From MDMA to Lady Gaga: Expertise and contribution behavior of editing communities on Wikipedia

Louis J. Dijkstra\textsuperscript{1,3*} and Lisa J. Krieg\textsuperscript{2}

\textsuperscript{1} University of Amsterdam (UvA) Computational Science Lab, Amsterdam, The Netherlands
\textsuperscript{2} University of Amsterdam (UvA) Department of Anthropology, Amsterdam, The Netherlands
\textsuperscript{3} ITMO University, Saint Petersburg, Russia

Abstract

In this paper we present a methodology for gaining a better understanding of the contribution behavior, interests and expertise of communities of Wikipedia users. Starting from a list of core articles and their main editors, we identify which other articles (outside of the initial list) they contributed to ‘significantly’. The ordering is based on (empirical) Bayesian estimates of the contribution probabilities for each of the articles. By constructing a co-contribution network, we can identify the general themes the community expresses exceptional interest (or disinterest) in. In order to show what type of insights one might gain from employing the proposed method, we use the editors that contributed to the articles on designer drugs as a case study. We find that the users in this community contribute significantly to articles on pharmaceuticals, popular party drugs, chemistry, mental illnesses, diseases, medicine and cell biology.

Keywords: Contribution community, Wikipedia, designer drugs, empirical Bayesian shrinkage

1 Introduction

The online encyclopedia Wikipedia\textsuperscript{1} has developed into a more and more reliable knowledge source that rivals traditional printed encyclopedias [12]. Founded in 2001, it today contains over 40 million articles in almost 300 languages, and is based on the principle of open knowledge and user contribution; basically, everyone can contribute to articles or raise issues of discussion. This process is meant to ensure that articles are of high quality and up-to-date. In reality however, a large percentage of Wikipedias content is actually contributed from a rather small community of active users [12].

Wikipedia, its communities and the way in which information and knowledge are generated and maintained on this public platform received quite some attention within the fields of sociology and anthropology over the last few years. People explored a variety of aspects and viewpoints, such as the hierarchy, expertise and (paradoxical) nature of the editor community

\*Corresponding author: l.j.dijkstra@uva.nl.
\textsuperscript{1}https://en.wikipedia.org/wiki/Wikipedia

doi:10.1016/j.procs.2016.11.013
Expertise and contribution behavior of editing communities on Wikipedia

L.J. Dijkstra and L.J. Krieg

[8], the way in which knowledge is created in a collaborative fashion and how conflicts are resolved [18], the concept of Wikipedia as a global memory place [11], and how the platform can be seen as a cultural reference [15]; just to name a few.

In this article, we propose a methodology to access communities of Wikipedia contributors and to get an improved understanding of their contribution behavior, using a case study of the contributors of 547 articles on designer drugs, taken from the Wikipedia article [2] ‘List of designer drugs’. With the methodology suggested here we aim to gain deeper insights into the contribution behavior of the editors; more precisely, into the significant topics of a particular community of contributors, and into the degree of their specialization.

The methodology involves the following steps: First, we select a number of articles that characterize the community of interest. Secondly, we obtain a list of all the Wikipedia users that contributed (substantially) to at least one of these articles, and who thus form the community who is most responsible for creating the knowledge in question.

The third step is to collect all the other articles (outside of the initial list) to which these users contributed as well. These articles reflect the users’ expertise and other interests, and can shed light on the structure of the semantic field in Wikipedia, i.e. which topics users tend to specialize in. The fourth and final step is to determine which of these articles the users contributed to ‘significantly’, i.e., to which articles the users contributed much more in comparison to the rest of the active Wikipedia editors. In other words, we are interested in finding those articles that are most characterizing or specific for our community of interest.

Our contributions in this paper are two-fold: 1) we present an empirical Bayesian approach for obtaining the articles that characterize the community of interest the most, and 2) the code for the Wikipedia scraper and the subsequent statistical analysis are publicly available as a Python package for anyone to use.

To exemplify what types of insights one might gain from performing such an analysis we will perform a case study on what we will refer to as the designer drug community. Designer drugs are synthetic analogues of restricted or prohibited substances. Their chemical structure and function resemble the illicit drug form, but circumvent the law, since the chemical compound is not illegal per se [2, 17]. Even though the act of developing and experimenting with designer drugs is legal, designer drugs are surveilled by the European Monitoring Centre for Drugs and Drug Addiction [6] and exist in a legal gray zone. Knowledge about designer drugs is frequently exchanged online, by people who experiment with effects of these novel psychoactive substances [1, 2, 17].

