Discussion Paper No. 12-053

Centrality and Content Creation in Networks – The Case of German Wikipedia

Michael E. Kummer, Marianne Saam,
Iassen Halatchliyski, and George Giorgidze

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 12-053

Centrality and Content Creation in Networks – The Case of German Wikipedia

Michael E. Kummer, Marianne Saam,
Iassen Halatchliyski, and George Giorgidze

First Version: 14.08.2012

This Version: 10.10.2012

Download this ZEW Discussion Paper from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/dp12053.pdf

Non-Technical Summary

The free online encyclopedia Wikipedia represents a prototypical case of peer production of an information good on a large online platform. This production mode is nowadays widely spread on the Internet. Peer production is governed neither by the market nor by a firm. A mass of producers usually contributes small fragments of the overall output without remuneration. In the absence of market signals and hierarchical decisions, it is important for platform administrators to understand how producers decide where to contribute. On a complex and dynamic platform like Wikipedia, this decision is expected to depend on the way the content is organized. One main organizing principle for content on wikis are hyperlinks, i.e. links that allow to browse from one article to another.

We study how the position of an article in the hyperlink network is related to how much content is provided by users, and which role the network position of an article plays in attracting the contributions of new authors. The network we consider is defined by incoming hyperlinks on articles within German Wikipedia. We chose a sample of more than $7,000$ articles belonging to a particular category ("Oekonomie" - "Economics") observed over a period of 153 weeks. For this sample, we compute centrality measures within the category and on the entire German Wikipedia. Thus we can compare links from articles that are semantically close to links coming from articles that are on average less closely related.

We find that increases in the number of links from the category are strongly associated with increases in page length. In particular, greater centrality of an article is associated with *new* authors contributing to the article. Evidence for a relation between links from outside the category to page length turns out to be rather weak. Social network analysis reveals that the category "Economics" is, like many networks, constituted by one large cluster and other single articles or small network components that are disconnected from it. Getting connected to the large cluster raises the page length and its rate of change sizeably in the following weeks. The size of contributions associated with new links is in the order of magnitude of several words to one or two sentences. While this may seem not very large, many weekly changes on Wikipedia articles are of this size.

Das Wichtigste in Kürze

Die frei zugängliche Onlineenzyklopädie Wikipedia ist ein prototypisches Beispiel für Peer Production eines Informationsgutes auf einer großen Onlineplattform. Diese Form der Produktion hat im Internet weite Verbreitung gefunden. Peer Production wird weder vom Markt noch von Firmen koordiniert. Gewöhnlich trägt eine Vielzahl von Produzenten kleine Fragmente zum Produktionsergebnis bei, ohne dafür eine Entlohnung zu erhalten. Da die Produktion nicht durch Marktsignale oder hierarchische Entscheidungen gesteuert wird, ist es für Plattformadministratoren wichtig zu verstehen, wie die Mitwirkenden entscheiden, was sie beitragen. Auf einer komplexen und dynamischen Plattform wie Wikipedia ist zu erwarten, dass diese Entscheidung davon abhängt, wie die Inhalte zueinander angeordnet sind. Ein wesentliches Anordnungsprinzip in Wikis sind Hyperlinks, die es ermöglichen, von einem Artikel zum anderen zu navigieren.

Wir untersuchen, wie die Netzwerkposition eines Artikels im Hyperlinknetzwerk mit der Textmenge zusammenhängt, die Plattformnutzer zu diesem Artikel beitragen. Besonders interessant ist dabei die Rolle, die die Verlinkung bei der Gewinnung von neuen Autoren für einen Artikel spielt. Wir betrachten das Netzwerk, das durch auf die Artikel zeigende Hyperlinks in der deutschen Wikipedia entsteht. Wir wählen eine Stichprobe von mehr als 7000 Artikeln aus der Kategorie "Ökonomie" über einen Zeitraum von 153 Wochen hinweg. Für diese Stichprobe berechnen wir Zentralitätsmaße basierend auf der Verlinkung innerhalb der Kategorie und der Verlinkung mit der gesamten deutschen Wikipedia. Somit können wir Links von Artikeln, die inhaltlich verwandt sind, mit Links von solchen Artikeln vergleichen, die inhaltlich im Schnitt entferntere Themen behandeln.

