Exploring gender differences in member profiles of an online dating site across 35 countries

Slava Kisilevich and Mark Last

Department of Computer and Information Science
University of Konstanz
slaks@dbvis.inf.uni-konstanz.de[***]
Department of Information System Engineering
Ben-Gurion University of the Negev
mlast@bgu.ac.il[†]

Abstract. Online communities such as forums, general purpose social networking and dating sites, have rapidly become one of the important data sources for analysis of human behavior fostering research in different scientific domains such as computer science, psychology, anthropology, and social science. The key component of most of the online communities and Social Networking Sites (SNS) in particular, is the user profile, which plays a role of a self-advertisement in the aggregated form. While some scientists investigate privacy implications of information disclosure, others test or generate social and behavioral hypotheses based on the information provided by users in their profiles or by interviewing members of these SNS. In this paper, we apply a number of analytical procedures on a large-scale SNS dataset of 10 million public profiles with more than 40 different attributes from one of the largest dating sites in the Russian segment of the Internet to explore similarities and differences in patterns of self-disclosure. Particularly, we build gender classification models for the residents of the 35 most active countries, and investigate differences between genders within and across countries. The results show that while Russian language and culture are unifying factors for people's interaction on the dating site, the patterns of self-disclosure are different across countries. Some geographically close countries exhibit higher similarity between patterns of self-disclosure which was also confirmed by studies on cross-cultural differences and personality traits. To the best of our knowledge, this is the first attempt to conduct a large-scale analysis of SNS profiles, emphasize gender differences on a country level, investigate patterns of self-disclosure and to provide exact rules that characterize genders within and across countries.

Key words: Social Networking Sites, Self-disclosure, Gender differences, Classification trees, Multidimensional Scaling, Hierarchical Clustering

[***] http://www.informatik.uni-konstanz.de/arbeitsgruppen/infovis/mitglieder/slava-kisilevich/

[†] http://www.bgu.ac.il/~mlast/

1 Introduction

Rapid technological development of the Internet in recent years and its worldwide availability has changed the way people communicate with each other. Social Networking Sites such as Facebook or MySpace gained huge popularity worldwide, having hundreds of millions of registered users. A major reason for the increased popularity is based on social interaction, e.g. networking with friends, establishing new friendships, creation of virtual communities of mutual interests, sharing ideas, open discussions, collaboration with others on different topics or even playing games. The key component of all SNSs is the user profile, in which the person cannot only post personal data, e.g. name, gender, age, address, but also has the opportunity to display other aspects of life, such as personal interests, (hobbies, music, movies, books), political views, and intimate information. Photos and videos are equally important for a self-description. All SNSs allow the user to upload at least one photo. Most mainstream SNSs also feature video uploading.

Various research communities have realized the potential of analysis of the SNS-phenomenon and its implication on society from different perspectives such as law [1], privacy [2–4], social interaction and theories [5–9]. Many hypotheses and social theories (gender and age differences, self-disclosure and self-presentation) have been raised and tested by social scientists using the context of Social Networks. Statistical analysis is the widely used instrument for analysis among social scientists and rely on the sampling rather than on data collected from an entire population segment.

The common approach to perform Social Network analysis is to analyze a sample of available user profiles or to conduct a survey using convenience samples (e.g. students in a particular university) by presenting descriptive statistics of the sample data and performing significance tests between dependent variables [2, 10, 3, 8]. The major drawback of such approach with respect to Social Networks is that in light of the large population of SNSs, which can vary from tens to hundreds of millions of users, living in all parts of the world, the results of the statistical analysis cannot be generalized for the whole population of users or a single nation or a single culture or genders, while theories can hardly be validated using only small samples. Moreover, Social Networks are heterogeneous systems in a sense that people may form closed sub-groups on different levels like country of living, national, or cultural with minimum interaction with other sub-groups and develop communication and self-presentational styles that are completely different from others due to cultural or national differences. For example, due to cultural differences, a theory of self-disclosure tested on students from American universities may be not be applicable to information obtained from students of Chinese universities, even if both groups use the same Social Networking Site. To the best of our knowledge, the state of the art Social Science research of Social Networks does not take into account the spatial or cultural components for the analysis of self-presentation differences (presumably due to the lack of sufficient data and difficulty involved in conducting cross-country studies).

Although, the problem and the importance of space and place in the Social Sciences was already highlighted a decade ago [11], this knowledge gap was not closed until to date. Therefore, in order to improve our understanding of social behavior, to analyze, to find hidden behavioral patterns not visible at smaller scales, and to build new theories of large heterogeneous social systems like Social Networks, other approaches and computational techniques should be applied [12].

In this paper, we answer the following hypothetical question: “Can we find some hidden behavioral patterns from user profiles in the large-scale SNS data beyond mere descriptive statistics?” We answer this question by applying a classification algorithm to the data obtained from more than 10 million profiles having more than 40 different attributes extracted from one of the largest dating sites in the Russian segment of the Internet. Specifically, we build gender classification models for most active countries and investigate what are particular differences between genders in one country and what are the differences in patterns of self-disclosure across countries. Self-disclosure can be defined as any information about himself/herself which a person communicates to others [13]. In the context of the current study, it refers to the information communicated by means of a person’s online profile.

Dating sites can be considered as a special type of social networks where members are engaged in development of romantic relationship. Information revealed in the users’ profiles is an important aspect for the assessment of potential communication, for maximizing the chances for online dating for the owner of the profile, and for minimizing the risks (e.g. misrepresentation) of online dating for the viewer of the profile. For this reason, in the broad context, assuming that the goal of the member of a dating site is to find a romantic partner, we investigate patterns of self-presentation that can vary from country to country and differ for both genders.

The preliminary results suggest that the classification model can successfully be used for analysis of gender differences between users of SNSs using information extracted from user profiles that usually contain tens of different categorical and numerical attributes.

To the best of our knowledge, this is the first attempt to conduct a large-scale analysis of SNS profiles for comparing gender differences on a country level using data mining approaches in the Social Science context.

2 Related Work

Gender differences have been studied long before the Internet became widely available. However, with the technological development of the Internet and proliferation of Social Networks, the research has focused on the analysis of online communities and differences between their members. Many studies were performed in the context of Internet use [14, 15], online relationships [5], ethnic identity [8], blogging [10], self-disclosure and privacy [2–4]. Though we could not find any related work on large-scale analysis of gender differences in social networks (except for [4]), we are going to review here some of the recent, mostly small-scale studies and findings about gender differences in social networking sites.

