we care what people think and not just what they do. P NATL ACAD SCI USA 112(6): 1727-1732. (2015)
12. Hogg, M., Turner, J.: Interpersonal attraction, social identification and psychological group formation. EUR J SOC PSYCHOL 15(1): 51-66. (1985)
13. Lukinova, E., Myagkov, M., Shishkin, P.: The value of sociality. Foresight 16(4): 309-328. (2014)
14. McClelland, D.: How motives, skills, and values determine what people do. AM PSYCHOL 40(7): 812. (1985)
15. McGrath, J., Arrow, H., Berdahl, J.: The study of groups: Past, present, and future. PERS SOC PSYCHOL REV 4(1): 95-105. (2000)
16. Moreland, R., Levine, J., Wingert, M.: Creating the ideal group: Composition effects at work. In Witte, E., Davis, J. (eds) Understanding group behavior: Small group processes and interpersonal relations. Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 2, pp 11-35. (1996)
17. Ryan, R., Deci, E.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. AM PSYCHOL, 55(1): 68. (2000)
18. Schutz, W.: FIRO: A three dimensional theory of interpersonal behavior. Holt, Rinehart Winston, New York (1958)
19. Tajfel, H. E.: Differentiation between social groups: Studies in the social psychology of intergroup relations. Academic Press (1978)
20. Tajfel, H., Turner, J.: An integrative theory of intergroup conflict. The social psychology of intergroup relations, 33(47): 74. (1979)
21. van der Werff, L., Buckley, F.: Getting to know you a longitudinal examination of trust cues and trust development during socialization. J MANAGE, 0149206314543475. (2014)

Modelling Human-like Behavior through Reward-based Approach in a First-Person Shooter Game

Ilya Makarov [1], Peter Zyuzin [1], Pavel Polyakov [1], Mikhail Tokmakov [1], Olga Gerasimova [1], Ivan Guschenko-Cheverda [1], and Maxim Uriev [1]

National Research University Higher School of Economics,
School of Data Analysis and Artificial Intelligence,
3 Kochnovskiy Proezd, 125319 Moscow, Russia,
iamakarov@hse.ru, revan1986@mail.ru
http://www.hse.ru/en/staff/iamakarov peter95zyuzin@gmail.com
polyakovpavel96@gmail.com matokmakov@gmail.com olga.g3993@gmail.com
vania1997qwerty@gmail.com maximuriev@gmail.com

Abstract. We present two examples of how human-like behavior can be implemented in a model of computer player to improve its characteristics and decision-making patterns in video game. At first, we describe a reinforcement learning model, which helps to choose the best weapon depending on reward values obtained from shooting combat situations. Secondly, we consider an obstacle avoiding path planning adapted to the tactical visibility measure. We describe an implementation of a smoothing path model, which allows the use of penalties (negative rewards) for walking through "bad" tactical positions. We also study algorithms of path finding such as improved I-ARA* search algorithm for dynamic graph by copying human discrete decision-making model of reconsidering goals similar to Page-Rank algorithm. All the approaches demonstrate how human behavior can be modeled in applications with significant perception of intellectual agent actions.

Keywords: Human-like Behavior, Game Artificial Intelligence, Reinforcement Learning, Path Planning, Graph-based Search, Video Game

1 Introduction

The development of video games always face the problem of creating believable non-playable characters (NPC) with game artificial intelligence adapted to human players. The quality of NPC's model in terms of game behavior extremely depends on an interest in the gameplay as in-game interaction of human players with game environment and NPCs. The main entertainment of many games consists of challenging enemy NPCs, so called, BOTs. Human players, on one hand, estimate BOTs to behave like humans, on the other hand, there should be high probability to mine BOT's patterns finding its weaknesses. Human players always estimate the possibility to overcome computer player through intelligence

supremacy. The combination of such beliefs is what makes a gameplay interesting and satisfying humans ambitions, but also providing new cognitive field of learning through reward based winning policy.

A first-person shooter game is a special genre of video games simulating combat actions with guns or projectile-based weapons through a first-person perspective. The human player experiences virtual world and action gameplay through the eyes of player's human-like model placed in virtual 3D scene, which is shown at the Figure 1. The problem aroused from the player's expectations

Fig. 1. First-Person Shooter Game

of computer players to obtain information from virtual world in a similar way. From the human point of view it is unfair to have an access to special features and information about game environment, which could not be available and processed by human players during a game. In [1] authors stated the principle that it is better to play against BOTs "on equal terms", rather than against "God-mode" undefeatable opponents. Thus, we aim to make behavior of BOTs similar to human players' behavior in first-person shooter (FPS).

The main criterion of evaluating the quality of a game artificial intelligence is the level of compliance for NPC actions with respect to ability of human experts to distinguish computer-controlled and human players in common and specific in-game situations. One of approaches consists of interpretation of such quality based level of BOT humanness through Alan Turing test for computer game BOTs [2]. In the competition, computer-controlled BOTs and human players that are also judges take part in combat actions during several rounds, whereby the judges try to guess which opponents are human. In a breakthrough result, after five years[1] of attempting from 14 international research collectives, two teams have succeeded in breaking through 25% human-like player behavior barrier. Researchers believed that methods developed for a game A. Turing test

[1] http://botprize.org/publications.html

should eventually be useful not just in developing intelligent games but also in creating virtual training environments. Both teams separately cracked test with two prototypes of human-like BOTs that try to mimic human actions with some delays and use neuro-evolution model under human gameplay constraints [3]. The disadvantage of such an approach consists of the fact that such models only imitate human intellect but do not give BOT its own cognitive model. In such a case we still do not know what are the reasons for human actions and how BOT could retrieve new information from human gameplay.

However, the most common ways to implement game AI are still finite-state machines and rule-based systems applied to BOTs behavior [4,5]. The cheapest way for game developing company is to script behavior of NPCs with respect to restricted game situations fully describing most common NPC actions but not giving it a freedom of choice or enough quantity of randomness in decision making. However, this approach has several serious disadvantages: developers can not script or predict all the possible game situations which may arise during the game, so it is impossible to write all patterns of the rules or the states for NPC behavior. As a result, in a certain game situations BOTs do not act optimal and become recognizable for the wrong decision templates from its scripted model, which significantly reduces the quality of gameplay. This could also lead to BUGs' appearance (semantical and technical errors in BOT's actions).

The idea of selecting script parameters via machine learning are now interesting for the researchers, which could study evolved systems based on rule-based systems [6]. Still, even the BOT model tweaked behavior can not be modified during online game testing without decreasing its quality and stability. The problem also appears when such programmed behavior seems to be static and is not sensitive to changes in the environment and game strategies of other players and their skills' levels.

The authors of [7] present another method for online interactive Reinforced Tactic Learning in Agent-Team Environments called RETALIATE. The system take fixed individual BOT behaviors (but not known in advance) as combat units and learns team tactics rather coordinating the team goals than controlling individual player's reactive behavior. Another real-time behavioral learning video game NERO was presented in [8]. The state-of-art researches on evolution approach can be found in [9,10,11,12].

Following empirical study of machine learning and discrete optimisation algorithms applied to modeling player behavior in a first-person shooter video game[2] we focus on some aspects of human decion-making, such as *weapon selection*, *path planning* and *incremental path finding*. Each section of the paper contains one example of AI improvement based on human behavior, thus creating intensified cycle of applying human behavioral patterns to model them in game.

[2] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/geisler.thesis.pdf

2 Weapon Selection

Considering the methods of machine learning, such as supervised, unsupervised and reinforcement learning, the latter one gives us the most suitable way to implement BOT's behavior in FPS game. During reinforcement learning BOT receives an award for each committed action, which allows him to accumulate an experience of various game situations and to act in accordance with the collected knowledge, constantly modifying its tactical and strategy decisions [1]. The disadvantage of such an approach is that reinforcement learning methods require to remember each specific pair of state-action.

A weapon selection tactics for the BOT should be similar to human player's. In real life we often could not predict the result of an action that we are going to perform. Humans' decisions are based on their personal experience. So, the agent interacts with the environment by performing some actions and then receiving reward from the environment. The purpose of this method is to train the agent to select actions in order to maximize reward value dependently on environment states. In such a model BOTs will choose the most effective weapons with respect to computed parameters of the game environment.

We apply a weapon selection model that is based on neural network from [13]. FALCON (Fusion Architecture for Learning, Cognition, and Navigation) is a self-organizing neural network that performs reinforcement learning. The structure of this neural network comprises a cognitive field of neurons (it can be also named a category field) and 3 input fields: sensory field, motor field and feedback field shown at the Figure 2. Sensory field is designed for representing

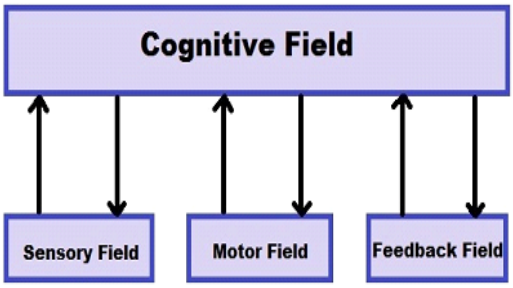


Fig. 2. FALCON Architecture

states of environment. Motor field is designed for representing actions that agent can perform (in our case, an action is selecting the most suitable weapon). Feedback field is designed for representing reward values. Neurons of input fields are

connected to neurons of a cognitive field by synapses. FALCON enables BOT to remember value of the reward that was received by the BOT when it used some weapon in a particular environment state and uses this information to select effective weapons in future.

As of today, we use the values of distance between the BOT and the enemy and enemies current velocity as state parameters; the set of weapons that accessible to BOT includes rifle, shot-gun, machine-gun and knife. Each of the weapons has advantages and disadvantages. Reward value is calculated using the following formula:

$$r = (a + b * distance) * damage,\ a = 1,\ b = 9 \text{ was found optimal,}$$

where distance is a normalized value of the distance between the BOT and the enemy, and damage is taken as normalized value of damage that the BOT inflicts to the enemy.

We add new features to the FALCON algorithm to improve it with respect to human's decision patterns:

- We remove the neuron when the number of its unsuccessful exceeded the number of successful;
- We remove the neuron if its consecutive activity brought zero reward;
- We limit the size of cognitive field for the case of network retraining by removing the neurons with the minimum average award;
- We change weighting coefficients of network only if we receive a positive reward.

The latter condition differs from what humans do. We always try to create a new strategy to overcome negative rewards, but a BOT simply try to forget all negative experience and try to make its significant part better with obtaining positive reward.

The results of experiments for one hundred of weapon usages are shown in the Table 1.

Table 1. Original/Modified FALCON

Weapon	Successes, %	Average Range	Average Enemy Velocity	Average Reward
Machine Gun	82/81	.29/.28	.21/.17	.44/.45
Shot Gun	48/72	.26/.18	.28/.24	.24/.36
Sniper Rifle	95/92	.35/.39	.12/.21	.57/.6

As a reader can see, a Sniper Rifle was used more efficiently for long ranges with enemy's higher velocity. A Shot Gun was used more optimal for short range distances and with greater amount of reward increasing it by 50%. A Machine Gun was used efficiently only for a decreased distance, which means that our aiming techniques do not work quite well for a rapid-firing weapon. The example of a modified FALCON showed us that neural network based on the FALCON

can be applied to human-like selecting effective weapons by BOTs during the battle in first person shooter.

3 Path Planning and Path Finding

Path planning and path finding problems are significant in robotics and automation fields, especially in games. There is a major number of approaches for path planning, such as [14], [15], [16], [17], [18].

The first strategy of path planning is connected with providing believable trajectory of BOT motion to a fixed goal under some constraints. In game programming, Voronoi diagrams (k-nearest neighbour classification rule with $k = 1$) are used to make a partition of a navigation mesh to find a collision free path in game environments [14,17,19]. Smooth paths for improving realism of BOT motion are made through splines [20,21] or by using Bezier curves [15,17,22,23]. We used combined approach of both smoothing methods following the works of [15,24,25].

The second strategy of path planning consists of computing tactical properties of a map as a characteristic of Voronoi regions areas. We compute offline tactical visibility characteristics of a map for further path finding penalties and frag map usage to transform paths found by the first strategy to optimise certain game criteria.

The navigation starts with BOT's query to navigation system. Navigation system uses path finding algorithm I–ARA* anytime algorithm from [26] to obtain a sequence of adjacent polygons on navigation mesh. Then a sequence of polygons is converted into a sequence of points. Finally, BOT receives a sequence of points and build a collision free path to walk. We design the interface for an interaction between a querier and the navigation system at each iteration of A^* algorithm. We use region parameters to manage penalties for path's curvature, crouching and jumping at the current Voronoi cell. There is also a special method for querier to influence the navigation with respect to previous movement direction, similar to Markov's chains in path planning [27]. We also used general penalties, such as base cost, base enter cost and no way flag, which can be dynamically modified by any game event.

Now we describe a family of path finding algorithms and how we could use modeling human behavior to reduce their complexity. In contrast to the Dijkstra's algorithm, A* target search uses information about the location of a current goal and choose the possible paths to the goal with the smallest cost (least distance obtained), considering the path leading most quickly to the goal.

Weighted A* as it was presented in [28] was a modified algorithm of A* search with the use of artificially increased heuristics, which leads to the fact that the found path was not optimal. The improvement of these algorithms is ARA* [29]. The purpose of this algorithm is to find the minimum suboptimal path between two points in the graph under time constraints. It is based on iterative running of weighted A* with decreasing to 1 heuristics values. If it decreases exactly to 1, then the found path is optimal.

Algorithm I-ARA* works as well as repeated ARA*, with the only difference that it uses the information from the previous iteration [30]. The first search made using I-ARA* is simple ARA* search.

We present a modification of I-ARA* as human discrete optimisation decision-making: rather than looking at each step for a new path to the target we simply walk proposed suboptimal path until we passed a certain part (partial path length) from the previously found path. The larger the distance, the longer the last iteration I-ARA*, so most of the time-consuming iterations of this algorithm could be omitted. As a result, we found that the number of moves in modified and original I-ARA* algorithms differs not greater than 10% in average but time for computation has been significantly reduced by 5-20 times when labyrinth has not extremely dense wall structure.

For proper work of I-ARA* algorithm, each penalty is jammed to a limited range, so the resulting penalty is not less than the Euclidean distance, which is used as heuristics in our implementation. Once a path is found, it should be converted into a point sequence.

We generated 2D mazes with sizes of 300 by 300 and 600 to 600 with density of free cells equaled to 0.1, 0.2, 0.3, 0.4. For every field size 100 trials have been conducted. During each test, we choose 30 pairs of random free cells and test the value of the heuristic P as percentage of a path length to go until next computation will be needed. In the Table 2 we presented the results of searching path time decreasing (%) and path length increasing (%) for modified I-ARA*. It is easy to see that for dense mazes our modification significantly wins in time with path length stabilizing or even shortening. For sparse mazes increasing of P leads to the error increasing.

Table 2. Time decreasing/Path Length Increasing Comparison

Sparseness	P=0.05	P=0.1	P=0.15	P=0.2	P=0.25	P=0.3	P=0.35
0.1	785/-.06	1194/-.40	1483/0.07	1744/-.64	1965/0.33	2238/0.83	2354/0.87
0.2	293/0.20	410/0.72	526/1.54	578/2.55	666/2.55	725/2.37	785/4.64
0.3	283/2.20	398/2.25	476/2.11	540/4.24	610/6.53	624/7.53	664/10.71
0.4	221/0.06	309/0.39	346/0.62	379/1.75	406/1.52	395/7.72	419/11.94

When developing a BOT navigation, *smoothing* is one of the key steps. It is the first thing for a human to distinguish a BOT from a human player. Several approaches can be used to smooth movements. *Bezier curves* seem to be the most suitable because they could be represented as a sequence of force pushes from obstacles guaranteeing that BOT will not be stuck into an obstacle.

In practice, the contribution of visibility component to remain undetected during BOT motion is very low if we are not taking into account the enemies' movements. We consider the relative dependence of the smooth low-visibility path length with the length of the shortest path obtained by Recast navigation mesh. The resulting difference between the smooth paths with and without a visibility component does not exceed 10–12% [31], so taking into account tactical

information seems to be a useful decision. The difference in 15–25% between smooth path length from our algorithm and the results from [24,25] is not too significant because we mainly focus on constructing realistic randomized paths for BOTs. We also create OWL reasoner to choose whether we have to use smoothing or piece-wise linear path to answer query for current combat situation like it is shown at the Figure 3. When implementing such an algorithm in 3D first-

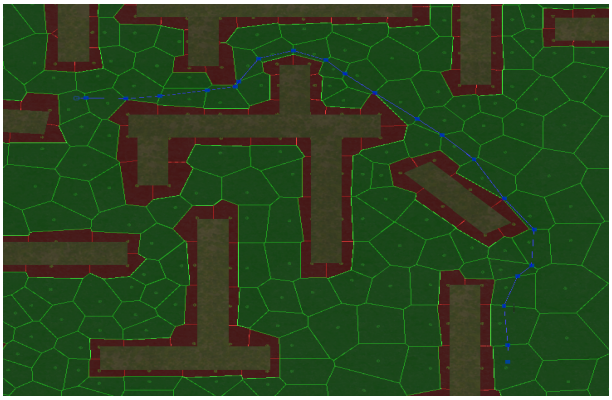


Fig. 3. Path finding

person shooter, we obtained more realistic motion behaviours than the minimized CBR-based path, while saving the property of the path to be suboptimal.

4 Conclusion

We started our research with stating the thesis that modeling human behavior in video games could be presented as game artificial intelligence problem that should be implemented by algorithms with human patterns of discrete optimisation. We used obvious assumptions on neuron to be useful in terms of short memory usage to balance neural network. Smoothing path trajectory was obtained through a native obstacle avoidance model supporting enough degree of randomness. Path finding algorithm with reduced time computations was obtained from discrete choice model used by human players (firstly implemented as the first and the simplest game AI for ghost-BOT in computer game PACKMAN). We hope that idea to use the simplest optimisation criteria from the Occam's razor to model human behavior in video games is a key to understanding correct reasoning of models containing information about evolution of decision-making models while increasing its game experience.

References

1. Wang, D., Tan, A.H.: Creating autonomous adaptive agents in a real-time first-person shooter computer game. IEEE Transactions on Computational Intelligence and AI in Games **7**(2) (June 2015) 123–138
2. Hingston, P.: A turing test for computer game bots. IEEE Transactions on Computational Intelligence and AI in Games **1**(3) (Sept 2009) 169–186
3. Karpov, I.V., Schrum, J., Miikkulainen, R. In: Believable Bot Navigation via Playback of Human Traces. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 151–170
4. van Hoorn, N., Togelius, J., Schmidhuber, J.: Hierarchical controller learning in a first-person shooter. In: 2009 IEEE Symposium on Computational Intelligence and Games. (Sept 2009) 294–301
5. da Silva, F.S.C., Vasconcelos, W.W. In: Rule Schemata for Game Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg (2006) 451–461
6. Cole, N., Louis, S.J., Miles, C.: Using a genetic algorithm to tune first-person shooter bots. In: Evolutionary Computation, 2004. CEC2004. Congress on. Volume 1. (June 2004) 139–145 Vol.1
7. Smith, M., Lee-Urban, S., Muñoz-Avila, H.: RETALIATE: learning winning policies in first-person shooter games. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada, AAAI Press (2007) 1801–1806
8. Stanley, K.O., Bryant, B.D., Miikkulainen, R.: Real-time neuroevolution in the nero video game. IEEE Transactions on Evolutionary Computation **9**(6) (Dec 2005) 653–668
9. Veldhuis, M.O.: Artificial intelligence techniques used in first-person shooter and real-time strategy games. human media interaction seminar 2010/2011: Designing entertainment interaction (2011)
10. McPartland, M., Gallagher, M.: Reinforcement learning in first person shooter games. IEEE Transactions on Computational Intelligence and AI in Games **3**(1) (March 2011) 43–56
11. McPartland, M., Gallagher, M.: Interactively training first person shooter bots. In: 2012 IEEE Conference on Computational Intelligence and Games (CIG). (Sept 2012) 132–138
12. McPartland, M., Gallagher, M. In: Game Designers Training First Person Shooter Bots. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 397–408
13. Tan, A.H.: Falcon: a fusion architecture for learning, cognition, and navigation. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. Volume 4. (July 2004) 3297–3302 vol.4
14. Bhattacharya, P., Gavrilova, Marina L.: Voronoi diagram in optimal path planning. In: 4th IEEE International Symposium on Voronoi Diagrams in Science and Engineering. (2007) 38–47
15. Choi, J.w., Curry, Renwick E., Elkaim, Gabriel H.: Obstacle avoiding real-time trajectory generation and control of omnidirectional vehicles. In: American Control Conference. (2009)
16. Gulati, S., Kuipers, B.: High performance control for graceful motion of an intelligent wheelchair. In: IEEE International Conference on Robotics and Automation. (2008) 3932–3938
17. Guechi, E.H., Lauber, J., Dambrine, M.: On-line moving-obstacle avoidance using piecewise bezier curves with unknown obstacle trajectory. In: 16th Mediterranean Conference on Control and Automation. (2008) 505–510

18. Nagatani, K., Iwai, Y., Tanaka, Y.: Sensor based navigation for car-like mobile robots using generalized voronoi graph. In: IEEE International Conference on Intelligent Robots and Systems. (2001) 1017–1022
19. Mohammadi, S., Hazar, N.: A voronoi-based reactive approach for mobile robot navigation. Advances in Computer Science and Engineering **6** (2009) 901–904
20. Eren, H., Fung, C.C., Evans, J.: Implementation of the spline method for mobile robot path control. In: 16th IEEE Instrumentation and Measurement Technology Conference. Volume 2. (1999) 739–744
21. Magid, E., Keren, D., Rivlin, E., Yavneh, I.: Spline-based robot navigation. In: International Conference on Intelilgent Robots and Systems. (2006) 2296–2301
22. Hwang, J.H., Arkin, R.C., Kwon, D.S.: Mobile robots at your fingertip: Bezier curve on-line trajectory generation for supervisory control. In: IEEE International Conference on Intelligent Robots and Systems. Volume 2. (2003) 1444–1449
23. Škrjanc, I., Klančar, G.: Cooperative collision avoidance between multiple robots based on bezier curves. In: 29th International Conference on Information Technology Interfaces. (2007) 451–456
24. Ho, Y.J., Liu, J.S.: Smoothing voronoi-based obstacle-avoiding path by length-minimizing composite bezier curve. In: International Conference on Service and Interactive Robotics. (2009)
25. Ho, Y.J., Liu, J. S.: Collision-free curvature-bounded smooth path planning using composite bezier curve based on voronoi diagram. In: IEEE International Symposium on Computational Intelligence in Robotics and Automation. (2009) 463–468
26. Koenig, S., Sun, X., Uras, T., Yeoh, W.: Incremental ARA*: An incremental anytime search algorithm for moving-target search. In: Proceedings of the Twenty-Second International Conference on Automated Planning and Scheduling. (2012)
27. Makarov, I., Tokmakov, M., Tokmakova, L.: Imitation of human behavior in 3D-shooter game. In Khachay, M.Y., Konstantinova, N., Panchenko, A., Delhibabu, R., Spirin, N., Labunets, V.G., eds.: 4th International Conference on Analysis of Images, Social Networks and Texts. Volume 1452 of CEUR Workshop Proceedings., CEUR-WS.org (2015) 64–77
28. Pohl, I.: First results on the effect of error in heuristic search. Machine Learning **5** (1970) 219–236
29. Likhachev, M., Gordon, G., Thrun, S.: ARA*: Anytime A* search with provable bounds on sub-optimality. In Thrun, S., Saul, L., Schölkopf, B., eds.: Proceedings of Conference on Neural Information Processing Systems (NIPS), MIT Press (2003)
30. Sun, X., Yeoh, W., Uras, T., Koenig, S.: Incremental ara*: An incremental anytime search algorithm for moving-target search. In: ICAPS. (2012)
31. Makarov, I., Polyakov, P.: Smoothing voronoi-based path with minimized length and visibility using composite bezier curves. In Khachay, M.Y., Vorontsov, K., Loukachevitch, N., Panchenko, A., Ignatov, D., Nikolenko, S., Savchenko, A., eds.: 5th International Conference on Analysis of Images, Social Networks and Texts. CEUR Workshop Proceedings, CEUR-WS.org, In Print (2016)

Studying Family Formation Trajectories' Deinstitutionalization in Russia Using Sequence Analysis

Ekaterina S. Mitrofanova[1], Alyona V. Artamonova[1]

[1]National Research University Higher School of Economics

mitrofanovy@yandex.ru, alyona89152694371@yandex.ru

Abstract. This study focuses on changing family formation trajectories in the Russian Federation. In European countries, pathways to family ceased being stable several decades ago, while in Russia – as in any post-socialist country – such features of life course deinstitutionalization as postponement of marriage, rising cohabitation, and reordering of events were revealed only in the 1990s and explained from the perspective of the Second Demographic Transition (SDT). Our aim is to demonstrate how family formation trajectories of men and women from different generations were transforming with the incorporation of data mining. The three-wave panel data of the Russian part of the "Generations and Gender Survey" (2004, 2007, 2011; N=5321) and the retrospective data of the survey "Person, Family, Society" (2013; N=4477) are used for achieving this aim. Sequence Analysis shows that generations born after 1970 started to exhibit de-standardized family formation trajectories. As the proportion of Russians who raise children in cohabitation or while single rises, such models of behavior become more widely accepted and practiced in contemporary Russia. Women experience more events in the family trajectory, take steps toward family formation earlier, and stay alone with children more often than men. Matrimonial and reproductive behavior has become diverse, proving that Russia fully exhibits the SDT.

Keywords: family formation trajectories, matrimonial and reproductive behavior, Sequence Analysis

1 Introduction

People's family formation trajectories have considerably changed in recent decades. In many European countries, marital union with children has been the only acceptable method of family organization for a long time. Since the 1990s, a couple may be formed not only through marriage but also through cohabitation, people may postpone the birth of children or remain childfree, and a union may not be dissolved solely through divorce but also through separation; because of new freedom of thinking and behaving and people's orientation to individual self-development, this is one of the distinctive features of modern society [1].

The theorists of the Second Demographic Transition approach, headed by pioneers Lesthaeghe and Van de Kaa, explain the transformation in demographic behavior as the result of the broad and long-term changes in the norms and values that many countries witnessed between the mid-1960s and the end of the 1980s [2]. Mayer [3] claims that, since the 1960s, societies have embraced so-called "hedonistic individualism", which includes alternative lifestyles, emphasizing individual fulfillment and self-expression rather than sacrifices to the family, traditional values and altruistic orientations regarding children and the collective good. Instead of following the tradition of marriage, young people realize their own personal goals of self-expression and enjoyment [4].

All the SDT changes in paths to family formation started in Western European countries and followed the model of the European type of marriage prevailing west of Hajnal's line. Eastern European Russia displays demographic outcomes of the SDT in atypical fashion. Growing cohabitation rates alongside declining marital rates emerged in the Soviet Union in the middle of the 1980s, years before the fall of socialism [5]. Zakharov [6] revealed that Russians born after 1970s already started to demonstrate all features of SDT (e.g. the formation of partnerships outside marriage, the rise in non-marital childbearing, and the postponement of marriage). Mills showed that new pathways to family in Russia, contrary to SDT theory, are prevailing among less-educated people, reminiscent of a 'pattern of disadvantage' concept. It makes Russia look more like the United States than Europe with regards to life course deinstitutionalization.