Es zeigt sich, dass ein starker Zusammenhang zwischen der Entstehung von zusätzlichen Links innerhalb der Kategorie und der Zunahme der Artikellänge besteht. Insbesondere finden wir heraus, dass höhere Zentralität mit Beiträgen von neuen Autoren korreliert. Effekte von Links von außerhalb der Kategorie erweisen sich als schwach. Eine Netzwerkanalyse ergibt, dass die Kategorie "Ökonomie", wie viele andere Netzwerke, aus einem großen verlinkten Cluster und anderen Artikeln oder kleinen Netzwerkkomponenten besteht, die nicht mit dem Cluster verbunden sind. Eine Verlinkung mit dem großen Cluster erhöht die Artikellänge und auch die Rate ihrer wöchentlichen Veränderung deutlich. Die Länge der zusätzlichen Beiträge, die mit einem neuen Link verbunden sind, bewegt sich in der Größenordnung von wenigen Wörtern oder ein bis zwei Sätzen. Dies mag gering erscheinen, jedoch entstehen auf Wikipedia viele wöchentliche Textveränderungen diesen Umfangs.

Centrality and Content Creation in Networks *

- The Case of German Wikipedia

Michael E. Kummer
Centre for European Economic Research (ZEW)

Marianne Saam
Centre for European Economic Research (ZEW)

Iassen Halatchliyski
Knowledge Media Research Center (IWM-KMRC)

George Giorgidze
University of Tübingen

First Version: 14.8.2012
This Version: 10.10.2012

Abstract

When contributing content to large and highly structured online platforms like Wikipedia, producers of user-generated content have to decide where to contribute. This decision is expected to depend on the way the content is organized. We analyse whether the hyperlinks on Wikipedia channel the attention of producers towards more central articles. We observe a sample 7,635 articles belonging to the category economics on the German Wikipedia over 153 weeks and we measure their centrality both within this category and in the network of over one million German Wikipedia articles. Our analysis reveals that an additional link from the observed category is associated with around 140 bytes of additional content and with an increase in the number of authors by 0.5. The relation of links from outside the category to content creation is much weaker.

JEL-Classification: L14, D83

Keywords: user-generated content, network analysis, hyperlinks, spillovers

*Correspondence: Michael Kummer: Centre for European Economic Research (ZEW); L 7, 1; 68181 Mannheim; Germany; Email: Kummer@zew.de. We thank Thorsten Doherr for support with the Wikipedia data, and Frédéric Schütz for providing us with the data on page views. We benefitted from discussions with Irene Bertschek, Ulrike Cress, Benjamin Engelstätter, Avi Goldfarb, Francois Laisney, Jose Luis Moraga-Gonzalez, Martin Peitz, Philipp Schmidt-Dengler, Michael Ward, the participants of the ICT Conference 2012 at ZEW in Mannheim and of the annual conference of the EARIE 2012 in Rome. Benedikt Achatz, Sergiy Golovin, Burak Tuerkoglu and Fabian Trottner provided helpful research assistance. We acknowledge financial support from the WissenschaftsCampus Tübingen.

1 Introduction

User-generated content has proved to be a cheap and surprisingly accurate source of information. Still, little is known about how its producers select the content to which they contribute and how platform administrators may influence this choice. While Wikipedia has been the most successful prototype of a wiki, wikis in other contexts, e.g. private businesses, often struggle to encourage and manage activity. Administrators of platforms face three challenges: motivating potential first-time users, making them connect to the platform and encouraging the contribution of content that is useful to others (Lerner and Tirole (2002), Jian and MacKie-Mason (2012)).