Information revelation, privacy issues and demographic differences between Facebook users were examined in [2] and [3]. [2] interviewed 294 students and obtained their profiles from Facebook. The goal of the survey was to assess the privacy attitudes, awareness of the members of the SNS to privacy issues, and the amount and type of information the users revealed in their profiles. It was found that there was no difference between males and females with respect to their privacy attitudes and the likelihood of providing certain information. Likewise, there was no difference between genders in information revelation. If some information is provided, it is likely to be complete and accurate. However, female students were less likely to provide their sexual orientation, personal address and cell phone number. [3] interviewed 77 students to investigate different behavioral aspects like information revelation, frequency of Facebook use, personal network size, privacy concerns and privacy protection strategies. Again, there were almost no differences between female and male respondents in the amount and type of the information revealed in their profiles. [4] analyzed about 30 million profiles from five social networks of Runet and conducted a survey among Russian speaking population to cross-check the finding extracted from the profiles and assess privacy concerns of members of Russian social networks. It was shown that there were differences between type of revealed information between females and males and these differences conditioned on the reported country of residence (20 most populous countries were presented). Particularly, males disclosed more intimate information regardless of their country of origin. However, the country with the highest difference in the amount of disclosed intimate information was Russia (20.67%) and the lowest was Spain (5.59%). In addition, females from 17 countries revealed more information about having or not having children, economic and marital status, and religion. The only exceptions were females in Russia, Israel and England.

Social capital divide between teenagers and old people, and similarities in the use of the SNS were studied in [7] using profiles from MySpace social network. The results of the analysis indicate, among other criteria, that female teenagers are more involved in the online social interaction than male teenagers. Likewise, statistical tests showed that older women received more comments than older men. Additionally, linguistic analysis of user messages showed that females include more self-descriptive words in their profiles than males. Friendship connections, age and gender were analyzed in [6] using $15,043$ MySpace profiles. The results showed that female members had more friends and were more likely interested in friendship than males, but males were more likely to be interested in dating and serious relationships. In the study that analyzed emotions expressed in comments [9], it was found that females sent and received more emotional messages than males. However, no difference between genders was found with respect to negative emotions contained in messages.

Online dating communities are typically treated differently because goals of the dating sites are much more limited in terms of connection development and often bear intimate context, which for the most part shifts to the offline context. Issues such as honesty, deception, misrepresentation, credibility assessment, and credibility demonstration, are more important in the dating context than in the context of general purpose social networks. Researchers are particularly interested in the analysis of self-presentation and self-disclosure strategies of the members of dating sites for achieving their goal to successfully find a romantic partner. The authors of [5] interviewed 349 members of a large dating site to investigate their goals on the site, how they construct their profiles, what type of information they disclose, how they assess credibility of others and how they form new relationships. The study found that cues presented in the users' profiles were very important for establishing connections. These cues included very well-written profiles, lack of spelling errors and uploaded photos. The last time the user was online considered to be one of the factors of reliability. Most of the respondents reported that they provided accurate information about themselves in the profiles.

3 Data

The data used in this paper was collected from one of the largest dating sites in Runet: *Mamba*[1]. According to the site's own statistics (June 3, 2010), there are $13,198,277$ million registered users and searchable $8,078,130$ profiles. The main features of the service is the user profile and search option that allows searching for people by country, gender, age and other relevant attributes. The friend list is discrete, so other registered users cannot know with whom a user is chatting. The friend list is implicitly created when the user receives a message from another user. There are no means to block unwanted users before they send a message. However, users get a *real* status by sending a free SMS to the service provider and confirming his/her mobile phone number. This allows the users with the *real* status to communicate with and get messages only from the real people. The user may exclude his/her profile from being searchable, but most of the profiles are searchable and accessible to unregistered users.

The user profile consists of six sections, where each section can be activated or deactivated by the user. Table 1 shows the names of sections and attribute parameters available in every section. We excluded the *About me* section, in which the user can describe himself in an open form, some intimate attributes of the *Sexual preference* section and the option to add multimedia (photos or videos). The attributes are divided into two categories. In the first category, only one value can be selected for the attribute (denoted as "yes" in the Single selection column), other attributes contain multiple selections (denotes as "no" in the Single selection column). Most of the attributes also contain an additional free text field that allows the user to provide his/her own answer. If the user decides not to fill in some field, the attribute won't be visible in his/her profile. The user

[1] http://www.mamba.ru/

can extend his/her main profile by filling two surveys. The one survey is provided by MonAmour site[2], owned by Mamba and contains about 100 different questions that estimate the psychological type of the respondent according to four components scaled from 0 to 100: *Spontaneity, Flexibility, Sociability, Emotions.* Another survey is internal and contains 40 open questions like *Education, Favorite Musician, etc*[3].

In order to collect the data, we developed a two-pass crawler written in C#. In the first pass the crawler repeatedly scans all searchable users which results in a collection of a basic information about the user such as *user id, profile URL, number of photos in the profile, and country and city of residence.* In the second pass, the crawler downloads the user's profile, checks if it is not blocked by the service provider and extracts all the relevant information, which is described in Table 1.

Within a two-month period, between March and June 2010, we extracted information from 13,187,295 millions users, where 1,948,656 million profiles were blocked, leaving us with 11,238,639 million valid profiles.

[2] http://www.monamour.ru/

[3] At the time of writing this paper, the structure of the profile and some of the fields were changed by the service provider