Taking this complexity of matrimonial and reproductive behavior into consideration, we decided to trace the family formation trajectories' deinstitutionalization in Russia based on gender-generational differences using Sequence Analysis.

2 Hypotheses

The standardized trajectory of "Soviet" generation Russia starts from singlehood and includes universal marriage with at least one child. The proportions of those single with children and those secondly married were minimal. From the middle of the 1980s until the collapse of the Soviet Union, Russians turned to Western European countries' family lifestyles [7]. The average ages of marriage have been rising since the early 1990s. In 1993, the ages for men and women were 23.9 and 21.8 years, respectively. In 1999 and 2004 they consisted of ages 25.0 and 26.1 for men and ages 23.1 and 23.3 for women [1].

According to Mills and her co-authors, there is a high proportion of single parents in Russia (even higher than in some Western European countries),

which may be caused by a high divorce rate and particularly high adult male mortality, which is largely due to alcohol-related deaths [7].

Taking into consideration the information above, we decided to verify two groups of hypotheses.

Group 1. Gender:

- Women take steps to family earlier than men;
- Women stay alone with children more often than men;
- Women experience more family formation events than men;

Group 2. Generations:

- De-standardization of family formation trajectories was demonstrated first by representatives of the first "Modern" generation (1970-79 birth cohort);
- "Modern" generations experience more varied matrimonial and reproductive events than the representatives of "Soviet" generations.
- To test these hypotheses, we decided to apply Sequence Analysis, which requires longitudinal or retrospective data.

3 Data

We used the panel data of the Russian part of the Generations and Gender Survey (GGS-panel: 2004, 2007 and 2011) and retrospective data of the "Person, Family, Society" survey (PFS: 2013). We choose these surveys because their designs apply the Life Course approach, which tends to understand different types of demographic events as a chain of interconnected processes. The questions about life course events were asked in a very accurate and detailed way. Most of the dates contain not only years but also months of starts and ends of events. We should mention that the questions about children were asked so as to show our interest in the biological children of respondents.

To work correctly with sequences, it was necessary to constrain the ages of events. 15 years as the lower age point was chosen because it is the beginning of possible reproductive behavior. Obviously, there were respondents who enter into their first union or have their first child before reaching this age but such atypical cases are outside the scope of our study. In the samples of used datasets there are respondents who, at the time of the survey, were 25 years old (GGS-2011, third wave) and even 18 years old (PFS-2013). Marriages in Russia were early and universal for a long time, and almost all representatives of the Soviet generations started their unions by the age of 25. We supposed that younger generations demonstrate a delay in the start of their first unions in comparison with the Soviet generations. That is why, if we want to trace the change in the age of the first union formation, we should analyze a wide range of ages. However, the representatives of the older generations have lived longer lives than the youth, and some unique cases of the first unions at ages over 40 years can shift the average age. Moreover, it is not correct to compare the full

matrimonial biographies of people who reached the age of final celibacy and people who only started their union histories. Taking into account all these arguments, we decided to impose a limit on the age of matrimonial and reproductive events occurring. After considering several options, we limited the age of entry into first union at 35 years, no matter whether not all respondents finished the transition to family life by the age of 35.

The final GGS and PFS datasets contain 5321 and 4477 cases, respectively.

In order to analyze the generational aspect of matrimonial behavior, we divided our samples into two key groups: the "Soviet" generations (1930-39, 1940-49, 1950-59, 1960-66 in GGS and 1960-69 in PFS), who socialized before the collapse of the Soviet Union, and the "Modern" generation (1970-79, 1980-86 in GGS and 1970-79, 1980-89, 1990-95 in PFS), who socialized after it [8]. The proportions of men and women in different generations of GGS and PFS can be found in the Table 1.

Table 1. Proportions of men and women in Russian generations

Generation	Gender	GGS		PFS	
		Absolute numbers	Percentages	Absolute numbers	Percentages
1930-1939	Men	192	25%	-	-
	Women	585	75%	-	-
1940-1949	Men	214	28%	-	-
	Women	552	72%	-	-
1950-1959	Men	387	30%	-	-
	Women	923	70%	-	-
1960-1969	Men	423	36%	-	-
	Women	761	64%	-	-
1970-1979	Men	325	36%	798	48%
	Women	585	64%	855	52%
1980-1986(89)	Men	158	42%	939	49%
	Women	216	58%	988	51%
1990-1995	Men	-	-	473	53%
	Women	-	-	424	47%

4 Methodology

In recent years, there has been a strongly growing interest in the study of life course trajectories to describe life trajectories, to classify individuals according to them by using the Sequence Analysis (SA) method [9, 10, 11, 12]. SA is

based on data mining approaches, namely on the measures of dissimilarity or distance between individual trajectories. It is entirely non-parametric.

The majority of papers devoted to SA highlights both certain socio-demographic phenomena and the methodological development of the method. There are some papers about the deinstitutionalization of the life course [13], starting events are postponed [14, 15, 16], women are more proactive in social life and they are postponing maternity [17], and the number of social roles are growing for both sexes [18]. Matrimonial trajectories are becoming more diverse and less predictable [19, 20, 21].

The development of methods goes in two directions: development with the sources of mathematical statistics and Data Mining [22, 23, 24]. The researchers not only discover typical sequences for different classes, but also cluster them [25, 26], evaluate their resemblance [27], creat classifiers [28], define the transaction costs [16], and build the decision trees [14].

The representation of life course trajectories in SA is similar to the code of DNA molecules [9]. It focuses on a time window with chosen ages of start and finish, inside of which studied events (e.g. entry to first and second cohabitations (P1 and P2), marriages (M1 and M2), and birth of first and second child (C1 and C2)) can occur. As was explained above, in our research, the first point of the time window is 15 years (when the majority of Russians do not have any matrimonial (i.e. single – S) or reproductive (i.e. childless – C0) events) and the last point is 35 years. We deal with so-called 'non-recurrent sequences', where an event may not repeat at all.

As individual life courses can be represented as a sequence of events, we are able to code every event with a letter and build the "word" that describes the state of an individual at every point of a chosen time window. Table 2 shows all possible states of partnership and fertility trajectory.

Table 2. Alphabet of partnership and fertility states

Code	State	Code	State
SC0	Single, no children	M1C0	First marriage, no children
SC1	Single, 1 child	M1C1	First marriage, 1 child
SC2	Single, 2 children	M1C2	First marriage, 2 children
P1C0	First cohabitation, no children	M2C0	Second marriage, no children
P1C1	First cohabitation, 1 child	M2C1	Second marriage, 1 child
P1C2	First cohabitation, 2 children	M2C2	Second marriage, 2 children
P2C0	Second cohabitation, no children		
P2C1	Second cohabitation, 1 child		
P2C2	Second cohabitation, 2 children		

In our study, we used TraMineR (R-package) to mine and visualize sequences of matrimonial and reproductive events [29]. The first tool we used

was chronograms. A chronogram is the representation of all the sequences of a group at each age. It is a summary of individual trajectories. We used the graphs representing the entropy – the measure of disorder of sequences – at each time period. We calculated the mean time spent in statuses which, that is, how long every member of each group, on average, was in each status. And finally, we calculated the number of family formation events, which mean how many events each member of each group experienced in his or her life.

5 Empirical Results

We first show the results of the first group of tested hypotheses and then move to the second group.

Gender. In order to prove that women take steps to family earlier than men, we compared distributions of partnerships and fertility statuses, all sequences, and entropy by gender.

On the horizontal axis of the plots, there are the ages of the respondents between 15 and 35 years. The youngest respondents have not yet reached the upper age limit: this is why we had to work with censored data (indicated in gray). On the vertical axes of the first and third plots, the proportions of individuals belonging to each state at a given age are shown. On the vertical axis of the second plot there are respondents, so we can observe individual family formation trajectories of men and women.

The plots reveal that either in GGS or in PFS, men start to experience family formation events at the age of 17, while women do it earlier. In fact, 80% of women have at least one event at 23, while among men this proportion is reached at 26.

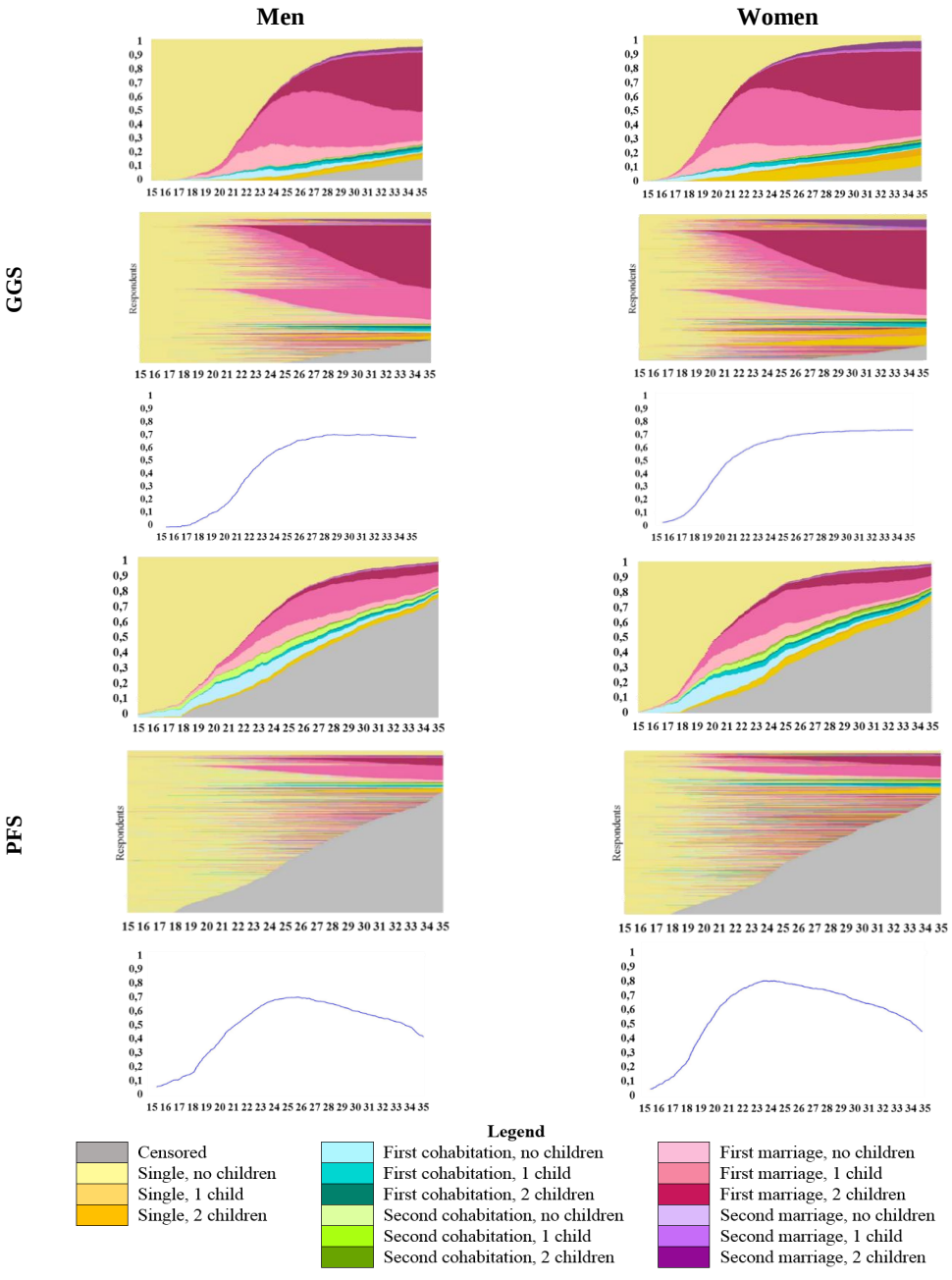


Fig. 1. Family formation trajectories of Russians

One more evidence for our hypothesis is the mean time spent in singlehood and without children (Figure 2). Men spend about 100 months after 15 years in this status, while women spend only about 80 months.

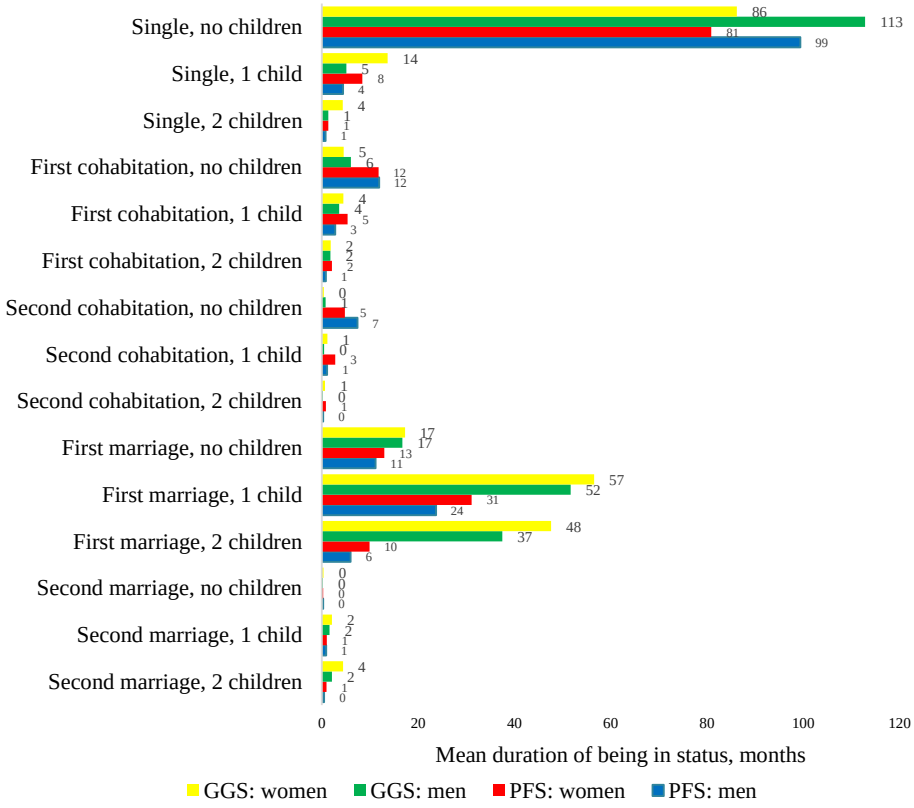
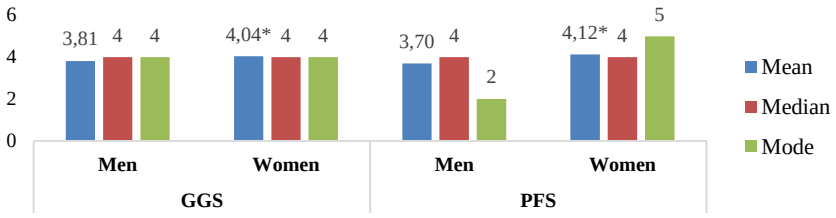


Fig. 2. Mean time spent in status

In order to prove that women stay alone with children more often than men do, we should look at Figures 1 and 2. The distribution of partnerships and fertility statuses plot demonstrates that the proportion of single women with children at the age of 35 (25%) is more than twice the proportion of such men (10%). Mean time spent in these two statuses is higher for women (about 14 months) than for men (about 5 months) as well.

In order to prove that women experience more family formation events than men, we compare mean, median, and mode number of family formation events for men and women. The mean demonstrates that women have significantly more events than men but, according to two other figures, the numbers are the same.



*The difference is statistically significant (p<0,001)

Fig. 3. Number of family formation events by gender

Generations. In order to prove that the de-standardization of family formation trajectories was demonstrated first by representatives of first "Modern" generation (1970-79 birth cohort) we compared the entropy of different generations (Figure 4) and the distribution of partnerships and fertility statuses by gender and generation (Appendix 1).

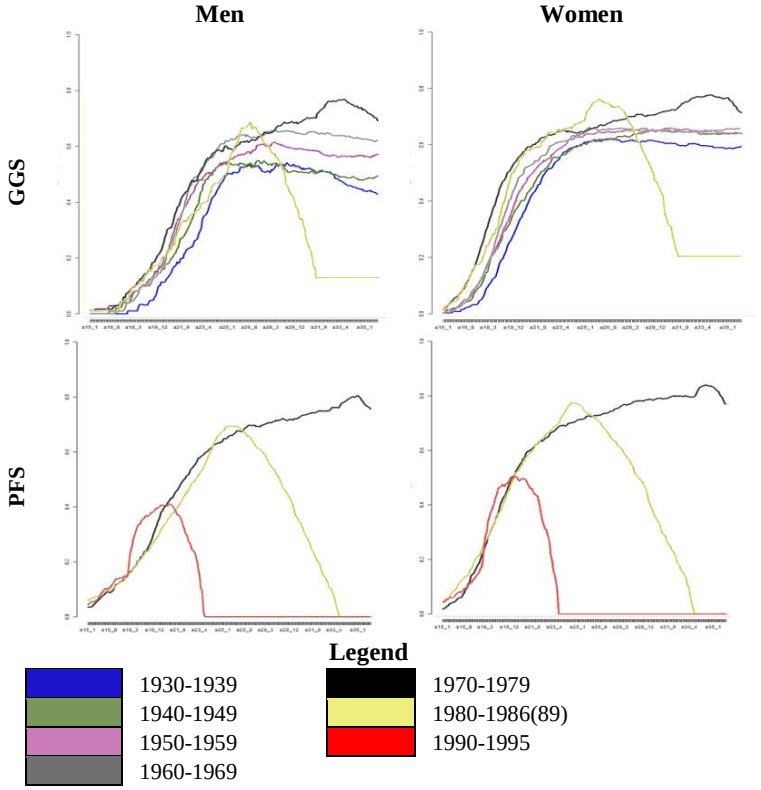


Fig. 4. Entropy by generations

It is apparent from the Figure in Appendix 1 that the proportion of married people with at least one child decreased while the proportions of cohabited (blue pallet) and single people with children (yellow pallet) have increased dramatically. The visible changes started with the generations born after 1970.

In order to prove that "Modern" generations experience more varied matrimonial and reproductive events than the representatives of "Soviet" generations, we counted mean, median, and mode number of family formation events for different generations (Table 3).

Table 3. Number of family formation events by generation and gender

	Men			**Women**		
	Mean	**Median**	**Mode**	**Mean**	**Median**	**Mode**
GGS						
1930-1939	3.70	4	4	3.68	4	4
1940-1949	3.82	4	4	3.84	4	4
1950-1959	3.81	4	4	4.07	4	4
1960-1969	3.89	4	4	4.07	4	4
1970-1979	4.10	4	4	4.54	4	4
1980-1986	3.15	3	2	3.95	4	4
PFS						
1970-1979	4.12	4	4	4.42	4	4
1980-1989	3.94	4	2	4.44	5	5
1990-1995	2.51	2	2	2.79	2	2

The figures demonstrate that the number of events for men and women in generations do not differ.

6 Conclusions

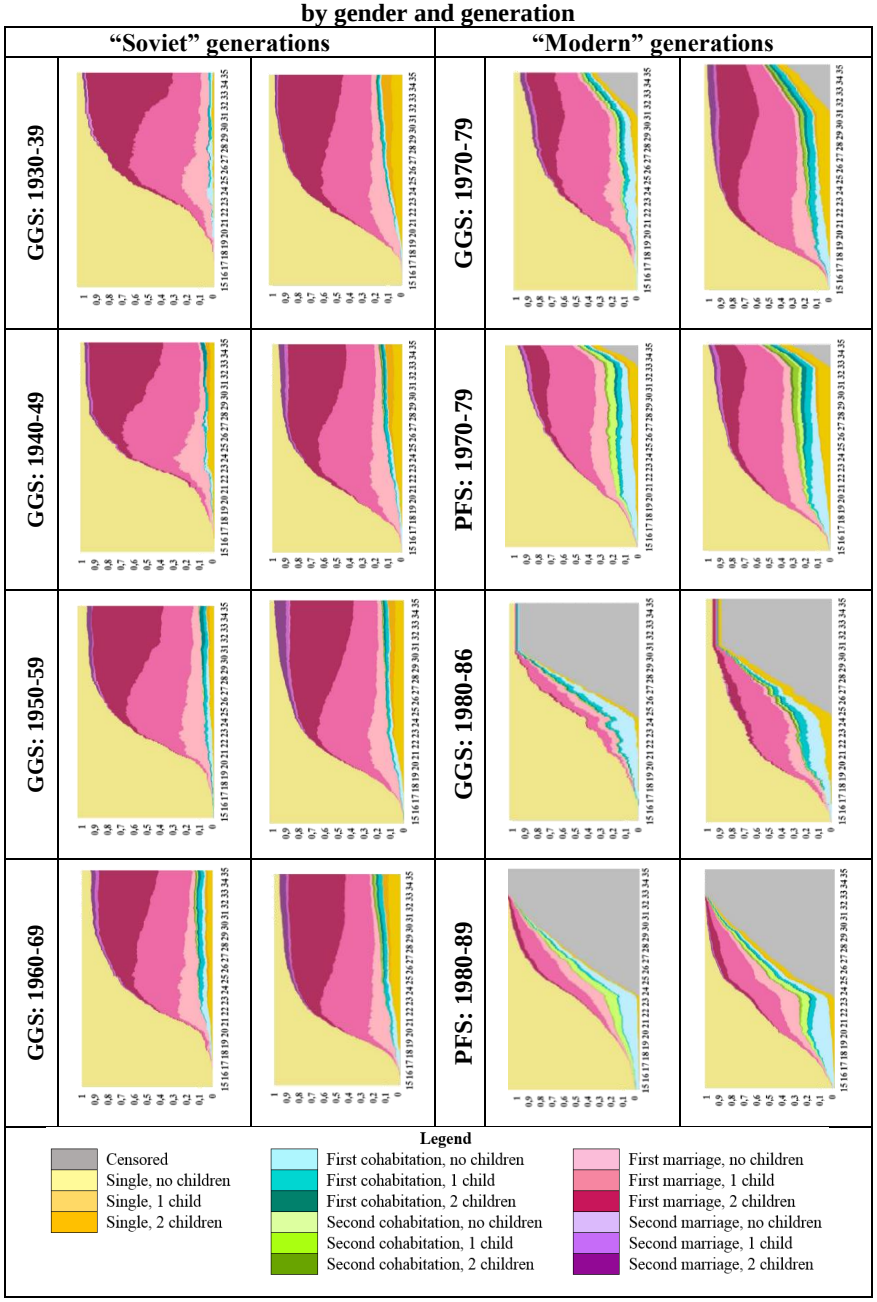
In this paper we revealed several points about family formation trajectories of Russians:

- women start to entry into first matrimonial events earlier than men;
- women stay alone with children more often than men do;
- women and men experience equal number of family formation events;
- generations born after 1970 started to exhibit de-standardized family formation trajectories;
- the number of events for men and women in different generations remains stable.

Matrimonial and reproductive behavior is becoming diverse, proving that Russia fully displays Second Demographic Transition.

Acknowledgements. The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 (grant № 16-05-0011 "Development and testing of demographic sequence analysis and mining techniques") and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. The authors also want to thank Thomas H. Espy for inestimable help in the preparation of this paper.

Appendix 1. Distribution of partnerships and fertility statuses by gender and generation



References

1. Avdeev, A., Monnier, A.: Marriage in Russia: A Complex Phenomenon Poorly Understood. Popul. Engl. Sel. 12, pp. 7–49 (2000)
2. Kaa, D.J. van de, Lesthaeghe, R.: Two Demographic Transitions? Population: Growth and Decline, pp. 9–24 (1986)
3. Mayer, K.U.: Whose Lives? How History, Societies, and Institutions Define and Shape Life Courses. Research in Human Development 1 (3), pp. 161–187 (2004)
4. Gerber, T.: Changing family formation behavior in post-socialist countries: Similarities, divergences, and explanations in comparative perspective. Draft for presentation at the "1989: Twenty Years After Conference" (2009)
5. Gerber, T.P., Berman, D.: Entry to Marriage and Cohabitation in Russia, 1985–2000: Trends, Correlates, and Implications for the Second Demographic Transition. Eur. J. Popul., pp. 26: 3–31 (2010)
6. Zakharov, S.: Russian Federation: From the first to second demographic transition. Demographic Research, 19, pp. 907–972 (2008)
7. Mills, M., Lesnard, L., Potarca, G.: Family Formation Trajectories in Romania, the Russian Federation and France: Towards the Second Demographic Transition?. European Journal of Population, 29, pp. 69-101 (2013)
8. Levada, Y.: Generations of XX Century: Opportunities of Studies. In Y. Levada, T. Shanin. M. Fathers and Children: Analysis of Contemporary Russian Generations. Moscow, pp. 39-60 (2005)
9. Billari, F.C.: Sequence Analysis in Demographic Research. Special Issue on Longitudinal Methodology. Canadian Studies in Population Vol. 28(2), pp. 439-458 (2001)
10. Ignatov, D., Mitrofanova, E., Muratova, A., Gizdatullin, D.: Pattern Mining and Machine Learning for Demographic Sequences. KESW 2015, pp. 225-239 (2015)
11. Blockeel, H., Fürnkranz, J., Prskawetz, A., Billari, F.C.: Detecting Temporal Change in Event Sequences: An Application to Demographic Data. PKDD, pp. 29-41 (2001)
12. Low-Kam C., Raïssi C., Kaytoue M., Pei J.: Mining Statistically Significant Sequential Patterns. ICDM, pp. 488-497 (2013)
13. Worts, D. et al.: Individualization, Opportunity and Jeopardy in American Women's Work and Family Lives: A Multi-State Sequence Analysis. Advances in Life Course Research, 18, pp. 296–318 (2013)
14. Billari, F. C., Fürnkranz, J. and Prskawetz, A.: Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. European Journal of Population, 22, pp. 37–65 (2006)
15. Mouw, T.: Sequences of Early Adult Transitions: How Variable Are They, and Does It Matter. Frontier of Adulthood: Theory, Research, and Public Policy, pp.256–91 (2005)
16. Billari, F.C., Rosina, A.: Italian "Latest-Late" Transition to Adulthood: An Exploration of Its Consequences on Fertility. Genus (2004)
17. Aassve, A., Billari, F.C. and Piccarreta, R.: Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. European Journal of Population, 23, pp. 369–88 (2007)
18. Jackson, P.B. and Berkowitz, A.: The Structure of the Life Course: Gender and Racioethnic Variation in the Occurrence and Sequencing of Role Transitions. Advances in Life Course Research, 9, pp. 55–90 (2005)
19. Elzinga, C.H. and Liefbroer, A.C.: De-Standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis // European Journal of Population/Revue Européenne de Démographie, 23, pp. 225–50 (2007)
20. Oris, M. and Ritschard, G.: Sequence Analysis and Transition to Adulthood: An Exploration of the Access to Reproduction in Nineteenth-Century East Belgium. Advances in Sequence

Analysis: Theory, Method, Applications. Springer International Publishing, pp. 151–167 (2014)
21. Piccarreta, R., Billari, F.C.: Clustering work and family trajectories by using a divisive algorithm. J. R. Stat. Soc. Ser. A Stat. Soc. T. 170. № 4, pp. 1061–1078 (2007)
22. Dong, G., Pei, J.: Sequence Data Mining. New York: Springer (2007)
23. Zaki, M.J. and Meira, W. Jr.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, New York, NY, USA (2014)
24. Han, J., Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2006)
25. Elzinga, C.H., Studer, M.: Spell sequences, state proximities and distance metrics. Sociol. Methods Res. T. 44. № 1 (2015)
26. Elzinga, C.H.: Distance, Similarity and Sequence Comparison. Advances in Sequence Analysis: Theory, Method, Applications. Springer, pp. 51–74 (2014)
27. Elzinga, C.H.: Sequence Similarity: A Nonaligning Technique. Sociol. Methods Res. T. 32. № 1, pp. 3–29 (2003)
28. Barban, N., Billari, F.C.: Classifying life course trajectories: A comparison of latent class and sequence analysis. J. R. Stat. Soc. T. 61. № 5, pp. 765–784 (2012)
29. Gabadinho, A., Ritschard, G., Muller, N. S., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software, 40(4), pp. 1-37 (2011) URL: http://www.jstatsoft.org/v40/i04/

QAIDS Model Based on Russian Pseudo-Panel Data: Impact of 1998 and 2008 Crises [1,2]

Maria D. Ermolova[a], Henry I. Penikas[b]

[a] International Laboratory of Decision Choice and Analysis, National Research University Higher School of Economics, Moscow, Russia.
mermolova@hse.ru

[b] Department of Applied Economics, National Research University Higher School of Economics, Moscow, Russia.
penikas@hse.ru

Abstract: The aim of this work is to compare shifts in the consumer behaviour of Russian households since the mid-nineties till nowadays. The research considers the consumer behaviour of the Russians over almost the maximum possible available data RLMS period, focusing on the crisis years. Special attention is paid to analysis of the effects of crises in 1998 and 2008. To reveal effects as shifts in consumer behaviour in the aftermath of two crises panel data analysis is used to estimate QAIDS model. Due to the complete sample attrition observed in RLMS dataset since 1994, pseudo-panel approach is used.