In order to encourage contributions, it is important to understand how authors select articles. In this paper, we study one mechanism that possibly channels their activity. We start from the hypothesis that the hyperlink network between Wikipedia articles attracts the attention of authors towards more central articles. In particular, we analyze how the position of an article in the network is related to the amount of content contributed and to the number of new authors joining the article. This question is situated in the more general context of understanding how producers in peer production of information goods select their tasks.

On Wikipedia, there are three main possibilities for finding articles of interest: categories, text search and hyperlinks. Frequent authors use additional devices such as lists of new articles, watchlists or lists of articles classified as needing improvement. Hyperlinks constitute an organizing principle that is indispensable to online peer production of a vast amount of information. They enable a non-hierarchical access and a nonlinear reading experience that are characteristic for wikis (Greenstein and Devereux (2009)).

Meanwhile little research has been undertaken on the question how hyperlinks influence contributions in wikis. Wikipedia's rules determine hyperlinks between articles to be semantic links, that means links that are set according to important connections in the attributes of the two subjects. The links need not be reciprocal and the guidelines on the German Wikipedia stipulate that an article must be readable without the information from the linked pages. It is not compatible with Wikipedia's rules to set links just to attract attention to an article or without embedding its subject into the text pointing to it. Finally, within Wikipedia, links should point only to pages about technical terms or to pages that contain further information on topics that might be of particular interest to readers of the originating article.[1] Hyperlinks on Wikipedia are generally regarded as a reliable source of information on semantic relations between words. They have been used extensively in linguistic research (see for example Medelyan et al. (2009)). Adafre and de Rikjke (2005) propose a procedure that automatically detects missing links between

[1] http://de.wikipedia.org/wiki/Wikipedia:Verlinken, accessed on July 23, 2012.

pages that should be linked given their relevance to each other. Taken together, this research suggests that hyperlinks on Wikipedia are generally set in accordance with the guidelines (see also Priedhorsky et al. (2007) on rapid detection of vandalism), but that the topics of articles on Wikipedia do not completely predetermine their link structure. The actual links depend on the dynamic content of an article and on the accuracy of linking. This implies that variations in centrality occur regularly and affect the navigation of readers and potential authors on a given set of articles. Our main hypotheses are that higher centrality is positively related to (i) the length of an article's content and (ii) the number of new authors joining the article.

Economic research considers spillovers to be a central feature of knowledge production. They arise when the production of new knowledge relies on existing knowledge, which can be used without paying for it and without diminishing anyone else's use of it (see for example Romer (1990) in the context of growth theory). Studies on R&D have highlighted that the strength of spillovers depends on the distance between the knowledge that is available and the knowledge that is being produced. This distance may be defined in various ways, for example geographically or according to sectors of economic activity (Griliches (1992), Audretsch and Feldman (1996)).

In the context of Wikipedia, we consider the centrality of an article in the hyperlink network as a measure of distance from other articles and thus as a factor affecting the strength of spillovers in knowledge production between articles. The knowledge that we expect to spill over from article A to article B is, however, not necessarily the knowledge codified in article A. Rather it is the effort and knowledge of the authors who pay attention to article B because a link in article A is pointing to it. The existence of an additional link may trigger the contribution of authors who might have not contributed in its absence.

When analysing the relation between centrality and content provision, we exploit different dimensions of proximity that exist between articles. In particular, we compare the links from articles that are semantically close to links which are on average less close. We also compare direct links to an article, measured by the number of incoming links (the indegree), to indirect links, measured by the closeness centrality. We chose a sample of more than $7,000$ articles belonging to a particular category (economics; German: "Wirtschaft"). For this sample, we compute centrality measures both within the category and on the entire German Wikipedia.