The paper is structured as follows: in Section 2 we describe the methodology in detail. We discuss both its rationale, introduce the statistical model and introduce ways to explore the results. In Section 3 we present the results for the designer drug community and show the type of insights one might gain by employing the proposed methodology. We end with our conclusions, a discussion and some pointers for future research.

1.1 Nomenclature

We will adopt the following nomenclature throughout the rest of the paper to avoid any potential confusion. Registered Wikipedians, i.e., with a specified user name, are simply referred to as a user or an editor. The core articles are the articles that are thought to characterize the community of interest the most. We will refer to the users that contributed (substantially) to the core articles simply as the user community or the core users. The main idea behind the presented approach is that this user community is an appropriate proxy for the original

---

community of interest. The Wikipedia articles the user community contributed to as well, so outside of the list of core articles, are referred to as peripheral articles. From these articles, we identify those articles that are most ‘characterizing’ for the user community; we will refer to this list as the characterizing or defining articles.

2 Methods

The analysis starts with an initial list of articles that are thought to characterize the community of interest. For each article on the list, we collect all registered Wikipedia users (so all users that were not anonymous during their edit) that contributed to that particular article using the online Wikimedia Tools Labs service\(^3\). These users form the user community that we are interested in. Subsequently, for each of these core users, we collect all other articles to which they contributed as well (outside of the initial list) which results in a list of \(q\) ‘peripheral’ articles. The idea is that these articles tell us something about the users’ other interests and expertise and can help in gaining a better understanding of the community of interest.

For each peripheral article \(i\) on the list, we count

1. the number of core users that contributed, denoted by \(s_i\), and
2. the total number of contributors (so including the core users), which we will denote by \(n_i\).

The collected data for every peripheral article can be represented as a \(2 \times 2\) contingency table as shown in Table 1, where \(m\) represents the size of the user community, \(s_i\) and \(n_i\) are defined as above, and \(N\) is the total number of (active) Wikipedia users.

<table>
<thead>
<tr>
<th>User community</th>
<th>Rest of Wikipedia users</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributed to article (i)</td>
<td>(s_i)</td>
<td>(n_i - s_i)</td>
</tr>
<tr>
<td>Did not contribute to article (i)</td>
<td>(m - s_i)</td>
<td>(N - m - n_i + s_i)</td>
</tr>
<tr>
<td>Total</td>
<td>(m)</td>
<td>(N - m)</td>
</tr>
</tbody>
</table>

If one wants to decide whether an article is of particular interest to the user community, one would normally explore the level of association between whether or not someone belongs to the user community and the probability of contributing to the article at hand. Common approaches to this are applying a test of association to the \(2 \times 2\) table (e.g., Pearson’s \(\chi^2\) or Fisher’s exact test), and, subsequently, employing a false discovery rate (FDR) control procedure to identify a subset of articles thought to ‘significantly’ differ between the core users and the rest of the Wikipedia community (see, for example, [4]).

In our case, however, it is hard to employ this approach, since it is unclear how large exactly the rest of the active Wikipedia community is (or even which users to consider as a part of the ‘population’). In other words, we do not know \(N\). One could potentially estimate \(N\) reliably using a (repeated) capture-recapture procedure [10], but the final list of articles thought to be significant can (heavily) depend on the estimate of \(N\). We, therefore, deviate here from the common approach and tackle the problem of obtaining a list of interest in a different, empirical Bayesian approach that does not depend on the choice of \(N\).

\(^3\)http://tools.wmflabs.org (last accessed on the 15th of August, 2016)
Expertise and contribution behavior of editing communities on Wikipedia
L.J. Dijkstra and L.J. Krieg

The approach we advocate here is inspired by a question commonly posed in sabermetrics [3, 9], a field of statistics concerned with the analysis of baseball\(^4\): Given the number of hits and the times players have been at bat, who would you consider the better hitter? For example, player A had 4 hits out of 10 times at bat, while player B had 350 hits out of 1,000. When we would base our decision solely on the observed hitting rate, we would have to conclude that player A is better than player B, since 40% > 35%. However, it might be that player A normally has a much lower hitting rate and performed unusually well in those 10 recorded attempts. In contrast, there is strong evidence that player B is a good hitter indeed, since he/she can boast a solid high hitting rate of 35% in 1,000 attempts; taking this into consideration, it might be, on the basis of the available evidence, more sensible to select player B as the better hitter.