Keywords: QAIDS, RLMS, pseudo-panel, consumer behaviour, crisis

JEL codes: D12, E21

1 Introduction

Economic recessions change consumer behaviour through consumers' expectations that can be also formed by economic policy, economy structure or distribution of households. Structural or temporal shifts determine the subsequent economic policy, whose efficiency, in turn, also evaluated by the change in the welfare of different households. Therefore, analysis of shifts in consumer behaviour needs to be determined accurately. However, any research about life quality is highly dependent on the data used. Data may not always be suitable for study for the following reasons: selection bias (no poorest or richest people), distrust of statistical authorities (respondents often refuse to answer questions or deliberately distort the data), the lack

[1] The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics.

[2] The authors are grateful to Rustam Zakirov for conducting initial calculations and to Sergey Vinjkov for research assistance.

of representativeness relative to the general population, and a depletion of the sample over a long period of time.

The aim of this work is to compare shifts in the consumer behaviour of Russian households since the mid-nineties till 2011. In this paper, the data from the survey "Russian Longitudinal Monitoring Survey HSE" (hereafter RLMS) is used.[3] The paper shows that descriptive statistics or model using panel data do not provide enough information about whether 1998 or 2008 crises leads to structural or temporal effect on consumer behaviour. Using pseudo-panels it was found that the effect of 1998 crisis was stronger in consumer behaviour than the crisis of 2008.

The paper is organized in the following way. Section 1 is introduction. Section 2 discusses the works based on QAIDS model briefly. In Section 3 the theoretical demand model is described. Section 4 explains data processing. Model estimation is given in section 5. Section 6 provides the conclusion about the effects of economic shocks and the evolution of consumer behaviour in Russia.

2 Literature review

As far as the authors know, there are no works devoted to the study of consumer behaviour of Russian households for such a long time interval (due to the problem of sample depletion).[4]

There are articles covering a relatively short period of time 2000 – 2005 [Penikas, 2008] or focusing on specific aspects of consumer behaviour (differentiation of real incomes of the population on the basis of consumer choice) [Matytsin et al., 2012]. In foreign literature the number of publications on consumer behaviour is much higher, because it is closely associated with the doctrine of welfare of the population.

The article [Deaton et al., 1980] firstly provides a theoretical description of the Almost Ideal Demand Model (AIDS). AIDS has proven its viability and vitality using the British data from 1954 to 1974. [Gardes et al., 2005] concludes using AIDS model that the estimates obtained using the pseudo-panel approach is less biased compared to the cross-section data the usage of cross-section data.

[Tovar et al., 2012] pseudo-panel estimation takes into account the time dependence of the different cohorts, because the same households may be in different households over time.

3 Quadratic almost ideal demand model (theoretical model)

Dynamics of consumer behaviour by consumption group and by various consumption directions is considered from the perspective of analysis of coefficients of income

---

[3] Source: "Russia Longitudinal Monitoring survey, RLMS-HSE", conducted by the National Research University Higher School of Economics and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS. URL: `http://www.cpc.unc.edu/projects/rlms-hse`

[4] A gradual decrease in the number of observations.

elasticity derived from QAIDS (Quadratic Almost Ideal Demand System) (Banks et al., 1997):

$$w_{iht} = \alpha_i + \sum_{j=1}^{J} \gamma_{ij} \ln p_j + \beta_i \ln(x_{ht}/P_t) + c_i \left(\ln(x_{ht}/P_t)\right)^2 / b(p) + Z_{ht} d_i + u_{iht} \qquad (1)$$

Where w_{iht} – share of household's expenses h for sets of goods $i = 1,2,3$ in the moment t, P_t – Stone Price Index ($ln\, P = \sum w_k \, ln\, p_k$), x_{ht} – household's income (the costs are usually used as an equivalent because respondents in surveys tend to understate their own revenues), Z_{ht} – the matrix of socio-economic characteristics, $b(p) = \prod_k p_k^{\beta_k}$ price index, ensuring the integrability of the entire system, u_{iht} includes both individual effect and random error. To estimate the elasticities it is necessary to take derivatives of the above equation $ln\, x$ and $ln\, p_j$:

$$u_i = \partial w_i / \partial \ln x = \beta_i + 2c_i \left[\ln(x_{ht}/P_t) - \ln(b(p))\right] \qquad (2)$$

$$u_{ij} = \partial w_i / \partial \ln p_j = \gamma_{ij} - u_i \left(\alpha_j + \sum_{k=1}^{K} \gamma_{jk} \ln p_k \right) - c_i \beta_j \left(\ln(x_{ht}/P_t)\right)^2 / b(p) \qquad (3)$$

The income elasticity for each household will be defined as $e_i = \frac{u_i}{w_i} + 1$ (4), and compensated price elasticity for good j as $e_{ij}^c = e_{ij}^u - w_j$ (5), where $e_{ij}^u = \frac{u_{ij}}{w_i} - \delta_{ij}$ is uncompensated elasticity (δ_{ij} is the Kronecker symbol, that is equal to 1 when $i = j$ and 0 in all other cases).

4 Data

This work is based on the second phase RLMS data, covering the period of 1994–2011. The RLMS surveys constitute an unbalanced panel, i.e. a household can vary from year to year in the survey (sample attrition). Only 35 % of the household (1 366 / 3 975 of observations) that took participation in the survey of 1994 remain in the polls by 2011.[5] 77% of households in the survey of 2008 are presented in the sample of 2011. The observations are placed in the same income group after data processing. It is the prevalent challenge in studying the effects of crises on different groups of households. In connection with these problems, the paper proposes to use pseudo-panels, namely to generate quasi-households on the basis of real data. Further, data processing will be described.

[5] Only 7.5% of households remain after data processing.

4.1 Outliers

If the difference between a one-year distribution of costs and incomes is more than 45 percentiles for a household, then the observation is recognized as atypical and removed. The threshold of 45 percentiles has been chosen in such a way to eliminate the problem of underestimating revenues, but at the same time to keep most of the sample. For example, 30 percentiles are not applicable because the sufficient sample part (16.3% of outliers) is removed in comparison with 45 percentiles (6.1 % of outliers).

4.2 OECD equivalence scale

The welfare of individuals of the household can be measured either "per capita" or "per consumption unit". The first approach is not applicable due to economies of scale. Two people do not consume two times more goods, because they have both public (car, refrigerator) and private goods (food) within a family. Therefore, it is necessary to implement the concept of "per consumption unit" that will depend on public-to-private goods ratio in the household.

The public-to-private goods ratio varies depending on time and country. Time is introduced through the function that depends on the age of household members. The function is a linear combination of the number of family members belonging to different groups. In our research the Oxford modified equivalence scale [Lubrano, 2010] will be used, as it is the most popular in research on consumer behaviour (see [Banks et al., 1997] and [Penikas, 2008]).

Table 1. Modified equivalence scale

Family member	**Coefficient**
The head of the household	1.0
All others, age > 14	0.5
All others, age < 14	0.3

Different measurement scales can lead to different estimates of elasticities by income.

Table 2. The effects of equivalence scales

	Equivalence scale			
Composition of household	**per capita**	**Oxford scale**	**Modified Oxford scale**	**Household income**
1 adult	1	1.0	1.0	1
2 adults	2	1.7	1.5	1
2 adults, 1 child	3	2.2	1.8	1
2 adults, 2 children	4	2.7	2.1	1
2 adults, 3 children	5	3.2	2.4	1
Elasticity	1	0.73	0.53	0

4.3 Welfare

Fig. 1 makes clearly visible the recession of 1998 and 2008 (effect is delayed by a year in 2009) in terms of real income and expenditures. The dynamics of real income for the RLMS sample is in line with trends in real income, represented by Federal State Statistics Service, which is in favour of the representativeness of the study sample.

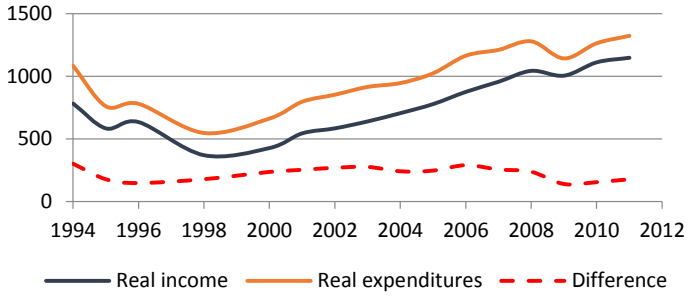


Fig. 1. Real income and expenditures (in 1994 prices)

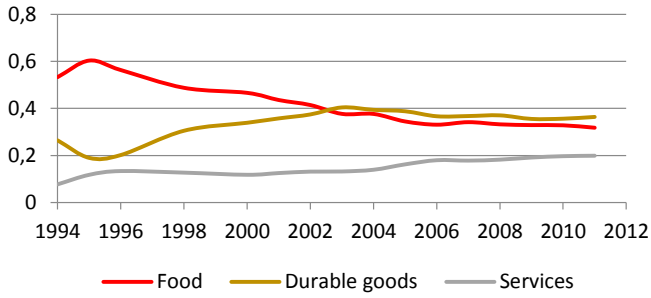


Fig. 2. Weights of expenditures by the groups of goods

The effects of the crises do not appear explicitly if weights of expenditures by good classes are examined. **Fig. 2** shows that the share of expenditures on food decreased over time, which is consistent with the growth in real incomes because the proportion of expenditure on food is a common first approximation of living standards. Over the beginning of the two thousandth's the share of durable goods was actively growing. In recent years an increase in the relative costs of services exceeded all other.

4.4 Homogeneous groups of income using cohort identification

Both multi-criteria index of poor-rich (IMPR) and cluster analysis (k-means) are used to identify homogeneous groups by material welfare. The hypothesis of statistical independence of the IMPR and k-means approach is rejected at 1 % significance level, because sample correlation coefficient of quadratic conjugacy [Ayvazian et al., 1983] is $X^2 = 29\,841$ ($p - \text{value} = 0.0000$). Further, it was decided to abandon the use of k- means. Using k-means there were two cases: (1) insignificant coefficients or (2) their sign does not coincide with the sign of model coefficients based on IMPR and with the sign of the correlation.

Multi-criteria index of poor-rich.

This method was proposed in [Gardes et al., 1999]. It was used in [Penikas, 2008]. Unlike simpler methods, it includes three main factors, each of which is assigned a score from 1 to 3 depending on the poverty group (1 – poor, 2 – average, 3 – rich):

Table 3. IMPR factors

	Population		
Factors	**Poor (score = 1)**	**Rich (score = 3)**	**Medium (score = 2)**
Non-satiated preference relation	Food costs > 4/3 average	Food costs < 2/3 average	Otherwise
Marginalization	Total costs < 2/3 average	Total costs > 4/3 average	Otherwise
Insufficiency of financial resources	Below 25 percentile of cost distribution	Above 75 percentile of cost distribution	Otherwise

Note: average = group averages for each year; distribution of costs is adjusted by modified equivalence scale.

Example: the total score of observation with marked characteristics (gray cells) equals 6 (=1+3+2).

Scores for each factor were assigned within one year. Based on the obtained ratings for each criterion, IMPR takes 5 different values: 3 if poor, 4 if quasi-poor, 5–8 if middle class, 8 if quasi-rich, and 9 if rich. Using IMPR scoring the following distribution of households by its types was obtained. Each class has a fairly constant weight (12%, 19%, 47%, 8% and 14% respectively for poor, quasi-poor, average, quasi-rich and rich). Also, clear differences in consumption are visible for the classes. The expenditure on food exceeds all other expenses for the two poorest groups over whole time period, while the main item of expenditure is durable goods for the richer households. The highest share of expenditure on services is observed for the two poorest groups. Probably, it is the cost of housing services. All households pay for the housing services, but the impact of these services is stronger for the poor population.

Splitting the sample by income groups allows to identify the impact of the crises on income per consumption unit for the richest three groups (**Fig. 3**). The constant sample (presented over whole period since 1994) is added to show that the constant sample differs insignificantly from quasi-poor households, but absolutely not identified two richest groups for which changes in consumer behavior are obvious particularly. It suggests that the original observations cannot provide sufficient variability to examine differences in consumption behavior deeply. It stresses the relevance of using pseudo-panels.[6]

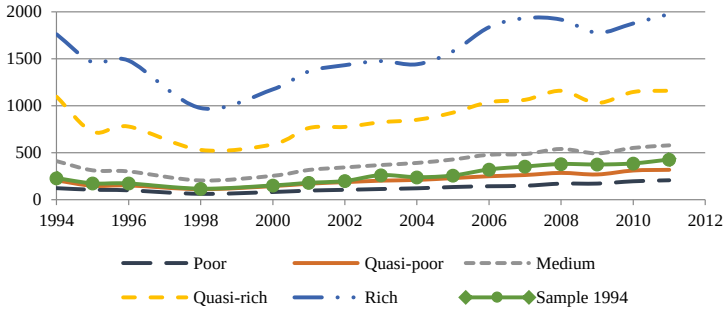


Fig. 3. Income per unit of consumption in 1994 prices

4.5 Quasi-households developing

The basic idea of pseudo-panel [Deaton, 1985] is the formation of cohorts that meet certain constant characteristics, such as belonging to a certain income group. Each cohort represents a quasi-household with the average values for cohort:

$$\bar{y}_{ct} = \alpha_c + \bar{x}'_{ct}\beta + \bar{u}_{ct}, \qquad c = 1, \dots, C; t = 1, \dots, T, \tag{4}$$

Where $\bar{y}_{ct}$ is the average of dependent variable in cohort c in time t, $\bar{x}'_{ct}$ is the average of explanatory variables, α_c is fixed effect for each cohort, c is cohort's number, and t is time. For QAIDS model y_{ct} means w_{ict} (the average weight of good i for cohort c at time t). Average revenue and descriptive statistics of the cohort are included in the vector of explanatory variables.

In the current paper the type of settlement and the average age of the household are used to identify cohorts. The optimal number of groups is formed in such way that the number of households in each group must be positive and the variation should not exceed a reasonable limit.

RLMS surveys indicate 4 main types of settlements: regional center, town, urban-type settlement, and village (44%, 27.7%, 5.7%, and 22.5% of observations, respectively). Two categories are combined into one to obtain approximately constant weights over time for each quasi-households. Urban-type settlement (town) and vil-

[6] The budget coefficients also differ for the three groups of goods insignificantly.

lage were joined, because there is no fundamental difference to interpret consumer behaviour (in both types there is a possibility of employment in agriculture). The weights of each settlement type are relatively stable over time (in average 44%, 28%, and 28% for the regional center, cities, and towns/villages, respectively). Expected differences concerning consumer behaviour are:

- the average income level for each settlement type is different. The larger the settlement, the greater the expected income that affects welfare. Then people who live in cities are rich people, and they need to be differentiated;
- the food expenditures are less in the rural population due to agriculture. The majority of expenditures are on services for the urban dwellers.

Three age groups are formed (**Table 4**). One can observe relatively stable weight for each group over time but still with a slight tendency to increase the percentage of senior households.

Table 4. Age groups

	Households		
	Young	**Older**	**Mature**
%	28 % households	33 % households	39 % households
Age	<28 years	28 – 45 years	>45 years

Two criteria (settlement type and age) create 45 quasi-households (45 = 5 income groups * 3 settlement types * 3 age groups) in each wave. Totally, there are 720 observations for the entite period (720 = 45 quasi-households in a year *16 years). The average composition of quasi-households is 87 real households.

5 Model estimation

Quadratic Almost Ideal Demand Model is estimated for quasi-households both based on income groups (5 quasi-households) and joint groups taking into account income, settlement type, and age of households (45 quasi-households). The model presented in Section 3 also includes the number of consumption units for each household and dummy variables for each year to account for time effect (1994 as a base).

5.1 5 quasi-households

The model with the fixed effect is the most preferred model, since every quasi-household is unique and cannot be regarded as the result of a random selection from the general population. Although, consumer behavior is influenced by psychological factors, then the random effect model may be more preferred.

According to the results of F-test the model with a fixed effect is more preferred than pool model for all goods at 1% significance level. Lagrange multiplier test con-

firms the model with a random effect is chosen for food and durable goods, while a final choice for services is the model with the fixed effect. Hausman test identifies that the model with a random effect is the most preferred specification for food and durable goods consumption. The results are in **Table 5**.

Table 5. Model specification choice (5 quasi-households)

5 quasi-households	Fixed vs. Pooled		Random vs. Pooled		Fixed vs. Random	
Model	F-statistics	P-value	χ^2- stat	P-value	χ^2- stat	P-value
Foods	57.87	0.00	187.16	0.00	1.51	1.00
Durable goods	25.82	0.00	115.39	0.00	0.98	1.00
Services	66.86	0.00	0.27	0.30	-	-

Hausman test used to check the null hypothesis about whether the variables are exogenous shows that the instrumentation of variables is not necessary (no evidence to reject the null hypothesis).[7]

Table 6. Hausman's statistics for testing endogeneity (5 quasi-households)

Model	χ^2-stat	P-value
Foods	6.46	0.97
Durable goods	14.66	0.48
Services	3.32	0.99

5.2 Analysis of structural changes

The homogeneity of three time periods (before the first crisis, between crises and after the second crisis) was studied using correlation analysis and Chow test. The dynamic of correlations of the basic model factors shows that correlation has changed over time. There is the probability to identify a structural change. Chow test rejects the null hypothesis, i.e. there is heterogeneity, and there are two structural breaks that confirms the potential impact of crises of 1998 and 2008. The result is consistent for both 5 and 45 quasi-households.

Table 7. Chow's test for 3 subsamples

Model	χ^2-stat	P-value
Foods	119.9	0.00

[7] It should be noted that covariance matrix for every type of goods was not positive definite. It makes difficult to make a strong conclusion.

Model	$\chi^2 - stat$	P-value
Durable goods	92.88	0.0000
Services	60.91	0.0000

The change in the model coefficients is observed for each group of goods in the dynamics. The coefficients of demand model for services have the greatest variation.

Table 8. The dynamics of estimates for 5 quasi-households.

Period	Total sample	Before 1998	Over 1998–2008	After 2008
Foods				
lnexp	-0.39	-0.09	-0.50	-0.37
s.e.	0.07	0.23	0.06	0.13
t	-5.32	-0.40	-8.13	-2.89
lnexp2	0.01	-0.01	0.02	0.01
s.e.	0.01	0.02	0.01	0.01
t	1.95	-0.61	4.80	1.25
Durable goods				
lnexp	0.30	-0.43	0.35	0.00
s.e.	0.07	0.18	0.07	0.16
t	4.38	-2.34	5.36	-0.02
lnexp2	-0.01	0.04	-0.01	0.01
s.e.	0.01	0.01	0.01	0.01
t	-1.87	3.25	-2.66	1.09
Services				
lnexp	0.19	0.06	0.25	0.42
s.e.	0.05	0.05	0.03	0.11
t	4.24	1.07	9.82	3.82
lnexp2	-0.01	-0.01	-0.02	-0.03
s.e.	0.00	0.01	0.00	0.02
t	-4.95	-1.92	-7.79	-1.74

5.3 Elasticity analysis

Below the results of income elasticities analysis are presented using the model outputs for 5 quasi-households. The income elasticity for 45 quasi-households repeats the case with 5 quasi-households for the respective income groups.

All income groups perceive food as basic necessity goods. However, the richer the group, the smaller the elasticity of food by income (that is, the less necessary the goods become). For the income group "Rich" there is a negative elasticity for the

period up to 1998. However, the estimated coefficients are statistically insignificant. Then we can argue of perfectly inelastic demand on food before 1998, i.e. a change in income has no effect on the food bought ("sticky good"). However, after 2008 a negative elasticity (calculated using the statistically significant coefficients) confirms the conclusion made previously that food products are inferior goods for rich groups.

The income elasticity of demand for durable goods has increased strongly in 1998 and then gradually decreased until 2011 mostly for poor and quasi-poor households. Over the most period durable goods are luxury goods for all income groups. The model estimates for durable goods after 2008 are statistically insignificant, then the durable goods can be recognized as "sticky good" after 2008 until 2011.

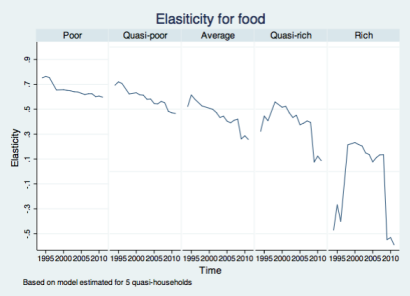


Fig. 4. Elasticity for food (5 quasi-households)

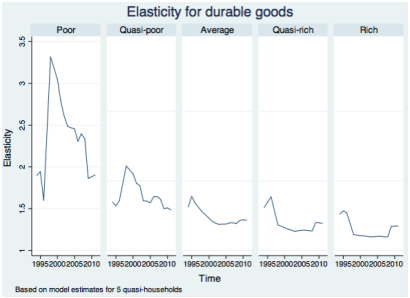


Fig. 5. Elasticity for durable goods (5 quasi-households)

The calculation with structural shifts for services showed that up to 2000 the demand for services was inelastic (statistically insignificant model estimates for the period before 1998). The calculations taking into account the structural changes also show that the services were luxury goods for the three poorest groups and normal goods for the others. Changes of the elasticities become visible at the moment of the 1998 crisis, namely the increase of elasticities in a time of crisis. The households spent their additional income on services less.

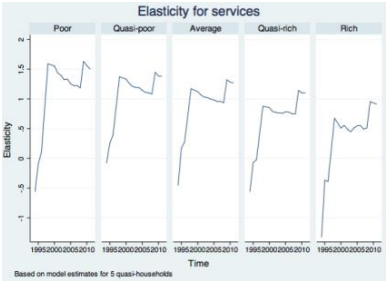


Fig. 6. Elasticity for services (5 quasi-households)

6 Conclusion

[Varian, 2014] draws attention to the need to explore new methods of data analysis for economics. Such necessity is explained by the fact that many modern solutions, including economic policies, require more complex data analysis tools than using only ordinary linear regressions. Our article provides an example of real data analysis problems motivated by the problem of the variability of consumption.

The work is aimed to study the effects of the crises of 1998 and 2008 on the consumer behaviour of Russian households. The research is based on pseudo-panels, which allowed to get rid of the sample attrition effect (a gradual decrease in the number of observations). Pseudo-panels have allowed us to examine the evolution of consumer behaviour for different groups of households according to two classifications: only by income group; and by income group, type of settlement and age of household members.

Descriptive statistics does not provide any evidence of significant impact of crises 1998 and 2008 on Russian consumption (costs weights have not changed significantly), although there was a decline in real income. The elasticity analysis and structural breaks identification shows that some effects are observed for the 1998 crisis, and there were no significant influence by the crisis of 2008. The estimation of the coefficients of dummy variables demonstrates that the effect of 1998 is the highest compared with all other years: a negative value for food products suggests that these goods became more and more necessary for the Russians while the remaining goods become relatively more luxurious. 1998 was preceded by unfavourable years after the collapse of the Soviet Union, when the population practically had no savings. Therefore the crisis affected consumer behaviour. In 2008 and 2009, the Russians have sufficient savings after favourable period for the economy during the period of 2000–2008.

References

1. Banks, J., Blundell, R., Lewbel, A. (1997). Quadratic Engel curves and consumer demand. Review of Economics and statistics, 79(4), 527-539.
2. Deaton, A. (1985). Panel data from time series of cross-sections. Journal of econometrics, 30(1), 109-126.
3. Deaton, A., Muellbauer, J. (1980). Economics and consumer behavior. Cambridge university press.
4. Deaton, A., Muellbauer, J. (1980). An almost ideal demand system. The American economic review, 70(3), 312-326.
5. Gardes, F., Duncan, G. J., Gaubert, P., Gurgand, M., Starzec, C. (2005). Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: The case of us and polish data. Journal of Business & Economic Statistics, 23(2), 242-253.
6. Gardes, F., Gaubert, P., Langlois, S. (2000). Pauvrete et convergence des consommations au Canada. CRSA/RCSA, 36(3), 1-27.
7. Lubrano M. The econometrics of inequality and poverty. Lecture 7: Equivalence scales. 2010. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.3739&rep=rep1&type=pdf
8. Tovar, A. O., Zulaica, I. G., Núñez-Antón, V. (2012). Analysis of pseudo-panel data with dependent samples. Journal of Applied Statistics, 39(9), 1921-1937.
9. Aivazyan S.A., Mkhitaryan V.S. (1983). Practical Statistics. Finances and Statistics.
10. Matytsin M.S., Yershov E.B. (2012). Research of Real Income Differentiations of Russians. Economics Journal of the Higher School of Economics, 16(3), 318–340
11. Penikas H.I. Analysis of Consumer Behaviour Evolution in Russia throughout 2000–2005. (2008). Economics Journal of the Higher School of Economics, 12(4), 512–542.
12. Varian H.R. Big Data: New Tricks for Econometrics. (2014). Journal of Economic Perspectives, 28(2), 3–28.