We find that an increase in the number of links from within the category is strongly associated with an increase in page length. It is also associated with new authors contributing to the article. The strongest relation between centrality and content generation is found for direct links from the category network. The relation to links from other pages of German Wikipedia is weaker and insignificant in our main specification. The additional influence of indirect links appears negligible. Social network analysis reveals

that the category economics is, like many networks, constituted by one large cluster and single articles or small network components that are disconnected from it. We find that getting connected to the large component raises the page length and its rate of change sizeably in the following weeks.

Our research is inspired by two strands of work on user-generated content: First, we are interested in knowing whether evidence generated inside a limited category of a network (e.g., Kittur and Kraut (2008), Ransbotham et al. (2012)) holds when taking into account links outside the category. Second, we follow work by Fershtman and Gandal (2011) and Claussen et al. (2012) on direct and indirect knowledge spillovers in networks of software producers. Contrary to these strands of work, we do not consider the network of authors but the hyperlink network of user-generated content.

2 Related Research

The empirical analysis of (social) networks has been of interest to scientists of different disciplines for several decades, resulting in a vast literature and in an established methodology based on the analysis of graphs. This tool has been widely used in empirical applications that are relevant to economics, so that we are forced to restrict ourselves to discussing only large overarching themes.[2] Some studies center around the existence and the structure of social networks, applying a variety of formally defined network measures. Other applications have analyzed the prevalence of homophily in networks, the importance of weak ties and social capital (for example in job-market outcomes), or the benefits associated with filling structural holes in networks.

Social networks have since then been at the heart of a variety of theoretical and empirical studies in economics. Diffusion in networks was originally studied in medicine and biology, but the methods can also be used in economics to study technology adoption or viral marketing. Moreover, economists became interested in citation networks. Goyal et al. (2006) analyze the evolution of the collaboration network of economists from the 1970s until the 1990s. They find that a structure of separated 'small islands' of researchers is increasingly replaced by a 'small world' network where every pair of nodes (authors) is connected by a short path. In fact, citation networks of scientific papers had been analyzed as early as the 1960s.[3] More recently, Albert et al. (1999) have undertaken a similar endeavor for web pages.

Particularly relevant to this paper are studies focussing on knowledge spillovers in production through social networks. Fershtman and Gandal (2011) analyse direct and

[2]For a more detailed summary of the literature (until 2008), cf. Jackson (2008).

[3]Without using the more recently developed measures of network position, de Solla Price (1965) evaluates citation data and provides several interesting statistics on average references and citations in the network.

indirect knowledge spillovers in the production of open source software and Claussen et al. (2012) in the electronic gaming industry. Both papers consider the relationship between developers' network position and the success of the project they are working on. Our research considers a different network in a similar context, namely the hyperlink network of articles. Thus, we borrow from the approach used by Halatchliyski et al. (2010) who analyze authors' contributions in two related knowledge domains considering the article network.

Several previous papers have studied collaboration between authors on Wikipedia. Denning et al. (2005) discuss problems associated with the collaboration of volunteers in Wikipedia, such as the unknown quality of articles or accidental inaccuracies. Focusing on a non-monetary reward tool at Wikipedia, "Barnstars", which can be awarded to hard working authors, and its contribution to content creation, Kriplean et al. (2008) offer a theoretical lens for understanding how wiki software can be designed to support the contribution of good work. In his dissertation, Soto (2009) reviews further existing research based on Wikipedia data and (among other things) quantitatively analyzes the ten largest Wikipedias finding that the patterns concerning the composition of authors on the platform as well as production patterns are highly similar.

Other empirical analyses focus on the determinants of the quality of articles. Kittur and Kraut (2008) examine how the number of collaborating editors in Wikipedia and the coordination methods they use affect article quality measured by peer evaluations in Wikipedia's quality assessment project. Their empirical results show that adding more editors to an article improves article quality only when the editors use appropriate coordination techniques. Zhang and Zhu (2011) empirically examine the potentially inverse relationship between the incentives to contribute and the size of the group of contributors. Based on exogenous variation in group size at the Chinese Wikipedia due to access blocks issued by the government, their analysis shows that contributors receive social benefits increasing with both the amount of contribution and group size. Accordingly, the result confirms that the more contributors value these social benefits, the more they tend to reduce their contributions after the block.

