Russian Computer Scientists, Local and Abroad: Mobility and Collaboration

Vincent Lepinay
EUSP St. Petersburg, Russian Federation and Sciences Po, Paris, France

Jean Philippe Cointet
INRA Sens, Paris, France

Lionel Villard
ESIEE, IFRIS, Paris, France

Andrei Mogoutov
IFRIS, Paris, France

ABSTRACT

In this paper we present the first results of the first comprehensive study of a population that has drawn attention over the past few years, Russian computer scientists (CS) and IT specialists. We collected data from digital platforms were CS and IT leave either signatures or digital traces. The difference between signatures and traces is the difference between intentional scientific claims (an article or a vitae) and by-products of activities that take place on the web. Digital signatures are a digital mode of existence of objects that exist otherwise; digital traces only exist on digital platforms.

Categories and Subject Descriptors

J.4 SOCIAL AND BEHAVIORAL SCIENCES - Sociology; K.4 COMPUTERS AND SOCIETY K.4.2 Social Issues – Employment, K.4.3 Organizational Impacts – Employment; K.7 THE COMPUTING PROFESSION - K.7.0 General, K.7.1 Occupations

General Terms

Management, Economics

Keywords

Diaspora, Russian computer scientists, Scientometrics, Web of Science, GitHub, LinkedIn, PatStat

1. INTRODUCTION

The institutional background of our paper is a collaboration between historians, sociologists and anthropologists at European University at Saint Petersburg. The project « Russian Computer Science (RCS) » is designed to study the strategies of Russian computer scientists who circulate within the Russian federation and beyond when they emigrate. The bulk of the research at EUSP belongs to the genre of scientific diaspora studies. Russian graduate students are sent to sites of intense Russian computer science activities, both academic and entrepreneurial, in Russia and abroad.

This is a mostly qualitative research project meant to understand the professional strategies of the (now) one and unique scientific disciplines where Russia can boast some success worldwide. Celebrated mathematicians are mostly all abroad (Perlman at NYU Courant Institute and Smirnov in Geneva, although also in Moscow through a Ministry of Education megagrant); Physics is no longer solidly funded, Biology is also heavily under-funded. In that context, Russian computer science and information technology (CS/IT from here on) thrives, but in ways that surprise all. Just consider the 3 following facts.

A) Russian teenagers win all the hacking prizes organized annually by Facebook or Google; Russian academic CS publish less than Portuguese CS.

B) Russian IT specialists are praised for their unique understanding of programming languages; no standard language has been designed by a Russian computer scientist.

C) One of the up and coming security firms challenging the economic dominance of Norton or Symantec is Kaspersky Lab, a venture launched by a Moscow-based computer scientist; recent worldwide and massive hacking schemes have been traced back to Russia or ex-soviet republics, and not to Chinese or Indian hackers.

The qualitative surveys launched in Fall 2013 and continuing through Fall 2015 will yield precious insights as to the organization of communities of Russians CS/IT. One of the puzzles we expect to clarify is the role of migration out of a still largely totalitarian environment on the culture of trust. The CS population is highly interesting for us as it is torn between 2 imperatives, deeply embedded in the computing culture:

- sharing, collaboration

- awareness of privacy and culture of codes

Recent anthropological studies of populations involved in the development and deployment of computing languages and protocols have shown how the ethos of hackers is a complex formula of technical expertise animated by the unique feature of a transparent language. Computer codes are the only fully explicit language but they are also the vectors of the most opaque operations, as the recent NSA ventures into cracking personal information owned by telephone and web operators have amply demonstrated. This tension between the most public and sharable and the most hidden and arcane takes an interesting turn for the population that we elect to study. Of the Russian CS, we focus on those who cross the threshold of academia and engage in entrepreneurial activities. Being Russian and born at the end or right after the communist experiment, crossing that boundary and cultivating the style of entrepreneurship does not come naturally:

- the scientific ethos - in its clearest mertonian form - rules out the cultivation of hybrid virtues that are so characteristic of the American university/industry partnerships.