Using Emotional Markers' Frequencies in Stock Market ARMAX-GARCH Model*

Alexander V. Porshnev[1], Valeriya V. Lakshina, and Ilya E. Red'kin[2]

[1] National Research University Higher School of Economics,
603155, Bolshaya Pecherskaya St., 25/12, Nizhny Novgorod, Russia
[2] STM LLC,
603024, 5 Kazanskaya naberezhnaya, office 25, Nizhny Novgorod, Russia
{aporshnev,vlakshina}@hse.ru, ilya-redkin@yandex.ru

Abstract. We analyze the possibility of improving the prediction of stock market indicators by adding information about public mood expressed in Twitter posts. To estimate public mood, we analysed frequencies of 175 emotional markers - words, emoticons, acronyms and abbreviations - in more than two billion tweets collected via Twitter API over a period from 13.02.2013 to 22.04.2015. We explored the Granger causality relations between stock market returns of S&P500, DJIA, Apple, Google, Facebook, Pfizer and Exxon Mobil and emotional markers frequencies. We found that 17 emotional markers out of 175 are Granger causes of changes in returns without reverse effect. These frequencies were tested by Bayes Information Criteria to determine whether they provide additional information to the baseline ARMAX-GARCH model. We found Twitter data can provide additional information and managed to improve prediction as compared to a model based solely on emotional markers.

Keywords: Twitter, mood, emotional markers, stock market, volatility

1 Introduction

Mood, emotions and decision making are closely connected. Modeling decision making process [1] report that psychological states invoked by reading stories can affect the evaluation of risk level. Positive moods lead individuals to make more optimistic choices and, vice versa, negative moods lead to pessimistic choices, see [2], [3].

Positive and negative moods influence the decision making process by invoking different heuristics. For example, individuals in positive mood tend to spend less time on decision making by referring more rarely to already reviewed alternatives and ignoring information they believe is irrelevant according to [4].

* We thank our colleagues of the International Laboratory of Intangible-driven Economy (National Research University Higher School of Economics, Perm, Russia, 614070, 38 Studencheskaya Ulitsa. E-mail: info@hse.perm.ru), who provided valuable comments and expertise that greatly assisted the research.

[5] expresses the idea that general level of optimism/pessimism in society can be connected with economic activity. Nofsinger also supposes that the stock market itself can be a direct measure of social mood. Following Nofsinger, we will regard the economy not as a physical system, but as a complex system of human interactions, in which moods and irrationalities can play a significant role. This point can be supported by observing the informational cascades phenomenon in the stock market.

Regarding the stock market and Twitter as two possible measures of social mood, we can assume their correlation and the possibility of using analyses of moods expressed in tweets to increase prediction accuracy for stock market indicators.

Experiments in psychology and behavioral economics show how moods and emotions influence decision making [6], [7], [8]. The role of moods and emotions in decision making grows in situations of uncertainty incidental to the stock market. Behavioral researchers found a trader's decision to demonstrate a wide set of human cognitive biases and influence of emotional factors [9], [10]. Publicly expressed emotions in Facebook and Twitter draw attention of many researchers [11], [12], [13]. A relation between Facebook's Gross National Happiness Index and 20 international markets is shown in [14]. They also demonstrated that negative sentiments are related to increases in trading volumes and return volatility. We propose to use an alternative measure of sentiment based on posts published by user marked their location in US in Twitter.

Noteworthy is that people tend to often use abbreviations and emoticons in Twitter, so we extended the list of words with such signs and termed them emotional markers.

Another important question we raise in our research is whether frequencies of emotional posts from Twitter add information according to the Bayes Information Criteria, see, for example, [15]. In their detailed review of methods and models applied in textual sentiment analysis in the financial field [16] note that volatility models have rarely been used. For example in recent paper by Nofer and Hinz the returns are modelled by a linear regression without taking into account autoregressive and conditional heteroskedasticity effects [12]. In our research we tested the hypothesis that Twitter could provide additional information to increase the fit of the ARMAX-GARCH econometric model. We expected information about Twitter users' sentiment to be a significant regressor in complex ARMAX-GARCH models for S&P500 index. ARMAX-GARCH model was chosen as one of the most widespread models in time series analysis, which allow autocorrelation and heteroskedasticity to be taken into account [17].

Over the last five years social media and sentiment analysis have drawn attention of many researchers in economics. According to [16], most of 38 studies run in this area in 2004-2013 were concerned with the usage of news articles, annual reports, earnings press or other financial-related information and only one dealt with information from Internet messages. On the one hand, the approach based on financial data looks more relevant, but it may fail to recognize faked or historical news as was the case in 2013 [18], [19]. We also expect public moods

and emotions to provide additional information and influence investors' response to certain news and events.

Although in several preprints [11], [20], [21], [22] the authors report that Twitter mood could be used to enhance the quality of stock market forecasts, the validity of the conclusions made by the authors remains doubtful. Regretfully, in the first preprints concerned with this topic there was either a short (less than 40 days) out-of-sample testing period or the authors only compared the results of using moods to a simple econometric model. In our research we extended the out-of-sample period to 100 days and applied a more complicated ARMAX-GARCH model.

Assuming that some words, abbreviations and emoticons can be more related to emotions, we verified the hypothesis that emotional marker frequencies can be indicators of stock prices movement.

The rest of the article is organized as follows. Section 2 describes the methods employed in the research, Sect. 3 describes the data and their preprocessing, Sect. 4 contains the results and Sect. 5 concludes.

2 Methodology

2.1 Emotional Markers

One of the simplest and most intuitive way of textual analysis is word counting [23], so we use the frequencies of words from a specially drawn up list instead of combining them into one or several mood indexes[3]. The results obtained reveal a high correlation between words expressed in Twitter and the S&P500 index, but those correlations did not always ensure that information was added to the ARMAX-GARCH model.

We compile our list of emotional markers using a Brief Mood Introspection Scale with 8 scales and 2 adjectives representing each mood as the starting point in creating dictionaries [24]. We extend this list with all the synonyms of the adjectives selected from the WordNet dictionary [25]. For example, we measure the presence of an energetic state in tweets by the occurrence of the following words: animate, animated, athletic, brisk, chipper, emphatic, enterprising, exuberant, fresh, lusty, passionate, robust, sprightly, spry, strenuous, strong, tireless, trenchant, warming party, honor, and vote. We also add the possibility of recognizing derived words, such as "happyyy" or "happppppyyyyyyy" and count them using regular expressions.

We do not include negations, because after analyzing a testing sample of 9000 tweets we found that negations were not common. For example, the testing sample with 51 words "happy" contains the negation "not happy" only once. The same is the case with "but" and sentences expressing desires, e. g. "wanna

[3] A similar approach is used by [22]. They analyze frequencies of several words (e.g. "worry", "hope", "fear" etc.) and find high correlation between the frequencies of emotional posts and S&P500, DJIA, and VIX indexes.

be happy". The probable reason for that is the small number of words allowed for a Twitter message (140 words).

[26] show that emoticons[4] have a very good classification power and that accuracy of emoticon-based sentiment classification exceeds 90% for tweets with emoticons. Impressed by this result we extend our list with emoticons used in [27]. It should be mentioned that we distinguish different types of smiles. For example, ":)", ":-)", and ":D" are not synonyms.

Importantly, that Twitter lexicon contains a lot of abbreviations and slang words, such as "LOL", "WTH"[5]. At the final stage we add abbreviations expressing emotional states from [28].

Our list of emotional markers contains totally 175 items. We count the number of posts with each emotional marker per day and consider it as emotional marker frequencies. Before that all the tweets are transferred to the lower case. The frequencies are included in (4) and (5) as additional regressors.

2.2 Granger Causality

We examined the predictive causality[6] relations between sentiment and log returns, using the idea of the Granger test (see, for instance, [29]). Following the methodology described in [30], we estimate (1) and (2).

$$R_t = a_0 + \sum_{i=1}^{L} \alpha_i R_{t-i} + \sum_{j=1}^{L} \beta_j X_{t-j} + \varepsilon_t, \tag{1}$$

$$X_t = \widetilde{a_0} + \sum_{i=1}^{L} \widetilde{\alpha_i} X_{t-i} + \sum_{j=1}^{L} \widetilde{\beta_j} R_{t-j} + \widetilde{\varepsilon_t}, \tag{2}$$

where R_t is asset's returns, X_t — emotional marker, a_0, α_i, β_j and their tilde counterparts are parameters, ϵ_t and $\widetilde{\epsilon_t}$ are uncorrelated error terms. We found the optimal lag of each sentiment series X_t by varying the L parameter from 1 to 30, whereas in the works undertaken the lags for Granger test do not commonly exceed 7 days [11, 22].

The estimation of (1) and (2) allows us to select those emotional markers which Granger-cause returns and simultaneously are not Granger-caused by them. In fact we leave only those markers for which (1) is significant and (2) is insignificant on 5% level according to F-test. Thereby we prevent the reverse causality problem, described in [31].

[4] Emoticon means "emotional icon" and usually denotes some combination of printed symbols expressing person's feelings or mood.

[5] "laughing out loud" and "what the hell"

[6] "True causality" relations is rather a philosophical question, here we explore the relations of preceding one time series to another, which are useful in establishing the predictability.

2.3 ARMAX-GARCH Model and Model Testing

To examine the impact of Twitter mood on the returns of stocks and stock market indexes, this study uses the well-known ARMAX-GARCH model, controlling for autocorrelation and conditional heteroskedasticity, see, for example, [32]. The resulting ARMAX(p,q)-GARCH(r,m) model can be written as in (3).

$$x_t = E(x_t|\mathcal{F}_{t-1}) + y_t, \tag{3}$$

where $E(x_t|\mathcal{F}_{t-1})$ is a conditional mean of daily return x_t at time t conditional on all available at $t-1$ information F_{t-1}, y_t are innovations. Returns x_t are calculated as a logarithm of today price divided by yesterday price: $x_t = log(\frac{P_t}{P_{t-1}})$. Conditional mean $E(x_t|\mathcal{F}_{t-1})$ is modelled as ARMA(p,q), (4).

$$E(x_t|\mathcal{F}_{t-1}) = a_0 + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \sum_{k=1}^{n} \gamma_k X_{k,t}, \tag{4}$$

where parameter α_i and β_j are the ith-order autoregressive (AR) and jth-order moving average (MA) terms respectively; parameter γ_k measures the impact of additional regressor X_k on the index return. In our research emotional marker frequencies play a role of the additional regressors X_k.

Innovations y_t are modeled as GARCH(r,m), (5).

$$\begin{gathered} y_t = \sigma_t \cdot \eta_t,\ \eta_t \sim f(\theta), \\ \sigma_t^2 = c_0 + \sum_{i=1}^{r} \kappa_i \epsilon_{t-i}^2 + \sum_{j=1}^{m} \mu_j \sigma_{t-j}^2, \end{gathered} \tag{5}$$

where parameters κ_i and μ_j account for ARCH and GARCH effects of ith and jth orders respectively; σ_t^2 — volatility, η_t — error term, distributed according to some distribution f with parameter set θ It is also possible to add Twitter mood X_k to GARCH equation in order to measure the influence of Twitter mood on volatility.

Traditional specifications of ARMAX-GARCH imply normal or Student-t distribution of the error term. These distributions is that they cannot capture asymmetry in returns distribution. In order to eliminate this drawback, we implemented the skewed normal and skewed Student's distributions for error term. The distributions are modeled as special cases of generalized hyperbolic distribution by [33]. We also estimate ARMAX-GARCH with normal errors as a benchmark.

We choose the parameters p, q, r and m by means of Bayesian information criteria (BIC) — the best specification corresponds to the minimal BIC. Estimation is carried out by means of *rugarch* package by [34]. We employ Vuong test for comparing models. Our choice is caused by the fact that this test can be used for non-nested models in contrast to traditional likelihood ratio test. The null hypothesis implies the equal goodness of fit for the comparing models. Since the observations in financial time series are not typically independent we use heteroskedasticity and autocorrelation consistent version of Vuong test [35].

We use the mean squared error (MSE) and directional accuracy (DAC) as measures of out-of-sample performance. The latter shows the percentage of matches between returns and their forecast.

3 Data Description

The data about eight assets, including S&P500 and DJIA indexes; Apple, Facebook, Google, JP Morgan Chase, Pfizer and Exxon Mobil stocks are obtained from [36]. The period under consideration spanned 521 trading days and lasts from February 13, 2013 to April 22, 2015.

By making use of Twitter API, we downloaded 2,349,036,300 tweets over the considered period. It's in average 3,098,992 tweets per day. The only restriction made on downloaded posts is that they should be published by people located in US. All the tweets were sorted by days and analyzed automatically in the created JAVA application. For each day we calculated frequencies of posts with each item from the emotional markers list, described in Sect. 2.1, and normalize them by the number of tweets downloaded on each day.

Most of the frequencies exhibited non-stationary behavior. On the other hand, some frequencies (approximately 10% out of all) are difference stationary, i. e. have a unit root, which is confirmed by the augmented Dickey-Fuller test [37].

If non-stationary regressors present in the ARMAX-GARCH model, then conventional statistical measures, such as t-statistics or R-squared, are inapplicable [38]. The non-stationary emotional markers' frequencies are brought to stationary series by means of either detrending (for trend stationary series) or taking the first difference (for difference stationary series). The repeated ADF test rejects non-stationarity in all cases.

The whole data set was divided into two subsamples: for in-sample and out-of-sample testing. We choose to use 100 days period for out-of-sample testing, what gives approximatly 400 days to find a model with a optimal fit. It is worth to mention that each emotional marker has its own optimal L parameter in (1) and (2), meaning the time lag on which Granger causality takes place. Therefore unique subsample, cut due to L, corresponds to each emotional marker and estimation of baseline model for an asset is conducted on these subsamples.

4 Empirical Results

Firstly we evaluate the causality relations between emotional markers and returns by Granger test as explained in Sect. 2.2. Secondly we define three groups of assets: indexes, emotion sensitive stocks and emotion insensitive stocks. The groups include S&P500 and DJIA; Apple, Facebook and Google; JP Morgan Chase, Pfizer and Exxon Mobil correspondingly. For each group two ARMAX-GARCH models are estimated: a sentiment model, which contains emotional marker in the mean equation (4), and a baseline model without additional regressor in the mean equation.

The estimation of (1) and (2) results in 17 emotional markers. We excluded emotional markers which appeared very seldom and had many zeros.

Almost each asset is Granger caused by *gloom*, except from JPM and PFE. *Hope* and *bad* are Granger-valid for the half of assets, however they don't occur among emotion sensitive assets. The other markers have frequencies two or one and can be considered as specific for some asset. For example, PFE is Granger caused by *cancer*, XOM — by *richer*.

We explored different specifications of the ARMAX-GARCH model with p, q, r and m, ranging from zero to three (except r, which cannot be smaller than one).

4.1 In-sample

Firstly, we study the exploratory power of the emotional markers in financial time series modeling.

The group of indexes consist of S&P500 and DJIA has three common emotional markers: *hope*, *bad* and *gloom*. The dynamics of DJIA is also affected by *alas*. Although, *alas* provide additional information by AIC and HQIC criteria, but not according to a BIC. The emotional marker *bad* add information to models with normal and skewed Student distributions. Specifications which provide better fitting according to BIC in most cases have skewed normal distribution for error term. It's worth mentioning that indexes have no GARCH-effects in this specifications, because m parameter is equal to zero for the optimal models.

The coefficients in the optimal models are significant on 5% level. Although, emotional marker *gloom* have a positive effect on the dynamics of the both indexes', it affects DJIA almost ten times stronger than S&P500, see Table 1. *alas* has substantial negative impact on DJIA log returns.

The next group of assets, that we call emotional sensitive stocks, includes AAPL, FB and GOOG. The group has more Granger-valid emotional markers than the previous one, for example, *awful*, *fear*, *frighten* and already mentioned *bad* and *gloom*.

For emotional sensitive stocks sentiment models with skewed normal distribution again performs better than baseline models. Normal distribution is also presents among optimal specifications. Skewed Student's distribution is included in optimal specifications only for *ok*[7] marker for AAPL stock returns. Vuong test supports the alternative hypothesis that sentiment models has better fit than baseline ones, Table 1.

Emotional markers associated with fear, i.e. *fear* itself and *frighten* exhibit strong negative impact on returns. The same behavior is demonstrated by *awful*. Interestingly that *ok* also has negative effect but the size of the effect is much smaller than for *fear* marker.

Looking forward to marker has substantial positive impact on AAPL and GOOG stocks' returns. *bad* that is likewise among AAPL and GOOG Granger-valid markers has minor positive effect on returns in specifications with normal

[7] *ok* means "only kidding".

and skewed normal errors. *gloom* being common for all stocks in the considered group is insignificant on 5% level in optimal specifications.

As for the last group of emotion insensitive stocks normal and skewed normal distributions similarly demonstrate better fit, verified by Vuong test. Already mentioned *hope* and *frighten* have significant positive and negative impact correspondingly. For Exxon *hope* turns out to be insignificant on 5% level.

We found that emotion insensitive stocks have specific emotional markers, discovered by Granger test (1) and (2). They are *cancer* for Pfizer, *br*[8] for JP Morgan and *richer* for Exxon Mobil. Although *cancer* marker is insignificant it helps to improve the predictive power comparing to the baseline model. JP Morgan's specific marker slightly decreases the returns. As opposed *richer* is one of the strongest determinants of Exxon's returns growth. The other important emotional markers for XOM are positively affecting *gloom* and *dark* and negatively affecting *sad*.

Our hypothesis is that emotion sensitive group of stocks is more affected by emotional markers than emotion insensitive group. The results evidence that stocks in both groups are influenced by emotional markers. Although, the tweets we analyze are not restricted to those that regard to the economy, business climate, world affairs, specific businesses and, for example, includes tweets by teenagers talking about regular things or events, we found that suggested sentiment measurement do add information to ARMAX-GARCH model. One of the possible ways to further research in this area is to organize filtering of downloaded messages to measure sentiments of a more inclusive group, based on context published in their posts (business or economics related).

Table 1. Summary for sentiment models which significantly outperform baseline

	Asset	DJI	SNP	AAPL	JPM
Baseline model parameters	Distribution	snorm	sstd	sstd	sstd
	p,q,r,m	0,0,3,0	2,1,1,1	0,0,3,0	0,0,1,0
Sentiment model parameters	Distribution	snorm	sstd	sstd	sstd
	p,q,r,m	2,3,1,1	2,2,1,1	3,2,1,0	3,2,1,0
	Emotional marker	gloom	gloom	ok	hope
	Coefficient	0.634*	0.065*	-0.016	0.053*
	Lag	8	8	29	8
BIC	Baseline	-7.309	-7.192	-5.750	-6.030
	Sentiment	-7.314	-7.203	-5.764	-6.031
AIC	Baseline	-7.397	-7.280	-5.811	-6.079
	Sentiment	-7.421	-7.310	-5.875	-6.138
Vuong test		1.042*	0.407*	0.885*	0.053*

* means significant on 1% level.
p,q,r,m are corresponding parameters in (4) and (5).

[8] *br* means "best regards".

4.2 Out-of-sample

100 observations are retained to evaluate the out-of-sample performance of emotional markers. We calculate MSE and DAC (refer to Sect. 2.3 for details) as measures of emotional markers' predictive power.

The optimal models for index group demonstrate less successive predictive performance, comparing to baseline models. *hope* marker is an exception, providing smaller MSE for both indexes. In addition *hope* and *bad* markers with normal and skewed Student's errors allow to increase directional accuracy of DJIA and S&P500 returns to 58% and 54% correspondingly.

Directional accuracy for emotional sensitive group of assets is higher than for the index group even in BIC selected models. The obtained DAC for optimal models starts from 50% and peaks on 63% for *frighten* marker.

It should be noted that k[9] being insignificant on 5% level yield outstanding out-of-sample results with smaller MSE and DAC equal to 57%–58%. It confirms our suggestion of that the emotional markers which provide poor in-sample performance can be successfully used in prediction models.

Models directional accuracy in the last group of emotional insensitive stocks varies from 47% to 56%. The same distributions, namely normal and skewed normal, provide enhanced prediction comparing to baseline models. Emotional markers which contribute to the out-of-sample performance most are *br* for JPM, *cancer* for PFE and *richer*, *hope* and *dark* for XOM.

Models which exhibit poorer performance in-sample demonstrate promising out-of-sample results. We consider this as a motivation to find optimal specifications by some predictive criteria, such as MSE or DAC, to obtain models with increased predictive power.

It's important to add that emotional markers being included in volatility equation (5) are insignificant on any reasonable significance level. We also controlled the mean equation for day effects and found no evidence of their presence or their impact on the prediction ability.

5 Conclusion

We started our research with a question: can Twitter data bring additional information to the ARMAX-GARCH model? Being positive the answer is based on the thoroughly elaborated methodology (see Sect. 2 for more details), which includes collecting and preprocessing Twitter posts, applying some textual analysis to the tweets, defining the Granger causality relations and implementing the output to the ARMAX-GARCH modeling. We wish to make the textual analysis stage transparent and simple, thus we use parsimonious word count technique to create so called emotional markers, which subsequently are used as the determinants of the log returns dynamics in the ARMAX-GARCH model. We form three groups of assets, namely indexes, emotion sensitive stocks and emotion insensitive stocks.

[9] k is short for "ok".

Studying the explanatory power of constructed models we show that emotional markers demonstrate smaller BIC and provide significant positive increment to the likelihood function subject to Vuong test. In order to capture higher order effects of returns, such as skewness associated with the third moment of returns' distribution, we implement skewed versions of normal and Student's distributions for errors. In some cases, including Facebook, Google and JP Morgan, the third moment effects turn out to be insignificant so normal distribution also works well for these stocks.

We find evidence of that such emotions as fear and sorrow, represented by markers *fear*, *frighten* and *alas*, have substantial negative impact on the dynamics of both stocks and indexes. Negative influence of sorrow is also confirmed by *sad* for XOM. *Looking forward to* marker which corresponds to anticipation has substantial positive impact on stocks' returns from emotional sensitive group. The analysis of emotion insensitive group reveals the existence of specific emotional markers for the members of this group. Being negligible in the exploratory sense they allow to increase the predictive ability of the ARMAX-GARCH model.

We expect that the relationship between emotional markers and returns can change over time. Firstly, it could happen because of some fundamental factors. In a period of financial stability, for example, emotions may play a smaller role than during a downturn. And vice versa, the market response to similar financial news may be different depending on the mood prevailing in society.

Secondly, the behavior of stock market players could change if they would take in account information from emotional markers. Since we detect two kinds of emotional markers (see 4.2) — those, which explain the returns well, and those, which predict the returns, we expect that the changes in investors' behavior should be based on the emotional markers of the second kind. On the other hand "explaining" emotional markers should not change the behavior of stock market players, because they seem to be an intrinsic characteristic of the market and are unlikely to generate profitable trading strategy.

In our further research we plan to move in two directions. Firstly, we will distinguish periods when the stock market is emotional driven and news driven. Secondly, we will monitor Twitter posts to see if there will be significant changes in emotional marker frequencies, which could be a sign of manipulation.

References

1. Johnson, E.J., Tversky, A.: Affect, generalization, and the perception of risk. Journal of personality and social psychology **45**(1) (1983) 20
2. Isen, A.M., Patrick, R.: The effect of positive feelings on risk taking: When the chips are down. Organizational Behavior and Human Performance **31**(2) (1983) 194–202
3. Schwarz, N., Clore, G.L.: Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. Journal of personality and social psychology **45**(3) (1983) 513
4. Isen, A.M., Means, B.: The influence of positive affect on decision-making strategy. Social cognition **2**(1) (1983) 18–31

5. Nofsinger, J.R.: Social Mood and Financial Economics. Journal of Behavioral Finance **6**(3) (September 2005) 144–160
6. Ding, T., Fang, V., Zuo, D.: Stock Market Prediction based on Time Series Data and Market Sentiment. (2013)
7. Loewenstein, G.: Emotions in Economic Theory and Economic Behavior. The American Economic Review **90**(2) (2000) 426–432 ArticleType: research-article / Issue Title: Papers and Proceedings of the One Hundred Twelfth Annual Meeting of the American Economic Association / Full publication date: May, 2000 / Copyright 2000 American Economic Association.
8. McFarland, C., White, K., Newth, S.: Mood acknowledgment and correction for the mood-congruency bias in social judgment. Journal of Experimental Social Psychology **39**(5) (September 2003) 483–491
9. Saunders, Jr., E.M.: Stock Prices and Wall Street Weather. The American Economic Review **83**(5) (December 1993) 1337–1345
10. Hirshleifer, D., Shumway, T.: Good day sunshine: Stock returns and the weather. The Journal of Finance **58**(3) (2003) 1009–1032
11. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2011) 1–8
12. Nofer, D.M., Hinz, P.D.O.: Using Twitter to Predict the Stock Market. Business & Information Systems Engineering **57**(4) (jun 2015) 229–242
13. Rao, T., Srivastava, S.: Using twitter sentiments and search volumes index to predict oil, gold, forex and markets indices. (2012)
14. Siganos, A., Vagenas-Nanos, E., Verwijmeren, P.: Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organization **107, Part B** (nov 2014) 730–743
15. Javed, F., Mantalos, P.: Garch-type models and performance of information criteria. Communications in Statistics-Simulation and Computation **42**(8) (2013) 1917–1933
16. Kearney, C., Liu, S.: Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis **33** (may 2014) 171–185
17. Horv, L., Kokoszka, P.: Garch processes: structure and estimation. Bernoulli **9**(2) (2003) 201–227
18. Prezioso, J.: Yom Kippur War Tweet Prompts Higher Oil Prices. `http://www.huffingtonpost.com/2013/10/10/yom-kippur-war-tweet-oil-prices-traders_n_4079634.html` (2013) Accessed: 2014-01-22.
19. Selyukh, A.: Hackers send fake market-moving AP tweet on White House explosions | Reuters. `http://www.reuters.com/article/2013/04/23/net-us-usa-whitehouse-ap-idUSBRE93M12Y20130423` (2013) Acessed: 2013-09-17.
20. Chen, R., Lazer, M.: Sentiment analysis of twitter feeds for the prediction of stock market movement. workpaper, Stanford (2013)
21. Porshnev, A., Redkin, I., Shevchenko, A.: Improving Prediction of Stock Market Indices by Analyzing the Psychological States of Twitter Users. SSRN Scholarly Paper ID 2368151, Social Science Research Network, Rochester, NY (December 2013)
22. Zhang, X., Fuehres, H., Gloor, P.A.: Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". The 2nd Collaborative Innovation Networks Conference - COINs2010 **26**(0) (2011) 55–62
23. Loughran, T., McDonald, B.: The use of word lists in textual analysis. Journal of Behavioral Finance **16**(1) (2015) 1–11

24. Mayer, J.D., Gaschke, Y.N.: The experience and meta-experience of mood. Journal of personality and social psychology **55**(1) (1988) 102
25. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11) (1995) 39–41
26. Boia, M., Faltings, B., Musat, C.C., Pu, P.: A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets. In: Social Computing (SocialCom), 2013 International Conference on, IEEE (2013) 345–350
27. Schnoebelen, T.: Do you smile with your nose? Stylistic variation in Twitter emoticons. (2012)
28. : The Online Slang Dictionary | Urban Thesaurus (aug 2016)
29. He, Z., Maekawa, K.: On spurious granger causality. Economics Letters **73**(3) (2001) 307–313
30. Kim, S.H., Kim, D.: Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior & Organization **107** (2014) 708–729
31. Brown, G.W., Cliff, M.T.: Investor sentiment and the near-term stock market. Journal of Empirical Finance **11**(1) (2004) 1–27
32. Francq, C., Zakoian, J.M., et al.: Maximum likelihood estimation of pure garch and arma-garch processes. Bernoulli **10**(4) (2004) 605–637
33. Barndorff-Nielsen, O.: Hyperbolic distributions and distributions on hyperbolae. Scandinavian Journal of statistics (1978) 151–157
34. Ghalanos, A.: rugarch: Univariate GARCH models. (2014) R package version 1.3-5.
35. Calvet, L.E., Fisher, A.J.: How to forecast long-run volatility: regime switching and the estimation of multifractal processes. Journal of Financial Econometrics **2**(1) (2004) 49–83
36. : Yahoo! Finance. `http://finance.yahoo.com` (2015) Accessed: 2014-01-22.
37. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association **74**(366a) (1979) 427–431
38. Zhou, Z., Shao, X.: Inference for linear models with dependent errors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(2) (2013) 323–343

Finding the Sweet Spot in the City: a Monopolistic Competition Approach

Elizaveta Bespalova, Alim Moskalenko, Alexander Safin, Constantine Sorokin*, and Andrey Yagolkovsky

National Research University Higher School of Economics,
20 Myasnitskaya ulitsa,
Moscow 101000 Russia.
{csorokin}@hse.ru
http://www.hse.ru

Abstract. We propose a general equilibrium model to study the spatial inequality of consumers and firms within a city. Our mechanics rely on Dixit and Stiglitz monopolistic competition framework. The firms and consumers are continuously distributed across a two-dimensional space, there are iceberg-type costs both for goods shipping and workers commuting (hence firms have variable marginal costs based on their location). Our main interest is in the equilibrium spatial distribution of wealth. We construct a model that is both tractable and general enough to stand the test of real city empirics. We provide some theoretical statements, but mostly the results of numerical simulations with the real Moscow data.