We will approach our problem of deciding whether an article \(i\) is more or less ‘characterizing’/‘important’ for the user community than article \(j\) by comparing the number of core users that contributed to the articles (\(s_i\) and \(s_j\); these will correspond to the number of ‘hits’) and the total number of users that contributed to the article (\(n_i\) and \(n_j\); the number of times being ‘at bat’). Informally, we want to select the article associated with

1. a high fraction \(s/n\), since it implies that a large part of the content of the article comes from core users and, therefore, might be of particular interest to them, and
2. a fairly large number of contributors to the article, \(n\). When \(n\) is low (say 3 or 4), even when the fraction \(s/n\) is high, the article can hardly be called very characterizing or defining for the community.

The approach we suggest here is commonly referred to as empirical Bayesian shrinkage towards a Beta prior and, as we will show, possesses these two properties; higher fractions are preferred, while the possibility of ‘being lucky’ is accounted for.

2.1 The Model

We can model the number of core users that contributed to a particular article \(i\) naturally as a Binomial distribution:

\[ s_i \sim \text{Binomial}(n_i, \pi_i) \quad \text{for } i = 1, 2, \ldots, q \quad (1) \]

where \(n_i\) is, again, the total number of users that contributed to article \(i\) and \(\pi_i\) is the probability of a core user to contribute. (We will refer to this probability as the contribution probability). We assume each of these parameters to follow a Beta prior

\[ \pi_i \sim \text{Beta}(\alpha, \beta) \quad \text{for } i = 1, 2, \ldots, q \quad (2) \]

where \(\alpha\) and \(\beta\) are the so-called hyperparameters. The Beta prior for probabilities such as \(\pi_i\) is a rather common choice, since its support is the unit interval and is (in this case) conjugate to the likelihood function.

The posterior distribution for the article contribution probability \(\pi_i\) is then given by

\[ g(\pi_i | s_i; n_i) = \frac{[h(\pi_i) \cdot f(s_i | \pi_i; n_i)] / f(s_i; n_i)} / f(s_i; n_i) \quad (3) \]

where \(h(\pi_i)\) is the Beta prior, \(f(s_i | \pi_i; n_i)\) is the likelihood function, and \(f(s_i; n_i)\) is the marginal probability mass function, given by

\[ f(s_i; n_i) = \int_0^1 \pi_i f(s_i | \pi_i; n_i) d\pi. \quad (4) \]

\(^4\)See, for example, http://varianceexplained.org/r/bayesian_ab_baseball (last accessed on the 15th of August, 2016)
As mentioned before, the posterior, $g(\pi_i | s_i; n_i)$, and the prior, $h(\pi_i)$, are conjugate distributions, i.e., the posterior distribution follows a Beta distribution as well:

$$\pi_i | s_i; n_i \sim \text{Beta}(\alpha + s_i, \beta + n_i - s_i).$$

(5)

We can estimate the contribution probabilities by maximizing the respective posterior distributions. For this, we first need to choose the hyperparameters $\alpha$ and $\beta$ appropriately.

Note that we are faced with many similar estimation problems; one for every peripheral article, which can be very large indeed. We, therefore, employ an empirical Bayesian approach in which we will estimate the hyperparameters of the prior on the basis of the data.

Let $s = (s_1, s_2, \ldots, s_q)$ and $n = (n_1, n_2, \ldots, n_q)$ denote the observed data, i.e., the number of editors from the user community and the entire Wikipedia community for the $q$ peripheral articles. The number of editors from the user community follows, when the contribution probabilities are Beta distributed, a (compound) Beta-Binomial distribution. The likelihood function $L(\alpha, \beta)$ is given by

$$L(\alpha, \beta) := f(s | \alpha, \beta; n) = \prod_{i=1}^{q} \frac{1}{B(\alpha, \beta)} \int_0^1 p^{s_i + \alpha - 1}(1 - p)^{n_i - s_i + \beta - 1} dp$$

(6)

where $B(\cdot)$ is the Beta function. The log-transform of this function can be numerically maximized which yield the maximum likelihood estimates (MLEs) of $\alpha$ and $\beta$:

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{(\alpha, \beta)} \log L(\alpha, \beta | s; n).$$

(7)