- the form of collaboration and the trust that are needed in early entrepreneurial ventures are not the virtues cultivated under the soviet era.

Both conditions make the observation of RCS switching to economic ventures particularly fruitful to us. We ask the following research question :

(1) What are the patterns of collaboration and trust among Russian CS/IT professionals?

(2) What are the consequences of the intangible nature of CS/IT's production on the possibility of collective ventures?

We use the natural experiment of a population divided into residents and expatriates to factor the geographic mobility into the decision to bridge science and collective enterprises, for profit or else.

Entrepreneurs, coming from the soviet and post-soviet background who focus on CS and IT offer a new sandbox for trust studies. Distrust is not abnormal; rather it is the default or fall back position in a population for whom private ownership is still an unstable position, absent a solid legal framework. The intangible nature of IT and CS productions and the difficulty of tracking them beyond their lines of codes add an additional reversal to a culture in which value was solidly equated to material goods.

Objective: disentangle the series of factors that weigh in the circulation of highly skilled professionals across territories with radically different political and legal characteristics (St. Petersburg and Moscow vs Silicon Valley) and dealing with highly mobile and intangible products. The experience of migration in the dual context of this politico-legal differential AND products' high level of compatibility creates a new research question for migration scholars. Most promising among these questions is that of the political models of production that are contained in the programming discipline. Coding skills allow mobility to an extent that other skills do not: as much as programming languages are specific, they always have family resemblance with other languages so that coders' skills have an underlying generality. And as much as skills of medical doctors could and should be universal and easily transportable from one region to the next, they are always regulated and controlled by national authorities. Not so with coding skills. The Soviet Union and subsequently the Russian federation have never let go of the dream of controlling their population but simultaneously they have also been the sites of extraordinary development of the most un-authoritarian forms of skill – a kind of skill prone to subvert the very apparatus of control and hierarchy that it simultaneously establishes. Coding and programming empower mobility and professional versatility; how do they fare in relation to trust and the construction of durable associations?

2. BACKGROUND AND METHODS

Studies of entrepreneurship or simply collaboration are hardly new. Our proposed research draws on the past 30 years of STS but it works on a population and with methods that enrich the field, as well as related fields, primarily migration studies, sociology of profession and economic sociology.

We are interested in patterns of collaboration that may lead to partnering up into business or not-for-profit association. Our target population has a unique characteristic that our project leverages: Russian computer scientists are present, in some capacity, on the web. Whether one is a freelancer, working from India for an American IT company, or an academic computer scientist publishing in Russian or English journal, some visibility singularizes this population, as opposed to other professionals.

This characteristic is both an analytic resource and a source of puzzles for our research project. The traces left by professionals allow us to capture their presence, yet their identity may not be unique when they leave traces. An individual can show up under ALEX_BELOV on GitHub; he can become BelovSasha on some other website.

We explore sources of information on which CS and IT professionals will be expected to "express themselves" and to leave digital traces. We use following complementary data sources for the IT & CS diaspora detection and description:

- Scientific publications in the field of Computer Science (source: Web of Science)

- Patents - IT related industrial applications (source: PatStat)

- Professional networks (source: LinkedIn)

- Open source software repository (source: GitHub)

2.1 Scientific publications in the field of Computer Science

As a platform that enables global monitoring framework we choose the Web of Science - the largest database of scientific publications, relevant primarily for the exact sciences and engineering and is one of the possible sources to identify CS involved in scientific research.

We extracted the data set from the Web of Science using a query combining the subject categories and publication years. We have selected formally all computer science related categories: COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE, COMPUTER SCIENCE HARDWARE ARCHITECTURE, COMPUTER SCIENCE INFORMATION SYSTEMS, COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS, COMPUTER SCIENCE SOFTWARE ENGINEERING, COMPUTER SCIENCE THEORY METHOD

Total number of selected publications is 1499127 for a period of 1985 – 2012

We parsed and transformed the data to a Sqlite database using CorText platform (cf. [1], [2], [3]). For each record we extracted the countries of affiliation of all authors and their names and surnames. We obtained general statistical distributions for authors and affiliation related fields of the database.