Ransbotham et al. (2012) analyze the relation between the network of authors associated with the collaborative writing of articles and the content value measured as article views. Their results based on social network analysis reveal a curvilinear relationship between the number of distinct contributors to user-generated content and viewership. They conclude that network effects are stronger for newer articles. Gorbatai and Piskorski (2012) and Piskorski and Gorbatai (2010) also test hypotheses related to the author network underlying Wikipedia. They ask whether the density of their individual social networks is related to both norm violations of authors and the likelihood of their easy discouragement after deletions and reverts of their work.

Ransbotham and Kane (2011) analyze the duration until an article on Wikipedia is promoted to a featured article or demoted. They find that an article is most likely to be promoted if the average experience of authors is close to the mean. Articles written by relatively "young" and relatively "old" teams face a longer time span until they are promoted. Halatchliyski et al. (2010) analyze contributions of authors that contributed to articles in two related but different domains of knowledge. They find that the authors that are most central in the author network also contribute to integrating the two fields. Greenstein and Zhu (2012a and 2012b) investigate the language bias of articles and its evolution over time. Comparing articles in the English Wikipedia to two reference corpora taken from publications of the Democrat and the Republican party in the U.S. Congress, they find that an early bias of Wikipedia towards Democrat language has gradually disappeared over time. Yet, this erosion of the overall bias comes from new articles, which use Republican vocabulary, while articles which used to be biased appear to stay biased. Gorbatai (2011) employs data from Wikipedia to highlight how demand and supply can be aligned in the absence of market prices. She shows that "professional" editors of Wikipedia strongly react to (attempted) contributions of "unexperienced" users, as they are a sign of increased demand.

3 Data

3.1 Preparation of the Data and Definition of the Category Economics

We downloaded a full-text dump of the German Wikipedia from the Wikimedia toolserver. The data had to be parsed in order to construct the weekly history of the articles' content including the hyperlink network for the entire encyclopedia. From the resulting tables, we constructed the time-varying graph of the article network and computed the weekly measures of an article's network position, which lie at the heart of our analysis. We extracted more information about the articles, such as the number of authors who contributed up to a particular point in time or the existence of a section with literature references. Before computing those numbers we accounted for the revisions that were made by small programs, so-called "bots", which automatically make small formal changes to ensure that a consistent style is maintained throughout Wikipedia. We did not consider the revisions that were carried out by bots and we also excluded bots from the author count. In our analysis, we use data on 153 weeks between December 2007 and December 2010. While articles have been selected from one category, network measures account for links between these articles and the entire German Wikipedia.

Because of the scale of the data in the order of magnitude of terabytes, it would be

unthinkable to conduct the data analysis using only in-memory processing. We stored the data in a disk-based, relational database and queried the data using Database Supported Haskell (DSH) (Giorgidze et al. (2010)), a novel high-level language allowing for formulation and efficient execution of queries on nested and ordered collections of data. DSH queries are automatically translated into efficient lower-level query languages that the underling database system understands. For this study, we utilised DSH's capability of translating high-level queries on nested and ordered collections of data to efficient bundles of SQL queries. For comparison, we have formulated several DSH queries used for the Wikipedia data analysis directly in SQL and found that the equivalent DSH queries were more concise, easier to write and easier to maintain. This was mostly due to DSH's support for order, nesting, abstractions for query reuse and concise comprehension notation (Giorgidze et al. (2011)).

