NetSci-X 2016 Proceedings of the 12th International Conference and School on Advances in Network Science - Volume 9564

Studying the Role of Diversity in Open Collaboration Network: Experiments on Wikipedia

Katarzyna Baraniak

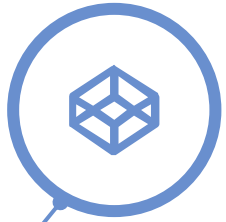
Marcin Sydow

Jacek Szejda

Dominika Czerniawska

01 | Introduction

Q: What leads to coherence and efficient collaboration?



or

MODULARITY

- Co-dependence between components should be eliminated
- No need knowledge for the whole system
- Higher specialization
- Less diversity

INTEGRITY

- Smoother adaption to new environment
- Require diverse knowledge and skills
- Lower specialization
- More diversity

Q: Which collaboration pattern is more successful in creating high quality Wikipedia articles?

Studying the Role of Diversity in Open Collaboration Network: Experiments on Wikipedia [1]

Presents some empirical study towards understanding the role of diversity of individual authors and whole teams of authors on the quality of the article they co-edit in open collaboration environment like Wikipedia.

02 | Related Work

The effects of diversity on group productivity and member withdrawal in online volunteer groups[2]

- Examined the effects of group diversity on the amount of work accomplished and on member withdrawal behaviors in the context of WikiProjects
- The measure of Interest variety is different (Blau index, $1 - \sum P_i^2$)

Consequences of content diversity for online public spaces for local Ccommunities[3]

- Hypothesized that the diversity of the purposes and content affects future content volume and frequency of information exchanges within a system
- Focused on diversity of categories such as culture, ethnicity, age and etc

02 | Related Work

Gender and Tenure Diversity in GitHub Teams[4]

- Studied how gender and tenure diversity relate to team productivity and turnover on GitHub dataset
- The measure of diversity is different (Blau index)

The Impact and Evolution of Group Diversity in Online Open Collaboration[5]

- Analyzed how tenure diversity and interest variety affect group productivity and member withdrawal and how the two types of diversity evolve over time
- Not directly address the issue of how diversity impacts the quality of the resulting articles

03 | Main Concept

- Let X denote a group of Wikipedia editors, who participate in editing Wikipedia articles.
- Each article can be mapped to one or more categories from a pre-defined set of categories $C = \{c_1, \dots, c_k\}$ that represent topics.
- Each editor $x \in X$ in our model is characterized by their editing activity.

- Let $t(x)$ denote the total amount of textual content (in bytes) that x contributed to all articles editor co-edited and let $t_i(x)$ denote the total amount of textual content that editor x contributed to the article belonging to a specific category c_i.

03 | Editor's Interest Profile

- Let $p_i(x) = t_i(x)/t(x)$ denotes x's interest in category c_i.
- Then we define the interest profile of the editor x, as the interest distribution vector over the set of all categories:

$$ip(x) = (p_1(x), p_2(x), \dots, p_k(x))$$

Assume that a set of categories C consists of 8 categories and editor x has contributed $t(x) = 10kB$ of text in total. 8kB of text has been contributed to articles in category c_2, which means $t_2(x) = 8kB$. The rest of text in c_5.

Therefore, the interest profile of this user is:

$$ip(x) = (0, \frac{4}{5}, 0,0, \frac{1}{5}, 0,0,0)$$

03 | Editor's Versatility Measure

- Define diversity of interests (or versatility) of x, $V(x)$, as the entropy of interest profile of x:

$$V(x) = H\big((p_1, p_2, \dots, p_k)\big) = \sum_{1 \le i \le k} -p_k \lg(p_k)$$

The value of entropy ranges from 0 (extreme specialization, i.e. total devotion to a single category) to $\lg(k)$ (extreme diversity, i.e. equally interest in all categories).