After estimating $\alpha$ and $\beta$, the posterior distributions for the contribution probabilities for the articles are simply given by $\hat{\pi}_i \sim \text{Beta}(\hat{\alpha} + s_i, \hat{\beta} + n_i - s_i)$. The empirical Bayes estimates for the contribution probabilities are then simply the expectation of the posterior distribution:

$$\hat{\pi}_i = \frac{\hat{\alpha} + s_i}{\hat{\alpha} + \hat{\beta} + n_i}.$$ 

(8)

We can create a ranking of the articles on the basis of their empirical Bayes contribution probability estimates; the article with the highest estimated rate is considered to be the most characterizing for the community, with the second highest estimate the second most characterizing etc. Note that such an ordering possesses the two properties we specified earlier. First, articles with a higher contribution rate $s/n$ are preferred. Suppose we compare two articles, $i$ and $j$, with both the same total number of contributors, i.e., $n = n_i = n_j$. If we base the ordering on the estimates from eq. (8), we will put the article with the higher number of core users contributions over the other. In addition, when the total number of contributions $n$ is small, the estimate is not strongly influenced by $s$; it will stay close to the mean of the prior, $\hat{\alpha}/(\hat{\alpha} + \hat{\beta})$, and will not be considered of particular interest to the user community. (This latter property is the reason why the method is often referred to as *Bayesian shrinkage*; the estimate is ‘shrunk’ towards the mean of the prior).

On the basis of the posterior distribution, we can also create a $(1 - \gamma)$-credible interval to get a sense of the amount of uncertainty around the contribution probability estimate. There are many ways in which to construct credible intervals (since any interval containing $1 - \gamma$ percent

---

5This might feel rather unconventional; the name ‘prior’ suggests that it reflects our believe before we are faced with any data. For a good introduction to empirical Bayes and its rationale, we refer the interested reader to [5, 13, 14].
of the distributions ‘weight’ would be valid). Here we will use the highest probability density (HPD) interval, which is defined as the smallest interval that contains the mode of the density at hand. More precisely, the HPD credible interval is that interval \([a, b]\) for which

\[ g(a \mid s_i; n_i) = g(b \mid s_i; n_i) \]

and

\[ \int_a^b g(x \mid s_i; n_i) \, dx = 1 - \gamma. \]

The code and scripts for performing the scraping of Wikipedia and the subsequent statistical analysis presented in this section can be found in the Git repository at https://github.com/louisdijkstra/wikiscraper published under the Apache 2 open source license, and is, therefore, free for anyone to use.

2.2 Visualizing the Results: the Co-contribution Network

An ordering of articles on the basis of their estimated contribution probabilities can help in gaining a better understanding of what other articles the user community contributes to significantly; still, it is hard to get a sense of the general topics the users show interest in, especially when the list of articles is long. Visualizing the articles and ‘clustering’ them into topical groups can help in understanding the global picture.

To this end, we propose to construct a co-contribution network; let \( G = (V, E) \) be a undirected graph where \( V \) is the set of peripheral articles and \( E \) is the set of edges between them. Let \( A = \{a_{ij}\} \) be the \( q \times q \) adjacency matrix, where the weight of the edges is equal to the number of core users that contributed to both the articles:

\[ a_{ij} = \#\{\text{core users that contributed to both } i \text{ and } j\}, \]

The rationale behind using such a co-contribution network is that users that contribute to a particular article are more likely to edit another article as well when they are closely related topic-wise, e.g., a user that contributed to a chemistry page is more likely to contribute to other chemistry-related pages as well. By grouping vertices with strong connections together, we can gain insight into the different themes in the list of articles. By appropriately coloring or scaling the nodes on the basis of their estimated contribution probability, we can visualize which articles and clusters are more characterizing/defining for the user community of interest, and which less. In the Results section, we show the co-contribution network for the designer drug community.

3 Results

To exemplify what type of insights one might gain from the analysis and visualization presented in the previous section, we will look at the users interested in designer drugs. As initial list of articles, we selected all 547 internal links to designer drug pages on the Wikipedia page ‘List of designer drugs’. We scraped all the users that contributed to at least one of these pages, which yielded a list of \( m = 4,573 \) distinct editors. Subsequently, we collected all the articles these core users contributed to as well yielding a collection of \( q = 200,842 \) distinct Wikipedia pages. For each of these peripheral articles, we counted the number of core users \((s_i)\) and the total number of users that contributed \((n_i)\). The data was downloaded in the first week of August, 2016.
The complete scraped data set is publicly available via https://github.com/louisdijkstra/wikiscraper; one can find the list of initial designer drug articles, the list of core users and their contributions and the list of peripheral articles, together with the counts.

