Does a “Renaissance Man” Create Good Wikipedia Articles?

Jacek Szejda[3], Marcin Sydow[1,2] and Dominika Czerniawska[4]
[1]*Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*
[2]*Polish-Japanese Institute of IT, Warsaw, Poland*
[3]*Educational Research Institute, Warsaw, Poland*
[4]*Digital Economy Lab, Warsaw, Poland*

Keywords: Diversity Of Interests, Article Quality, Open Collaboration, Wikipedia.

Abstract: We introduce a concept of *diversity of interests* or *versatility* of a member of an open-collaboration environment such as Wikipedia and aim to study how versatility influences the work quality. We introduce versatility measure based on entropy. In preliminary experiments on Wikipedia data we indicate the positive role of editors’ versatility on the quality of the articles they co-edit.

1 INTRODUCTION

Open-collaboration environments like Wikipedia produce outcome of varying quality. It is important to study what properties of community members increase chances for high-quality results of their work. Such studies can help in future in developing tools that improve and support open-collaboration team-building process.

For example, it is interesting to study whether editors that have *diverse* interests tend to create better Wikipedia articles.

Diversity has prooved to play important role in multiple fields of applications: text summarisation, web search, databases, recommender systems and semantic entity summarisation. Recently, the concept of diversity has attracted interest also in the domain of open collaboration research (e.g. (Aggarwal, 2014)).

In this paper we introduce a quantitative measure of *diversity of interests* of a member of an open-collaboration environment such as Wikipedia and aim to study how versatility influences the work quality. The measure is based on the information-theoretic concept of entropy. We demonstrate on Wikipedia data that versatility of editor seems to be correlated with the quality of articles they co-edit.

1.1 Sociological Background

Team diversity is one of the fundamental issues in social and organisational studies that has been broadly researched on free software communities. Wikipedia has a similar workflow where the community members can edit any article. It rises analogous issues concerning team’s coherence vs efficiency. There are two competing theories describing efficient team organisation: modularity and integrity. The first was introduced by David Parnas who suggested that codependence between components should be eliminated by limiting the communication (Parnas, 1972). In our approach, a module corresponds to a task of creating an article on Wikipedia. Participation in a module does not require knowledge about the whole system or other modules, e.g. Wikipedia users can co-author articles about social science without knowing anything about life sciences or mathematics. It leads to higher specialisation and less diversity in individual performance. Modular approach enables more flexibility and decentralized management (Sanchez and Mahoney, 1996). On the other hand, integral approach to organisation is easier to adapt to new environments, to change the cooperation rules and gives better results when it comes to fine-tuning of the system (Langlois and Garzarelli, 2008). In an integral mode team members have diverse knowledge and skills. We aim to study whether modular/specialized or integral collaboration pattern is more successful in creating high-quality Wikipedia articles.

1.2 Related Work

The potentially positive role of diversity was noticed very early in the beginnings of Information Retrieval a few decades ago (Goffman, 1964). One

Szejda J., Sydow M. and Czerniawska D..
Does a “Renaissance Man” Create Good Wikipedia Articles?.
DOI: 10.5220/0005155804250430
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 425-430
ISBN: 978-989-758-048-2

of the earliest successful applications of diversity-aware approach was reported in (Carbonell and Goldstein, 1998) in the context of text summarisation. Recently, diversity-awareness has gained increasing interest in other information-related areas where the actual user's information need is unknown and/or the user query is ambiguous. Examples range from databases (e.g.(Vee et al., 2008)) to Web search (e.g. (Agrawal et al., 2009)) or very recently to the quite novel problem of graphical entity summarisation in semantic knowledge graphs (Sydow et al., 2013). From the open collaboration point of view, diversity can be considered from many perspectives, for example as a team diversity vs homogeneity or a single editors's diversity of interest vs specialisation. For example, the positive role of team diversity was studied in (Chen et al., 2010), but the used definitons of diversity and its measures (e.g. Blau index) are different than in our paper, where it is based on the concept of *entropy*. Most importantly, in contrast to our work, the mentioned work studies the influence of diversity on amount of accomplished work and withdrawal behaviour rather than the work quality that is considered here. In contrast to our work most of previous works focus on diversity of editor teams in terms of categories such as culture, ethnicity, age, etc. A very recent example, with a special emphasis on adhoc "swift" teams where the members have very little previous interactions with each other is (Aggarwal, 2014). (López and Butler, 2013) studies how the content diversity influences online public spaces in the context of local communities.