Table 1. Profile sections and attributes

Section	Attributes	# of options	Single selection	Possible choice
Personal	Age	-	yes	20
	Gender	2	yes	Male/Female
	Zodiac	12	yes	Capricorn...
Acquaintances	Looking for	5	no	Man/Woman/Man+Woman/Man+Man/Women+Woman
	Partner's age	8	no	16-20/21-25/26-30/31-35/36-40/41-50/51-60/61-80
	Aim	5	no	Friendship/Love/Sex/Marriage/Other
	Marriage	4	yes	Married/Live separately Sham marriage/No
	Material support	3	yes	Want to find a sponsor/Ready to become a sponsor/No sponsor is required
	Kids	4	yes	No/I'd like to have/Live together/Live separately
Type	Weight	1	yes	70 kg.
	Height	1	yes	180 cm.
	Figure	8	yes	Skinny/Regular/Sportive...
	Body Has	2	no	Tattoo, Piercing
	Hair on the head	7	yes	Dark/Grey-haired...
	Smoking	4	yes	No/Yes/Seldom/Drop
	Alcohol	3	yes	No/Yes/Seldom
	Drugs	8	yes	No/Yes/Drop/Dropped
	Profession	-	-	Open field
	Economic conditions	4	yes	Occasional earnings/Stable and small income/Stable and average income/Wealthy
	Dwelling	6	yes	No steady place/Apartments/Dorm/Live with Parents/Friend/Spouse
	Languages	87	no	English/German...
	Day regimen	2	yes	Night owl/Lark
	Life priorities	8	no	Carrier/Wealth/Family/Harmony/Sex/Self-realization/Public activity/Other
	Religion	7	no	Christianity/Atheism/Other...
Sexual preferences	Orientation	3	yes	Hetero/Homosexual/Bi
	Heterosexual experience	4	yes	Yes/No/Little/Other
	Excitement	18	no	Smell/Latex/Tattoos/Piercing...
	Frequency	6	yes	At least once a day/Other Several times per Day/Week/Month Not interested in sex
Interests	Leisure	14	no	Reading/Sport/Party...
	Interests	19	no	Science/Cars/Business...
	Sports	12	no	Fitness/Diving...
	Music	11	no	Rock/Rap...
Other	Car	76	yes	Nissan...
	Mobile Phone	50	yes	Ericsson...

4 Methodology

In this section we describe the data mining process that includes data selection, data transformation and model construction.

4.1 Data selection

The data preparation and selection is very crucial for the data mining process. If sampled data is not a representative of the whole dataset, the data mining process will fail to discover the real patterns. Another aspect of data preparation is related to user profiles. As was already discussed in Sections 1 and 2, the ultimate goal of members of the dating site is to find a romantic partner. Since this kind of activity may involve elements of intimacy, persons employ different strategies to balance the desire to reveal information about themselves and stay anonymous (for example, the profile without a photo). Moreover, many people may run several user profiles for different purposes.

In order to minimize the impact of fake profiles (e.g. empty profiles or profiles containing the minimal amount of information) on the pattern mining, we employed a four level filtering process. First, the profiles of persons who filled the external survey on the MonAmour site (described in Section 3) were retrieved. Since the respondent should answer about 100 questions, it is unlikely that the person has non-serious intentions on the dating site. Second, we retrieved profiles who filled additional external survey that includes about 40 questions. Next, the users with the status "real" were retrieved and finally, the users who uploaded at least one photo and no more than one hundred photos were extracted. Table 2 shows the demographic statistics by country and gender. It also shows how many profiles were selected for mining and the resulted percentage of females and males in the selected instances. The selected age range was from 18 to 73.

4.2 Data transformation

Almost all the attributes described in Table 1 were selected for inclusion into the model (except for *Weight*, *Height*, and *Mobile Phone*). Numerical attributes such as age, number of photos and number of words used in the "About me" section were discretized. The age was discretized into ten equal size bins. We analyzed the distribution of photos and words in "About me" section individually for females and males. Based on the data distribution, the number of photos was divided into three categories: *none* if no photo was uploaded by the user, *normal* if the number of photos was between 1 and 8 for females and between 1 and 6 for males, *high* if the number of photos was between 9 and 16 for females and between 7 and 10 for males, *very high* if the number of photos was larger than 16 for females and larger than 10 for males. The number of words used in the "About me" section was divided into three categories: *none* if nothing was written, *normal* if the number of words was between 1 and 24 for females and between 1 and 22 for males, *high* if the number of words was between 25 and

Table 2. Demographic statistics of the 35 most active countries and statistics related to the sampled data

Country	**Total**	**Females %**	**Males %**	**# instances**	**Sampled Females %**	**Sampled Males %**
Russia	7,844,969	65	35	3,039,762	45	55
Ukraine	1,257,890	52	48	711,586	49	51
Kazakhstan	456,940	57	43	201,775	46	54
Belarus	310,819	45	55	217,143	47	53
Germany	128,168	43	57	79,140	41	59
Azerbaijan	102,726	31	69	44,150	15	85
Uzbekistan	86,010	22	78	40,485	25	75
Moldova	78,835	40	60	54,561	44	56
Armenia	68,334	43	57	22,382	18	82
Georgia	67,554	20	80	33,022	21	79
Latvia	53,433	59	41	29,512	53	47
Estonia	48,243	52	48	26,731	48	52
USA	47,111	40	60	30,517	41	59
Israel	42,627	37	63	27,296	37	63
England	35,938	62	38	14,989	35	65
Turkey	35,001	16	84	23,884	14	86
Lithuania	34,795	59	41	16,481	48	52
Kyrgyzstan	32,798	36	64	16,592	38	62
Italy	18,389	42	58	13,635	43	57
Spain	18,220	38	62	11,503	40	60
France	11,988	36	64	7,187	33	67
Turkmenistan	11,609	31	69	5,952	34	66
Canada	10,623	36	64	6,604	35	65
Greece	10,092	30	70	7,088	30	70
Tajikistan	9,879	14	86	3,917	14	86
Czech	9,401	42	58	6,443	43	57
Poland	9,376	65	35	3,171	36	64
Finland	7,186	41	59	4,460	40	60
Sweden	6,348	32	68	4,045	28	72
Norway	5,994	28	72	3,437	28	72
Belgium	5,849	34	66	3,102	29	71
Bulgaria	5,649	31	69	3,719	28	72
Ireland	5,603	35	65	4,061	37	63
Austria	5,474	36	64	3,065	35	65
China	5,277	34	66	3,453	41	59

260 for females and between 23 and 243 for males, *very high* if the number of photos was larger than 260 for females and larger than 243 for males.

Attributes such as *Car, Languages, Religion, Leisure, Interests, Sports, Music* and attributes describing body characteristics whose exact values are not important for classification but only the fact of their disclosure in a profile, were encoded as binary attributes: if the information about any of these attributes was revealed, it was encoded as *True*, otherwise it was treated as *False*. On the other hand, attributes, whose values are used for classification were encoded as multi-valued categorical attributes. For example, the *Marriage* attribute has four explicit options and one implicit *no answer*. In this case the four options were encoded like *1,2,3,4*, and *0* in the case of non-disclosure. Another group of attributes that may take more than one value (when the user chooses more than one answer) was decomposed into separate binary attributes representing distinct categories. For example, the user can select any of the 5 different categories related to the aim on the site (*Aim* attribute). In case a person selects some category, a binary *True* is assigned to that attribute, otherwise *False* is assigned (*Aim* not disclosed). Two binary attributes that were composed from the *Looking for*, namely *Looking for a man* and *Looking for a woman* were removed since they are found in the majority of profiles, highly correlated with the opposite gender and trivial in terms of gender classification.