Countries/Territories	records	% of 1788013
USA	444557	24.863
PEOPLES R CHINA	250229	13.995
JAPAN	90587	5.066
GERMANY	89700	5.017
ENGLAND	87211	4.878
FRANCE	72904	4.077
CANADA	69522	3.888
ITALY	57385	3.209
SOUTH KOREA	50031	2.798
SPAIN	49856	2.788
TAIWAN	48939	2.737
AUSTRALIA	40867	2.286
INDIA	38458	2.151
NETHERLANDS	31143	1.742
BRAZIL	19737	1.104
SINGAPORE	19446	1.088
POLAND	19275	1.078
SWITZERLAND	19046	1.065
ISRAEL	18891	1.057
GREECE	17605	0.985
BELGIUM	16515	0.924
SWEDEN	16328	0.913
AUSTRIA	15126	0.846
IRAN	14921	0.835
FINLAND	13989	0.782
TURKEY	12893	0.721
PORTUGAL	12321	0.689
SCOTLAND	11903	0.666
RUSSIA	11370	0.636

Figure 1. Distribution of countries by number of publications in the field of computer science for the period of 1985-2013

The rank of Russian CS is very low- the 29th place for the overall number of publications for the period of 1985-2012.

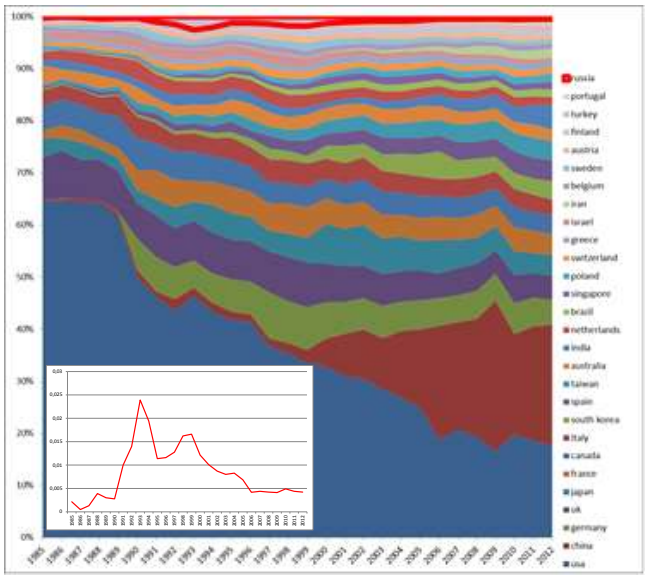


Figure 2. Number of publications by country and year. Inserted chart shows the evolution of relative part of publications with Russian affiliations compared with worldwide CS publications

A part of publications with Russian affiliations rapidly grows starting from 1990, attends a peak in 1993 and crashes in 1995, grow up during the period of 1997-1999, slows down starting from 2000 and stabilize around 0.4% starting from 2006.

We can also see that China become a biggest player of the field of CS starting from 2006.

2.2 How to find "Russians" outside of Russia? Diaspora detection algorithm

One of the goals of our study should take into account not only professionals with Russian citizenship, but also immigrants from the so-called "post-Soviet space." We applied a Naïve Bayesian classifier for evaluation of possible "ethnicity" of names and surnames (similar methodology exposed in e (4)). As a training dataset we use a subset of publications indexed in the Web of Science with explicit affiliations in Russia and countries of former Soviet Union. First, every name and surname are transformed to a vector of attributes - the last letter of the surname, the last two letters, the last three letters and the full surname, for example - Andrei Petrov will be considered as a vector of substrings EI REI DREI ANDREI OV ROV TROV. Second, the Bayesian classifier (Library scikit-learn, in particular naïve Bayer module http://scikit-learn.org/stable/modules/naive_bayes.html) is applied to the whole list of all names and surnames of authors. For each name the possible ethnicity is evaluated. Manual check of the names shows that the false positives represent around 7% and the false negatives represents around 5% of names.

We applied simple and formal criteria of the diaspora definition: the authors having county of actual affiliation which differs from the evaluated from name are considered as a part of diaspora.