With this tool, we sampled all the articles belonging to the categories and subcategories of economics ("Wirtschaft" - which may mean both "economy" and the discipline of economics in German) from this relational database. The choice of articles sampled was based on Wikipedia's category tree. Even though the ordering is not purely hierarchical, articles that belong to a category are usually allocated among specific subcategories. The more general category is often not reported on the article page. Therefore we had to account also for subcategories if we wanted to ensure that our definition of a category is not too narrow. Consequently, to sample the pages belonging to economics, we extracted a list of the subcategories of that category and eliminated those which were too remotely related to economics. This procedure left us with a list of 380 subcategories. We then proceeded to identify all pages that were linked to one of the categories on the list during at least one week that lies within our period of observation, which resulted in a sample of roughly $19,000$ articles.

Sampling articles based on categories of content is an approach that is used in previous papers dealing with large content networks like Wikipedia (cf. Halatchliyski et al. (2010)). However, when evaluating the information from the network formed by directed hyperlinks between articles, we do not rely exclusively on the subset of articles that we sampled. While we compute the social network measures only for the articles inside the category (i.e., for roughly 10,000 nodes), we use the links from all pages in the entire network (i.e., more than one million nodes) to compute them. This is different from previous work, where network measures are often computed only on subnetworks, that means abstracting from the existence of all the other articles. We therefore consider it to be of methodological interest to see whether estimating the effect of the network position on such a reduced network leads to a big or a small error. Hence, we define the category network as the set of nodes that remain within the category economics and the global network as the one that is set up by the entire German Wikipedia.

As we are interested in the network position within the entire Wikipedia, we have to handle the large mass of more than a million articles. We compute the number of incoming links and, using the igraph library by Csardi and Nepusz (2006), the closeness centrality for each article at every week. We do this for both the network corresponding only to the pages in the category and for the global network of all articles in the German Wikipedia. It is important to note that we carry out the entire analysis using the *directed* network formed via incoming hyperlinks. These links are observed and edited on those pages from which they direct away, but considered in our analysis as features of the pages which they are pointing to. On the latter pages they are only visible when using a tool provided on the side bar.

Wikipedia collects all content about a topic on one single article and creates "redirect pages" for widely used synonyms that users might be looking for. These pages redirect users, who search for synonyms of the Wikipedia entry, almost silently to the main page. Before computing the network measures, we accounted for the existence of redirect pages, by counting a link to a redirect page also as a link to its target page.

3.2 The Anatomy of the Data Set

In the data set we find approximately $7,000$ articles that were inexistent at the beginning of our period of observation or ceased to exist before the end and are, hence, excluded from the analysis. Using network analysis we identify one large cluster within the category that can be reached via the directed network of incoming links. Following a typical classification that Capocci et al. (2006) apply to Wikipedia, we observe that these pages are either part of the one strongly connected component (set of pages mutually reachable via hyperlinks) or of the out-component (pages reachable from the strongly connected component) of the subnetwork formed by pages associated to the category of economics. We observe $7,635$ pages that are always part of this cluster, which we refer to as the "connected component in the category economics" (or just "connected" or "reachable" articles). The other pages could not always be reached via the category network. During the period of observation, $1,237$ of these pages received an incoming link from the connected component in the category economics, and thus became part of that component.

Consequently we use two data sets for our analysis. The first data set is a balanced panel observing the $7,635$ articles that remain in the connected component during 153 weeks. It contains in total $1,168,155$ observations.[4] In the second data set we use only those articles that get connected to the economics category during the period of observation. In total we count $1,237$ such pages and observing them weekly results in $203,031$ observations of this group. In this sample we discarded a small portion of articles

[4]In ongoing research we analyze articles that come to existence during the period of observation.

that are not only disconnected (in the sense of not linked to the major cluster in the network) from the economics category but also from the entire German Wikipedia at some point in time.