4.3 Model construction

Our research hypothesis is that specific gender differences exist on the country level as well as there are differences between the same-genders in different countries. The differences should be expressed in specificity of attributes and values that describe the gender. In other words, we hypothesize that profiles of females and males living in the same country have unique characteristics, which characterize the gender of the owner of the profile. In addition, we hypothesize that, although the main characteristic of the users of the featured dating site is Russian language, cultural and national differences impact the characteristics of user profiles even for people of the same gender across countries. In our study, the data mining process that can capture unique characteristics of the genders is based on decision tree learning, which constructs a classification model using input variables for prediction of the target class value (gender in our case).

We applied C4.5, a popular decision tree induction algorithm [16] to the sampled data for every country with the *gender* as a binary class attribute, using Weka data mining package [17].

Here is a general outline of the algorithm:

- Tree is constructed in a top-down recursive divide-and-conquer manner.
- At start, all the training examples are at the root.
- Attributes are categorical or continuous-valued.
- Examples are partitioned recursively based on selected attributes.
- Split attributes are selected on the basis of a heuristic or statistical measure (in our experiments, we have used information gain).
- The complete tree can be post-pruned to avoid overfitting.

A decision tree can be easily converted into a set of classification rules (one rule per each terminal node). Tables 5 and 6 show examples of classification rules extracted from the induced decision trees. We left all the options in the default state namely: the minimum number of instances per leaf was 2, pruned decision tree, 0.25 pruning confidence factor. Table 3 shows the total number of classification rules and the number of rules by gender generated for every country.

Table 3. The total number of rules by country and the number of generated rules by gender

Country	**All Rules**	**Female**	**Male**
Russia	70,719	34,605	36,114
Ukraine	21,181	10,315	10,866
Kazakhstan	5,863	2,815	3,048
Belarus	8,343	4,062	4,281
Germany	3,221	1,482	1,739
Azerbaijan	781	338	443
Uzbekistan	945	418	527
Moldova	1,754	818	936
Armenia	490	191	299
Georgia	649	267	382
Latvia	1,721	812	909
Estonia	1,433	692	741
USA	1,581	723	858
Israel	1,024	453	571
England	784	350	434
Turkey	350	149	201
Lithuania	1,150	534	616
Kyrgyzstan	656	292	364
Italy	699	327	372
Spain	791	382	409
France	483	209	274
Turkmenistan	234	106	128
Canada	404	175	229
Greece	334	156	178
Tajikistan	134	48	86
Czech	587	280	307
Poland	256	127	129
Finland	314	149	165
Sweden	307	135	172
Norway	193	91	102
Belgium	225	94	131
Bulgaria	158	79	79
Ireland	267	124	143
Austria	227	106	121
China	226	104	122

5 Analysis

The purpose of this section is to analyze the data and the model described in Section 4. We apply a number of analytical steps to test our hypotheses that there are differences between genders and that these differences are country-dependent.
The analytical steps are:
(1) Analysis of the sampled data
(2) Analysis of the quantity of rules that classify females and males
(3) Cross-country similarity
(4) Gender characterization

5.1 Data Analysis

As was mentioned in Section 2, we applied four filtering steps to minimize the effect of false profiles. By inspecting the initial and resulting number of females and males (Table 2), we can deduce cross-country differences on the gender level.

Russia, Poland, England, Latvia, Lithuania, Kazakhstan, Ukraine, and Estonia are countries in which the number of female users outnumber male users. The difference is as large as 30% for Russia and Poland and as small as 4% in Ukraine and Estonia. After applying the four filtering steps, only Latvia remains a single country among the eight mentioned above where the number of females still outnumber male users, however, the difference decreases from 18% to 6%. Since the number of people that do not have photos in their profile is much larger than the number of people who do not meet the requirement of the first three filtering steps, we may conclude that more females do not have photos in their profiles. As the photo is one of the important components of a dating site, we can also assume that male users apply more efforts to find romantic partner than females or that female users would likely establish a relationship independent of physical appearances.

Uzbekistan, Georgia, Turkey and Tajikistan are the four countries that stand out in the difference between the number of male and female users: Uzbekistan (56%), Georgia (60%), Turkey (68%), Tajikistan (72%). This number does not almost change after applying the filtering step. Armenia, Poland, Russia, Tajikistan, England, Azerbaijan, Kazakhstan, Uzbekistan, Lithuania, and Georgia lose more than 50% of users, while Greece, Ireland, and Italy lose less than 30% after applying the filtering process.

5.2 Model Analysis

We use several different metrics to analyze gender differences in homogeneity and heterogeneity as well as variability of information revealed in user profiles by analyzing classification rules. The metrics, presented in Table 4, include the average amount of male/female members per rule (larger numbers indicate higher homogeneity), the number of male/female rules that cover 90% of the instances in the sampled dataset (larger numbers indicate higher heterogeneity), and the ratio of the number of male/female rules to the entire male/female population (larger numbers indicate higher variability). We also compute the difference between the male and the female rule ratios.

The inspection of the average number of female and male users that are classified per one rule (Table 4), shows that there are only three countries Latvia (4%), Lithuania (1%), and Ukraine (1%), in which the average number of females classified per rule is larger than in the other 32 countries. Such countries as Tajikistan (28%), Uzbekistan (33%), Armenia (40%), Georgia (42%), Azerbaijan (65%), and Turkey (80%) are countries with the largest difference between the average number of female and male users classified per rule out of 32 countries where the average number of males per rule outnumber females. However, in 31 countries the number of rules that cover 90% of the population is larger for females with the greatest difference in Azerbaijan (71%), Bulgaria (66%), Turkey (66%), Uzbekistan (65%), Georgia (65%), Poland (53%), Greece (51%), while Finland (4%), Estonia (9%), Lithuania (10%), Latvia (29%) are the only four countries where the number of rules that cover 90% of the population is larger for males. This finding may suggest that female users are more creative in profile construction and provide more heterogeneous information about themselves, while males reveal more homogeneous information to describe themselves. This is also supported if we inspect the amount of rules generated for females and males relative to the number of females and males in the data set (Table 4). The amount of rules in the percentage relationship is higher in 32 cases for females. The highest relative amount of rules for female population is in Sweden (10.92%), Poland (11.13%), Belgium (10.45%), Czech Republic (10.11%) and the lowest in Ukraine (2.96%) and Russia (2.53%). For the male population, the highest relative amount of rules is observed in Czech Rupublic (8.36%), Lithuania (7.19%), Latvia (6.55%), and the lowest in Azerbaijan (1.18%) and Turkey (0.98%).