Third, we have extracted a subset of publications with at least on of co-author having "Russian" name.

We have found 19094 Russian names for the total number of publications with their participation equal to 42938. Taking in account the post-USSR diaspora, the overall production of "Russian" or more precisely post-USSR CS local and abroad would be found at the 12th on the distribution of countries by number of publications.

We extracted a subset of data and obtained some descriptive statistics by country, year of publication and profile of co-authorship

Non-Russian and non-post-Soviet affiliations	N Publications	Mixed affiliations	N Publications	Russian and post-Soviet	N Publications
usa	11700	russia	1978	russia	3571
germany	2722	lithuania	483	ukraine	1071
uk	2589	ukraine	308	lithuania	891
canada	2029	usa	227	latvia	294
france	1503	germany	184	estonia	179
israel	1420	latvia	160		
netherlands	988	france	99		
australia	923	sweden	97		
italy	796	uk	91		

Figure 3. Places of the diaspora. Number of publication by country for following subgroups: papers with exclusively Russian affiliations, mixed Russian and non–Russian affiliations and exclusively non-Russian affiliations

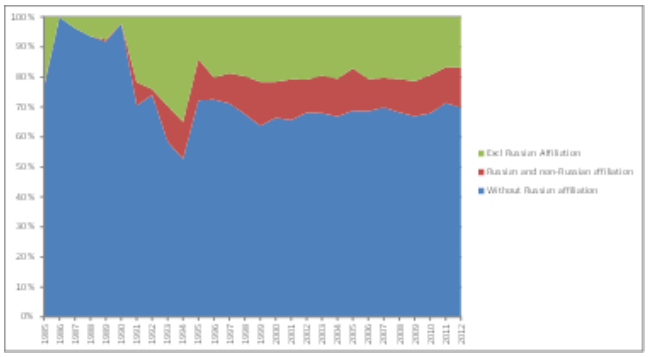


Figure 4. Number of publication per year with exclusively Russian affiliations, mixed Russian and non –Russian affiliations and exclusively non-Russian affiliations.

The obtained results show that:

The number of publications with Russian affiliations is very low, Russia ranks 29th.

The Russian (and post-Soviet) diaspora plays an important role. More than 68% of the overall number of publications having co-authors with Russian names is produced by diaspora. 12% have mixed affiliations and only 20% have exclusively Russian affiliations.

The trend is also negative for the domestic Russian computer science: starting from 1999 the part of diaspora related publications is growing constantly.

2.3 Forms of collaboration and appropriation: Patents

Russian CS offers an interesting case for the study of technological entrepreneurship. CS and IT fields are allegedly the least equipment-intensive of Big Sciences. The threshold leading to entrepreneurship is easier to cross in CS than in most other sciences if one assesses the cost of crossing on the basis of the infrastructures needed to conduct research. Center of this transition from research to entrepreneurship is usually a piece of code, at times a software with direct marketing possibilities. Codes are problematic objects at the intersection of computer science and economies of information technologies.

Here we leverage the existence of databases pointing to two different strategies of appropriation of codes: patenting and sharing.

We use the data base PatStat for extracting patents with IT related industrial applications. We formally apply a search criteria based on the IPC code (International Patent Classification) G06 : COMPUTING; CALCULATING; COUNTING for the period of reference 1985-2011; We obtain 1358098 priority patents, the data set is parsed and transformed to a database with the CorText platform. We apply a procedure of the name-ethnicity evaluation as for the scientific publications.