Keywords: spatial distribution, linear city, circular city, monopolistic competition

1 Introduction

We propose a general equilibrium model to study the spatial inequality of consumers and firms within a city. Our mechanics rely on Dixit and Stiglitz monopolistic competition framework [1]. The firms and consumers are continuously distributed across a two-dimensional space, there are iceberg-type costs (as in Krugman's models of trade[2]) both for goods shipping and workers commuting (hence firms have variable marginal costs based on their location, we borrow some of Melitz's ideas here [6]). Our main interest is in the equilibrium spatial distribution of wealth. We construct a model that is both tractable and general enough to stand the test of real city empirics. We provide some theoretical statements, but mostly the results of numerical simulations with the real

* The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'

Moscow data. Our model is somewhat similar to [7], however, we rely on a different framework, adjusting the ideas from the international trade models to the scale of a city.

Our theoretical modelling process consists of two parts. First, we developed a one-dimensional model of the linear city on the $[0, 1]$ interval. This model is an adaptation of the Dixit-Stiglitz-Krugman model of trade to the case of spatial distribution of firms and consumers within the city. Although it is far from reflecting the real city structure, it is easier to solve and interpret the results. Second, we demonstrate how the unidimensional model can be upgraded to the empirics-friendly two-dimensional setup.

2 One-dimensional City

Consider a city with workers distributed uniformly on $[0, 1]$ and firms distributed uniformly on $\left[\frac{1}{2} - \frac{N}{2}; \frac{1}{2} + \frac{N}{2}\right]$, where N is an endogenous parameter for the total number of the firms within the city. Each firm offers a single variety of a composite differentiated good. We assume that the demand for that goods is not only due to workers spending their wages, but also due to firm owners spending their profits, so the consumers (both "firms" and workers) are distributed on $\left[\min\left\{0, \frac{1}{2} - \frac{N}{2}\right\}; \max\left\{1, \frac{1}{2} + \frac{N}{2}\right\}\right]$, which is hereinafter denoted by χ.

Consumer located at point x (or just consumer x) has an endowment of $e(x)$. Firm located at point y offers its own variety of composite differentiated good at the price $p(y)$. So here the product differentiation coincides with spatial differentiation of firms — there is just one firm standing on the head of a pin. Each consumers preferences are given by a CES-type utility function, thus they exhibit the preference for variety. By solving the consumer's problem we obtain the demand $q(x, y)$ — the amount of good that consumer x wants to buy from firm y.

Our approach to firm's production differs from the classical one in the structure of marginal costs. We assume that a firm has an equal probability of hiring any worker in the city — firms can't discriminate workers by their location, thus there is a uniform equilibrium wage w. So in our model workers do not choose place, where they want to work and wage, that they want to earn - place, where they will work chooses randomly and wage is the same for all workers. Workers commuting costs are paid by the firms, so that distant commuting results in low productivity. Thus we have the following formula for labor requirements to produce a unit of good for a firm located at point y:

$$c(y) = a \int_0^1 (1 + \theta\rho(x, y))dx,$$

where $\rho(x, y)$ is a distance from point x to point y. Firms also have to pay a fixed cost of $fc(y)w$.

The equilibrium in our model is characterized by the following conditions:

1. Full employment: all the labor supplied is used in production. Let $C(y)$ be the total labor used by a firm located at y, so that we have:

$$\int_{\frac{1-N}{2}}^{\frac{1+N}{2}} C(y)dy = 1.$$

2. Free entry: firms located on the borders of the city have zero profits. Let $Q(y)$ be the overall amount of goods sold by firm y, therefore we have:

$$\left(p\left({}^{N-1}/_{2}\right) - wc\left({}^{N-1}/_{2}\right)\right) Q\left({}^{N-1}/_{2}\right) - wc(y)f = \\ \left(p\left({}^{N+1}/_{2}\right) - wc\left({}^{N+1}/_{2}\right)\right) Q\left({}^{N+1}/_{2}\right) - wc(y)f = 0.$$

3. Finally, the budget balance: the total expenditure at some point equals the sum of worker's wage and firm's profits — of course, if they are located there. Thus:

$$e(y) = \begin{cases} 0, & y \notin [0;1] \cup \left[\frac{1}{2} - \frac{N}{2}; \frac{1}{2} + \frac{N}{2}\right], \\ w, & y \in [0;1] \setminus \left[\frac{1}{2} - \frac{N}{2}; \frac{1}{2} + \frac{N}{2}\right], \\ \left(p\left(y\right) - wc\left(y\right)\right) Q\left(y\right) - wc(y)f, & y \in \left[\frac{1}{2} - \frac{N}{2}; \frac{1}{2} + \frac{N}{2}\right] \setminus [0;1], \\ w + \left(p\left(y\right) - wc\left(y\right)\right) Q\left(y\right) - wc(y)f, & y \in [0;1] \cap \left[\frac{1}{2} - \frac{N}{2}; \frac{1}{2} + \frac{N}{2}\right]. \end{cases}$$

3 Results for One-dimensional City

We are able to prove the equilibrium existence result for the model, also we do some natural comparative statics. However, for illustrative purposes we created a MATLAB program to find the equilibrium given all the exogenous parameters and illustrate all the comparative statics of interest. For example, for the following values $f = 0.2,\ a = 1,\ \sigma = 2,\ \tau = 0.1,\ \theta = 0.1, Ł = 0.1$ the total number of firms is 2.25 and the corresponding interval is $[-0.625, 1.625]$.

For these values of parameters we can look at the changes in the profit of firms $\pi(y)$ and the optimal number of firms N according to some changes in exogenous parameters. For all of the following graphs horizontal axis shows the interval for firms and vertical axis shows the profits value. Blue lines here in after depict small values of the changing parameter, yellow lines — highest values.

For instance, figure 1 illustrates that when the fixed cost for the firms f rises, the optimal number of firms in the city falls, but the profit level of existing firms $\pi(y)$ in some times increases, but then falls with the rise of fixed costs. This is happening because when the fixed costs rise for sufficiently low number of firms, the effect from rising costs exceeds the one from "killing" competitors, so the profit falls.

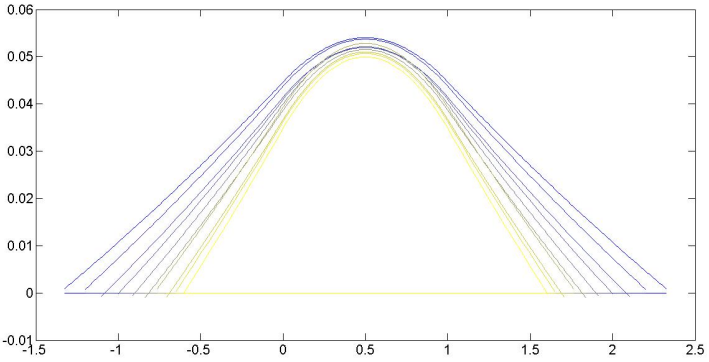
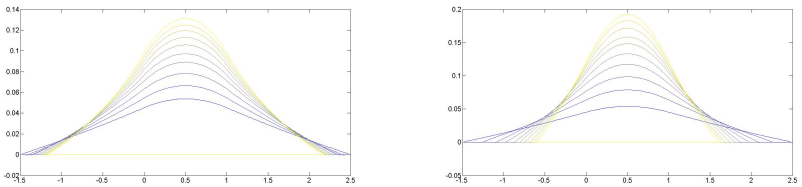


Fig. 1. Firm profits as f changes from 0.1 to 0.2.

As we can see from figure 2, increase in the sensitivity to changes in distance for the transportation of goods (τ) and workers (θ) affect the profit of firms differently. Although the number of firms N decreases only slightly with the rise of both τ and θ, it causes a significant increase in profit, which is 1.5 times greater for θ.



τ θ

Fig. 2. Firm profits as iceberg costs coefficient changes from 0.1 to 1.

Figure 3 shows us, that increase in elasticity of substitution (σ) leads to decrease in number of firms and in their profit. It happens because in this case preference for variety also decreases, so consumers prefer to buy goods from neighbours and the farthest firms lose their profit.

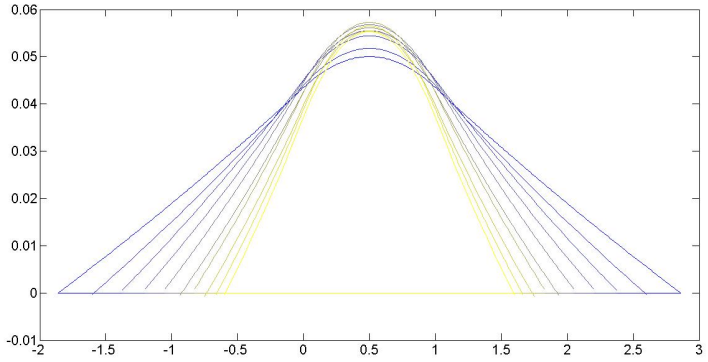


Fig. 3. Firm profits as σ changes from 1.5 to 3.5.

As we can see from figure 4, increase in number of workers gives us decrease in profit for firms. It happens because bigger part of all money in the city goes to wages for workers instead of going to profit for firms (total amount of money in the city is constant and equals to 1)

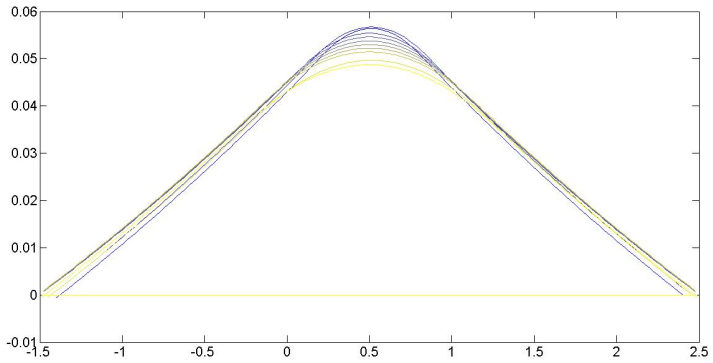


Fig. 4. Firm profits as L changes from 0.5 to 1.5.

4 Two-dimensional City

The generalisation of the model to the 2 dimensional case is straightforward — one just needs to replace the unit interval with some compact set in R^2 and the Euclidean distance with some other metric that reflects the structure of transportation within the city. The assumption that each point can host just one firm cam be relaxed with some more general capacity constraint, but then we will have also take into account that several varieties might be produced in a single point, instead of just one — technically, it adds just one more dimension to our model.

To adjust the model to the empirical estimation we assume that city is divided into M districts, residents are distributed among districts and so do the firms; each district has its own number of workers and a capacity constraint for firms. Also, firms may be located on the radial highways spanning outside from the borders of the city — we need this assumption to be able balance the number of firms with free-entry condition. All the calculations are rather similar to the continuous model, however, we use sums across districts to approximate the integrals across space.

The main difference lies in the form of Melitz-inspired cutoff level condition. We assume that firms, as they enter the market, first fill all the vacant offices inside the city, from best to worst, and then the remaining ones locate along the highways. So, if we have an exogenous value of the maximum number of firms in each district N_j, then we can write the cutoff level condition for each of the highways in linear form as in the one-dimensional model.

5 Finding real data for two-dimensional city

In our model we need to get some statistics about the city, such as distribution of workers, distribution of firms, distances between objects in the city. We get real data about Moscow, making detalization for districts. Distribution of workers we get from official statistics about population in Moscow. Distribution of firms we get by extrapolation of data from one non-official source, and distances between districts we estimated, using specially designed algorithm.

The results of model will be distribution of profit among all firms in city, value of wage and number of firms, which in model means length of highways. To compare these results with real situation, we also get statistics about profit for firms in all districts, using data about tax on profit from tax service.

6 Results for Two-dimensional City

Our next goal was to find out the values of exogenous parameters (elasticity of substitution, for example) such that the results of the model closer to real data.

We were unable to run the classical OLS, instead, we varied our exogenous parameters to make ratio between total profit of firms and total income of workers equal to the real ratio in Moscow. It happens for a set of parameters, so,

after that we tried to match the length of "occupied" highways and maximise the Spearman coefficient that shows correlation between lists of districts, sorted by profit for one firm, from model and from real data.

In the end, we get that in all sets f and σ are inversely dependent, τ and θ are close to 1, Spearman coefficient is close to 0.62 and length of highways are at most 25 km.

Also our results can be illustrated by figure 5, where districts are marked with blue, if profit for conventional unit of firms there is bigger than income for conventional unit of workers, else district is marked with yellow.

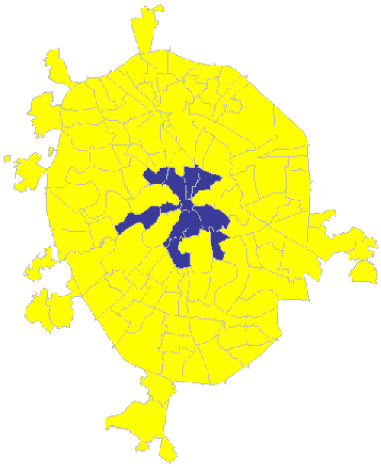
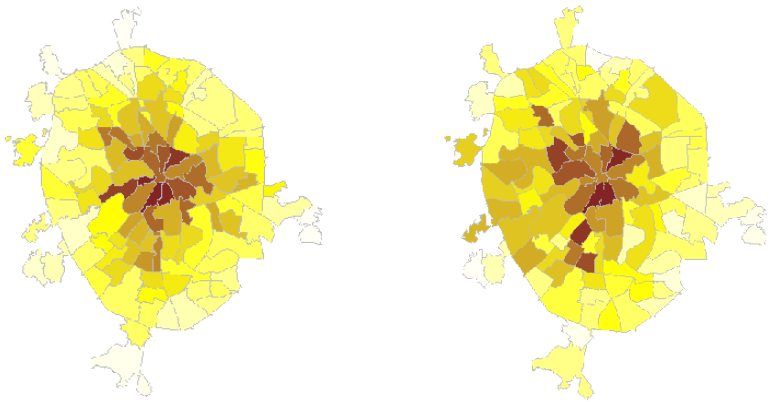


Fig. 5. Comparing profits of firms and income of workers

Fig. 6. Length of the highways

In figure 6 we illustrate length of highways that we get from the model.

And in figure 7 we compare distribution of profit for one firm between results from model and real data.



Model Real data

Fig. 7. Distribution of profit for firms

Conclusion

To summarize, we see our main contribution in developing a model that is both simple enough to be tractable and scalable enough to allow estimations using real city data. The key feature is our ability to incorporate real city travel time costs into a classical new economic geography mathematical framework. This model can be used to address a variety of questions, from estimating economic advantages of a particular location to evaluating the best directions of a long-term city development. Though in this paper we mainly demonstrate the results of model simulations, our efforts lead way to testing the model against a reality of a modern city.

References

1. Dixit, Avinash K.; Stiglitz, Joseph E. Monopolistic Competition and Optimum Product Diversity // The American Economic Review. 1977, 67, 297–308.
2. Krugman P. R. Scale Economies, Product Differentiation, and the Patterns of Trade. // American Economic Review. 1980, 70, 950–959.
3. Krugman P. R. Increasing Returns, Monopolistic Competition, and International Trade. // Journal of International Economics. – 1980. – Vol. 9. – P. 469479
4. Fujita M., Krugman P., Venables A. J. The Spatial Economy: Cities, Regions, and International Trade. - Cambridge, Massachusetts: The MIT Press, 1999. - p.367
5. Marshall A. Principles of Economics. 1890 - 1891.
6. Melitz M. The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. // Econometrica. 2003, 71, 1695–1725.
7. Picard, P. M.; Tabuchi T. On microfoundations of the city. Journal of Economic Theory. 2013, 148, 2561–2582.
8. Picard P. M., Tabuchi T. City with forward and backward linkages. IEB Discussion Paper, 2010. – 34.
9. Weber A. Theory of Industrial Location. – M., 1926.
10. Samuelson P., Nordhaus W. Economics, 19 e. – M.:Williams, 2014.
11. Combes P.-P.., Mayer T., Thisse J.-F. Economic Geography. The Integration of Regions and Nations. // Princeton University Press. 2008, p.82.

Gift Ratios in Laboratory Experiments

Rustam Tagiew[1] and Dmitry I. Ignatov[2]

[1] POLAREZ ENGINEERING, Dresden, Germany,
Alumni of TU Freiberg and Uni Bielefeld, Germany
rustam.tagiew@polarez.com, http://www.polarez.com
[2] National Research University Higher School of Economics, Moscow, Russia
dignatov@hse.ru, https://www.hse.ru/en/staff/dima

Abstract. This paper presents statistics of a controlled laboratory gift-exchange-game experiment. These numbers can be used for assumptions about human behavior in analysis of noisy web data. The experiment was described in 'The Impact of Social Comparisons on Reciprocity' by Gächter et al. 2012. As already shown in relevant literature from experimental economics, human decisions deviate from rational payoff maximization. The average gift rate was 31%. Gift rate was under no conditions zero. Further, we derive some additional findings and calculate their significance.

1 Introduction

As our experience shows [1], extraction of knowledge from noisy industrial datasets requires reasonable assumptions. Data analysis results extracted from clean data of laboratory experiments can help to create these assumptions. Market leaders in Big Data, as Microsoft, Facebook, and Google, have already realized the importance of experimental economics know-how for their business [2][3][4].

Before vast data and computational power were available, classical economists used game theory to predict outcomes of human interactions. People were assumed to be intelligent and autonomous, and to act pursuant to their existing preferences. It is important to underline that game theory is a mathematical discipline, whose task was never to define human preferences, but to calculate based on their definition. A preference is an order on outcomes of an interaction. One can be regarded as rational, if one always makes decisions, whose execution has referred to subjective estimation the most preferred consequences [5,6]. The level of intelligence determines the correctness of subjective estimation. Beyond justifying own decisions, rationality is a base for predictions of other people's decisions. If the concept of rationality is satisfied, and applied mutually, and even recursively in a human interaction, then the interaction is called strategic. Game is a notion for the formal structure of a concrete strategic interaction [7].

A definition of a game consists of a number of players, their preferences, their possible actions and the information available for the actions. A payoff function can replace the preferences under assumed payoff maximization. The payoff function defines each player's outcome depending on his actions, other players' actions and random events in the environment. The game-theoretic solution of a game is a prediction about the

behavior of the players also known as an equilibrium. The basis for an equilibrium is the assumption of rationality. Deviating from an equilibrium is outside of rationality, because it does not maximize the payoff according to the formal definition. There are games, which have no equilibria. At least one mixed strategies equilibrium is guaranteed in finite games [8].

In common language, the notion of game is used for board games or video games. In game-theoretic literature, it is extended to all social, economical and pugnacious interactions among humans. A war can be simplified as a board game. Some board games were even developed to train people, like Prussian army war game 'Kriegspiel Chess' [9] for their officers. We like it to train in order to perform better in games [10]. In most cases, common human behavior in games deviates from game-theoretic predictions [11,12]. One can say without any doubt that if a human player is trained in a concrete game, he will perform close to equilibrium. But, a chess master is not necessarily a good poker player and vice versa. On the other side, a game-theorist can find a way to compute an equilibrium for a game, but it does not make a successful player out of him. There are many games we can play; for most of them, we are not trained. That is why it is more important to investigate our behavior while playing general games than playing a concrete game on expert level.

Although general human preferences are a subject of philosophical discussions [13], game theory assumes that they can be captured as required for modeling rationality. Regarding people as rational agents is disputed at least in psychology, where even a scientifically accessible argumentation exposes the existence of stable and consistent human preferences as a myth [14]. The problems of human rationality can not be explained by bounded cognitive abilities only. '... people argue that it is worth spending billions of pounds to improve the safety of the rail system. However, the same people habitually travel by car rather than by train, even though traveling by car is approximately 30 times more dangerous than by train!'[15, p.527–530] Since the last six decades nevertheless, the common scientific standards for econometric experiments are that subjects' preferences over outcomes can be insured by paying differing amounts of money [16]. However, insuring preferences by money is criticized by tossing the term 'Homo Economicus' as well.

The ability of modeling other people's rationality and reasoning as well corresponds with the psychological term 'Theory of Mind' [17], which lacks almost only in the cases of autism. For experimental economics, subjects as well as researchers, who both are supposed to be non-autistic people, may fail in modeling of others' minds anyway. In Wason task at least, subjects' reasoning does not match the researchers' one [18]. Human rationality is not restricted to capability for science-grade logical reasoning – rational people may use no logic at all [19]. However, people also make serious mistakes in the calculus of probabilities [20]. Even in mixed strategy games, where random behavior is of a huge advantage, the required sequence of random decisions can not be properly generated by people [21]. Due to bounded cognitive abilities, every human 'random' decision depends on previous ones and is predictable in this way. In ultimatum games [12, S. 43ff], the former economists' misconception of human preferences is revealed – people's minds value fairness additionally to personal enrichment. Our minds originated from the time, when private property had not been invented and social

values like fairness were essential for survival.

From the view point of data scientists fascinated by human behavior, the sizes of datasets originated from social networks predominate the ones from experimental economics by orders of magnitude [12]. Nevertheless, analyzing data from experimental economics has the same importance for understanding human psychology as studying Escherichia for understanding human physiology. Data from experimental economics has the advantage of originating from simple and controlled human interactions.

In current experimental economics, the models are first constructed by philosophical plausibility considerations and then are claimed to fit the data. In this work, we reverse the order of common research in experimental economics. We follow the slogan 'existence precedes essence' – the philosophical plausibility considerations follow after the correlations and regularities are found. For these needs, we analyze the dataset of the paper "The Impact of Social Comparisons on Reciprocity" by Gächter et al. [22]. The only assumption about human behavior is its determinism.

The next section summarizes related work on data mining approaches and economical models. Then, the experiment setup and the gathered data are introduced. Before extracting rules of behavior, we explain the reasons for the assumption of determinism. We also explain conceptual problems of using linear model on this data. The results and their interpretations follow afterwards. Then, a section is devoted to p-hacking. A suggestion for more efficient research on human behavior is made in future work. Summary and discussion conclude this paper.

2 Related Work

A similar approach is already explored on three datasets – a zero-sum game of mixed strategies, an ultimatum game and repeated social guessing game [23,24]. For these datasets, extracted deterministic regularities outperformed state-of-art models. It was shown that some regularities can be easily verbalized, what underlines their plausibility.

A very comprehensive gathering of works in experimental psychology and economics on human behavior in general games can be found in [25]. Quantal response equilibrium became popular as a model for deviations from equilibria [26]. It is a parametrized shift between mixed strategies equilibrium and an equal distribution. The basic idea for quantal response equilibrium is the concept of trembling hand – people make mistakes with certain probability. Unfortunately, the Akaike information criterion [27] is rarely calculated to judge the trade-off between fit quality and model complexity [28]. Another popular model is the linear regression. It is used in the original paper to model the dataset [22]. For linear regression, data is translated into real numbers.

3 Gift-Exchange-Game

Since Akerloff and Yellen published their leading work [29] on unemployment, gift-exchange-games (GEG) became standard for modeling labor relations. Such a game

involves at least two players – an 'employer' and an 'employee'. The 'employer' has to decide first, whether to award a higher salary or not. Then, the 'employee' has to decide, whether to put extra effort or not. Unfortunately, the experiment conducted by Gächter et al. did not implement a real-effort task. The 'employee' does not put real effort, but can decide to make a gift, which reduces his/er own payoff. Nevertheless, this game is not zero-sum. For what it's worth, real-effort tasks are already established in experimental economics – in works of Ariely e.g. [30]. Therefore, we refuse to draw any inferences from the behavior in the experiment to the behavior in real labor relations. The 'employer' is renamed to originator and 'employee' to follower. If the originator and the follower are both only interested in maximizing their payoff in a pure monetary case and it is a one-shot game, the actual gift exchange will not take place.

The experiment was conducted at University of Nottingham and consisted of one-shot games, whereby no subject participated twice. The participants were 20 years old in average and of both genders. Every one-shot game involves three players – one originator and two followers. The originator has the choice to award none, one or both followers. The followers have four levels of rewarding (including non-rewarding) the originator. In the original game description, the originator and every follower have to give at least a minimum gift, which we denote non-gift for simplicity. At the beginning, the originator gets £8.3 and every follower gets £11.1. The additional payoff of the originator is the sum of margins from gift exchange with both followers. The additional payoff of every follower is the gift of the originator minus reduction through own gifts. The originator can give a fixed amount of £1.6 to a follower. A follower can give £1, £2 or £3, whereby his/er payoff reduces by £0.5, £1 or £1.5 accordingly.