The versatility of user x in previous example is:

$$V(x) = -p_2 \lg(p_2) - p_5 \lg(p_5) = 0.8 \times 0.32 + 0.2 \times 2.32 = 0.72$$

04 | Data Preparation

Experiments on two dumps of Wikipedia: **Polish Wikipedia** and **German Wikipedia**

- Employed a method that sought main content categories iteratively among parents of categories directly describing any given page
  - If assign to multiple categories, split contribution size equally
  - If unable classify, exclude from dataset

Table 2. Wikipedia main content categories

Dataset	Main content categories
Polish Wikipedia	Humanities and social sciences, Natural and physical sciences, Art & Culture, Philosophy, Geography, History, Economy, Biographies, Religion, Society, Technology, Poland
German Wikipedia	Art & Culture, Geography, History, Knowledge, Religion, Society, Sport, Technology

- The quality of articles is modelled based on Wikipedia community, who evaluate articles as good and featured

05 | Experiments Results for Editors

- **Versatility:** Entropy of editor's partial contributions to each category
- **Productivity:** The number of bytes by which they modified Wikipedia content
- The goal is to experimentally study the **dependence between editors' versatility and the quality of articles they co-edit**

Table 4. Analysed groups of editors

Editor group	Co-edited
N	(regular) neither good nor featured article
G	(good) at least one good article
F	(featured) at least one featured article
GF	(good and featured) at least one good and one featured article