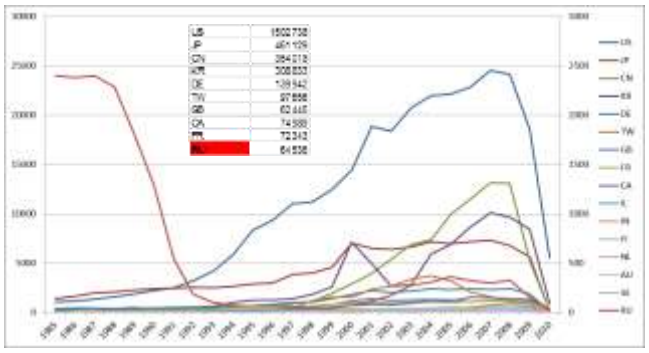


Figure 5. Number of publications by year and country of inventor (The red line corresponding to Russian patents plotted with the right scale)

We can see that Russia has a tenth place by total number of patents for the period 1990-2011

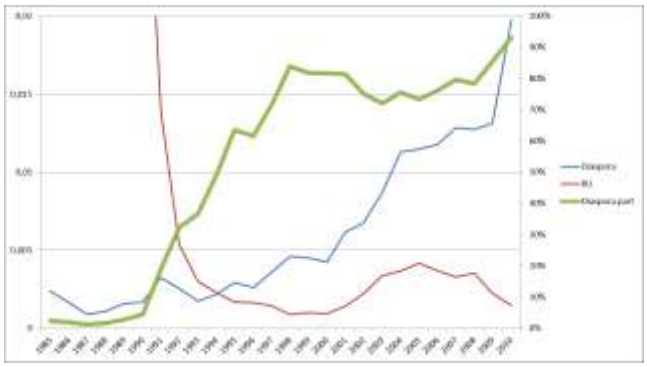


Figure 6. Number of patents by year for Russian domestic inventors (red line) and diaspora (blue line). A relative number of diaspora compared to overall number domestic and diaspora patents (Yellow line, right 100% scale)

The diaspora represents more than 90% of the overall patent production of domestic and diaspora.

The global trend have non-linear shape, one can observe a maximum of ration of diaspora vs. domestic around -1997, a minimal value around 2002. After 2005 the trend is negative for the domestic patents

Country of diaspora	Number of patents
US	12823
CA	1175
IL	1062
DE	854
BG	472
GB	508
UA	384
KR	268
IE	256
CN	232
FR	206
IN	205
BY	193
FI	186
JP	135
AU	98
RO	98
CH	94
GZ	94
SG	84
MD	66
NL	62
RS	45
PL	47

Figure 7. Places of the diaspora. Number of patents Russian diaspora by country of their location or affiliation

Country	Applicant	NbPatents
US	MICROSOFT CORP	1152
US	IBM	909
US	SUN MICROSYSTEMS INC	222
US	METROLOGIC INSTR INC	199
US	INTEL CORP	149
US	LSI LOGIC CORP	103
US	SYMBOL TECHNOLOGIES INC	84
US	MOTOROLA INC	74
US	HEWLETT PACKARD DEVELOPMENT CO	73
US	YAHOO INC	64
US	EMC CORP	58
US	ORACLE INT CORP	55
US	SYMANTEC OPERATING CORP	54
US	CITRIX SYSTEMS INC	54
CA	RESEARCH IN MOTION LTD	95
CA	IBM CANADA	27
CA	ONTARIO INC 2012244	22
CA	ATI TECHNOLOGIES INC	21
CA	COREL CORP	20
CA	SEMICONDUCTOR INSIGHTS INC	18
CA	ATI TECHNOLOGIES ULC	16
CA	COGNOS INC	11
IL	SANDISK CORP	10
IL	SANDISK IL LTD	10
IL	FREESCALE SEMICONDUCTOR INC	8
DE	SAP AG	173
DE	SIEMENS AG	35
DE	INFINEON TECHNOLOGIES AG	24
DE	THOMSON BRANDT GMBH	11
DE	FREESCALE SEMICONDUCTOR INC	10
DE	FRAUNHOFER GES FORSCHUNG	9
DE	NANOPHOTONICS AG	6
GB	ADVANCED RISC MACH LTD	18
GB	ACRONIS INC	14
GB	SWSOFT HOLDINGS LTD	9
GB	EBS GROUP LTD	8
KR	SAMSUNG ELECTRONICS CO LTD	107
KR	LG ELECTRONICS INC	7
KR	KORPORATSIJA SAMSUNG EHLEKTRON	4
KR	KORPORATSIJA S1	4

Figure 8. Diaspora Patents Assignees. Top companies selected for most representative countries

One could find big national or international companies as assignees for the diaspora patents.