2 MODEL DESCRIPTION

In this section we explain the model of editor's interest diversity that we apply in our approach. We will use Wikipedia terminology, to illustrate the concepts, however our model can be adapted to other, similar open-collaboration environments.

Let X denote the set of Wikipedia editors. Editors participate in editing Wikipedia articles. Each article can be mapped to one or more of some pre-defined *set of categories* $C = \{c_1, \dots, c_k\}$ that represent topics.

Each editor $x \in X$ in our model is characterised by their editing activity i.e. all editing actions done by x.

We assume that the interests of an editor x can be represented by the amount of work that x committed to articles in particular categories.

Let $t(x)$ denote the total amount of textual content (in bytes) that x contributed to all articles they co-edited and let $t_i(x)$ denote the total amount of textual content that editor x contributed to the articles belonging to a specific category c_i. [1]

Now, lets introduce the following denotation: $p_i(x) = t_i(x)/t(x)$ and interpret it as representing x's *interest in category* c_i. Henceforth, we will use a shorter denotation p_i for $p_i(x)$ whenever x is understood from the context.

2.1 Interest Profile

Finally, we define the *interest profile* of the editor x, denoted as $ip(x)$, as the *interest distribution vector* over the set of categories of the articles that x edited:

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

Notice that according to the definition the interest profile represents a valid distribution vector i.e. its coordinates sum up to 1.

2.1.1 Example

Assume that the set of categories C consists of 8 categories: $\{c_i\}_{1 \le i \le 8}$ and that editor x has contributed $t(x) = 10kB$ of text in total, out of which $t_2(x) = 8kB$ of text has been contributed to articles in category c_2, $t_5(x) = 2kB$ in category c_5 and nothing to articles that were not assigned to c_2 nor c_5. Thus the $x's$ interest in c_2 is $p_2(x) = t_2(x)/t(x) = \frac{4}{5}$, in c_5 is $p_5(x) = t_5(x)/t(x) = \frac{1}{5}$ and is equal to 0 for all other categories. The interest profile of this user is:

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

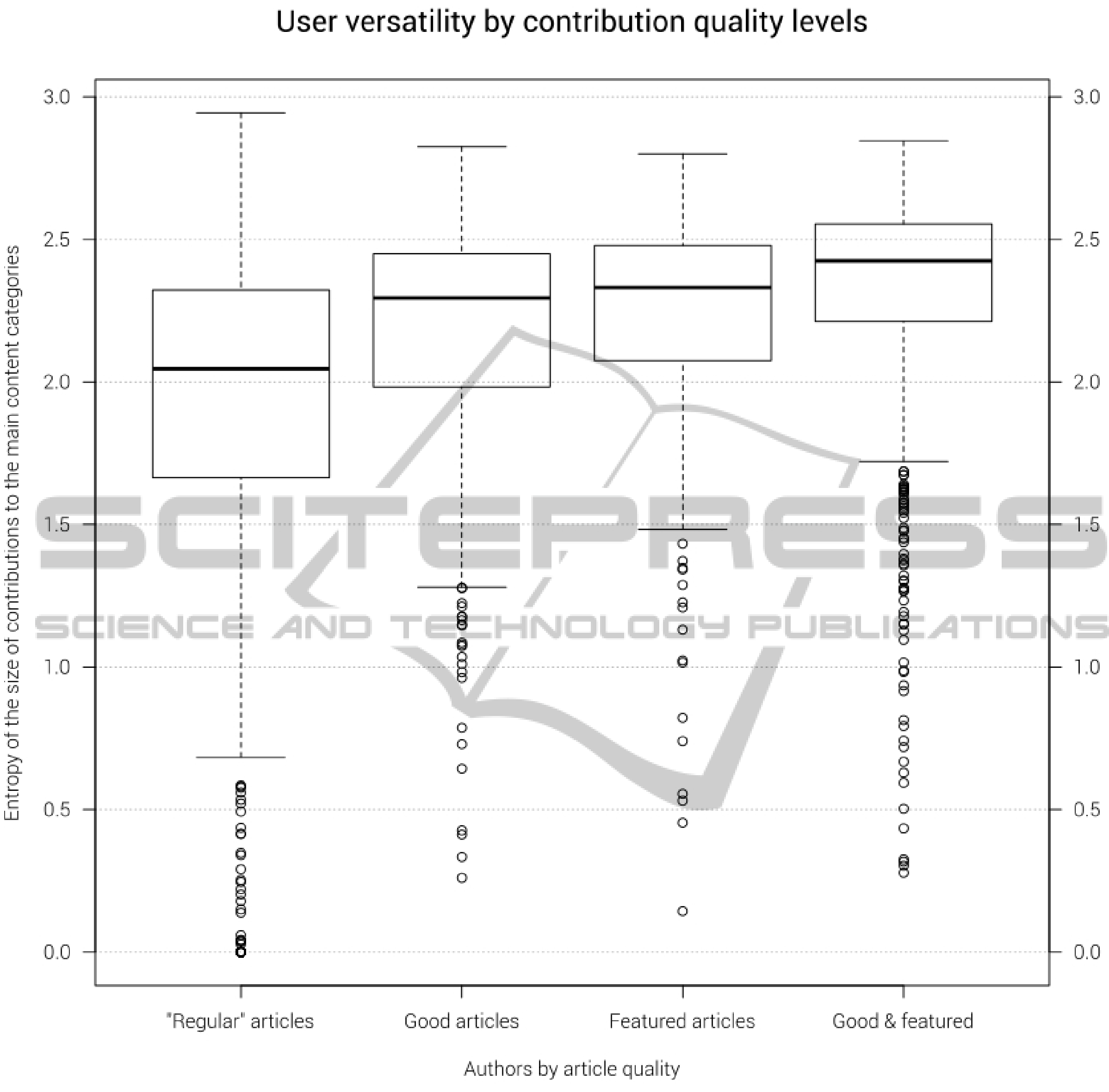
2.2 Measuring the Diversity of Interests

There are many possible ways of measuring diversity. Since the interest profile $ip(x)$ is modelled as a distribution vector over categories, we define *diversity of interests* (or equivalently *versatility*) of x, $V(x)$, as the *entropy of interest profile* of x:

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

Where lg denotes binary logarithm. The value of entropy ranges from 0 (extreme specialisation, i.e. total devotion to a single category) to $lg(k)$ (extreme diversity, i.e. equal interest in all categories).

[1] Since a single article can be assigned to multiple categories, we split the contribution equally for all the categories of the article



[Editors of German Wikipedia edition (de.wikipedia.org), 2013]

Figure 1: Versatility vs Quality.

2.2.1 Example

The versatility of user x from Example 2.1.1 has the following value:

$$V(x) = -p_2 lg(p_2) - p_5 lg(p_5) =$$

$$= 0.8 \times 0.32 + 0.2 \times 2.32 = 0.25 + 0.46 = 0.6$$

Now assume that another user x' has contributed equally to the four first categories, i.e. their interest profile is: $ip(x') = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0)$. The versatility value for this editor has the following value:

$$H(ip(x')) = -4 \times 0.25 \times (log_2(0.25)) = 2$$

Notice that the versatility measure of x' is higher than that of x and that this is according to the intuition since x' has similar interest in four different categories and x only in two (mostly in one). In other words, x' is more versatile while x is more specialised. Maximum versatility for eight categories would have value of 3, for an editor that is equally intested in all categories.