We split the 3-players game into two 2-players games. Fig.1 shows the 2-players game between an originator and a follower in extensive form. Extensive form is known in AI as game tree. The originator has to decide for two of such 2-players games. After the originator makes his choice, the followers make their choices either sequentially or simultaneously. Every follower can observe both of originators' decisions. In the sequential case, first follower's decision can be seen by the second follower. Besides mutual visibility, both 2-players games are independent. Adding both games, the originator's total payoff ranges between £5.1 and £14.3. The follower's total payoff ranges between £9.6 and £12.7.

4 Dataset

123 subjects participated in the game – 84 for the sequential case and 39 for the simultaneous case. $\frac{123}{3} = 41$ originators have made $41 \times 2 = 82$ decisions – two 2-players games per originator. The follower were asked to submit their decisions for every possible combination of others' observable decisions. There are 4 decision combinations for an originator. First followers in the sequential case submitted $4 * \frac{84}{3} = 112$ and all followers in the simultaneous case submitted $4 * 2 * \frac{39}{3} = 104$ decisions. Second followers in the sequential case submitted $4 * 4 * \frac{84}{3} = 448$ decisions. Therefore, we have a dataset of total 746 human decisions.

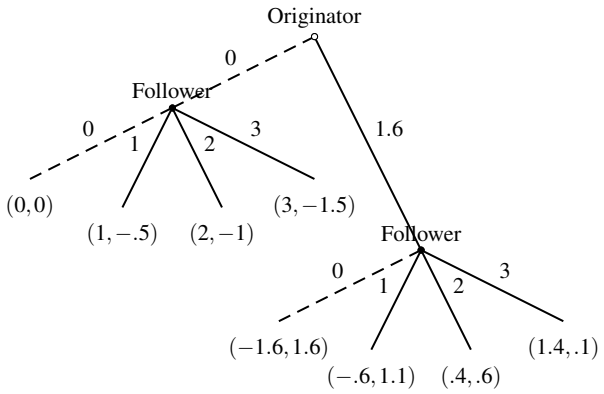


Fig. 1. Experimental non-zero-sum 2-players GEG in extensive form. (Originator's payoff, Follower's payoff) – payoffs are in £. Payoff maximizing equilibrium is marked by dashed lines.

5 Assumption of determinism

Modeling human behavior outside of game playing with human subjects should not be confused with prediction algorithms of artificial players. Quite the contrary, artificial players can manipulate the predictability of human subjects by own behavior. For instance, an artificial player, which always throws 'stone' in roshambo, would success at predicting a human opponent always throwing 'paper' in reaction. Otherwise, if an artificial player maximizes its payoff based on opponent modeling, it would face a change in human behavior and have to deal with it. This case is more complex than a spectator prediction model for an 'only-humans' interaction. This work is restricted on modeling behavior without participating.

Human behavior can be modeled as either deterministic or non-deterministic. Although human subjects fail at generating truly random sequences as demanded by mixed strategies equilibrium, non-deterministic models are especially used in case of artificial players in order to handle uncertainties.

'Specifically, people are poor at being random and poor at learning optimal move probabilities because they are instead trying to detect and exploit sequential dependencies. ... After all, even if people don't process game information in the manner suggested by the game theory player model, it may still be the case that across time and across individuals, human game playing can legitimately be viewed as (pseudo) randomly emitting moves according to certain probabilities.' [31] In the addressed case of spectator prediction models, non-deterministic view can be regarded as too shallow, because deterministic models allow much more exact predictions. Non-deterministic models are only useful in cases, where a proper clarification of uncertainties is either impossible or costly. To remind, deterministic models should not be considered to obligatory have a

formal logic shape.

6 Nominal, Ordinal or Numeric

The usage of right data types is essential for correct data analysis. There are basically three categories, in which variables can be classified – nominal, ordinal and numeric. Nominal variables assume values from a finite set, which has no order. Ordinal variables are like nominals plus ordering relationship over the set of values. Numeric variables assume real numbers $\mathbb{R}$ as values. Ordinal values can be projected into numeric under assumption about their distribution over the number axis. In contrast, nominal values can not.

Some variables, which impact human actions, are actions of other players. Since presuming human preferences over the outcomes has no base, an ordering relationship over the actions can not be presumed as well. In the addressed problem, all variables impacting human actions are actions of others. Preferences over outcomes in the earlier described GEG can not be presumed. For instance, the outcomes $(.4, .6)$ and $(1.4, .1)$ (Fig.1) are the total payoffs $(£8.7, £11.7)$ and $(£9.7, £11.2)$. An egoistic follower would prefer the first and altruistic one the second. Since the variables have to be nominal and not even ordinal, they can not be projected into real numbers. An application of a linear model as in the original paper [22] becomes therefore nonsense for this data.

7 Results

Originator's both decisions are nominal or rather boolean – it is either a gift or not. In average, originators gift in 36.6% of samples. We calculate Kappa [32,33] to measure the inter-rater agreement between these two decisions. Having zero Kappa as null hypothesis, the significance of the measured Kappa can be calculated. Tab.1 displays significant fair agreement between originator's both decisions in the sequential case. There is no significant agreement between them in the simultaneous case. Unfortunately, the data is too marginal and Fisher's test [34] does not show any significant difference between the frequencies in both cases – p-value is 0.4686. We can at least claim that both decisions are dependent in the sequential case.

Tab.2 shows absolute statistics for the follower's decisions, which did not observe another follower. Besides own received gift, there is the gift received by the other follower, which might have an impact on the observing follower's decision. If no own gift is received in the simultaneous case, Fisher's test results a p-value of 0.0496 for gifting $>£0$ depending on whether or not the other follower received a gift. Receiving less than the other follower is therefore significantly reciprocated in the simultaneous case only. The significance of this result is thoroughly discussed in section 8. In the sequential case, there is no significant difference between decision frequencies depending on the other's received gift. Obviously, the order between the follower delivers a reason for an unequal treatment.

Table 1. Originator's two decisions – absolute statistics and agreements.

	Sequential	Simultaneous	Sum
Gifts for 1st	9	6	15
Gifts for 2nd	10	5	15
All samples	28	13	41
Kappa	.4432	.2169	.3692
p-value	.017	.22	.014

Table 2. Follower's decision without observing another follower's decision – absolute statistics.

		Sequential				Simultaneous			
	Own gift	£0	£1	£2	£3	£0	£1	£2	£3
Received gifts									
Own	Other's								
£0	£0	21	6	1	0	19	7	0	0
£0	£1.6	21	6	1	0	25	0	1	0
£1.6	£0	13	6	5	4	16	6	4	0
£1.6	£1.6	13	6	4	5	16	3	4	3

Tab.3 lists agreements as well as their significances between subsets of own decisions and observed decisions. The subsets of own decisions are defined by thresholds on gift size. We use thresholds to define subsets, because based on decreasing frequencies by raising gift size (Tab.2), an order on the gift decisions can be derived. One can see that only own received gift has a significant influence on own decision in sequential as well as in simultaneous cases. Since only one variable has influence on the decision, the deterministic model is trivial – gift $>$£0 in the sequential case having received a gift, non-gift anywhere else.

As Fig.2 shows, non-gift covers $\geq 50\%$ of decisions for the second follower for all possible combinations of input variables and 71.2% in average. One can of cause assume that some hidden variables influence the gift decision. Since we do not see these variables, we can not build a useful valid deterministic model – none can be better than the null hypothesis suggesting non-gift everywhere. We restrict the analysis to agreements between the decision and the three observed variables.

Tab.4 shows the agreements between subsets of the second follower's decisions and the decisions of the originator – none of them is significant. It also shows the agreement between the decisions of the first and the second follower, whereby correspondence between the values' sets of both variables is assumed. This agreement is not significant. Therefore, the value sets of both variables have to be transformed. Tab.5 shows agreements between both variables transformed to booleans in different combinations. The highest agreement and the lowest p-value are achieved for the first follower's gift $>$£1

Table 3. Follower's decision without observing another follower's decision – agreements with originator's decisions.

	Sequential		Simultaneous	
	Kappa	p-value	Kappa	p-value
Own gift ...	... vs. received gift			
>£0	.2857	.0012	.2308	.0093
>£1	.2857	.0012	.1923	.0249
>£2	.1607	.045	.0577	.2781
	... vs. other's received gift			
>£0	0	.5		
<£1			.1154	.1197
>£1	0	.5	.0769	.2164
>£2	.0179	.4251	.0577	.2781

and the second one's >£0. Once the first follower is extra generous, the second one is also driven to gift the originator.

To summarize, the frequency of non-gift is 63.4% for the originator, 60.7% for the first sequential follower, 73% for the simultaneous follower and 71.2% for the second sequential follower. According to Fisher test, none of these frequencies significantly deviates from the rest. The average non-gift frequency is 69%.

8 On p-hacking

It is not a secret that p-values closely under 0.05 cause suspicion about the scientific methods used in research [35]. Although p-value was never thought to be an objective criterion for proof or disproof of a hypothesis, many researchers misunderstand it and conduct the so called 'p-hacking' on the data to archive significant results.

The results achieved through p-hacking might not be reproducible, since the a-priori probabilities of the hypotheses have to be incorporated as well. For instance, if the hypothesis is a long shot and has an a-priori probability of 5%, a p-value of 0.01 raises the chance of its validity to only 30%. The more hypotheses are tested on p-value, the higher the probability to achieve a p-value under 0.05. Obviously, the difference between long shots and good bets has to be derived from the researcher's expert knowledge, which is known to be absent in the case of an pure data scientist analyzing human data. Here, we suggest the data scientist to be 'agnostic' and use some background knowledge to advocate the result.

During the data analysis in this paper, we got a p-value of 0.0496 for the hypothesis that an unfairly treated simultaneous follower negatively reciprocates. Using the background knowledge about human reaction to unfairness [36], we can assume that it is a good bet. If the a-priori probability for a good bet is assumed to be about 90%, a

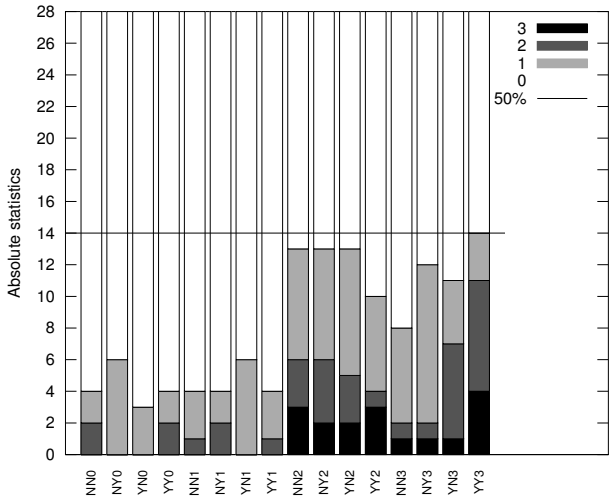


Fig. 2. Choice of the second follower – absolute statistics depending on other players' decisions encoded on x-axis as: other follower's received, own received and other follower's given. N is £0 and Y £1.6.

p-value of 0.05% raises its chance of validity to 96%.

9 Future work

During the work on this paper, we confronted the time consuming requesting, selection and reformatting of data. Unfortunately, there is no online portal, where most of the datasets are offered in a common format. This is an issue, which we will address in the future. Like in the field of bioinformatics, common formats are an important part of an interdisciplinary research infrastructure and are needed to accelerate the progress [37].

As for methodological aspects of Machine Learning in the context of Experimental Economics, we would like to use the advanced pattern mining techniques for economic game data analyses. For example, in papers [38,39] was made an attempt to use sequential patterns and similarity dependencies on pattern structures for video game players' behavior analysis, in particular sequential attribute dependencies might be a tool of choice. We will try to apply sequential pattern mining in a supervised task, where the outcome of a game (or a turn) is a target attribute [40,41] to see which patterns better generalize the user behavior. These experiments are able not only to broad the tools of experimental economics, but also help to reveal potentially new knowledge of human behavior in games based on sequential pattern description.

Table 4. Second follower's decision – agreements.

	Sequential case	
	Kappa	p-value
Own gift ...	... vs. received gift	
>£0	.0223	.3183
>£1	.0223	.3183
>£2	.0134	.3884
	... vs. other's received gift	
>£0	.0045	.4624
>£1	.0402	.1975
>£2	.0134	.3884
	... vs. other's gift	
	.0446	.0509

Table 5. First and second follower's decision – agreements between subsets.

1st follower	>£0		>£1		>£2	
	Kappa	p-value	Kappa	p-value	Kappa	p-value
2nd follower						
>£0	.1123	.0016	.2634	$1.238e-8$	.1445	.0068
>£1	.0564	.0365	.1563	.0005	.1345	.029
>£2	.026	.1826	.0759	.0541	.0456	.2789

10 Conclusion

First of all, the average non-gift frequency is only 69% in the studied one shot GEG. These are far away from the 100%, which an egoistic payoff maximization assumption would predict. But, it is also over 50% in almost all cases. There, it is impossible to create valid nontrivial deterministic models of human behavior without having access to the hidden variables, which determine the choice. Only if the first follower receives a gift in the sequential case, the frequency of gifts goes slightly over 50%.

Although first follower's decision depends only on his received gift and second follower's decision does not depend on originators' decisions at all, originators' decisions are interdependent in the sequential case. The order between the players obviously delivers a reason for the first follower to not mind differences in gifts. Having no order in the simultaneous case leads to significant negative reciprocation of receiving less than the second follower.

A curious finding is that not minimal but extra generosity is 'contagious' for the followers. Second follower reacts only on the first follower being extra generous and

with normal generosity.

Acknowledgment The authors would like to thank Martin Sefton for the friendly reply and the transfer of data. We also thank the people, who provided the Weka library, for their wonderful work, as well as Minato Nakazawa for the fmsb package.

References

1. Tagiew, R., Ignatov, D.I., Delhibabu, R.: Economics of internet-based hospitality exchange. In: WI-IAT 2015, Singapore. Volume I., IEEE Computer Society (2015) 493–498
2. Bailey, M.: What does an economist at facebook do? quora.com/What-does-an-economist-at-Facebook-do (2014)
3. Microsoft Inc.: Microsoft research new york city. research.microsoft.com/en-us/labs/newyork/ (2014)
4. Varian, H.R.: Big data: New tricks for econometrics. people.ischool.berkeley.edu/ hal/Papers/2013/ml.pdf (2013)
5. Russel, S., Norvig, P.: Artificial Intelligence. Pearson Education (2003)
6. Osborne, M.J., Rubinstein, A.: A course in game theory. MIT Press (1994)
7. Morgenstern, O., von Neumann, J.: Theory of Games and Economic Behavior. Princeton University Press (1944)
8. Nash, J.: Non-cooperative games. Annals of Mathematics (54) (1951) 286 – 295
9. Li, D.H.: Kriegspiel: Chess Under Uncertainty. Premier (1994)
10. Genesereth, M.R., Love, N., Pell, B.: General game playing: Overview of the aaai competition. AI Magazine **26**(2) (2005) 62–72
11. Pool, R.: Putting game theory to the test. Science **267** (1995) 1591–1593
12. Camerer, C.F.: Behavioral Game Theory. Princeton University Press (2003)
13. Stevenson, L., Haberman, D.L.: Ten Theories of Human Nature. OUP USA (2004)
14. Bazerman, M.H., Malhotra, D.: Economics wins, psychology loses, and society pays. In De Cremer, D., Zeelenberg, M., Murnighan, J.K., eds.: Social Psychology and Economics. Lawrence Erlbaum Associates (2006) 263–280
15. Eysenck, M.W., Keane, M.T.: Cognitive Psychology: A Student's Handbook. Psychology Press (2005)
16. Chamberlin, E.H.: An experimental imperfect market. Journal of Political Economy **56** (1948) 95–108
17. Verbrugge, R., Mol, L.: Learning to apply theory of mind. Journal of Logic, Language and Information **17** (2008) 489–511
18. Wason, P.C.: Reasoning. In Foss, B.M., ed.: New horizons in psychology. Penguin Books (1966) 135–151
19. Oaksford, M., Chater, N.: The probabilistic approach to human reasoning. Trends in Cognitive Sciences **5** (2001) 349–357
20. Kahneman, D., Slovic, P., Tversky, A.: Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press (1982)
21. Kareev, Y.: Not that bad after all: Generation of random sequences. Journal of Experimental Psychology: Human Percetion and Performance **18** (1992) 1189–1194
22. S. Gächter, D. Nosenzo, M. Sefton: The impact of social comparisons on reciprocity. The Scandinavian Journal of Economics **114**(4) (2012) 1346–1367

23. Tagiew, R.: Mining determinism in human strategic behavior. In: EEML, KU-Leuven (2012) 86–91
24. Tagiew, R., Ignatov, D., Amroush, F.: Social learning in networks: Extraction of deterministic rules. In: Experimental Economics and Machine Learning. ICDM Workshops, IEEE Computer Society (2013)
25. Plott, C.R., Smith, V.L., eds.: Handbook of Experimental Economics Results. North-Holland (2008)
26. Richard, M., Palfrey, T.: Quantal response equilibria for normal form games. Games and Economic Behavior **10** (1995) 6–38
27. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Int. Symp. on Information Theory, Akadmiai Kiad (1971) 267–281
28. Tagiew, R.: Strategische Interaktion realer Agenten: Ganzheitliche Konzeptualisierung und Softwarekomponenten einer interdisziplinren Forschungsinfrastruktur. PhD thesis, TU Bergakademie Freiberg (2011)
29. Akerlof, A.G., Yellen, J.L.: The fair wage-effort hypothesis and unemployment. Quarterly Journal of Economics **105** (1990) 255–283
30. Ariely, D., Kamenica, E., Prelec, D.: Man's search for meaning: The case of legos. Journal of Economic Behavior & Organization **67**(3-4) (2008) 671–677
31. Rutledge-Taylor, M.F., West, R.L.: Cognitive modeling versus game theory: Why cognition matters. In: ICCM. (2004) 255–260
32. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20** (1960) 37–46
33. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5) (1971) 378382
34. Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society **85**(1) (1922) 87–94
35. Nuzzo, R.: Scientific method: Statistical errors. Nature **506** (2014) 150–152
36. Rabin, M.: Incorporating fairness into game theory and economics. American Economic Review **83** (1993) 1283–1302
37. Nakamura, H., Date, S., Matsuda, H., Shimojo, S.: A challenge towards next-generation research infrastructure for advanced life science. New Generation Computing **22** (2004) 157–166
38. Low Cam, C., Rassi, C., Kaytoue, M., Pei, J.: Mining Statistically Significant Sequential Patterns. In George Karypis, H.X., ed.: IEEE International Conference on Data Mining (ICDM), IEEE Computer Society (December 2013)
39. Baixeries, J., Kaytoue, M., Napoli, A.: Computing similarity dependencies with pattern structures. In Ojeda-Aciego, M., Outrata, J., eds.: CLA 2013, Proceedings of the Tenth International Conference on Concept Lattices and Their Applications, University of La Rochelle (2013) 33–44
40. Kuznetsov, S.O.: Machine learning and formal concept analysis. In: Concept Lattices. Volume 2961 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2004) 287–312
41. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Rassi, C.: On projections of sequential pattern structures (with an application on care trajectories). In Ojeda-Aciego, M., Outrata, J., eds.: CLA 2013, Proceedings of the Tenth International Conference on Concept Lattices and Their Applications, University of La Rochelle (2013) 199–210

Churn Prediction for Game Industry Based on Cohort Classification Ensemble

Evgenii Tsymbalov[1,2]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Webgames, Moscow, Russia
etsymbalov@gmail.com
http://corpwebgames.com/en/

Abstract. In this paper, we present a cohort-based classification approach to the churn prediction for social on-line games. The original metric is proposed and tested on real data showing a good increase in revenue by churn preventing. The core of the approach contains such components as tree-based ensemble classifiers and threshold optimization by decision boundary.

Keywords: Churn prediction, ensemble classification, cohort-based prediction, on-line games, game analytics

1 Introduction

The churn prediction is a real problem, which can be found in businesses that deal with a permanent stream of customers using subscription services: banking [1,2], telecommunication [3], and entertainment industries, with increasing popularity of game analytics (e.g. [4]). The focus of business development shifts from the attraction of new customers to retention of the old ones. Therefore modeling users' outflow can be used to plan company's tactics and strategies. However, it is often important to not simply know the outflow indicators on a macro level, but to predict the churn probability for every customer to use the spot interventions.

The definition of churn and the churners varies depending on a specific problem. Here we define the churners as users who were absent for 30 days and more. Moreover, in this research we are interested in users who were absent for at least three days and made at least one transaction. Such restrictions are motivated by the low revenue of the late returners and recommendations from social platforms, e.g., Facebook does not recommend to send out notifications to the players with more than 28 days of absence.

In this paper, we propose a cohort-based classification approach to the churn prediction for social on-line games. Our data processing pipeline includes feature engineering and selection along with optimization of specific metrics. Thus, we introduce a meta-metric, which penalizes undesired outcomes of the cohort test procedure; it is designed to reflect real-life experience of using the prediction

models. All the steps of the model training pipeline include optimization of classifiers; the final step optimizes the whole ensemble on meta-metric values.

In Section 2, we introduce our cohort-based ensemble method for churn prediction. In Section 3, we provide the reader with the results of experimental evaluation. Section 4 concludes the paper.

2 Proposed Method

In a few words, the churn prediction problem can be stated as follows: given the data about users, who recently stopped playing, predict whether a particular user will abandon the game or not. This information is used for sending app-to-user notifications to those users, who are likely to stop playing our game at all. These notifications also offer some in-game goods to the user, usually some in-game currency, rare valuable items or discounts.

2.1 Meta-metric

This problem is rather different from the standard data mining classification problem due to several evaluation requirements, which result in so-called "cohort test". It is defined as follows (see Fig. 1 for short description):

1. We take the cohort of players who had been absent for 3 days by the given date D.
2. A model, that solves the classification problem for a given day of absence, predicts for every player, whether or not he or she becomes a churner.
3. Those, who were predicted as churners (both true and false churners), are eliminated from the next steps.
4. We take the cohort of players who had been absent have been absent for 4 days by the given date $(D+1)$ and had not been eliminated in the previous part.
5. Steps 2, 3, 4 are repeated until the date $(D+29)$ or no more users left in cohort.
6. We evaluate the meta-metric, using the information about successful and unsuccessful predictions at each step as well as users left after Step 5 (if there are any).

Final challenges are connected with the result of the cohort test and its influence on the game balance:

- We do not want to send prizes (bonuses) to people who are likely to return by themselves, because it may ruin the game balance.
- We want the players to be returned as soon as possible. The probability of the player's future transactions as well as the average revenue dramatically decreases with the growing number of days the player is absent for.
- We assume that we should not send notifications to the same player twice.

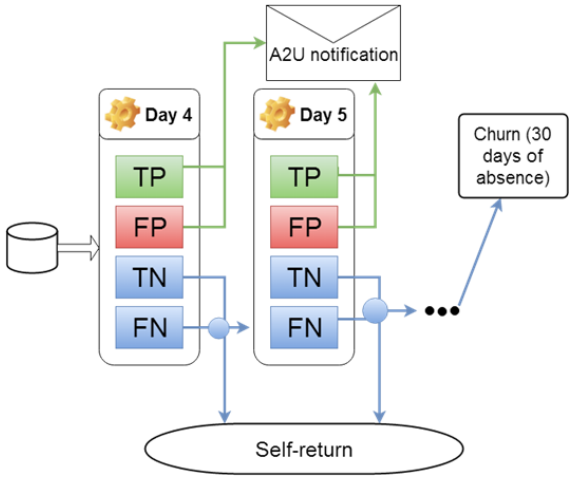


Fig. 1. The cohort test evaluation scheme. The users, defined by a classification algorithm as churners, are provided with app-to-user notification, while those users who remain go to the next step of the algorithm according to Eq. (2).

To answer these requirements, the cohort-based meta-metric is introduced:

$$MM = \sum_{day=4\ldots26} (\gamma^{day-4} TP_{day} - \alpha FP_{day}) - \beta(TP_{27} + FP_{27}), 0 < \gamma \leq 1 \quad (1)$$

This metric can be interpreted as a weighted number of returned players, with the reward for early return and the penalty for type I error (marking a user who will return as a churner). Note that different goods offered to a user with notifications, as well as various social platforms and projects require different parameters for the meta-metric.

The question of determining coefficients α, β, γ for a given project should be discussed at both levels, analytical and managerial. While α could be estimated by approximating the probability of a player being a payer after he or she returns, β and γ could be estimated by experts (like we do in Webgames now) or based on the data from the previous experiments (which is unavailable for this research).

2.2 Problem statement

Let $C^D_{date} = (P_1, \ldots, P_{N_{date}})$ be data of D-th day cohort of players on date $date$, where $D = 4 \ldots 26$, $P_j \in \mathbb{R}^{|F| \times D}$ are all player's features up to day $(D + date)$, F is the set of player's features at a particular day.

For every user in a cohort, our algorithm A makes a decision S: on which day $D = 4 \ldots 26$ send out the notification (or not to send, then $D = 27$):

$$A : (C^j_{date})_{j=4\ldots26} \longmapsto S = S(D_1, \ldots, D_N); \quad (2)$$

Our goal is to find the decision, which maximizes the meta-metric (1):

$$S = \arg\max_{S'} MM((C^j_{date})_{j=4...26}, S') \tag{3}$$

2.3 Ensemble of classifiers

To classify a user on a given day, we use the set of classifiers $\{C_4, C_5, \ldots, C_{26}\}$, one for the data for each day the user absents. So the classifier C_4 is trained on the users who have been absent for 3 days and evaluated on the 4-th day of absence; the classifier C_5 is trained on the users who have been absent for 4 days, etc. This results in total of 23 classifiers.

2.4 Optimization and greedy solution

Since the main problem requires the complex elimination procedure, therefore it seems to be complicated (if even possible) to optimize the meta-metric analytically. Instead, we use a greedy approach, with models for solving Problem (3) optimized by the cohort ensemble classifiers parameters in terms of ROC AUC metric, and these classifiers thresholds (decision boundaries) are optimized by the meta-metric.

3 Experiments

3.1 Data description

We used data from the Ghost Tales project on Facebook during the period 11.2015 – 03.2016, with the last month as a test subset. Hereby the features described below are mostly game-independent (for free-to-play games), so this feature set could be used for other projects as well.

Here we treat a player on his/her different days of absence as different players, which allows us to increase the number of objects in the dataset by more than 10 times, resulting in 2.7M records in total.

Typically, features could be divided into three main groups:

- Personal features: year of birth, country, etc.
- Behavioral features: the total number of active days for a player, the number of days a player is absent, the number of completed quests, etc.
- Transactional features (based on users payments).

3.2 Data processing

Our data processing includes the following steps:

1. Feature engineering. We have calculated pairwise ratios between all the features with the same measurement units (e.g. number of days with payments divided by total days for a given player) and added some other feature combinations based on expert knowledge.

2. Deletion of low variance features.
3. Selection of the best features based on F-test.
   The number of the best features selected may vary for different base classifiers. Here we used Extra Random Forest classifier [5] and logistic regression. We have also tried to use Gradient Boosting algorithms (XGBoost [6] and AdaBoost), but these classifiers shows similar or even worse results and do not support effective parallelization. We do not use algorithms based on neural nets due to their long training time, which is undesirable for weekly model retraining.