2.4 Open source repository - GitHub

A part of the Russian CS & IT community could be involved in open source projects. GitHub is considered as a major platform for open code repository. Analysis of the data available in GitHub can provide information on project participants, their names, place of residence and professional affiliations.

Rank	Country	N	%
1	United States	20631	35,9%
2	United Kingdom	4154	7,2%
3	Germany	3752	6,5%
4	France	2441	4,2%
5	Canada	2339	4,1%
6	China	1881	3,3%
7	Japan	1705	3,0%
8	Brazil	1683	2,9%
9	Russia	1521	2,6%
10	Australia	1448	2,5%
11	The Netherlands	1264	2,2%
12	Sweden	1040	1,8%
13	Spain	1021	1,8%
14	India	983	1,7%
15	Poland	768	1,3%
16	Switzerland	733	1,3%
17	Italy	687	1,2%
18	Ukraine	621	1,1%
19	Norway	532	0,9%
20	Belgium	514	0,9%
21	Denmark	439	0,8%
22	Czech Republic	423	0,7%
23	Finland	412	0,7%
24	Austria	410	0,7%
25	Argentina	405	0,7%

Rank	Country	N	%
1	Russia	805	42,6%
2	Ukraine	326	17,2%
3	United States	224	11,8%
4	Belarus	85	4,5%
5	Germany	63	3,3%
6	Canada	40	2,1%
7	United Kingdom	40	2,1%
8	Lithuania	23	1,2%
9	The Netherlands	20	1,1%
10	Bulgaria	19	1,0%
11	Latvia	17	0,9%
12	Australia	17	0,9%
13	China	14	0,7%
14	Sweden	13	0,7%
15	Switzerland	12	0,6%
16	Estonia	11	0,6%
17	Czech Republic	11	0,6%
18	Spain	10	0,5%
19	Georgia	9	0,5%
20	Brazil	9	0,5%
21	Finland	8	0,4%
22	France	8	0,4%
23	Japan	8	0,4%
24	Norway	7	0,4%
25	Israel	7	0,4%

Figure 9. Ranking countries of GitHub contributors (left table). Ranking countries of Russian diaspora found in GitHub (right table)

Rank of Russian GitHub contributors is quite high, the professional community is visible. We could also mention that the part of diaspora represents 31.5% of the overall Russian and post-USSR population of GitHub contributors. So the GitHub community is mostly domestic (68.5%) and visible for the international community numerically dominated by contributors from the US.

2.5 Professional network (LinkedIn)

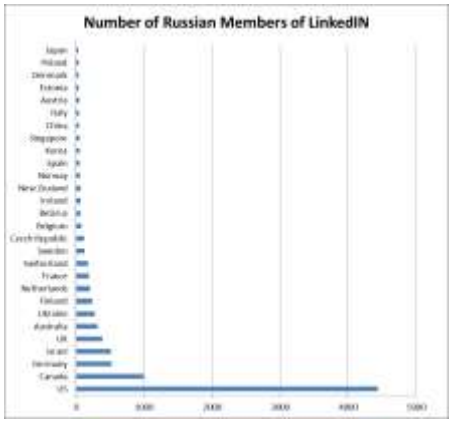
Another type of online sources that allow monitoring of IT professionals could be professional networks such as LinkedIn. Registration on LinkedIn and regularly updated information is becoming increasingly common practice among professionals as well as the resource becomes an important mechanism in the formation of the expert's image and reputation, and also allows you to install more informal contacts with colleagues and potential employers. Profile of LinkedIn members generally contains details about education and career path.

We apply a LinkedIn Premium/ Recruiting & Talent Solutions on LinkedIn as a research tool, growing personal network and using extended search capabilities of the platform.