3 EXPERIMENTS

In this section we report experiments made on data extracted from Wikipedia that reflects recorded activity of its editors.

The goal is to experimentally study the dependence between editors' versatility as defined in Section 2 and the quality of articles they co-edit. In the reported experiments the quality of articles is modelled based on the information available in the data. More precisely, we utilise two kinds of information regarding the articles' quality: some articles are marked as *featured* and, independently, some as *good*. We treat this information as "gold-truth" in our experiments.

3.1 Data

The data covers sample of 2714 contributors to German-language edition in 2013. We used the Wikipedia API for retrieving the list of contributors and their activity logs, and database dumps for the page (article) list and category graph.

Considering the *categories* mentioned in the Section 2, we utilise the fact that each Wikipedia article can be mapped to one of the eight main content categories: *Art & Culture*, *Economy*, *History*, *Knowledge*, *Religion*, *Society*, *Sport*, *Technology*. Technically, the mapping to categories was computed so that they were encountered by the algorithm traversing the category graph using given article as a root node and iterating over neighbors up to 1000 times. If the article was mapped to more than one category, contribution size was split equally among them, so that we could use valid totals after per-user aggregation.

3.2 Experimental Results

We analysed four groups of editors: N, G, F, GF that denote editors who co-edited: none good nor featured article, at least one good, at least one featured and at least one article that is both good and featured, respectively. Notice that the four groups represent a graded "hierarchy" of high-quality editors, with the GF representing the highest-quality editors in some way. For each of the four groups we computed some statistics concerning versatility measure $V()$ (Equation 1), including mean, median and quartiles. The results are presented on Figure 1, where one can observe a noticeable regularity that indicates clear positive connection between editors versatility and the quality of their work. More precisely, the aggregated versatility statistics for the groups N, G, F, FG are strictly increasing.

Furthermore, we observed that the distribution of user versatility has a negative skew (Figure 2), with median value at 2.29 bits (out of 3-bit maximum). Users co-authoring at least one featured article score 2.31 on mean versatility measure, compared to 2.00 of those who co-authored only non-featured articles.

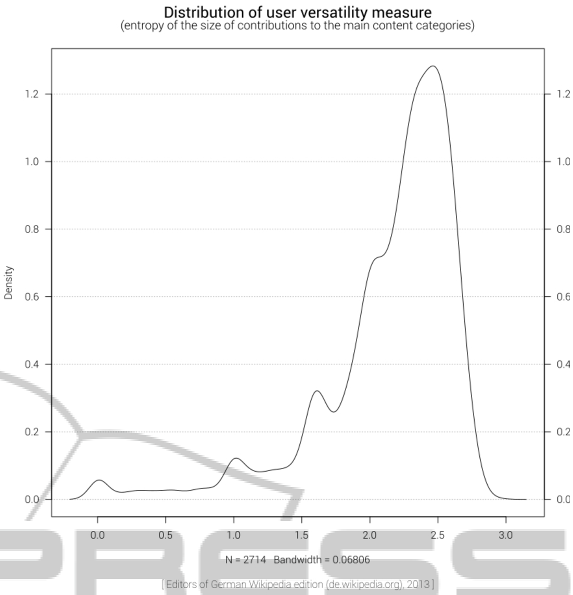


Figure 2: Distribution of Editor's Versatility.