Table 1 provides summary statistics of our variables for the balanced panel of articles that are always reachable from the category.[5] The unit of observation is an article in a given week and we observe the network position of each article in terms of incoming hyperlinks. We observe the length of a page in bytes, how many authors it has and when it was created. One byte corresponds roughly to one letter. The median length is 3630 bytes and the median article was written by 16 authors. Our main centrality measures are indegree and closeness centrality. Both are calculated for the entire Wikipedia and for the articles belonging the category economics. The indegree is calculated as the number of direct links pointing to a page from the entire German Wikipedia and from the category the article was drawn from. Since articles from the category are also contained in the entire Wikipedia, we report the difference of the two indegrees. By sample construction, every page is connected to the category and hence receives at least one link from it. The median page has eleven links from Wikipedia, four of which are from the category. Articles usually belong to more than one category, but we do not observe these additional categories.[6] The distributions of the centrality variables show that for many articles half or more of the links come from economics. Consequently we consider that this category is central to the majority of the articles we observe. Maximal values of page length, the number of authors and indegree lie far above the $90th$ percentile.

The closeness centrality measures are based on the inverse average distance of one article to all other articles in the relevant network.[7] Again, the directed centrality measure is computed on both the network made up by pages in the category and the entire German Wikipedia. We observe in our data that the original closeness measures are mainly driven by the variations in the share of disconnected articles and in the network size over time (not reported). In order to abstract from these effects, we compute the relative closeness ranks for our balanced panel. This procedure may be useful in work on dynamic networks in general. In the econometric estimation, we use age and dummies for redirect pages and pages containing a literature section as control variables. The presence of a

[5]Since many distributions are strongly left-shaped while having a long right tail, we prefer tables with percentiles to a graphical illustration

[6]Except for the category sociology that we use for sensitivity analysis.

[7] Closeness centrality in terms of incoming links for an article i on a network containing N articles is defined as the inverse of the sum of shortest paths (geodesic distances) D_{ij} to that article multiplied by the maximal path length $N-1$. Articles j from which no path leads to i ($j \notin M$) are assigned the distance N, which exceeds the longest possible distance by one:

$$C_i = \frac{N-1}{\sum_{j \in M} D_{ij} + \sum_{j \notin M} N}.$$

literature section usually points to an article that draws extensively on scientific, literary or journalistic sources outside Wikipedia and therefore tends to be longer. The median age of articles is 217 weeks, that is roughly four years. Only around ten percent of the articles are less than two years old, so the majority of articles in our sample are mature articles.

Table 2 shows the same summary statistics as Table 1, but for the sample of articles that get connected to the category of economics during the period of observation. We consider the sample over the entire 153 weeks. In the beginning, none of the articles can be reached from the main component but all become connected later on. The page length and the number of authors are generally a bit smaller, but otherwise show a rather similar distribution, except for the 90th percentile and the maximum. The median page length of $3,044$ bytes is about 600 bytes shorter than the median page length of articles that are always part of the connected component. The number of links within the category is smaller by sample design, since most of the articles are disconnected from the main component of the category for many weeks. The number of links from outside the category is similar in median in both samples but considerably smaller in the upper percentiles of the sample of articles that are initially disconnected. We do not report the closeness in this sample because it is mainly driven by the fact of being connected or disconnected. The articles are a bit younger than in the main sample, but the median age still lies far above three years.

Figure 1 shows the development of median values of page length, the number of authors and indegree over the 153 weeks observed. The figure documents the growth that articles experience over time and hence the need to control for time effects in our estimation.

To see how often the variables typically change for individual pages, we aggregate the frequency of changes in the network and content variables over time. This is shown in Table 3, where the unit of observation is a page observed throughout the 153 weeks and the table displays the number of changes in variables. The changes are reported for our main sample of articles that are always reachable from the large component of the category. Less than 25 percent of the pages never experience any change in their number of incoming links and less than ten percent are never edited nor receive any additional author. At the same time we see that most articles do not change in any given period, since the frequency of changes of 90 percent of the articles lies at or below 15 to 36 out of 153. An exception are the closeness measures, which change nearly every week for every page. They depend on the structure of the entire network, which is subject to almost permanent change, especially when the entire German Wikipedia is being considered.

