Diversity in the Quality of Team Work in Collaboration Network: Experiments on Wikipedia

Katarzyna Baraniak[1], Marcin Sydow[1,4], Jacek Szejda[2] and Dominika Czerniawska[3]

[1]Polish-Japanese Academy of Information Technology, Warsaw, Poland
[2]Educational Research Institute
[3]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
[4]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

AIM AND MOTIVATION OF STUDY

Common access to the Internet makes it possible that virtual open-collaboration environments became an important platform for massive collaborative work.
We study whether and how the interests diversity of editors and experience diversity of editor teams affect the quality of work on the Wikipedia example.

CONTRIBUTIONS

- the concept of editor's "interest versatility" and various measures of team diversity
- exploratory analysis of two dumps of Wikipedia (Polish and German), which indicate that diversity is positively correlated with quality of articles
- deepened statistical analysis of the studied datasets
- series of experiments with logistic regression, decision trees, Random Forest

MEASURES OF DIVERSITY

VERSATILITY (MEASURE OF INTEREST DIVERSITY)

Let X denote a group of Wikipedia editors.
editor x's interest in category :
$p_i(x) = t_i(x)/t(x)$

where t(x) denote the amount of textual content x contributed to all articles and $t_i(x)$ denote the total amount of textual content editor x contributed to a specific category

interest profile of the editor x, denoted as ip(x), as the interest distribution vector over the set of all categories:

$$ip(x) = (p_1(x), \ldots, p_k(x)) \tag{1}$$

Versatility as entropy of interest profile of x:

$$V(x) = H((p_1, p_2, \ldots, p_k)) = \sum_{1 \leq i \leq k} -p_k \log_2(p_k) \tag{2}$$

STANDARD DEVIATION

Standard deviation of numerical attribute X taking n values: $X_1, \ldots, X_n$ is defined as

$$\mathrm{sd}(X) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mathrm{avg}(X))^2},$$

where $\mathrm{avg}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an arithmetic mean of attribute X. Standard deviation $\mathrm{sd}(X)$ measures how much (on average) an attribute varies around its arithmetic mean.

DATA

DATASETS

Polish Wikipedia wiki-pl March 2015
German Wikipedia wiki-de September 2015

Table: Summary of Datasets wiki-pl and wiki-de

	wiki-pl dataset	wiki-de dataset
editors	126,406	555,355
articles	947,080	1,422,940
editions	16,084,290	61,266,990

MEANS OF MEASURING THE QUALITY OF WIKIPEDIA ARTICLES

quality of articles criteria defined by the Wikipedia community:

- GOOD article (G): “well-written, comprehensive, well-researched, neutral, stable, illustrated”
- FEATURED article (F): (in addition to the above) “length and style guidelines including a lead, appropriate structure and consistent citation”

Table: Analysed groups of editors

Editor group	co-edited
N	(normal) neither good nor featured article
G	(good) at least one good article
F	(featured) at least one featured article
G∪F	(good or featured) at least one good or one featured article
G∩F	(good and featured) at least one good and one featured article

TOPICAL CATEGORIES OF ARTICLES

Table: Wikipedia main content categories

Dataset	Main Content Categories	Dataset	Main Content Categories
wiki-pl	Humanities and Social Sciences Natural and Physical Sciences Art & Culture Philosophy Geography History Economy Biographies Religion Society Technology Poland	wiki-de	Art & Culture Geography History Knowledge Religion Society Sport Technology