3.3 Single classifier optimization

We have performed 10-fold cross-validation procedure on the training set with optimization by ROC AUC score. The parameters used during the grid search are summarised below.

The list of ensemble trees' parameters (taken according to the real-life resource constraints):

- number of trees in ensemble: 50 – 500;
- tree depth: 5 – 20;
- minimal sample split: 2 – 15.

Logistic regression's parameters:

- penalty type: l1 or l2;
- C (regularization constant): 0.1 – 300.

The performance summary for selected values of the parameters of Extra Random Forest classifier is given in Table 1.

Table 1. Optimized parameters and classification metrics on the test set for the extra trees based classifier for selected days.

Day	Number of trees	Tree depth	Minimal sample split	Precision	Recall	F1-score
4	500	10	6	0.73	0.75	0.74
7	500	10	10	0.66	0.66	0.66
10	300	15	6	0.64	0.65	0
14	200	15	6	0.68	0.7	0.69
18	500	15	10	0.63	0.79	0.7

3.4 Ensemble optimization

The simplest method to continue optimization is to set decision boundaries (thresholds) for all the classifiers to the same value (which can be found via cross-validation, for example). However, it is easy to see that we should reduce

the decision boundaries of classifiers for several first days, because the users, who returned on the first days, have more impact on the meta-metric. Since 23-parameter optimization (for every classifier) is a computationally exhaustive problem, we reduced the number of parameters by considering the threshold as a function of day. We optimized the thresholds for different types of such functions:

- linear: $threshold = A \cdot day + B$;
- exponential: $threshold = A \cdot exp(B \cdot day)$;
- quadratic: $threshold = A \cdot day^2 + B \cdot day + C$ (where A, B, and C are constants)

using differential evolution as a global optimization algorithm. However, the optimization on various cohorts shows that exponential and quadratic approximations tends to reduce to linear, see Table 2.

Table 2. Optimal parameters found by differential evolution algorithm for the cohort of users absent from 2016-02-21.

Approximation function	A	B	C
Linear	0.4921	0.0141	-
Exponential	0.5416	0.0207	-
Quadratic	0.0002	0.0138	0.5344

The optimal parameters found by linear optimization results in 3-7% increase of the meta-metric, compared to the optimized parameters for the constant decision boundary, see Fig. 2.

4 Conclusion

In this research, we have proposed a cohort-based ensemble model for the churn prediction. The final part of this model includes optimization of the original meta-metric, which is designed to reflect the real-life experience in usage of prediction models that rely on a cohort test; other steps are also constructed by taking real-life resource constraints into account. Various numerical experiments show the importance of the steps used in the model training pipeline.

During this research, the algorithm for the weekly model evaluation has been developed and implemented in the Webgames company. Mostly automated, this algorithm requires human assistance only at the step of the meta-metric parameters' choice.

We assume it can be used in various areas for churn prediction. Thus, the obtained results form the groundwork for model improvement and generalization in other areas such as telecommunication and banking industries.

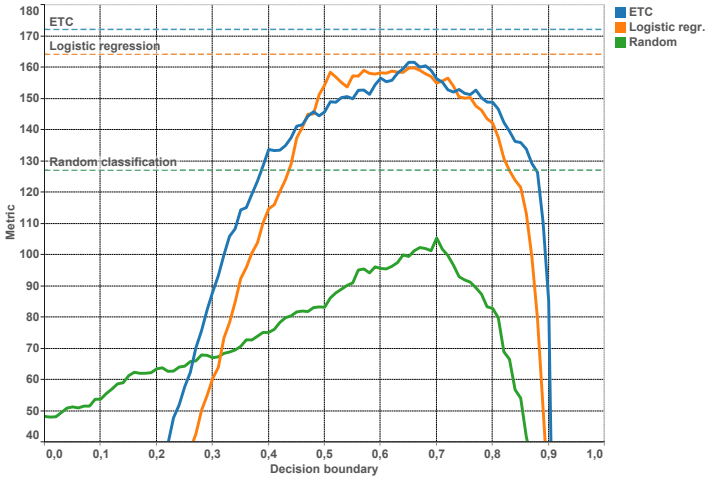


Fig. 2. Meta-metric values on α=0.8, β=1.2, γ=0.87. (The Ghost Tales on Facebook with diamonds prize) for the cohort absent from 2016-02-27. The solid line shows the meta-metric value for the case of the constant threshold, the dashed lines correspond to linearly optimized values of the metric.

Acknowledgment. We would like to thank Webgames' Analytics department for the provided data and feedback. We are also grateful to Mehdi Kaytoue from INSA-Lyon (France), and Evgeniy Sokolov and Peter Romov from Yandex Data Factory (Moscow, Russia) for their advice. Last but not least we would like to thank Dmitry Ignatov (HSE, Moscow) for his supervision.

References

1. Nie, G., Rowe, W., Zhang, L., Tian, Y., Shi, Y.: Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications **38**(12) (2011) 15273–15285
2. Mutanen, T., Ahola, J., Nousiainen, S.: Customer churn prediction-a case study in retail banking. In: Proc. of ECML/PKDD Workshop on Practical Data Mining. (2006) 13–19
3. Ahn, J.H., Han, S.P., Lee, Y.S.: Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy **30**(10) (2006) 552–568
4. Bosc, G., Kaytoue-Uberall, M., Raïssi, C., Boulicaut, J., Tan, P.: Mining balanced sequential patterns in RTS games. In: ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic. (2014) 975–976
5. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1) (2006) 3–42
6. Chen, T., He, T.: XGBoost: eXtreme Gradient Boosting. R package version 0.4-2 (2015)

Scientific Portal of University Department – Shaping Research Area of Users through their Behavior

Mikhail Navrotskiy, Nataly Zhukova

ITMO University, Saint Petersburg, 197101, Russian Federation,
m.navrotskiy@gmail.com

Abstract. Nowadays open access to scientific data in addition to obtaining information is an essential factor for educational and research processes. A lot of universities publish open data (include scientific and educational data), but not all of them is used.
The paper dwells on the scientific portal of a university department. It contains information about the university research activity (projects, publications, employees) that university employees could use for researches. To represent scientific and educational data the ontology model is developed. This model is based on the following vocabularies: FOAF, VIVO, BIBO and Teach. In addition, a case study of implementing the proposed solution at Computer Science and Applied Mathematics Department of ITMO University is presented.

Keywords: ontology, linked open data, data integration, data search, SPARQL.

1 Introduction

Researchers in their works often deal with the search of scientific information like scientific publications, research projects, educational courses and other, which is necessary for completing many tasks such as searching new scientific discoveries and colleagues progress. For this purpose they use resources such as Google Scholar, Academia.eu, BingAcademic or other publications bases, and these bases include a modern search function. On the other hand, it would be beneficial if researches had a single start point for research, and this system could find information in different sources. One of the solutions to this problem is the technologies of Semantic Web.

Semantic Web is a Web of machine readable data, where each data source can be used in different applications [1]. One of the best solutions for publishing related data is to use the technology of linked [2] data which is based on W3C standards and technologies such as RDF and OWL.

Universities play a major role in the development of this research area. Today leading universities of Europe and the USA are developing projects using Semantic Web principles and Linked Open Data (LOD):

- LODUM, a portal supported by the University of Munster, contains educational and scientific information in RDF format;
- Oxford University publishes the results of scientific activities;
- University of Southampton Open data portal contains the results of scientific activities;
- Harvard University produces publications;
- Aalto University develops the data portal which will store data about staff, courses, research projects, publications, buildings.

Researchers could use information from LOD universities portals for their researches but, it is necessary to develop a research portal that implements the search of scientific information on the open data portals of the universities. The accurancy of the result can be improved by using the description of researchers behaviour through their research interests.

2 Background

Leading European universities have already developed LOD portals (Linked Universities). Linked Universities is an alliance of European universities engaged into exposing their public data as linked data [3]. The organization includes universities such as The Open University, University of Southampton, University of Munster and others. Portals of linked universities publish scientific and educational data such as articles (results of research activities), research projects, open courses, events. In addition, a lot of data was published from libraries (access points to the libraries data are available at datahub.io). On the one hand, LOD portals publish data in a machine-readable format, but on the other hand, the scientific community does not have a single application for scientific and educational data search for the universities open portals and other open data portals.

Russian universities do not have LOD portals. Some universities publish open data but this information is published with non-Semantic Web formats.

Researchers, of course, can use full text search engines or databases such as Google Scholar, Scopus, Academia.eu (Russian researchers can also use elibrary[1]). However, these databases do not take into account the scientific interests of the user.

Thus, for developing the scientific search with the users research interests the authors should develop an ontology model. This model should include classes which will store links to data from data sources.

Portal's data sources include:

- university LOD portal (for this project it is a LOD of ITMO University [4]);
- external LOD portals (from other universities);
- external public libraries.

[1] http://elibrary.ru/default.asp

The Scientific portal under development has the following function: if the user finds the necessary content (a publication, a project, an employee or courses), the system displays similar content (publications, a project, an employee or courses) on the page with this content. The developed search algorithm simulates the behavior of a user. It implements search data in user's research interests and with keywords.

The search algorithm is the following:

1. the system defines the research interests of the current user;
2. the system defines a set of keywords of the current publication;
3. the system finds publications in the local ontology (maximum value of results is 5 but the user can change this value in settings);
4. if the number of results is lower than 5, the system finds publications in LOD portal of ITMO University (maximum value of results is 5 too);
5. if the number of results is lower than 5, the system finds publications in other data sources like in step 4.
6. the user can apply or cancel each publication which was found in steps 3 - 5. If the user selects to apply, relevance value is incremented by 1 otherwise relevance value is not changed. If this publication is not located in the local ontology, the system will add it.

3 Ontology Model for Scientific Portal

The developed ontology should, firstly, describe a domain, it should include classes of scientific publications, research projects, educational courses and university employees. Secondly, it should implement storage results of searching (links to data from other data sources). In addtition, it should implement requests for searching publications, projects, courses, employees by the user research area and keywords.

The authors used modern vocabularies for developing the ontology model: FOAF [8–10], VIVO [11, 12], BIBO [13, 14] and Teach [16].

- FOAF ontology was used to describe employees and research projects;
- VIVO ontology was used to describe a research area, an academic status, an academic degree and an academice qualification;
- BIBO onotology describes publications;
- Teach ontology describes educational courses.

In addition, the authors described classes, properties for searching, and this onotlogy was called LODIFMO.

In fig. 1 the authors have shown part of the ontology model. The part (fragment) shows classes and properties for the process of searching publications. A full ontology includes similar classes, properties for searching projects (foaf:Project), courses(teach:Course) and employees (foaf:Person).

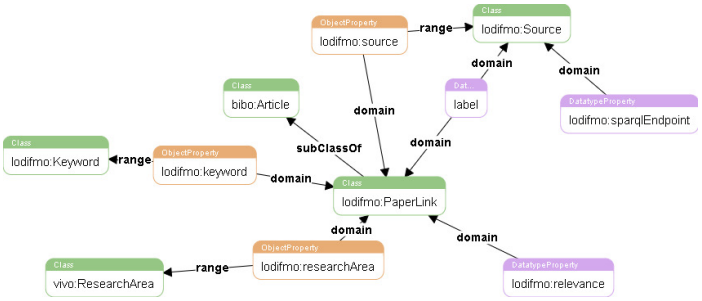


Fig. 1. Ontology model of scientific portal (Fragment)

A fragment of the ontology describes the following:

"Keyword" is a class which describes a keyword of a publication, a project, a course from the LODIFMO ontology.

"Article" is a class describing a scientific paper from the BIBO ontology.

"ResearchArea" is a class describing a research area of the user, a publication, a project, a course from the VIVO ontology.

"PaperLink" describes a scientific publicaton from the external data source (LODIFMO ontology).

"Source" is a class of a data source (e.g. LOD portal of ITMO University) from the LODIFMO ontology.

"label" Data Property stores a link to a publication (URI).

4 Profile of the Scientific Interests of the User

A model of the user is developed by the university corporate portal and includes general information, a description of his research activity and scientific interests. It is designed [15]:

$$P =< D, P_a, P_p, M, K, Apr > \tag{1}$$

where D is set of descriptive characteristics of the user including an academic degree, a status, a major, competence in languages, education, experience of academic advising, P_a is a set of publications, P_p is set of research-oriented projects the user has been taking part in or has supervised, M is a set of scientific events the user participated in, K is a set of users research interests, Apr is a set of scientific profiles.

The authors profile from scientometric databases is expressed in the following way:

$$Apr =< K_{ape}, P_{pr}, Ind_{pr} > \tag{2}$$

where P_{pr} is a set of publications in the authors profile, Ind_{pr} is an identifier of the authors profile.

The model of a publication is the following:

$$At =< Iz, Aat, Kat > \tag{3}$$

where Iz is a journal profile, Aat is a set of publications authors, Kat is a set of thematic charactersitics of publications.

The user imodel s filled with scientific interests in the following way:

$$K = K^p \smile K^a \tag{4}$$

where K^p is a set of keywords defined by the user, K^a is a set of the automatically identified keywords in compliance with their showing frequencies in the information system.

5 System Prototype

The System will perform search until the sufficient number of results is found to meet the users needs or until all the data sources are checked. The user can set the result size; a default value equals 5. For example, the users research interests include Semantic Web and Open Data. When the user searches publications on ontology, the system will not display publications related to philosophy, but only to Semantic Web.

If the results contain new objects not stored in the local ontology, the links to these objects are added to the local ontology with keywords, links to the research area and with URI from data source.

The first example is a request for search publications by a keyword and scientific interest in the local ontology:

```
SELECT *
WHERE
{
  ?publication a lodifmo:PaperLink .
  ?publication rdf:label ?link .
  ?publicaiton lodifmo:researchArea ?ra .
  ?ra rdf:label ?ra_label .
  ?publicaiton lodifmo:keyword ?keyword .
  ?keyword rdf:label ?keyword_label .
  FILTER(STR(?ra_label)="Semantic Web") .
  FILTER(STR(?keyword_label)="linked open data") .
}
```

Then system uploads data about a paper using a programming language.

The second example is a request to add a new link to the scientific object with a keyword:

```
INSERT DATA INTO GRAPH
<http://www.semanticweb.org/mikhailnavrotskiy/ontologies/2016/3>
```

```
{
  <bibo:Publication_137> lodifmo:link
    "http://svn.aksw.org/papers/2015/SEMANTICS_ITMOLOD_DEMO/public.pdf" .
  <bibo:Publication_137> lodifmo:keyword
    <lodifmo:Keyword_linked_open_data> .
  <bibo:Publication_137> vivo:researchArea
    <http://lod.ifmo.ru/ResearchArea_SemanticWeb > .
}
```

The last example is a request to search a paper by a keyword in an external source:

```
SELECT *
WHERE
{
  ?publication a <http://purl.org/spar/fabio/PosterPaper> .
  ?publication rdf:label ?title .
  FILTER(regex(str(?title), ".*(ontology)|(semantic\sweb).*")) .
}
```

The features of the scientific portal include searching different LOD sources such as AKSW LOD portal, LOD portal of ITMO University and extending datasets of local ontology with new data from external data sources.

The developed scientific portal presents a web application. The client is an HTML5 application. The server is a Rack application [6, 7] written with Sinatra (ruby web framework).

The server consists of the following modules:

1. Search Module is a module giving access to ontologies with SPARQL requests;
2. Ontology Module is a module giving access to ontology (the authors use INSERT SPARQL requests);

In Fig. 2 the system architecture is presented.

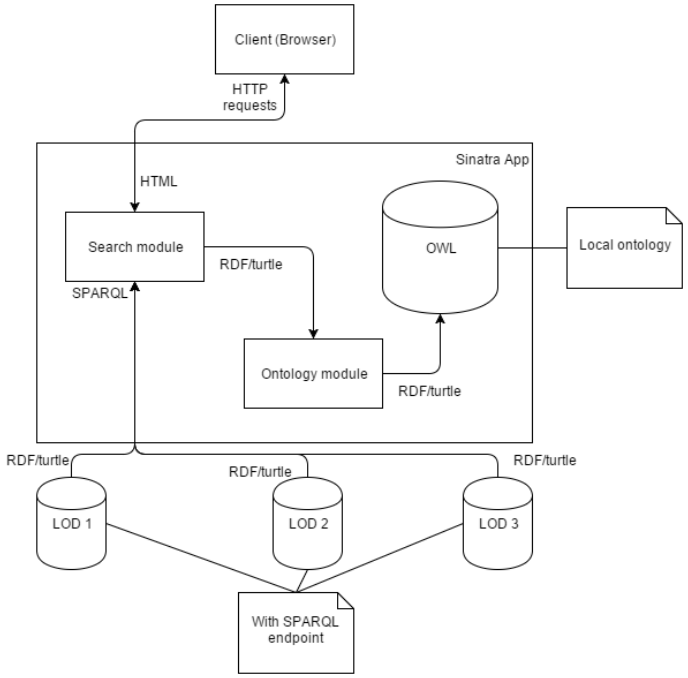


Fig. 2. Scientific portal architecture

For example, the user (researcher) searches a publication about linked open data. And they find an article about linked open data in ITMO university (Fig. 3). The system searches an article with the keywords from the current article: ontology, RDF, linked open data, data integration, data publishing; and with the user research interests: linked data, open data, sematic web. Thus, the input data are: research interests of the user (linked data, open data, ontology) and the article keywords (ontology, RDF, linked open data, data integration, data publishing) and output data are: two articles ("Towards the Linked Russian Heritage Cloud: Data Enrichment and Publishing and Metadata Extraction from Open edX Online Courses Using Dynamic Mapping of NoSQL Queries"). The Figure presents the following:

- (A) The current publication description area;
- (B) A description card with publications found through the user research areas and current publication keywords;
- (C) If a publication is correct, then the user can "apply" this result and relevance value of this artilce will be incremented;
- (D) If a publication is not correct, then the user can "cancel" this result and relevance value of this article will not be changed.

Fig. 3. Example output of search

The authors developed the ontology model of the scientific portal which is used for the search process based on keywords and the users scientific interests. The developed ontology is used to store links to scientific and educational objects (projects, publications, employees, courses).

This portal is used in the work of graduate students and employees of the Computer Science and Applied Mathematics Department of ITMO University and the portal is in the status of beta testing now.

6 Conclusion and Future Work

The system was deployed on ITMO University server. Now posgraduate students of Computer Science and Applied Mathematics Department of ITMO University are using this system in their research work. It is a closed beta testing process. In the future the authors aspire to deploy this system for all the university departments, laboratories.

Some problems raised in the project require additional research and further development. The most challenging problems are:

1. Now the portal performs searching scientific publications but the authors want to add searching research projects, educational courses and researchers.
2. Increasing the number of search sources.
3. Development of a better search algorithm (for example, case based reasoning).
4. Development of more complex queries.

References

1. Leinberger M., Scheglmann S., Lammel R., Staab S., Thimm M., Viegas E. Semantic web application development with LITEQ // Lecture Notes in Computer Science. 2014. V. 8797. P. 212227.
2. Heath T., Bizer C. Linked Data: Evolving the Web into a Global Data Space. 1st ed. Morgan & Claypool Publ., 2011. 136 p. doi: 10.2200/S00334ED1V01Y201102WBE001
3. Halaç, Tayfun Gökmen and Erden, Bahtiyar and Inan, Emrah and Oguz, Damla and Gocebe, Pinar and Dikenelli, Oguz, Publishing and Linking University Data Considering the Dynamism of Datasources, I-SEMANTICS 2013, (2013)
4. Mouromtsev D. I., Lehmann J., Semerkhanov I. A., Navrotskiy M. A., Ermilov I. S. Study of current approaches for Web publishing of open scientific data. Scientific and Technical Journal of Informatics Technologies, Mechanics and Optics, 2015, vol. 15, no. 6, pp. 1081-1087.
5. Ermilov, I., Hffner, K., Lehmann, J., Mouromtsev, D. kOre: Using Linked Data for OpenScience Information Integration, SEMANTiCS 2015, (2015)
6. Getting data from the Semantic Web. http://semanticweb.org/wiki/ Getting_data_from_the_Semantic_Web. accessed: 12-06-2016.
7. Engblom, J., Published Linked Data in Ruby on Rails. Royal Institute of Technology. 2012.

8. Mumtaz M. Ali Al-Mukhtar et al, International Journal of Computer Science Engineering and Technology (IJCSET), January 2014, Vol. 4, Issue 1, 10-14
9. Vacura, M., Svatek, V. Ontological Analysis of Human Relations for Semantically Consistent Transformations of FOAF Data. KIELD-2010, CEUR-WS, 2010, s. 15-27
10. FOAF Vocabulary Specification 0.99., Dan Brickley, Libby Miller, http://xmlns.com/foaf/spec/, 2014
11. Ding, Y., Yan, E., Ghazinejad, A., & Jia, H. (2013). Extending the VIVO ontology to iS-chools: Enabling networking of information scientists. iConference 2013 Proceedings (p. 905-908).
12. Mitchell, S., Chen, S., Ahmed, M., Lowe, B., Markes P., Rejack, N., Corson-Rikert, J., He, B., Ding, Y. and VIVO Collaboration. 2011. The VIVO Ontology: Enabling Networking of Scientists.
13. Ruiz-Iniesta A., Corcho O. A review of ontologies for describing scholarly and scientific documents.SePublica2014. May 2014
14. Peroni S and Shotton D (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. J. Web Semantics: Science, Services and Agents on the World Wide Web. Available online 13 August 2012
15. Varenikov D.A., Shley M.D., Muromtsev D.I. Building scientific profiles for scientific and educational process participants in it system of the university. Modern problems of science and education, 2015, vol. 2-2, pp. 178
16. Teaching Core Vocabulary Specification, Tomi Kauppinen, http://linkedscience.org/teach/ns/, 2013

Big Data and Machine Learning in Government Projects: Expert Evaluation Case

Nikita Nikitinsky[1], Sergey Shashev[2], Polina Kachurina[3], Aleksander Bespalov[4]

[1]NAUMEN, Russia
[2]Tsentr Razrabotki, Russia
[3]DocSourcing, Russia
[4]Saint Petersburg Electrotechnical University "LETI", Russia

Abstract. In this paper, we present the Expert Hub System, which was designed to help governmental structures find the best experts in different areas of expertise for better reviewing of the incoming grant proposals. In order to define the areas of expertise with topic modeling and clustering, and then to relate experts to corresponding areas of expertise and rank them according to their proficiency in certain areas of expertise, the Expert Hub approach uses the data from the Directorate of Science and Technology Programmes. Furthermore, the paper discusses the use of Big Data and Machine Learning in the Russian government project.

Keywords: government project · Big Data · Machine Learning · expert evaluation · clustering

Introduction

Big Data projects for the government sector embody several prerequisites that expert believe are the hallmarks of fast analysis based on effective resources of information. Machine learning helps to build the hierarchy of importance of different parts of this information and gives a possibility to design semi-automated or completely automated services. World practices in this field are diverse. Currently there is a high degree of uncertainty as to which extent it is possible to use automated systems and where only human subjective evaluation works. However even conservative view on the issue allows using Dig Data and machine learning in prior analysis – it reduces the scope of the study area.

Information gathering and evaluation of heterogeneous distributed sources in experience and skills evaluation has previously been a manual process. At the same time the complexity of the operational environment increases due to the increase of labor mobility. Even in classical sociology the studies of social mobility were engaged in comparative inquiries. Nowadays retrospective questions and the use of cohort approach (comparing data with the early stage mobility studies) are not that useful: contemporary society created a new paradigm of existence. In these circumstances, further compara-

tive and longitudinal mobility studies have little point. However fast services for expertise evaluation, especially for collecting data on experience and expertise of professionals who evaluate technological projects seeking for state financing, are of great demand.

The framework outlined above lead to specific methods of research, used in this paper. First, it is a descriptive research on the worldwide experience. Secondly, in order to bring a cross-field study we try to analyze state-of-the-art technology and its use in a narrow sphere of e-government. Thirdly, we try to go in a very detail in description of the Russian Expert Hub system and as a conclusion – to compare it with the best world practices.

Current state-of-the-art technology and projects regarding information collection, fusion and analysis have a clear focus on Big Data and machine learning. The main goal of this article is to study the major international cases of government experience of the use of these technologies – which are a part of e-government, and then to depict the Russian case of expert evaluation. By comparing several clustering algorithms. The main method of our studies is the experimental method. Conclusions should derive from comparisons and be useful for further cases of Big Data and machine learning deployment for government projects. Also there might be a possibility to use this experience in other government projects. Russia is one of the leaders in software development and its market players are interested in fast development and potentially even in the export of technological products and solutions.

Machine Learning and Data Analysis for government projects in Russia

First thing we need to understand is that Data analysis field and e-governance in Russia are phenomena of completely different nature. They intersect during specific cases and amount of such cases is growing but both of them have their own features and specific history. In addition, both of this fields progress rapidly and information sources older than 10 years are almost outdated and have only historical interest.

Official history of Russian e-governance begins in 2000 with Okinawa charter of global information society [1,2,3] which was signed by Russia. The initial position of Russia in these matters was quite weak. In 2003, IT minister Reyman L. stated that only 1% of federal government workers use internet [4].

In 2002 governmental 2.57 billion dollars program “Electronic Russia” (E-Russia) began [2,3], [5,6]. Mostly it was covering the problem of delivering municipal services and information by internet. Results of this program were evaluated very diversely. In 2005 year Putin V. stated that IT market grew from 2% to 5.3% of total GDP, however 40 thousand of localities in Russia have no internet access [8]. Informational and service coverage showed growth from 2000 to 2005 but service coverage was only 6% in 2005 [2]. In 2012 year from 10 basic UN E-Gov objectives Russia targeted only 5 and had some success only in 4 of them [6]. Whole program was widely criticized for ineffectiveness [2,3], [5,7].

Next big governmental attempt in these matters was State programme: "Information Society 2011-2020" which was issued by government in 2010 year [5], [8,9]. Significant growth e-governance services of was stated. Public opinion poll showed that 66% of internet users are ready to user e-gov and according to official statistics 10.6% of Russians have interactions with electronic services at least once[9].

Under governmental patronage large business accelerator, IIDF was founded in 2013[10]. One of its goals was to deliver high quality IT and e-gov services to the people. However, only 7 of 152 successful projects are somehow connected with e-governance [11]. Moreover, very few of IIDF successful projects exploited machine learning or data analysis. In the end of 2015 IIDF representative stated that 500 million of rubles will be invested in big data soon[12].

History of data analysis and machine learning is less dramatic. Yandex Company, which is 4-th largest search engine in the world, started big data analysis trend in the middle of 2000s [13]. In 2007 it opens the school of data analysis[14]. Trend was picked up by many educational institutions ITMO [15], HSE [16], MIPT [17] and so on. It was stated on governmental level, that data analysis and machine learning are development priorities for Russia and country can become competitive in this fields [18]. Data analysis market is quite small ($340 m. in 2014) but its growth rate is almost 40% per year[19]. Main buyers of analytical solutions are banks and telecom. Advertising firms use big data storages most intensively compared to other business directions but absence of world famous successful business stories in this field slows the growth down [20].