05 | Experiments Results for Editors

- Clear **positive** connection between editors versatility and quality

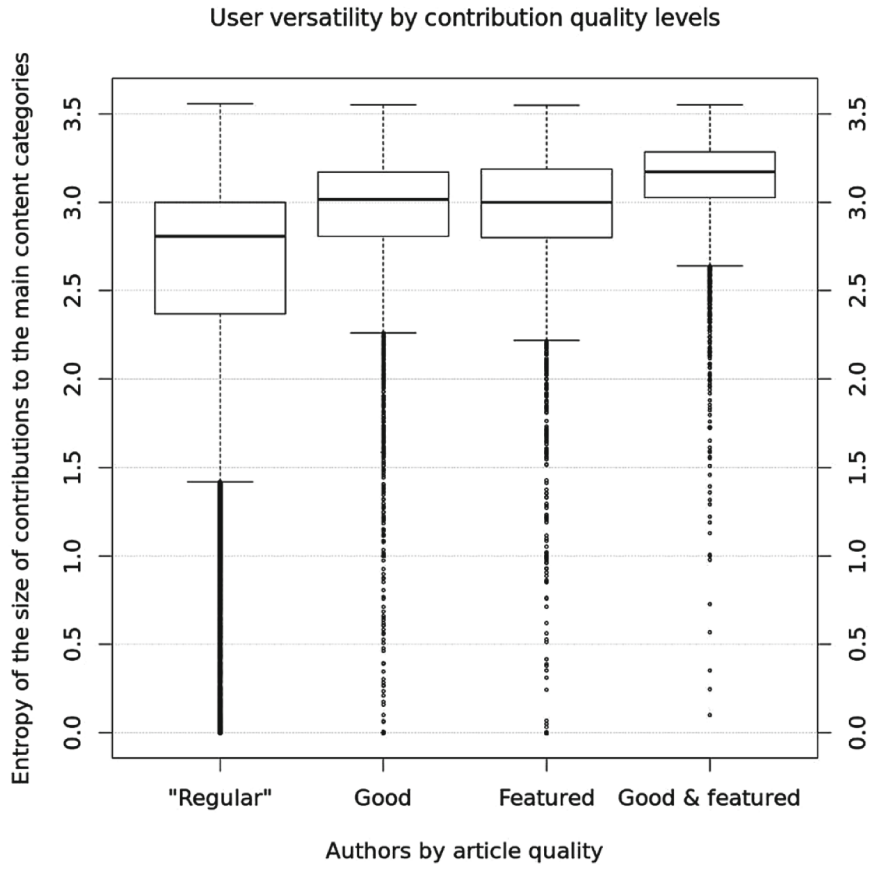


Fig. 1. Versatility vs quality for Polish Wikipedia

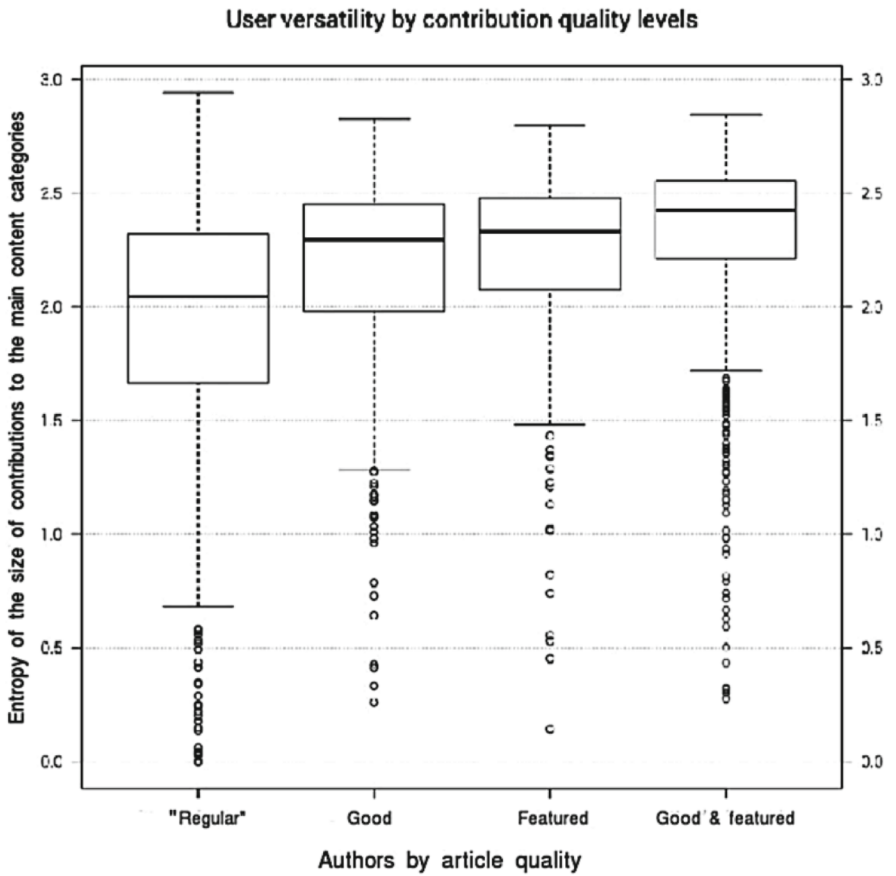


Fig. 2. Versatility vs quality for German Wikipedia (denotations as on Fig. 1)

06 | Experiments Results for Teams

- How productivity and diversity of teams impact the quality of articles they create ?

Table 9. Logistic regression model for teams on Polish Wikipedia

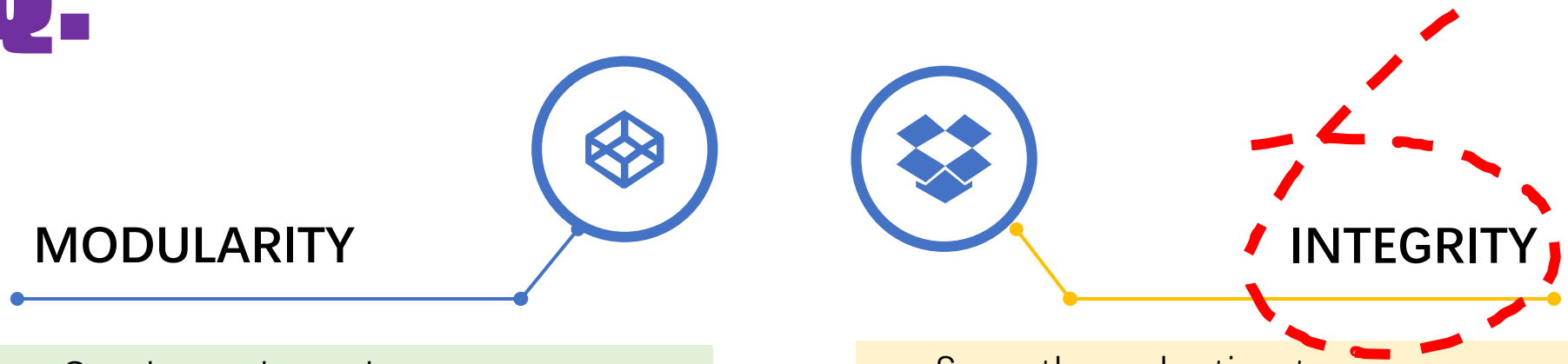
	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	−7.753e+00	5.333e−01	−14.539	<2e−16	***
Versatility	**7.984e−01**	1.668e−01	4.787	1.69e−06	***
Mean productivity in article	−2.502e−04	1.685e−05	−14.851	<2e−16	***
Mean total productivity	2.832e−08	9.191e−09	3.081	0.00206	**
Size of team	**1.179e−02**	5.052e−04	23.336	<2e−16	***
Mean tenure in article	**−1.242e−02**	5.198e−04	−23.896	<2e−16	***
Mean tenure in wikipedia	−3.169e−04	5.974e−05	−5.304	1.13e−07	***
Sd productivity in art	1.638e−04	6.122e−06	26.754	<2e−16	***
Sd total productivity	−9.191e−08	9.522e−09	−9.652	<2e−16	***
Sd tenure in article	7.450e−03	2.239e−04	33.272	<2e−16	***
Sd tenure in wikipedia	−6.709e−04	8.746e−05	−7.672	1.70e−14	***