3.1 Estimating the Prior

Figure 1a shows a histogram of the contribution rates, \( s/n \), for all the 200,842 scraped peripheral articles. After fitting the observed rates to the Beta-Binomial distribution by maximizing the likelihood function in (6), we found the MLEs of the hyperparameters to be

\[
\hat{\alpha} \approx 6.28 \quad \text{and} \quad \hat{\beta} \approx 16.93.
\] (12)

Their 95% (likelihood-based) confidence intervals are, respectively, [6.28, 6.34] and [16.76, 17.10].

The fit of the Beta distribution is shown in Figure 1a as a red dashed line. Note that the quality of fit is reasonable.

![Figure 1: (a) A histogram of the contribution rates, i.e., the fraction of users that contributed to a peripheral article that belong to the designer drug community. The fit of the Beta distribution is shown as dashed red line. The hyperparameters were estimated to be \( \hat{\alpha} \approx 6.28 \) and \( \hat{\beta} \approx 16.93 \). Their 95% (likelihood-based) confidence intervals are, respectively, [6.28, 6.34] and [16.76, 17.10]. (b) The posterior distributions of a number of Wikipedia articles. The prior distribution is depicted as the dashed black line. The articles ‘Aminorex’ and ‘Lady Gaga discography’ are the articles with, respectively, the highest and lowest contribution probability estimate.](image)

3.2 Posterior Densities

Having estimated the hyperparameters \( \alpha \) and \( \beta \), we can continue with estimating the posterior distributions for each of the peripheral articles.
Figure 1b shows, as an example, the posterior densities of 6 different articles (the prior is represented by a black dashed line for reference). The page ‘Aminorex’ is added since it is the article with the highest contribution probability estimate. Aminorex is a stimulant drug and a pharmaceutical formerly used as appetite suppressant, which has been removed from the market [7]. The article with the lowest estimate is the page on Lady Gaga’s discography. Of all pages the user community contributed to, they contributed to this page the least. The ‘Magical Half-Dozen’ refers to the 6 ‘most important’ phenethylamine compounds from a book titled ‘PiHKAL: A chemical love story’ by the couple Shulgin in which they discuss various designer drugs [16]. The page ‘Addiction’ is edited by the community as well, but does not seem to attract unusual interest, since it does not seem to deviate much from the prior. World War II appears to attract significantly less attention, while fish physiology does seem to have the interest of the community.

Figure 2a shows a larger number of articles. We added in addition to the articles from Figure 1a 50 randomly selected articles from the original list of peripheral articles. The point represents the empirical Bayes estimate; the whiskers denote the HPD 95% credible intervals. The counts next to the titles of the articles are the number of core users and the total number of users that contributed to that article, i.e., \( \frac{s_i}{n_i} \). The dashed red line represents the mean of the prior distribution, i.e., \( \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \approx 0.27 \). The entire list of peripheral articles, together with their contribution probability estimates and credible intervals can be found on https://github.com/louisdijkstra/wikiscraper.

3.3 The Co-contribution Network

Figure 2b shows the co-contribution network for the designer drug community (see Section 2.2). Recall that the nodes represent the peripheral articles and that the weight of the edges is equal to the number of core users that contributed to both of the articles it connects.

We reduced the original number of nodes (which is over 200,000) by selecting the top 250 articles that received most contributions from the designer drug community, which yielded a set of 7,745. (The discrepancy is because some articles had the same number of contributors). In addition, we filtered out nodes with a degree of 1; edges with a weight less than 3 were removed as well. The node size in the figure reflects the node’s degree. The color of the node represents the empirical Bayes contribution probability estimate; red reflects a low contribution probability, yellow is middle-ground, and blue corresponds to high contribution probabilities.