Another interesting observation are cross-country and cross-gender variabilities of the relative amount of rules. The difference between the highest (Sweden) and the lowest (Russia) relative amount of rules for females is 9.39%, while the difference between the highest (Czech Republic) and the lowest (Turkey) relative amount of rules for males is 7.38%. If we assume that information disclosed in the users' profile is a deliberate and considerate act that also reflects personal traits of a person (otherwise the profile would have been randomly filled) and the variability of rules shows the variability in different facets of personal traits then our observation of cross-country variability between females and males in relative amount of rules is orthogonal to previous studies. For example, [18] showed that the considerable gender differences in personality traits are among European and American cultures, whereas the miniscule differences are among African and Asian cultures. In our case, the highest cross-gender difference in the relative amount of rules is in Tajikistan (6.20%), which is an Asian country, Sweden (6.01%), and Norway (5.33%) while the lowest is in Estonia (0.06%) and Ukraine (0.04%). Other Asian countries such as Azerbaijan (3.92%) or Turkey

(3.48) are ranked on the seventh and tenth place, respectively among the countries with highest variability of the 35 countries we analyzed. Consequently, our results indicate that the gender differences are not emphasized by the Russian-speaking users in masculine countries [19]. The Masculine Index of scandinavian countries is very low according to Hofstede Masculine Index [20], while Sweden and Norway share the second and the third places in the magnitude in differences between females and males. However, these differences may be attributed to the fact that the Swedish and Norwegian Russian-speaking members of the dating site have a stronger influence of their original culture rather than the culture of their current residence.

Any decision tree construction algorithm builds rules by determining the best attributes that build up the tree. The attribute at the root of the tree is the first attribute selected and, thus, is the best in the classification model to discriminate between genders. Inspection of the root attributes of the models reveals four groups of countries:
(1) The majority of countries (14 in total) namely Spain, Kyrgyzstan, Lithuania, Italy, Ireland, Greece, Estonia, England, China, Moldova, Latvia, Kazakhstan, Israel, and Belarus are characterized by the attribute *AimSex* (the aim on the site is to find a partner for having sex).
(2) Turkmenistan, Poland, Norway, France, Czech Republic, Canada, Bulgaria, Austria, USA, Uzbekistan, Ukraine, and Russia are countries in which the classification tree is splitted according to the *Car* attribute.
(3) Turkey, Tajikistan, Georgia, Armenia, and Azerbaijan are characterized by *MinMaxAge*. This attribute holds the desired age range of a romantic partner.
(4) The remaining four countries: Sweden, Finland, Belgium, and Germany are characterized by the *kids* attribute, which specifies whether the person does or does not have kids, whether the kids live in family or separately or if the person wants to have kids.

5.3 Cross-country similarity

In Section 4.3 we applied a decision tree construction process to the user profiles from 35 countries, and generated models that contain a number of rules that discriminate between females and males in a specific country. As mentioned already, classification trees are used for predicting the target class value. Usually, in order to estimate a classifier's predictive performance, the model is evaluated on a separate test set. In the context of our analysis, we have applied the classification rules generated for each country to the data of other 34 countries. The high classification rate in this case should suggest that there is a high similarity between user profiles (including user information disclosure) across countries. As a result of the evaluation, we have a 35-dimensional vector of classification accuracies for each of the 35 countries (including the training accuracy of the model on the data that was used to induce the model). We applied Multidimensional Scaling (MDS)[21], a widely used data exploratory technique, on the 35×35 matrix of classification accuracies. MDS performs transformation of multidimensional space into a two-dimensional coordinates by preserving the

Table 4. The average amount of females and males classified per rule, the number of rules that cover 90% of the instances in the sampled dataset, percentage of rules relative to the female and male population, and difference between relative amount of rules

Country	**Females Per Rule**	**Males Per Rule**	**90% Rule Coverage Male**	**90% Rule Coverage Female**	**Female Rel. Amount Rules (%)**	**Male Rel. Amount Rules (%)**	**Difference**
Russia	40	46	5,707	3,815	2.53	2.16	0.37
Ukraine	34	33	2,060	1,951	2.96	2.99	-0.04
Kazakhstan	33	36	590	513	3.03	2.80	0.24
Belarus	25	27	1,094	1,006	3.98	3.72	0.26
Germany	22	27	498	442	4.57	3.72	0.84
Azerbaijan	20	85	179	52	5.10	1.18	3.92
Uzbekistan	24	58	165	57	4.13	1.74	2.39
Moldova	29	33	245	193	3.41	3.06	0.34
Armenia	21	61	93	47	4.74	1.63	3.11
Georgia	26	68	121	42	3.85	1.46	2.39
Latvia	19	15	267	374	5.19	6.55	-1.36
Estonia	19	19	235	259	5.39	5.33	0.06
USA	17	21	310	261	5.78	4.77	1.01
Israel	22	30	155	125	4.49	3.32	1.16
England	15	22	170	108	6.67	4.45	2.22
Turkey	22	102	73	25	4.46	0.98	3.48
Lithuania	15	14	224	251	6.75	7.19	-0.44
Kyrgyzstan	22	28	116	91	4.63	3.54	1.09
Italy	18	21	143	112	5.58	4.79	0.79
Spain	12	17	204	135	8.30	5.93	2.38
France	11	18	115	97	8.81	5.69	3.12
Turkmenistan	19	31	52	29	5.24	3.26	1.98
Canada	13	19	87	75	7.57	5.33	2.24
Greece	14	28	88	43	7.34	3.59	3.75
Tajikistan	11	39	32	19	8.75	2.55	6.20
Czech	10	12	157	138	10.11	8.36	1.75
Poland	9	16	75	35	11.13	6.36	4.77
Finland	12	16	68	71	8.35	6.17	2.19
Sweden	8	17	79	69	11.92	5.91	6.01
Norway	11	24	57	34	9.46	4.12	5.33
Belgium	10	17	56	48	10.45	5.95	4.50
Bulgaria	13	34	44	15	7.59	2.95	4.64
Ireland	12	18	66	46	8.25	5.59	2.66
Austria	10	16	63	50	9.88	6.07	3.81
China	14	17	52	42	7.35	5.99	1.36

relative distances (we used squared Euclidean distance measure) between original multi-dimensional vectors. Thus, the countries located close to each other on the two-dimensional graph are more similar in information disclosure between their residents than countries that are located farther away.