Finally, when we reached big data and machine learning in governmental activities and services we can see that government does not use them much by now. At most analytical firms directed to business and education. However, demand for these services is obvious and undeniable but somehow hidden from statistics.

Main problem is that there are three basic levels of governance in Russia:

- State level
- Regional level
- Municipal level

State level has several success stories in developing analytical solutions. All of them can be counted by one hand however, they are quite massive. Total revenue of the leading IT companies in the public sector of Russia in 2013 was $4321 m. This is 77% of their total revenue in Russia [21]. For instance – Federal Pension Fund created analytical services based on SAP HANA, Sberbank launched several complex solutions based on Teradata, Federal Tax service uses various instruments like Teradata, Oracle Exadata and SAP to create analytical layer and monitor tax payers' activities [21], [23]. Some of the state governmental projects are listed in table 1.

Regional level and municipal level are almost completely hidden from view. At analytical companies' sites there can be found proposals of analytical solutions for every level of governmental structure, but very few stats of success histories on regional or municipal levels can be found. It's obvious that some projects require analysis and it is done for them. For instance, in news there can be found that several students created algorithm for optimal car trafficking on municipal toll roads entrances [23]. Obviously,

it is a part of some municipal project, but this project was not tagged with "machine learning". It is very hard to evaluate real volume of analytical demand in regions. However surely it was rising in recent years [21,22].

Table 1. Recent state ML projects.

Governmental client	Implemented solutions
Sberbank	Marketing and sales, risk management scoring, CRM, anti-fraud
Federal tax service	Establishment of the analytical layer for federal data warehouse
Pension Fund	Analytics and reporting
Federal Compulsory Medical Insurance Fund	Analytics and reporting
Federal Road Agency	Traffic jams forecasting system
Ministry of Finance	Security system, civil service positions classification system
Ministry of Education and Science	Expert-analytical prediction system, automated e-learning resources examination system, financial analytical system
Central Bank	Automated support system for IT departments, real estate analytical system
Supreme Court of Arbitration	E-governance integration, HR system,
Federal Treasury	"Electronic Russia budget" system, security
Roscosmos State Corporation	Computing networks integration and control systems
Federal Service for Hydrometeorology and Environmental Monitoring	Forcasting system update
Ministry of Natural Resources and Environment	Decision Support System
Federal Service for State Registratio Cadastre and Cartography	Automation of real estate registration service, analytics
Federal Financial Monitoring Servic	Automated classification and clustering system
Federal Drug Control Service of Russia	Data storage and analytics

The overall technological progress dictates a shift towards the use of the latest solutions in database management, data processing and automation of prior services. Despite the differences in systems of government administrative entities, the new generation of clerks brought a renewed vision on automation and the use of technology in government projects. This in its turn stimulates emergence of new specific projects and demands. One of such projects – the Expert Hub system, will be presented in the next part of the article.

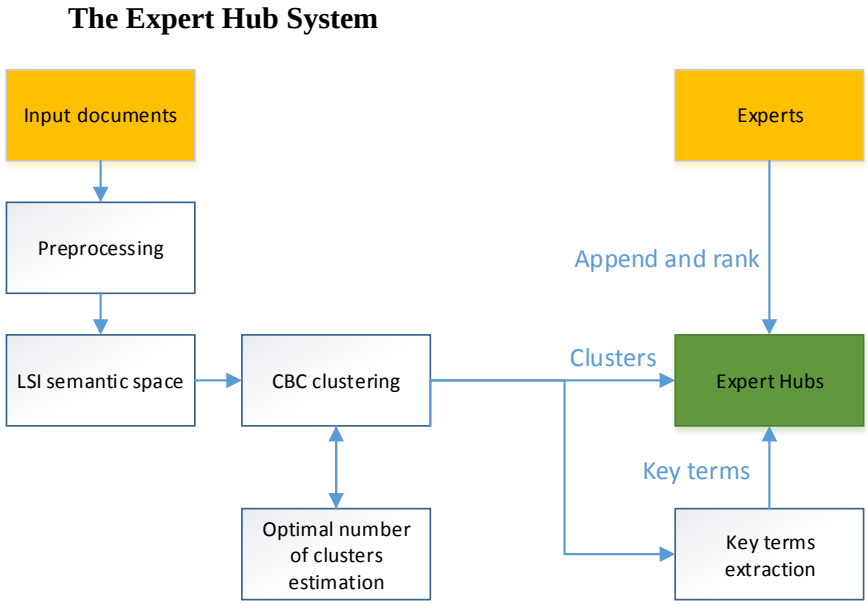


Fig. 1. Schema of the system

Concept. In this part, we will make a general description of the system and then – go into a more detail describing the algorithms that were used, focusing with a special attention on experiments with the algorithms and the way in which optimal variants were chosen.

To increase the implementation speed of innovational solutions in government the Xpir project was created [25]. Its main goal is to provide information support for Russian scientific and technical society. This platform contains science news, conferences information, data on Russian and international funds and organizations. The Expert Hub system prototype was originally created as a module for the Xpir project.

The idea of the Expert Hub approach is to use the documents from Examination System in order to define areas of expertise with semantic space construction and clustering, then to relate experts to the corresponding areas of expertise and rank them according to their proficiency in certain areas of expertise. The Examination System is an internal system in the Directorate of Science and Technology Programmes for evaluating research project proposals. In this system, invited or employed experts are reviewing incoming grant proposals and deciding whether a given research project should be or should not be awarded with a grant or other kind of benefit.

Data. The Directorate of Science and Technology Programmes provided us with 30 000 documents created by 13545 experts for the study.

Data preparation and preprocessing. First, we extract all the so-called metadata from the documents – author names, document titles etc. This data is later used for reference purposes. Then, we conduct the tokenization of the contents, and remove all the punctuation marks as well as the stop-words. We consider words having almost no

meaning, such as prepositions and conjunctions, stop-words. As the final step of data preprocessing, we lemmatize the contents in order to reduce the number of unique terms as different forms for one word by converting them into one conventional form.

Semantic space construction. After preprocessing the input documents, we create LSA term-document semantic space of all documents from data we have, where each row denotes document and each column is a word.

LSA (Latent Semantic Analysis) is a technique for Natural Language Processing, which is widely used for solving various tasks in information retrieval. The underlying idea of LSA is that words with similar sense tend to occur in similar contexts. Thus, this technique can deal with homonymy. We employ LSA as it is faster and can work with larger data sets compared to other approaches [26].

LSA is based on the well-known singular value decomposition technique (SVD):

$$M = U\Sigma V^* \tag{1}$$

where M is m × n matrix whose entries come from some field K, U is m×m matrix, Σ is m×n diagonal matrix with non-negative real numbers on the diagonal and V* is an n×n unitary matrix over K.

We apply Log Entropy weighting function for LSA as this function works well in many practical studies [27].

Particularly, each cell aij of a term-document matrix A is computed as follows:

$$p_{ij} = \frac{tf_{ij}}{gf_i}, \quad g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}, \; a_{ij} = g_i + \log(tf_{ij} + 1) \tag{2}$$

where n is total number of documents, g_i is the global weight, tf_{ij} is the number of occurrences of term i in document j, and gf_i is the total number of times the term i occurs in the corpus.

Semantic space clustering. We cluster the LSA semantic space (currently we use DBSCAN + CBC hybrid clustering algorithm for this task, see Experiments section to know why we used it). As the result of clustering, we obtain centroids and document vectors belonging to those centroids. The optimal number of clusters is computed with Silhouette score using Grid search hyperparameter optimization approach. We then call each cluster an Expert Hub. Since we know, which document belongs to which expert, we may estimate areas of expertise for every expert based on their documents - and the documents are already distributed to clusters.

CBC (Clustering by committee) is a centroid-based clustering algorithm, which was designed with motivation to cluster texts written in natural languages. The algorithm consists of three phases. In Phase I, each element's top-k similar elements are computed for some small value of k. In Phase II, a collection of tight clusters is constructed, using the top-k similar elements from Phase I, where the elements of each cluster form a so-called "committee". The algorithm tries to form as many committees as possible on the condition that each newly formed committee should not be equal or much similar to any already existing committee. All the committees violating this condition are simply discarded from further computing. In the final phase of the algorithm, each element e

is assigned to its most similar cluster or clusters if we apply soft clustering approach [28].

DBSCAN (Density-based spatial clustering of applications with noise) is a density-based clustering algorithm. Given a set of points in some space, it groups together points that are close to each other, marking as outliers points that lie alone in low-density regions. [29].

Silhouette score is an internal clustering validation measure, which is the measure that does not employ any external knowledge about the data e.g. known class labels. It just evaluates the quality of clustering based on the data used for clustering and the result of clustering. Silhouette coefficient compares the average distance from element to element within a cluster with the average distance to elements in other clusters, assigning highest scores to the algorithm producing dense clusters (with high similarity within the cluster) located far from each other (low similarity between clusters). [30].

Grid search hyperparameter optimization approach is a simple approach for selecting the best hyperparameters (e.g. parameters set by a researcher and not learned by algorithm itself) by generating candidate hyperparameters from a grid of possible hyperparameter values specified by a researcher [31].

Key terms extraction. We extract key terms (including n-grams) from LSA semantic space for each Expert Hub based on documents belonging to that Hub. The extracted key terms represent the Expert Hub making it possible for user to name the Expert Hub. In addition, we extract keywords for every area of expertise for every expert for the same purpose.

We experimented with two methods of key term extraction:

1. Computing research area vector for an expert as average vector of his or her documents belonging to the research area. Then, we select top-20 lemma vectors from the whole semantic space, which are similar to the research area vector. These top-20 lemma vectors are selected to represent the research area for the expert. This approach has a feature that among words representing research area for the expert there may occur words not presented in documents of the expert. Caveat of this approach is that we retrieve only unigrams as the semantic space consists of lemma vectors representing single words (bag-of-words approach)
2. We compute research area vector for an expert as average vector of his or her documents belonging to the research area as in first approach. Then we take top-20 lemma vectors, which are most similar to the research area average vector, only from documents of the expert. For the most similar words, we look for n-grams in documents based on rules, which we created. We estimate the LogEntropy weights of a bigram as maximum weight of unigram constituents of a bigram. This approach can retrieve n-grams up to trigrams and consider terms occurring only in this expert' documents.

As bigrams represent areas of expertise better than unigrams, we employ the second approach in our system.

Expert assigning and ranking. We then append experts to the corresponding Expert Hubs and rank them based on their impact weight to the Hub.

Impact weight is computed based on multiple factors:

- scientific background of an expert (from the expert's profile in the Examination System)
- information about the previous expert assessments of an expert
- similarity of the documents of the expert belonging to the certain Hub (the higher similarity to the Hub documents have, the more impact weight the expert obtains).

Since every expert may have multiple different areas of expertise, we apply soft clustering method allowing the experts be related to several Hubs with various impact weight.

Experiments

In this section, we conduct experiments in order to define the best approach to clustering the term-document semantic space built from the documents of the experts. We try three different types of clustering algorithms (i.e. centroid-based, agglomerative hierarchical and density-based) and their combinations.

The aim of the experiments is to select the optimal clustering algorithm or combination of clustering algorithms providing the best clustering results. To measure the quality of clustering, we use Silhouette score. With optimal clustering results, the Expert Hub System should maintain the optimal quality of experts' allocation to the hubs.

In every experiment, for every clustering algorithm, we iteratively select certain number of clusters and on every cluster number we measure Silhouette score. The number of clusters with the highest measured Silhouette score we consider optimal for the clustering algorithm. As we expected the number of Expert Hubs to be from 40 to 120 depending on the possible degree of fragmentation of scientific fields, we conducted iterative clustering on this possible distribution of the clusters. For experimental purposes, we cluster the LSA space constructed from all the 30 000 documents.

Experiment 1. In the first experiment, we clustered the LSA space with just CBC algorithm. The optimal Silhouette score value (0.131) was obtained on 45 clusters (table 2). The keywords extracted from the clusters contained much common lexis and words irrelevant to clusters, which indirectly indicated bad quality of clustering.

Experiment 2. In the second experiment, we clustered the LSA space with just Agglomerative Hierarchical Clustering algorithm (also known as AGNES and AHC).

AGNES (AGglomerative NESting) or AHC is a standard agglomerative hierarchical clustering algorithm, consisting of two phases. Phase I initially starts with n clusters each containing a different element, Phase II embraces the merge of two most similar clusters (repeated n – 1 times) [28].

The results of clustering were better than in previous experiment, however, not much: the highest Silhouette score value was 0.1457 for 43 clusters (table 2). Key terms extracted from the clusters contained common lexis. This makes us to conclude that Agglomerative Hierarchical Clustering is also irrelevant method for clustering in our case.

As we may see from the above experiments, agglomerative hierarchical and centroid algorithms worked not very well on the data. We supposed, that the cause of such results was that the distribution of data points in the LSA semantic space contained much

outliers and the shape of the resulting clusters could be arbitrary. Thus, we decided to try a density-based clustering algorithm DBSCAN as its advantages included robustness to outliers and ability to locate arbitrary-shaped clusters

Experiment 3. In the experiment number three, we first clustered the LSA space with DBSCAN algorithm (with the following hyperparameters: epsilon = 0.5, minPts = 10 and cosine distance function). As this algorithm discovers the appropriate number of clusters by itself and this number of clusters may not fit our predefined possible cluster distribution, we applied CBC to the resulting average vectors of DBSCAN clusters and iteratively measured Silhouette score. The highest Silhouette score value (0.697) was achieved on 42 clusters (table 2). The key terms, which we extracted, contained a lot of special lexis and almost no common lexis. This indirectly indicates a good clustering, i.e., documents with a large number of common lexis were assigned to separate clusters.

As the final trial, we also experimented with applying AHC to the average DBSCAN vectors, but ended up with lower Silhouette score of 0.579 on 46 clusters (table 2).

Conclusion. We consider hybrid clustering algorithm, consisting of DBSCAN and CBC, the most appropriate algorithm to cluster the LSA space constructed from the data. For the data under study, we found out that based on experimental study the optimal number of clusters was between 40 and 46 clusters with the most probable number of 42 clusters.

Table 2. Highest results for clustering algorithms

Clustering algorithm	Optimal number of clusters	Highest Silhouette Score
CBC	45	0.131
AHC	43	0.1457
DBSCAN + AHC	46	0.579
DBSCAN + CBC	42	0.697

Evaluation

To evaluate the Expert Hub System, we employed the expert analysis approach. First, we named the expert hubs with the appropriate names according to the key terms of those hubs (for example, Physical Chemistry, Biology etc.). Then, we selected the top-5 most highly ranked (i.e. relevant to the hub) persons from each of the 42 hubs ending up with the total number of 210 persons. After that, we asked our experts to check some bibliographic databases (i.e. RSCI, Scopus and Web of Science) to make sure that persons indeed should have been related to certain expert hubs. The criterion of the person relevance to the certain expert hub was the following: a person should have had more publications relevant to the topic of the hub he or she was assigned to than to every other topics.

The evaluation showed us, that 83.34% of experts (175 persons) met the criterion. Thus, we can suppose that the Expert Hub System prototype maintains relatively high accuracy in assigning experts to the corresponding hubs.

Results and discussion

In this study, we presented our attempt to create a system to automatically detect and rank experts in certain areas of expertise in order to provide governmental structures with the most highly qualified experts for reviewing incoming grant proposals and research projects. The Expert Hub System prototype operates well – the accuracy of assigning experts to the corresponding expert hubs is above 80%. The clustering algorithm with the best performance on the data we had for the study appeared to be the hybrid DBSCAN + CBC algorithm.

Furthermore, we described prior experiences of technologisation of government services and projects. The basic goals of the previous stage of government services development focused much on automation and storage of information, while nowadays it is possible to work with well-structured data, to shift from database management towards Data Mining, to use Big Data and Machine Learning for sophisticated projects.

Certainly, our study has many limitations. For instance, the evaluation of the system was not strictly formal and we evaluated only some aspects of the system. Moreover, the data we had for this study was relatively small in order to be applied to real processes in the Directorate of Science and Technology Programmes and other governmental structures.

Conclusion and future work

To conclude, we would like to say that the Expert Hub System prototype shows promising results and demonstrated decent performance during the evaluation. Product and market opportunities make the project scalable for other tasks, i.e. for HR solutions or for automated studies of competitors (especially for SME).

For a future study, we suggest:

- Conducting a more thorough and formalized evaluation of the system
- Applying other methods for creating the semantic space from documents in order to obtain better results. Currently, we consider word2vec and similar tools suitable for this.
- Conducting a research in order to better detect the optimal number of clusters, currently we think about applying semi-supervised approach to cluster analysis to handle the task.
- Carrying out usability studies of the system to discover its applicability to other tasks.

Acknowledgements Special thanks goes to the Directorate of Science and Technology Programmes for providing us with the data for this study. Ministry of Education and Science of the Russian Federation supported the research reported in this publication. Unique id of the research project is RFMEFI57914X0091.

References

1. Okinawa charter of global information society, http://www.iis.ru/library/okinawa/charter.en.html
2. McHenry W. and Borisov A. (2006) "E-Government and Democracy in Russia," Communications of the Association for Information Systems: Vol. 17, Article 48.
3. Pardo, Theresa, "Digital Government Implementation: A Comparative Study in USA and Russia" (2010). AMCIS 2010 Proceedings. Paper 330.
4. Reyman, L. (2003) "Information Technologies in the Work of Federal Governmental Agencies" (in Russian), Vestnik Svyazi International, 9, pp. 1-8
5. Vinogradovaa N., Moiseevaa A.,Open Government and "EGovernment" in Russia , Sociology Study, January 2015, Vol. 5, No. 1, 29-38 doi: 10.17265/2159-5526/2015.01.004
6. Bershadskaya, L., Chugunov, A., Trutnev, D. 2012. EGovernment in Russia: Is or Seems? In: Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance (ICEGOV2012, Albany, New York, United States, 22-25 October 2012). Ed. by J.Ramon GilGarcia, Natalie Helbig, Abegboyega Ojo. N.Y.: ACM Press, 2012. 79-82. DOI: 10.1145/2463728.2463747
7. Highlights on meeting of the Presidium of the State Council on the development of information and communication technologies in the Russian Federation, http://www.vestnik-sviazy.ru/news/16-fevralja-2006-goda-v-g-nizhnijj-novgorod-pod-predsedatelstvom-prezidenta-rf-vv-putina-sostojalos-zasedanie-prezidiuma-gossoveta-rf/ (in Russian)
8. State programme: Information Society 2011-2020, http://government.ru/en/docs/3369/
9. Vidiasova L., Chugunov A., Mikhaylova E., E-Governance in Russia: Toward New Models of Democracy, 2015. Proceedings of the 2015 2nd International Conference on Electronic Governance and Open Society: Challenges in Eurasia. ACM, New York, NY, USA, pp. 44-49
10. IIDF homepage, http://www.iidf.ru/ (in Russian)
11. IIDF projects analysis, https://megamozg.ru/post/20382/ (in Russian)
12. IIDF will invest 500 million in Big Data, http://www.vedomosti.ru/technology/articles/2015/11/24/618040-frii-nuzhno-bolshe-dannih (in Russian)
13. From startup to IPO: How Yandex became Russia's search giant, http://www.ewdn.com/2011/05/17/from-startup-to-ipo-how-yandex-became-russias-search-giant/,
14. The Yandex School of Data Analysis , https://yandexdataschool.com/about
15. ITMO extreme computing programme, http://en.ifmo.ru/en/viewjep/2/5/Big_Data_and_Extreme_Computing.htm
16. HSE Machine learning course overview, https://www.hse.ru/data/2015/09/18/1082472447/1Applied%20Machine%20Learning%20-%202015-2016.pdf
17. MIPT github Machine learning course sides, https://github.com/vkantor/MIPT_Data_mining_in_action_2015/tree/master/Slides

18. The order of the Russian Federation Government from November 1, 2013 N 2036-p Moscow, ,http://rg.ru/2013/11/08/texnologii-site-dok.html
19. Structure of the big data market in Russia , http://rusbase.com/howto/big-data-in-russia/ (in Russian)
20. Russian big data in early stages, http://www.computerweekly.com/news/4500259726/Russian-big-data-in-early-stages
21. Kuraeva A., Kazantsev N., Survey on big data analytics in public sector of russian federation, Information Technology and Quantitative Management (ITQM 2015), Procedia Computer Science 55 (2015) 905 – 911
22. Largest Big data projects in Russia, http://www.cnews.ru/tables/a9249186ccefd9e546774ec36da1970ba20ca212/ (in Russian)
23. Big data for governmental sector, http://www.cnews.ru/reviews/ikt_v_gossektore_2014/articles/bolshie_dannye_novye_vozmozhnosti_dlya_gossektora/ (in Russian)
24. Toll roads traffic optimization algorithm, https://www.mos.ru/news/item/7968073 (in Russian)
25. Xpir - platform for communication and cooperation between scientists and entrepreneurs. www.xpir.ru
26. Deerwester S., Dumais S. T., Furnas G. W., Landauer T.K., Harshman R.., "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41 (6), 1990, pp. 391-407.
27. T.Landauer, D.S. McNamara, S.Dennis and W. Kintsch., Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, 2007
28. Pantel, Patrick A. Clustering by committee. Diss. University of Alberta, 2003.
29. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231
30. Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
31. Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." The Journal of Machine learning Research 13.1 (2012): 281-305

Small Differences in Experience Bring Large Differences in Performance

Sheen S. Levine[1], Charlotte Reypens[1,2]

[1] University of Texas, Dallas, USA
[2] University of Antwerp, Belgium

Abstract. In many life situations, people choose sequentially between repeating a past action in expectation of a familiar outcome (exploitation), or choosing a novel action whose outcome is largely uncertain (exploration). For instance, in each quarter, a manager can budget advertising for an existing product, earning a predictable boost in sales. Or she can spend to develop a completely new product, whose prospects are more ambiguous. Such decisions are central to economics, psychology, business, and innovation; and they have been studied mostly by modelling in agent-based simulations or examining statistical relationships in archival or survey data. Using experiments across cultures, we add unique evidence about causality and variations. We find that exploration is boosted by three past experiences: When decision-makers fall below top performance; undergo performance stability; or suffer low overall performance. In contrast, individual-level variables, including risk and ambiguity preferences, are poor predictors of exploration. The results provide insights into how decisions are made, substantiating the micro-foundations of strategy and assisting in balancing exploration with exploitation.

Keywords: Exploration, Exploitation, Decision Making, Experiment, Protocol Analysis, Cross-culture

In many life situations – R&D investments, market entry, military campaigns, romantic choices – a decision-maker chooses an action, receives feedback, and then chooses again. The choice ranges from repeating a past action in expectation of a familiar outcome (exploitation) to a novel action whose outcome is largely uncertain (exploration). For instance, in each quarter, a manager can budget more advertising for an existing product, expecting familiar (but uncertain) sales figures; or he can spend to develop a revised version of the product, whose sales prospects are more uncertain; or he can invest in a completely new product, where prospects are even more uncertain. The optimal action is not obvious: Probabilities are unknown, feedback is ambiguous. Such decisions have been discussed across domains and species, from bees and birds foraging to organizations searching for innovations (for reviews, see [1,2]). We examined – empirically – how people decide in such situations. To do that, we created a behavioral task an oil exploration game in which participants earn money by searching an unfamiliar landscape (cf. [3]). A participant chooses a spot for

drilling and then discovers the quantity of oil it contains. The participant can keep drilling in the same spot, earning the same quantity of oil. Or he can choose a nearby location, which likely has a similar oil quantity. Or he can jump across the landscape to a faraway location, where the oil quantity is likely very different. The participant repeats the choice a fixed number of times. When the game ends, the oil he found is converted to cash, which is paid to him. The task faithfully represents the important features of an exploration-exploitation situation: The landscape is rugged, containing "peaks" and "valleys" of oil, but there is no map that describes the terrain a decision maker discovers it by experiencing it. Because probabilities are unknown, optimization is impossible [4,5,6,7]. And since information accumulates only with experience, initial steps are necessarily random [8]. The seeker has only limited resources, so he can sample only a fraction of the entire landscape ([9]). We studied how people decide in exploration-exploitation in four studies. First, we conducted one-on-one sessions, where we collected quantitative data as well as verbal accounts from the participants, describing their decision-making process [10]. Second, we conducted laboratory sessions in the U.S. using a web-based version of the game. Third, also using the web-based instrument, we collected data from workers in a labor market [11,12]. Fourth, to ascertain the robustness of the findings, we conducted laboratory sessions in Russia, a country whose history, culture, and institutions differ from those of the U.S. [13]. Across all studies, we find that exploration is driven more by immediate experience, less by individual characteristics. Three situations boost it: When a decision-maker falls below his or her top performance; when he or she experiences performance stability; or when he or she suffers low overall performance. Exploitation is increased by the reverse experiences: exceeding top performance, experiencing performance variance, and enjoying high overall performance. Individual traits, such as risk and ambiguity preferences, are poor predictors of exploration. These experiences have similar effects in all of the studies. Behavior is strongly influenced by experience, so two identical players that face the exact same landscape can undergo completely different experiences and end up with a wide gap in performance, all due to random differences in their early choices. In everyday life, we often attribute differences in performance to traits – "she is brilliant," "this manager is incompetent" – and not to experience. But the results here suggest otherwise: History matters.

References

1. Hills, T.T., Todd, P.M., Lazer, D., Redish, A.D., Couzin, I.D.: Exploration versus exploitation in space, mind, and society. Trends in Cognitive Sciences **19**(1) (2015) 46–54
2. Mehlhorn, K., Newell, B.R., Todd, P.M., Lee, M.D., Morgan, K., Braithwaite, V.A., Hausmann, D., Fiedler, K., Gonzaleza, C.: Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. Decision (Washington) **2**(3) (2015) 191–215
3. Mason, W., Watts, D.J.: Collaborative learning in networks. Proceedings of the National Academy of Sciences **109**(3) (2012) 764–769

4. Alchian, A.: Uncertainty, evolution, and economic theory. Journal of Political Economy **58** (1950)
5. Gittins, J.C., Gittins, J.C.: Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society, Series B (1979) 148–177
6. Kauffman, S., Levin, S.: Towards a general theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology **128**(1) (1987) 11 – 45
7. Nagel, R., Vriend, J.N.: An experimental study of adaptive behavior in an oligopolistic market game. Journal of Evolutionary Economics **9**(1) (1999) 27–65
8. Winter, S.G.: Purpose and progress in the theory of strategy: Comments on gavetti. Organization Science **23**(1) (2012) 288–297
9. March, J.G.: Exploration and exploitation in organizational learning. Organization Science **2**(1) (1991) 71–87
10. Ericsson, K.A., Simon, H.A.: Protocol analysis: Verbal reports as data (Revised ed.). Cambridge, MA: MIT Press (1993)
11. Horton, J.J., Rand, D.G., Zeckhauser, R.J.: The online laboratory: conducting experiments in a real labor market. Experimental Economics **14**(3) (2011) 399–425
12. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. Judgment and Decision Making (2010) 411–419
13. Henrich, J., Heine, S.J., Norenzayan, A.: Most people are not WEIRD. Nature **466**(7302) (July 2010) 29

Author Index