School:
Lomonosov Moscow State University (MSU)
Saint Petersburg State University
Bauman Moscow State Technical University
Moscow Institute of Physics and Technology (State University) (MIPT)
Novosibirsk State University (NSU)
Saint Petersburg State Polytechnical University
Saint Petersburg University of Telecommunications
Saint Petersburg State Electrotechnical University "LETI"
Moscow Power Engineering Institute (Technical University)
Moscow State Linguistic University
Saint Petersburg State University of Finance and Economics
Tomsk State University
Tomsk Polytechnic University
Moscow State Institute of Electronics and Mathematics (Technical University)
National University of Science and Technology "MISIS" (Moscow Institute of Steel and Alloys)
Moscow Pedagogical State University
Novosibirsk State University of Economics and Management (NSUEM)
Novosibirsk State Technical University (NSTU)
Saint Petersburg State University of Engineering and Economics
Tomsk State University of Control Systems and Radioelectronics
Moscow State University of Economics, Statistics and Informatics (MESI)
Moscow State University of Transport (MIIT)
Novosibirsk Institute of Economics and Management
Tomsk State Pedagogical University

Industry:
Information Technology and Services
Computer Software
Telecommunications
Electrical/Electronic Manufacturing
Information Services
Semiconductors
Computer Networking
Computer & Network Security
Computer Hardware

Figure 9. Search criteria with additional filters by country of actual affiliation or residence



Russia	27078
US	4433
Canada	994
Germany	522
Israel	516
UK	391
Australia	316
Ukraine	279
Finland	236
Netherlands	204
France	193
Switzerland	176
Sweden	128
Czech Republic	122
Belgium	78
Belarus	67
Ireland	64
New Zealand	63
Norway	57
Spain	54
Korea	54
Singapore	51
China	47
Italy	47
Austria	46
Estonia	43
Denmark	43
Poland	42
Japan	33

Figure 10. Number of Russian alumni - LinkedIn members by country of residence/affiliation

The size of the Russian diaspora in the sense of alumni of Russian universities working and living abroad is 9299 persons. Compared with the overall population of Russian CS and IT related members of LinkedIn the diaspora represents 25%.

CONCLUSION

In this paper we present preliminary results of detection and description of Russian Computer scientists, local and abroad. We developed a set of methods for data collection and analysis from multiple sources (scientific publications - Web of Science, industrial activity - Patents, professional network - LinkedIn, open source software projects - GitHub). We test and apply a method of diaspora detection using a name –ethnicity evaluation algorithm. We provide a descriptive statistics about visibility of Russian CS, about geographic location of the diaspora and about trends of their relative evolution. This paper pictures a rather dismal state of academic Russian computer science.

The rank of Russian CS publications is low- the 29th place for the overall number of publications for the period of 1985-2012. The role of the diaspora is important: during last 3 years diaspora

produce 68% of publications produced by post-USSR CS local and abroad. The trend is also negative for the domestic CS community. Starting from the 2000 the part of the papers published by diaspora continues to grow compared with domestic production.

We observe even more negative trends for the patenting activities starting from 2005. The diaspora represents more than 90% of the overall patent production of domestic and diaspora.

We also found that the community of contributors of GitHub platform is mostly domestic (68.5%) and quite visible for the international community numerically dominated by contributors from the US.

The LinkedIn CS and IT members is mostly domestic, 75% of alumni of Russian universities works in Russia.

3. ACKNOWLEDGMENTS

Project “Russian Computer Scientists at home and abroad”, EUSP funded by Mega Grant of the Russian Ministry of Education

4. REFERENCES

[1] http://manager.cortext.net

[2] Jones D, Cambrosio A, Mogoutov A Detection and characterization of translational research in cancer and cardiovascular medicine, Journal of Translational Medicine, 2011; 9: 57

[3] Cointet JP, Mogoutov A, Bourret P, El Abed R, Cambrosio A; The emergence and development of gene expression profiling: a key component of the 3B (bench, bedside, bytes) in translational research, Med Sci (Paris) 2012 ; 28: 7–13

[4] Garcia Flores J. J., Zweigenbaum P., Yue Z, Turner, W.A., Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis, Proceedings of the 8th International Conference on Natural Language Processing, October 22-24, 2012, Kanazawa, Japan: Springer Publishers, Lecture Notes in Computer Sciences