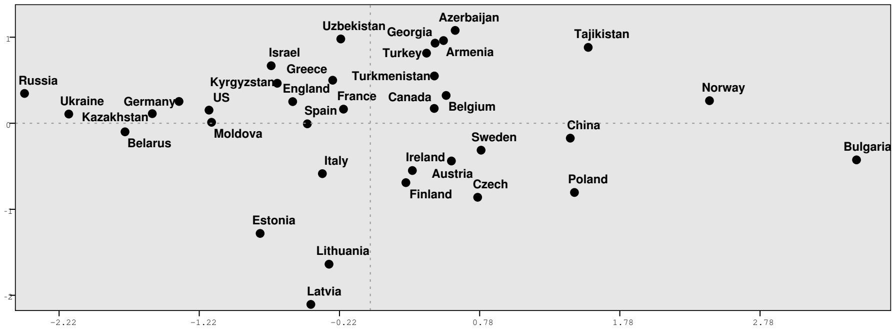

Fig. 1. MDS plot of similarity of information disclosure across 35 countries

Figure 1 shows the results of multidimensional scaling. It is possible to visually discern eight clusters according to the similarity in user profiles. Russia and Ukraine are located close to each other and can form the first cluster. Germany, US, Kazakhstan, Belarus, and Moldova are located close to each other and form the second cluster. Uzbekistan, Israel, Kyrgyzstan, Greece, England, Spain, and France are members of the third cluster. The fourth cluster includes Armenia, Turkey, Azerbaijan, Georgia, Turkmenistan, Belgium, and Canada. Sweden, Austria, Ireland, Finland and Czech Republic are in the fifth cluster. Italy is located equally distant from other countries and can be a single country in the sixth cluster. Estonia, Lithuania, and Latvia are in the seventh cluster. Tajikistan, China, Poland, Norway, and Bulgaria are located farther than other countries. We assign them to an eighth cluster. The first and the seventh cluster include countries that are located geographically close to each other. This may suggest that cultural similarities between those countries play a crucial role in the similarity of user profiles. Similar observations were reported in [22, 23] in the study of personality traits. Other clusters include a mix of close and faraway countries. For example, the fourth cluster contains five Asian countries geographically close to each other such as Armenia, Azerbaijan, Turkmenistan, Turkey, and Georgia as well as two countries situated in Europe and America. A notable feature of the cluster two is that Kazakhstan and Germany are located close to each other. While those countries are not located close geographically, it is known that a significant number of Russian Germans now living in Germany, immigrated from Kazakhstan during 1990s where their ancestors had lived in the late 19th Century.

A more consistent way to summarize the similarities is to explicitly cluster the countries according to the similarities and differences in member profiles.

We applied Farthest Neighbor (complete linkage) hierarchical clustering using cosine similarity between 35-dimensional vectors. Results of the clustering are shown in Figure 2. It can be clearly seen that some clusters are discerned by geographically close countries. For example, Armenia, Georgia, Turkey, Azerbaijan, and Tajikistan are linked together at the first level of hierarchical clustering. Likewise, Russia, Ukraine, and Belarus or Uzbekistan and Turkmenistan. On the highest level of clustering, Estonia, Lithuania, and Latvia are linked with all other countries suggesting the considerable difference in member profiles between the three countries and the rest of the countries.