3.3 Versatility, Quality and Productivity

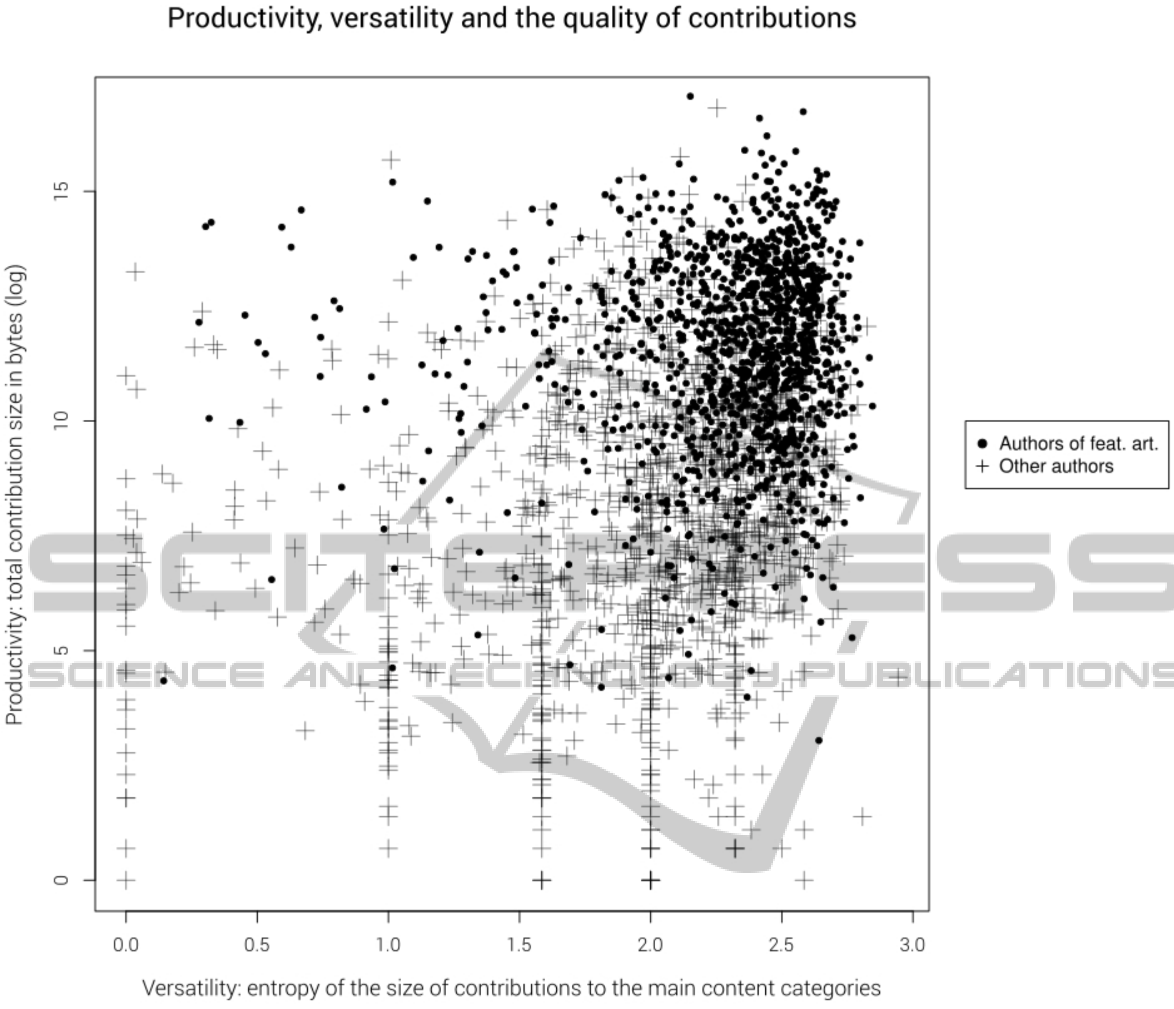
We also computed for each editor, their *productivity* defined as the total amount of text (in Bytes) committed to the articles they co-edited. We divided editors into two groups: F (at least one co-edited featured article) and $X \setminus F$ and made scatterplots of versatility vs productivity for these two groups (see Figure 3). Again, one can notice that the authors of featured articles are noticeably more versatile than others.

Since the results on Figure 3 might suggest that versatility and productivity are somehow correlated, we additionally repeated analogous (to that reported in Section 3.2) experiment on comparison of article quality and editors' *productivity* (Figure 4).