EXPERIMENTAL RESULTS FOR EDITORS

PRELIMINARY EXPLORATORY ANALYSIS OF THE DATA

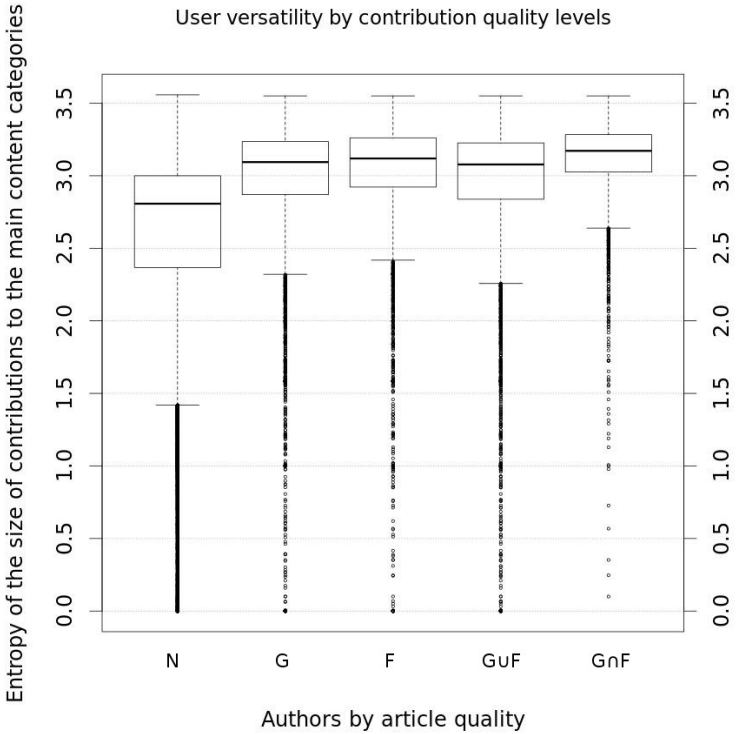


Figure: Versatility vs Quality for wiki-pl dataset

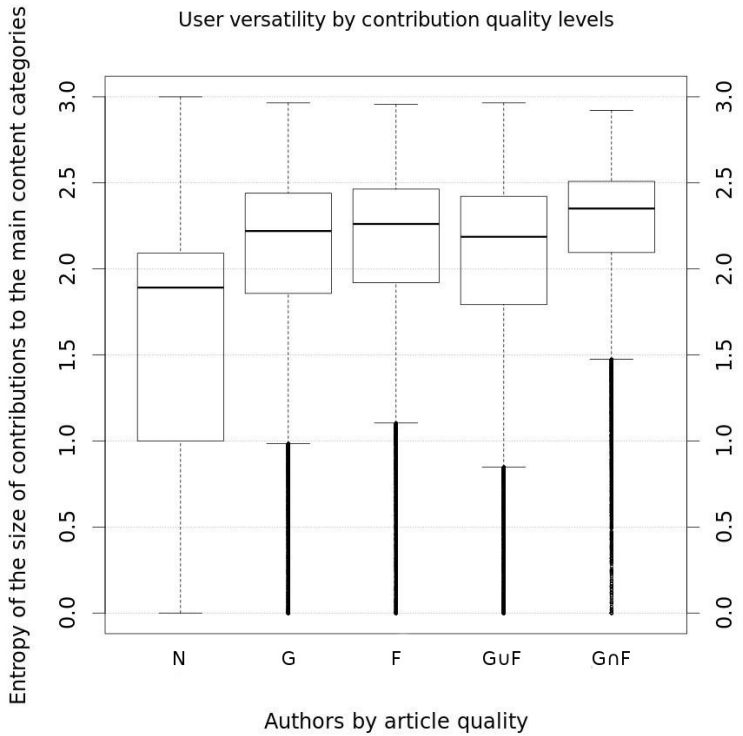


Figure: Versatility vs Quality for wiki-de dataset (denotations as on Fig. 1)

PRELIMINARY EXPLORATORY ANALYSIS OF THE DATA: CONTINUATION

Table: Median of versatility and productivity of editors vs. quality for wiki-pl and wiki-de dataset

	wiki-pl		wiki-de	
quality	versatility	productivity	versatility	productivity
G∩F	3.1720	159300	2.351	46080
G∪F	3.011	2992	2.064	1502
F:	3.000	2322	2.053	1283
G:	3.016	3347	2.070	1629
N:	2.807	237	1.891	264

Table: Editors gender vs versatility

wiki-pl				
	number of women	number of men	versatility of women	versatility of men
G∩F	1.73e+02	3.98e+02	3.25e+00	3.25e+00
G∪F	2.46e+02	5.69e+02	3.18e+00	3.20e+00
F:	2.00e+01	4.70e+01	3.01e+00	3.02e+00
G:	5.30e+01	1.24e+02	3.09e+00	3.06e+00
N:	1.81e+02	4.14e+02	2.87e+00	2.91e+00
wiki-de				
	number of women	number of men	versatility of women	versatility of men
G∩F	5.53e+002	1.03e+003	2.51e+000	2.41e+000
G∪F	6.43e+002	1.32e+003	2.46e+000	2.44e+000
F:	3.40e+001	8.00e+001	2.17e+000	2.14e+000
G:	5.60e+001	2.11e+002	2.07e+000	2.18e+000
N:	1.95e+002	5.29e+002	1.84e+000	2.00e+000

EXPERIMENTS WITH QUALITY PREDICTION FOR EDITORS

Two-class prediction problem, where:

- class $C = 1$ corresponds to G$\cup$F editors
- class $C = 0$ corresponds to the remaining ones

data randomly split:

- training set 50% observations
- testing set 50% observations

Classification models:

- logistic regression model
- tree model

EXPLAINING QUALITY WITH LOGISTIC REGRESSION MODEL

Table: Logistic regression model for editors on wiki-pl dataset

	Estimate	Std. Error	z-value	Pr(> |z|)
(Intercept)	-5.35e+000	1.11e-001	-48.115	<2e-16***
versatility	**9.32e-001**	**3.82e-002**	**24.384**	**<2e-16*** **
productivity	-5.96e-006	2.74e-006	-2.174	0.0297*
versatility:productivity	6.4e-006	9.18e-007	6.971	3.15e-012***
Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '				

Table: Logistic regression model for editors on wiki-de dataset

	Estimate	Std. Error	z-value	Pr(> |z|)
(Intercept)	-3.539e+00	2.183e-02	-162.110	<2e-16***
versatility	**7.879e-01**	**1.098e-02**	**71.767**	**<2e-16*** **
productivity	3.214e-06	5.829e-07	5.514	3.52e-08 ***
versatility:productivity	1.213e-05	3.317e-07	36.581	<2e-16 ***
Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '				

EXPLAINING QUALITY WITH TREE MODEL

Figure: Tree model for wiki-pl dataset

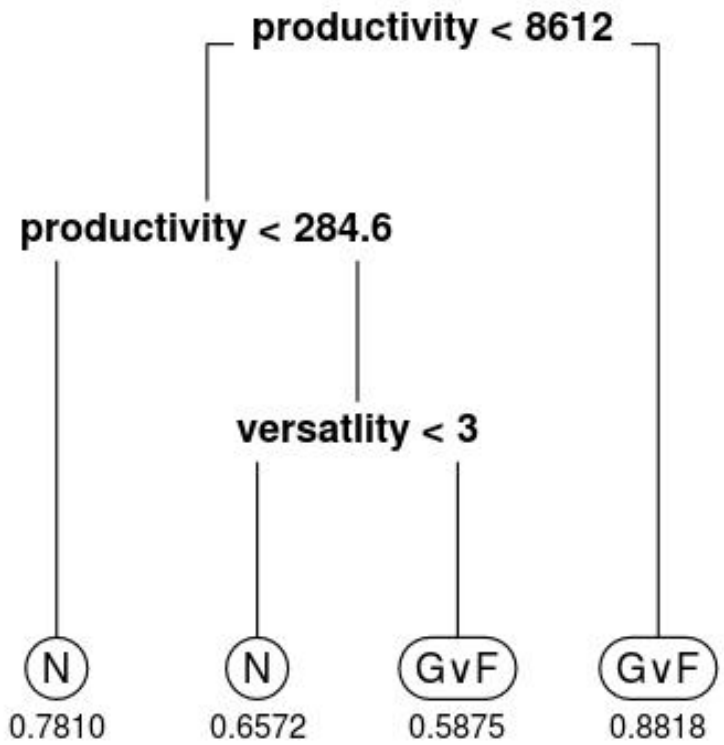
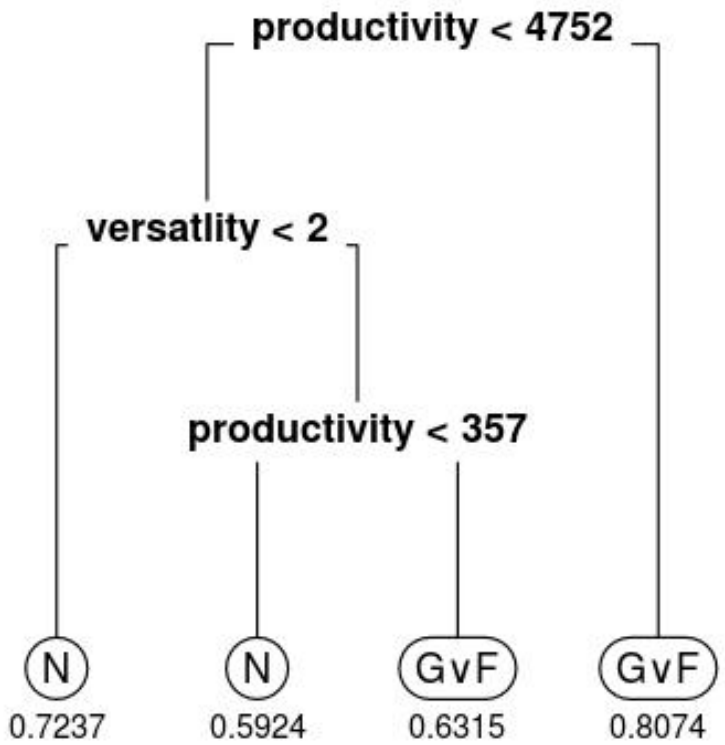