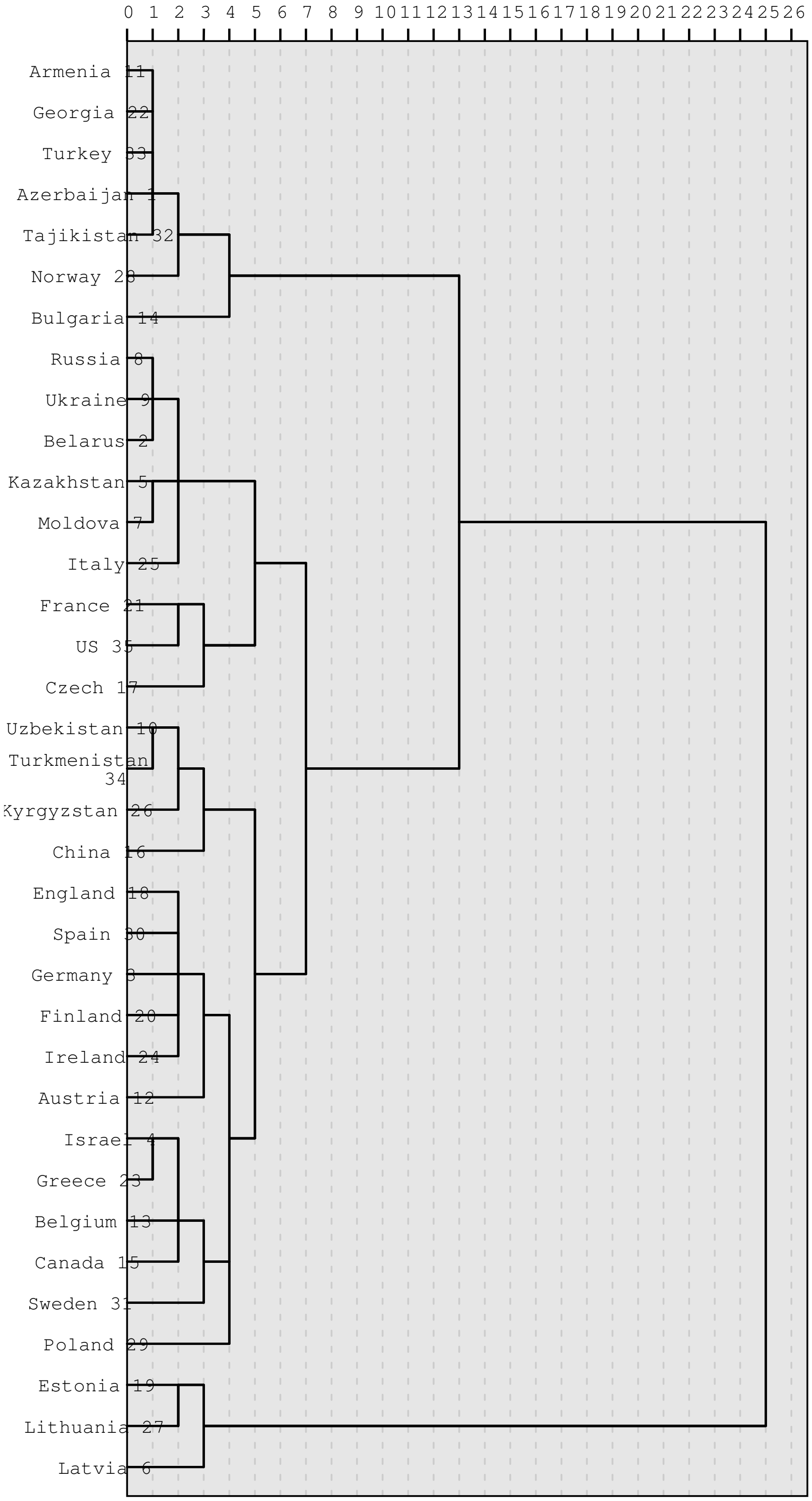

Fig. 2. Hierarchical clustering diagram of similarities of information disclosure accross 35 countries

5.4 Gender characterization

Since the space limitation does not allow us to present the whole list of rules generated for every gender and country, we provide a number of rule examples picked from the set of most frequent rules. We arbitrarily selected two representative countries from every visible cluster produced by the multidimensional scaling (Figure 1) described in the previous section. Tables 5 and 6 show frequent classification rules that distinguish between females and males across countries. The rightmost column shows the number of males (M) and females (F) that correspond to each rule.

The inspection of the rules show some clear-cut patterns. The sex and car components are dominated in "male" rules (rules that classify a person as a male). Information about kids and the desired age of a romantic partner is dominated in "female" rules. A noteworthy feature found in "male" rules is that they are relatively short. There are cases where a single attribute can classify a person as a male like in the case of Bulgaria and China.

6 Conclusions

In this paper we investigated gender differences and patterns of information dislosure between countries in the context of dating sites using the Data Mining approach.

We applied decision tree construction algorithm to the user profiles from 35 most active countries using more than 10 million profiles from one of the biggest dating sites in the Russian segment of the Internet. We analyzed the induced classification rules and outlined differences between genders within and across countries. We used Multidimensional scaling and hierarchical clustering to analyze similarities and differences in member profiles and information disclosure across countries. The Russian-speaking residents of some geographically close and culturally similar countries exhibit higher similarity in information disclosure between user populations living in those countries.

We showed that social phenomena can be investigated by applying data mining methods to large quantities of user profile data, and that statistical analysis alone is not enough for finding interesting patterns. Our research overcomes the limitations of most previous studies, where the analysis was performed on small, non-representative and non-generalizable samples of the user population. However, some uncertainty is associated with the large-scale analysis of real profiles mined from a social networking site, since the analyst cannot verify the real purpose of profile creation (whether it has a serious intention or was created for fun). At this point, we assume that the majority of SNS users have real profiles that reflect their real self. Automated cleaning of profile data may be a subject of future research.

Our study provided insights into the patterns of gender differences across countries. The reasons for such differences can be unlimited: influences of the

Country	**Rule**	Gender (F/M)
Russia	IF I have kids AND I am not married AND I specified what I'd do on my free day AND information about a car and smoking habits is not revealed AND I did not reveal that my aim is to have sex	F (21,712/160)
Ukraine	IF I have kids AND I am not married AND my age is between 23.5 and 29 AND information about a car is not revealed AND I did not reveal that my aim is to have sex	F (7,965/103)
Germany	IF I have kids AND I am not married AND I did not reveal that my aim is to have sex AND information about a car is not revealed	F (4,786/262)
USA	IF I have a car AND I do not have any photo AND I revealed information about my body AND I wrote nothing about myself AND I did not reveal that my life priority is sex AND information about my profession is not revealed	F (322/20)
Israel	IF I have kids AND I am not married AND I did not reveal that my aim is to have sex	F (2,853/223)
England	IF I am looking for a person between 41 and 60 AND information about a car is not revealed AND I did not reveal that my aim is to have sex	F (102/5)
Belgium	IF I have kids AND I have between 1 and 8 photos AND my life priority is to have a family AND I did not reveal that my aim is to have sex AND I am not looking for a friendship with a female or female couple	F (133/22)
Turkey	IF I am looking for a person between 31 and 40 AND my age is between 29 and 34.5 AND information about other aims is not revealed	F (104/6)
Czech	IF I revealed some of my life priorities but did not select that my life priority is self-realization AND I am looking for a person between 18 and 20 AND I revealed what I like in sex AND I earn well AND information about a car or how frequent I'd like to have sex were not revealed	F (276/1)
Sweden	IF I have kids AND information about a car or how frequent I'd like to have sex is not revealed AND I did not reveal that my aim is to have sex	F (242/31)
Italy	IF I am looking for a person between 18 and 20 AND harmony is my life priority AND I do not take drugs AND I did not reveal that my aim or life priority is sex AND I did not reveal my orientation, heterosexual experience or information about a car	F (330/64)
Lithuania	IF I have kids AND I am not married AND I did not reveal that my aim is to have sex AND information about a car is not revealed	F (1,001/59)
Latvia	IF I have kids AND information about a car is not revealed AND I did not reveal that my aim is to have sex	F (3,667/339)
Bulgaria	IF I wrote between 1 and 24 words about myself AND my hobby is music AND information about a car is not is not revealed AND I did not reveal that my aim it to have sex or how frequent I'd like to have sex	F (242/70)
China	IF harmony is my life priority AND I reveal some information about my body AND information about a car or economic condition is not revealed AND I did not reveal that my aim is to have sex	F (375/103)

Table 5. Frequent rules that classify females across countries. The rightmost column shows the number of females (F) in the dataset that are classified by the corresponding rule and the number of males (M) that correspond to the same rule