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

Table 8. Features of logistic regression model for teams on Polish Wikipedia

Name	Description
Versatility	for every article we count the mean amount of contribution in bytes of all team members (using data about editors interests distribution) to 12 main categories for Polish Wikipedia (Table 2). Then we count versatility as entropy of distribution vector over these categories
Mean productivity in article	mean amount of editors' contribution in bytes to individual article. Counted as sum of all bytes contributor change in one article
Mean total productivity	mean amount of editors' contribution in bytes to all articles on the Wikipedia. Counted as sum of all bytes contributor change in all articles in Wikipedia
The size of team	the number of editors who contributes in one article
Mean tenure in article	mean number of days spent on individual article, counted as the amount of days between first and the last contribution of editor to a given article
Mean tenure in Wikipedia	mean number of days spent on the Wikipedia, counted as the amount of days between the first and the last contribution of editor contribution to all articles on the whole 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 between the first and the last editors contribution to all articles on the Wikipedia

Q: What leads to coherence and efficient collaboration?



- Co-dependence between components should be eliminated
- No need knowledge for the whole system
- Higher specialization
- Less diversity

- Smoother adaption to new environment
- Require diverse knowledge and skills
- Lower specialization
- More diversity

07 | Final Project

- The goal of project is to analyze the effect of diversity of members on GitHub team productivity.
- Collect data by using GHTorrent on Google BigQuery
- In order to migrate this Wikipedia-based approach to GitHub dataset, I select following features to analyzing:

Features	Methods
Domain versatility	• Identity the domain of the projects by Latent Dirichlet Allocation(LDA)[6] • Calculate versatility in entropy approach
Language Versatility	• Using the data from GHTorrent language table • Calculate versatility in entropy approach
Productivity	• Commits could be a good feature to quantify the productivity of GitHub teams

08 | Reference

[1] M. Sydow, K. Baraniak, and P. Teisseyre, "Diversity of editors and teams versus quality of cooperative work: experiments on wikipedia," Journal of Intelligent Information Systems, vol. 48, no. 3, pp. 601–632, jun 2017. [Online]. Available: http://link.springer.com/10.1007/s10844-016-0428-1

[2] Jilin Chen, Yuqing Ren, and John Riedl. 2010. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). ACM, New York, NY, USA, 821-830. DOI: https://doi.org/10.1145/1753326.1753447

[3] Claudia A. López and Brian S. Butler. 2013. Consequences of content diversity for online public spaces for local communities. In Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13). ACM, New York, NY, USA, 673-682. DOI: https://doi.org/10.1145/2441776.2441851

[4] Vasilescu, B., Posnett, D., Ray, B., van den Brand, M. G. J., Serebrenik, A., Devanbu, P., & Filkov, V. (2015). Gender and Tenure Diversity in GitHub Teams. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 3789–3798). New York, New York, USA: ACM Press. https://doi.org/10.1145/2702123.2702549

[5] Ren, Y., Chen, J., & Riedl, J. (2015). The impact and evolution of group diversity in online open collaboration. *Management Science*, *62*(6), 1668-1686.