Figure: Tree model for wiki-de dataset



PREDICTION RESULTS FOR LOGISTIC REGRESSION AND TREE MODEL

Table: Evaluation measures on testing data for editors on wiki-pl and wiki-de datasets

measure	logistic regression wiki-pl dataset	logistic regression wiki-de dataset	tree model wiki-pl dataset	tree model wiki-de dataset
precision	87.73%	86.85%	74.50%	75.36%
recall	17.72%	17.91%	29.56%	26.04%
accuracy	93.40%	88.53%	93.73%	88.84%
F-measure	29.48%	29.70%	42.33%	38.70%

SUMMARY OF EXPERIMENTAL RESULTS FOR EDITORS

Versatility is the most significant variable according to logistic model and it is also useful for tree.

Both diversity and productivity allow to predict a quality of articles successfully.

EXPERIMENTAL RESULTS FOR TEAMS

Table: Attributes of Teams

Name	Description
versatility	entropy of distribution vector over main categories
mean productivity in article	mean amount of editors' contribution in bytes to individual article
mean total productivity	mean amount of editors' contribution in bytes to all articles on the Wikipedia
the size of team	the number of editors who contributes in one article
mean tenure in article	mean number of days spent on article
mean tenure in Wikipedia	mean number of days spent on the Wikipedia
std. dev. productivity in art	standard deviation of the number of editors' contribution bytes to individual article
std. dev total productivity	standard deviation of editors' contribution bytes to all articles on the Wikipedia
std. dev tenure in article	standard deviation of number of days between the first and the last editors contribution to individual article
std.dev tenure in wikipedia	standard deviation of number of days spent on the Wikipedia

Table: Median of team features vs. quality articles of wiki-pl dataset

quality	versatility	**mean productivity in articles**	mean total productivity	**sd productivity in articles**	sd total product.
G∪F	3.26e+000	1.80e+003	4.52e+006	6.84e+003	5.35e+006
F	3.26e+000	2.93e+003	4.31e+006	9.62e+003	5.42e+006
G	3.26e+000	1.73e+003	4.58e+006	6.10e+003	5.33e+006
N	3.53e+000	4.99e+002	5.88e+006	7.96e+002	5.96e+006
quality	**team size**	**mean tenure in article**	mean tenure in Wikipedia	**sd tenure in article**	sd tenure in Wikipedia
G∪F	2.00e+001	1.25e+002	1.81e+003	3.56e+002	8.46e+002
F	3.30e+001	1.44e+002	1.85e+003	4.11e+002	9.02e+002
G	1.70e+001	1.20e+002	1.80e+003	3.37e+002	8.20e+002
N	4.00e+000	7.71e+000	1.81e+003	4.39e+001	8.15e+002

Table: Median of team features vs. quality articles of wiki-de dataset

quality	versatility	mean product. in art.	mean total product.	sd product. in art.	sd total product.
GUF	2.65e+000	1.16e+003	5.94e+006	6.05e+003	1.31e+007
F	2.65e+000	1.44e+003	6.12e+006	8.09e+003	1.37e+007
G	2.65e+000	9.98e+002	5.82e+006	4.98e+003	1.27e+007
N	2.62e+000	4.07e+002	6.16e+006	9.10e+002	9.20e+006
quality	team size	mean tenure in article	mean tenure in Wikipedia	sd tenure in article	sd tenure in Wikipedia
GUF	7.45e+001	1.02e+002	2.09e+003	3.33e+002	1.05e+003
F	8.60e+001	1.01e+002	2.11e+003	3.30e+002	1.05e+003
G	6.60e+001	1.03e+002	2.08e+003	3.36e+002	1.04e+003
N	9.00e+000	4.38e+001	2.08e+003	1.33e+002	9.94e+002

EXPERIMENTS WITH QUALITY PREDICTION FOR TEAMS

Two-class prediction problem, where:

- class $C = 1$ corresponds to $G \cup F$ teams
- class $C = 0$ corresponds to the remaining ones

data randomly split:

- training set 50% observations
- testing set 50% observations

Classification models:

- logistic regression model
- random forest model

Table: Logistic regression model for teams on wiki-pl dataset

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-7.571e+00	7.565e-01	-10.008	< 2e-16 ***
versatility	7.718e-01	2.373e-01	3.253	0.00114 **
mean productivity in article	-2.401e-04	1.574e-05	-15.255	< 2e-16 ***
mean total productivity	2.157e-08	1.330e-08	1.622	0.10478
size of team	**1.205e-02**	**7.014e-04**	**17.186**	**< 2e-16 *****
mean tenure in article	-1.220e-02	7.373e-04	-16.550	< 2e-16 ***
mean tenure in wikipedia	-3.530e-04	8.435e-05	-4.185	2.86e-05 ***
sd productivity in art	**1.499e-04**	**7.349e-06**	**20.402**	**< 2e-16 *****
sd total productivity	-7.840e-08	1.353e-08	-5.797	6.75e-09 ***
sd tenure in article	**7.298e-03**	**3.180e-04**	**22.949**	**< 2e-16 *****
sd tenure in wikipedia	-7.214e-04	1.234e-04	-5.845	5.05e-09 ***
Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '				

LOGISTIC REGRESSION ANALYSIS: CONTINUATION

Table: Logistic regression model for teams on wiki-de dataset

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.408e+01	7.165e-01	-19.658	< 2e-16 ***
versatility	1.937e+00	2.578e-01	7.514	5.71e-14 ***
mean productivity in article	-5.218e-05	7.794e-06	-6.695	2.15e-11 ***
mean total productivity	-2.578e-07	1.205e-08	-21.395	< 2e-16 ***
size of team	**1.138e-02**	**1.948e-04**	**58.401**	**< 2e-16 *****
mean tenure in article	-1.602e-02	7.732e-04	-20.721	< 2e-16 ***
mean tenure in Wikipedia	1.495e-03	7.863e-05	19.018	< 2e-16 ***
sd productivity in art	2.782e-05	2.328e-06	11.950	< 2e-16 ***
sd total productivity	**9.789e-08**	**4.222e-09**	**23.184**	**< 2e-16 *****
sd tenure in article	**7.838e-03**	**2.722e-04**	**28.799**	**< 2e-16 *****
sd tenure in wikipedia	-1.626e-04	1.227e-04	-1.326	0.185

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '

IMPORTANCE OF DIVERSITY MEASURES IN QUALITY PREDICTIONS

Table: Random Forest importance for wiki-pl dataset

	Imp1	Imp2
versatility	**5.20e+001**	1.16e+002
mean productivity in article	3.25e+001	**1.33e+002**
mean total productivity	2.71e+001	1.16e+002
size of team	3.84e+001	1.01e+002
mean tenure in article	1.28e+001	8.07e+001
mean tenure in Wikipedia	2.23e+001	8.75e+001
sd productivity in art	3.13e+001	**1.73e+002**
sd total productivity	**4.38e+001**	**1.19e+002**
sd tenure in article	1.16e+001	8.35e+001
sd tenure in Wikipedia	**4.02e+001**	1.05e+002

IMPORTANCE OF DIVERSITY MEASURES: CONTINUATION

Table: Random Forest importance for wiki-de dataset

	Imp1	Imp2
versatility	**5.37e+001**	2.40e+002
mean productivity in article	2.50e+001	3.00e+002
mean total productivity	1.16e+001	1.91e+002
size of team	**3.43e+001**	**3.52e+002**
mean tenure in article	7.25e+000	1.97e+002
mean tenure in Wikipedia	**3.61e+001**	**3.14e+002**
sd productivity in art	2.51e+001	**3.97e+002**
sd total productivity	1.69e+001	1.95e+002
sd tenure in article	7.23e+000	1.96e+002
sd tenure in Wikipedia	1.42e+001	1.97e+002

Table: Evaluation measures on testing data for teams on wiki-pl and wiki-de datasets

measure	logistic regression teams wiki-pl dataset	logistic regression teams wiki-de dataset	random forest model wiki-pl dataset	random forest wiki-de dataset
precision	15.90%	27.50%	70.60%	52.80%
recall	1.10%	3.40%	5.68%	7.34%
accuracy	99.70%	99.60%	99.70%	99.60%
F-measure	2.06%	6.05%	10.50%	12.90%

SUMMARY OF EXPERIMENTAL RESULTS FOR TEAMS

The experiments clearly indicate that diversity of teams in combination with other properties of teams allows to predict the quality of articles very successfully.

CONCLUSIONS AND FUTURE WORK

- the interest diversity of single authors and teams has positive influence on their work quality
- it is possible to predict the quality of Wikipedia articles using diversity measures and some other properties of teams successfully
- take into account some other features of editors and teams
- develop an intelligent decision-support tool for suggesting how to build a successful editor team in order to produce high-quality article