Finally, since the results of this experiment also seem to indicate some positive influence of productivity on quality we finally decided to compare the influence of versatility and productivity on quality in a more quantitative way. For this reason we built the logistic model with versatility and productivity as explanatory variables. Table 1 shows no significant role

Table 1: Explaining quality with logistic model.

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-3.566e+00	2.720e-01	-13.111	$< 2e-16*$
versatility	1.434e+00	1.214e-01	11.820	$< 2e-16*$
productivity	4.822e-07	6.017e-07	0.801	0.423
vers. * prod.	5.474e-07	2.865e-07	1.911	0.056



[Editors of German Wikipedia edition (de.wikipedia.org), 2013]

Figure 3: Versatility – Activity Scatterplot.

of productivity in explaining the quality of contributions (the fact of authoring at least one featured article by a given user), however a significant, 4.2 odds ratio for one-bit increase in versatility measure.

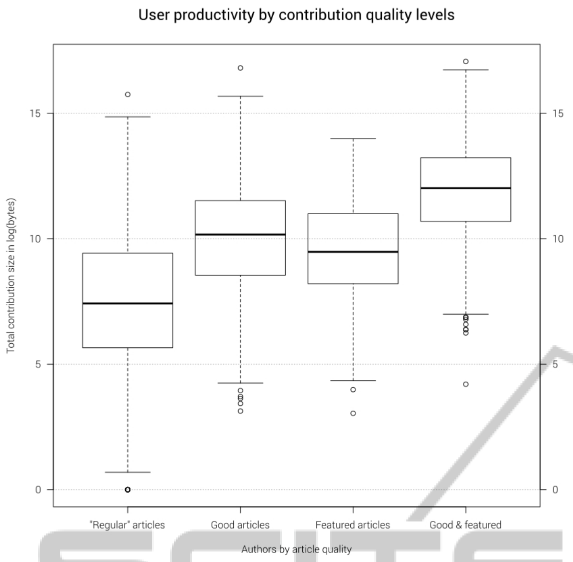
4 CONCLUSIONS AND FUTURE WORK

We proposed a model of user interests and entropy-based measure of interest diversity of a single Wikipedia editor. Preliminary experiments indicate that editors with more diversed interests seem to co-author better-quality content. On the other hand, despite an observed correlation between versatility and productivity, the latter one does not seem to explain article quality so well.

The continuation work would benefit from deepened and repeated experiments on other datasets and settings. For example, other choice of main thematic categories can be considered in next experiments. Also, other interest-diversity measures can be proposed. Since the reported preliminary experiments are promising, a natural future extension of this work would be to define *team diversity* based on some concepts introduced here and to extend the study on the issue of team-work quality.

ACKNOWLEDGEMENTS

The work is supported by Polish National Science Centre grant 2012/05/B/ST6/03364.



[Editors of German Wikipedia edition (de.wikipedia.org), 2013]

Figure 4: Productivity vs Quality. The denotations are analogous to those on Figure 1.

REFERENCES

Aggarwal, A. (2014). Decision making in diverse swift teams: An exploratory study. In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences*, HICSS '14, pages 278–288, Washington, DC, USA. IEEE Computer Society.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Chen, J., Ren, Y., and Riedl, J. (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, pages 821–830, New York, NY, USA. ACM.

Goffman, W. (1964). A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73 – 78.

Langlois, R. N. and Garzarelli, G. (2008). Of Hackers and Hairdressers: Modularity and the Organizational Economics of Open-source Collaboration. Working papers 2008-53, University of Connecticut, Department of Economics.

López, C. A. and Butler, B. S. (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, pages 673–682, New York, NY, USA. ACM.

Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. *Commun. ACM*, 15(12):1053–1058.

Sanchez, R. and Mahoney, J. T. (1996). Modularity, Flexibility, and Knowledge Management in Product and Organization Design. *Strategic Management Journal*, 17:63–76.

Sydow, M., Pikula, M., and Schenkel, R. (2013). The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, 41:109–149.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S. A. (2008). Efficient computation of diverse query results. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 228–236, Washington, DC, USA. IEEE Computer Society.