In the upper right corner, one can see a clear blue cluster consisting mainly of articles related to various forms of illicit substances and party drugs (such as methamphetamine and MDMA) and pharmaceuticals (e.g., paracetamol and ibuprofen). This cluster contains the peripheral articles with the highest empirical Bayes contribution probability estimates, and can thus be considered to be the most characterizing for the editors community. It shows that people who tend to write on designer drugs are more likely to contribute to other forms of chemical substances as well, and thus seem to be knowledgeable in the field of pharmacology. The cluster on the right contains a variety of articles on chemical elements and other chemistry-related topics, such as gold, ammonia, the periodic table etc. The blue/yellow clusters in the upper left corner entail a part on psychological, neurological, and psychiatric diseases and conditions, and a second part on the very left which relates to other diseases such as Tuberculosis, Multiple Sclerosis and forms of cancer. The large red cluster in the lower right corner comprises a large number of ‘popular topics’ such as countries, people, and TV shows; it appears that these topics are significantly less contributed to by the designer drug community. In the middle, there is a small blue cluster related to biology with articles such as Chromosome, Amino Acid and
Figure 2: (a) The empirical Bayes estimates for 50 randomly selected articles and the 6 articles from Figure 1a. The whiskers represent the highest probability density 95% credible intervals. The observed fraction, i.e., s/n, is added to the title of the articles. The mean of the prior ($\approx 0.27$) is shown as a red dashed line. (b) The co-contribution network of the designer drug community of the top 250 articles that received most contributions from the designer drug community (7,745 articles in total). Nodes with a degree of 1 and edges with a weight less than 3 are ignored. The node size reflects the node’s degree. The node’s color represents the contribution probability estimate; red stands for a low estimate, yellow for middle and blue for high.

Cell membrane. The most characterizing topics for the user community we are interested in, emerging from this graph, are pharmaceuticals, chemistry, mental illness and neurology and diseases.

4 Conclusions & Discussion

In this paper we presented a methodology to access communities of Wikipedia contributors and to get an improved understanding of their contribution behavior. We did this by collecting all articles they contributed to, and, subsequently, ordered them according to the (empirical) Bayesian estimate of the contribution probabilities, i.e., the chance for a user from the community of interest to contribute to the article at hand. By basing the ranking on these estimates, the resulting list satisfies two intuitive properties: articles with a higher contribution rate are thought to be more characterizing/important for the user community than pages with a lower rate, while, at the same time, we account for ‘chance’: articles with a small number of contributors to start with are less likely to be considered important.
We applied the proposed methodology to the community of Wikipedia users that contributed to designer drugs and found that they contribute significantly more to illicit drugs, pharmaceuticals, chemistry-related topics, neurological and psychiatric conditions, diseases (such as cancer) and cell biology. In addition, we observed that they contribute significantly little to a variety of pop culture related topics (e.g., Lady Gaga’s discography) and history (e.g., World War II). See Figure 2b for the co-contribution network of this community.

There are a few issues that future research should tackle. In the case of designer drugs, the list of ‘core articles’ used to define the user community was based on the Wikipedia page ‘List of designer drugs’. In some cases, it might make sense to create a novel list of articles as a starting point. As a way to tackle this problem we suggest to create the list of core articles in an iterative fashion. First, start with a set of articles related to the topic of interest. This list does not have to be comprehensive. Secondly, compile the list of peripheral articles and determine their contribution probabilities. One might find articles that have an extraordinarily high estimate and are directly relatable to the topic of interest. Add them to the list of core articles and repeat the process until one feels the list is complete.

Ordering the articles can help in gaining a better understanding of the content of the list, but, as mentioned before, when the list is long (as was the case with the designer drug community), it can be challenging to identify general themes. Using the fact that users tend to edit articles that are related topic-wise, as is done with creating a co-contribution network (see Section 2.2), helps, but there might be other ways to identify articles to belong to the same theme. For example, one might explore the possibility to use Wikipedia’s own categorization; rather than looking at which articles the users contributed to more significantly, one could assess which categories are edited more. The challenge here is that the number of categories associated with articles can be quite numerous; even to such an extent, that the number of categories might exceed the list of articles.

Here we only assess the contribution probability of individual articles, but it might be useful to get a sense of the ‘significance’ of groups/clusters of articles as well. Having observed clusters emerging in the co-contribution network, one might rather assess their ‘importance’ as a whole rather than a collection of single articles.

The code for scraping Wikipedia and to perform the subsequent statistical analysis is publicly available under the Apache 2 license, see https://github.com/louisdijkstra/wikiscraper. We hope that the presented methodology together with the availability of a working package will aid others in gaining a better, deeper understanding of Wikipedia, its communities and the way in which information and knowledge is created and maintained on such an open platform.

Acknowledgments

The methodology and the idea for using the designer drug community as a case study originated mainly during the Chemical Youth Data Sprint held at the University of Amsterdam this year. The authors would like to thank all participants and organizers for their input and ideas.

References

Expertise and contribution behavior of editing communities on Wikipedia

L.J. Dijkstra and L.J. Krieg