[6] B. Ray, D. Posnett, V. Filkov, and P. Devanbu, "A large scale study of programming languages and code quality in github," in Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering - FSE 2014. New York, New York, USA: ACM Press, 2014, pp. 155–165. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2635868.2635922

Thanks for listening

ZHANG YIFAN
1004245952

04 | Dataset

Experiments on two dumps of Wikipedia: **Polish Wikipedia** and **German Wikipedia**

For each contributor:

1. the number of pages he edited
2. the number of edits they made
3. the number of characters by which they modified Wikipedia articles
4. the number of both “good” and “featured” articles he edited.

Table 1. Data size for Polish Wikipedia

Measure	Size
The number of authors	126,406
The number of all articles	947,080
The number of regular articles	944,585
The number of good articles	1,889
The number of featured articles	606
The number of editions	16,084,290

05 | Experiments Results for Editors

- Compare the influence of interest versatility and productivity on quality in a quantitative way
- Build **multinomial model** with versatility, productivity, and their interaction as explanatory variables

- **Versatility has significant correlation with quality than productivity**
- Interaction of versatility and productivity has no positive correlation with quality for G, F and CF categories.

Model 1: Up to 5 edited pages

	Estimate	Sd. Err	P-val	Odds
"(Intercept)"				
G	-5.221	0	0***	0.005
F	-5.465	0	0***	0.004
GF	-10.157	0	0***	0
Versatility				
G	**0.312**	0	0***	1.367
F	**0.185**	0	0***	1.204
GF	**0.728**	0	0***	2.071
Productivity				
G	0	0	0.889	1
F	0	0	0.125	1
GF	0	0	0***	1
Interaction: vers.*prod.				
G	0	0	0.932	1
F	0	0	0.188	1
GF	0	0	0.005*	1

Model 2: 6 to 10 edited pages

	Estimate	Sd. Err	P-val	Odds
"(Intercept)"				
G	-3.343	0	0***	0.035
F	-4.170	0	0***	0.050
GF	-6.231	0	0***	0.001
Versatility				
G	**0.146**	0	0***	1.156
F	**0.229**	0	0***	1.256
GF	**0.279**	0	0***	1.322
Productivity				
Good art.	0	0	0.069	1
F	0	0	0.535	1
GF	0	0	0.230	1
Interaction: vers.*prod.				
G	0	0	0.251	1
F	0	0	0.634	1
GF	0	0	0.202	1

Model 3: 10 to 20 edited pages

	Estimate	Sd. Err	P-val	Odds
"(Intercept)"				
G	-2.737	0	0***	0.064
F	-3.479	0	0***	0.030
GF	-6.219	0	0***	0.002
Versatility				
G	**0.189**	0	0***	1.208
F	**0.205**	0	0***	1.227
GF	**0.695**	0	0***	2.005
Productivity				
G	0	0	0.997	1
F	0	0	0.857	1
GF	0	0	0.002*	1
Interaction: vers.*prod.				
G	0	0	0.719	1
F	0	0	0.821	1
GF	0	0	0.007*	1

Model 4: More than 20 edited pages

	Estimate	Sd. Err	P-val	Odds
"(Intercept)"				
G	-2.170	0	0***	0.114
F	-3.252	0	0***	0.038
GF	-4.331	0	0***	0.013
Versatility				
G	**0.357**	0	0***	1.429
F	**0.471**	0	0***	1.602
GF	**1.033**	0	0***	2.810
Productivity				
G	0	0	0.138	1
F	0	0	0***	1
GF	0	0	0***	1
Interaction: vers.*prod.				
G	0	0	0***	1
F	0	0	0***	1
GF	0	0	0***	1

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '
all values are approximated to 3 decimal places