Country	**Rule**	Gender (M/F)
Russia	IF I have a car AND I am a heterosexual AND I am looking for a person between 18 and 30 AND I am not looking for a friendship with a male and female couple	M (15,973/431)
Ukraine	IF I have a car AND I am a heterosexual AND my aim is to have sex AND I am not looking for a friendship with a male and female couple	M (18,461/902)
Germany	IF I do not have kids AND my aim is to have sex AND information about a car is not revealed AND I am a heterosexual	M (951/101)
USA	IF I have a car AND I have between 1 and 6 photos AND I am looking for a person between 18 and 20 AND my life priority is sex	M (688/66)
Israel	IF my aim is sex AND information about a sexual orientation is not revealed	M (986/140)
England	IF my aim is sex AND I am a heterosexual AND I revealed information about my interests	M (667/35)
Belgium	IF I do not have kids AND I have a car	M (345/42)
Turkey	IF I have a car AND I am looking for a person between 18 and 20	M (3807/152)
Czech Republic	IF I revealed how frequent I'd like to have sex AND information about a car is not revealed	M (344/74)
Sweden	IF I do not have kids AND my aim is to have sex	M (189/15)
Italy	IF I have a car AND I do not have kids AND my aim is marriage AND I did not reveal that my aim is to have sex	M (911/113)
Lithuania	IF I have a car AND I am looking for a person between 18 and 20 AND information about kids is not revealed AND I did not reveal that my aim is to have sex	M (1022/241)
Latvia	IF my aim is to have sex AND I am a heterosexual AND I wrote nothing about myself AND I am not looking for a friendship with a homosexual male, and heterosexual couples	M (748/61)
Bulgaria	IF I have a car	M (1156/132)
China	IF I have a car	M (245/30)

Table 6. Frequent rules that classify males across countries. The rightmost column shows the number of males (M) in the dataset that are classified by the corresponding rule and the number of females (F) that correspond to the same rule

hosting country' culture, immigration, spoken language, original culture, personal traits. Therefore, we could not provide exact explanations of such difference and did not attempt to speculate on possible reasons. Moreover, the meaning of gender differences could be explained by domain experts like anthropologists, culturalists, behaviorists or sociologists. Without a doubt further studies are necessary. Previous studies on gender differences [24, 15] have been carried out on a much smaller scale in the context of "digital divide"[4]. The results of such studies can affect design principles and guidelines and provide insights for the development of SNS and other information systems. However, these potential applications are beyond the scope of this paper.

The preliminary results provided in this paper are encouraging, though the work presented here is exploratory in nature. In our future work, we will apply more analytical methods to conduct all-embracing analysis of gender differences, user profiles, and information disclosure and work closely with social scientists to test hypotheses that so far have been evaluated on very limited amounts of user data.

Acknowledgements

This work was partially funded by the German Research Society (DFG) under grant GK-1042 (Research Training Group "Explorative Analysis and Visualization of Large Information Spaces")

References

1. Nelson, S., Simek, J., Foltin, J.: The Legal Implications of Social Networking. Regent University Law Review **22**(1) (2009) 2
2. Acquisti, A., Gross, R.: Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: Privacy Enhancing Technologies, Springer (2006) 36–58
3. Young, A., Quan-Haase, A.: Information revelation and internet privacy concerns on social network sites: a case study of facebook. In: Proceedings of the fourth international conference on Communities and technologies, ACM (2009) 265–274
4. Kisilevich, S., Mansmann, F.: Analysis of privacy in online social networks of Runet. In: Proceedings of the 3rd International Conference on Security of Information and Networks, ACM (2010)
5. Ellison, N., Heino, R., Gibbs, J.: Managing impressions online: Self-presentation processes in the online dating environment. Journal of Computer-Mediated Communication **11**(2) (2006) 415
6. Thelwall, M.: Social networks, gender, and friending: An analysis of MySpace member profiles. Journal of the American Society for Information Science and Technology **59**(8) (2008) 1321–1330

[4] The phrase "digital divide" has been used to refer to a wide variety of inequities, including differential access to, contact with, and use of ICTs cross-nationally as well as between social and demographic groups within individual nations [15]

7. Pfeil, U., Arjan, R., Zaphiris, P.: Age differences in online social networking-A study of user profiles and the social capital divide among teenagers and older users in MySpace. Computers in Human Behavior **25**(3) (2009) 643–654
8. Grasmuck, S., Martin, J., Zhao, S.: Ethno-Racial Identity Displays on Facebook. Journal of Computer-Mediated Communication **15**(1) (2009) 158–188
9. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology (2009)
10. Pedersen, S., Macafee, C.: Gender differences in British blogging. Journal of Computer-Mediated Communication **12**(4) (2007) 1472
11. Goodchild, M., Anselin, L., Appelbaum, R., Harthorn, B.: Toward spatially integrated social science. International Regional Science Review **23**(2) (2000) 139
12. Kleinberg, J.: The convergence of social and technological networks. Commun. ACM **51**(11) (2008) 66–72
13. Cozby, P.: Self-disclosure: A literature review. Psychological Bulletin **79**(2) (1973) 73
14. Golub, Y., Baillie, M., Brown, M.: Gender Differences in Internt Use and Online Relationships. American Journal of Psychological Research **3**(1) (2007)
15. Jones, S., Johnson-Yale, C., Millermaier, S., Pérez, F.: US College Students' Internet Use: Race, Gender and Digital Divides. Journal of Computer-Mediated Communication **14**(2) (2009) 244–264
16. Quinlan, J.: C4. 5: programs for machine learning. Morgan Kaufmann (1993)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter **11**(1) (2009) 10–18
18. Costa, P., Terracciano, A., McCrae, R.: Gender differences in personality traits across cultures: Robust and surprising findings. Personality and Social Psychology **81**(2) (2001) 322–331
19. Hofstede, G.: Masculinity and femininity: The taboo dimension of national cultures. Sage Publications (1998)
20. Hofstede, G.: Culture's consequences: International differences in work-related values. Sage Publications (1984)
21. Borg, I., Groenen, P.: Modern multidimensional scaling: Theory and applications. Springer Verlag (2005)
22. Allik, J., McCrae, R.: Toward a geography of personality traits: Patterns of profiles across 36 cultures. Journal of Cross Cultural Psychology **35**(1) (2004) 13–28
23. Schmitt, D., Allik, J., McCrae, R., Benet-Martinez, V., Alcalay, L., Ault, L., et al.: The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. Cross Cultural Psychology **38**(2) (2007) 173
24. Jackson, L., Zhao, Y., Kolenic III, A., Fitzgerald, H., Harold, R., Von Eye, A.: Race, gender, and information technology use: the new digital divide. CyberPsychology & Behavior **11**(4) (2008) 437–442