Finally, Table 4 displays the magnitude of changes for all observations with non-zero change. The reason not to keep the balanced panel here is to make the distribution of changes more visible, which is otherwise dominated by zeros. The median change in page

length is 18 bytes in a week, which corresponds to about two words. This highlights that small changes are frequent in the work that many authors contribute to Wikipedia in order to improve the quality of articles. The 75th and the 90th percentile lie at 70 and 309 bytes, which corresponds to a short sentence and a very short paragraph. The median and also most frequent change in incoming links per week is equal to one. The maximal values of changes in page length and links seem to correspond to reverts of entire articles and lie far above the 99th percentile. Changes in closeness are quite symmetrically distributed around zero, which is not surprising, since we use a relative closeness measure. As much as 80 percent of the changes amount to an increase of far less than one point (of 100) in the relative closeness position per week. The distribution of changes is important for interpreting the strength of the effects obtained in our regressions.

4 Relationships of Interest and Methodology

4.1 Network Position and User-Generated Content

We are interested in analyzing whether a higher centrality in the article network is associated with (i) more content being generated and (ii) contributions by new rather than by previous authors of a page. Our main explanatory variables are measures of centrality in the network of incoming hyperlinks. As described in the previous section, we have four centrality measures: the number of incoming links within the category economics (indegree within category) and from the entire German Wikipedia (global indegree) as well as the closeness rank in the network of the category and in the global network. As further control variables we add dummies for an article being a redirect, for the presence of a literature section and for article age. We assume that the relation between outcomes and indegrees may be linear or quadratic while the other variable enter our estimation only in a linear way. Data from Wikipedia pages are generated inside two network contexts, the authors network, analyzed in several previous studies, and the hyperlink network formed by the pages, which we are investigating. The skewness and the long tails in the distributions of the number of incoming links, the page length and the number of authors underline that the data show similar properties as other network data. Like with almost all dynamic network data, at least three sources of endogeneity play a role in potentially affecting our estimates.

Firstly, articles differ substantially in their relevance to the wider audience and in other unobserved dimensions. Particularly the difference in their relevance is likely to affect both the network position and the content generation in the same direction, thus generating correlation between these two variables. Secondly, Wikipedia is a collaborative site where the content matter of certain pages is subject to unobserved exogeneous shocks

and seasonalities. Sudden spikes of interest in certain issues might lead to more authors contributing to single pages or to the entire platform. Moreover, since contributions to Wikipedia continuously grow and inevitably generate some hyperlinks, page length and hyperlinks may both have a time trend. The third source of endogeneity stems from editors who simultaneously edit page B and set a link from page A to page B. Such activity will also lead to a correlation between the network position of a page and its content, but the author's attention will not have been attracted to editing page B via the link from page A. Note that measuring the position of articles based on a two-mode author-article network suffers from similar problems.

Like Kittur and Kraut (2008) and Ransbotham et al. (2012) we use the temporal structure of the data to track the variation within one and the same article by using article fixed effects. Moreover the data are rich enough to allow controlling for systematic temporal variation or particularities of singular weeks by employing time fixed effects. We estimate two-way fixed effects panel regressions based on the following equations:

$$(1) \qquad (page\ length)_{it} = \alpha_i + \alpha_t + \beta * (centrality_{it}) + \gamma * X_{it} + \epsilon_{it}$$

$$(2) \qquad (num.\ authors)_{it} = \alpha_i + \alpha_t + \beta * (centrality_{it}) + \gamma * X_{it} + \epsilon_{it}$$

where $centrality_{it}$ is a vector of the four centrality measures mentioned above. X_{it} includes the three control variables indicating redirects, literature sections and age (weeks since the first edit), i designates the article and t the week.

Since the data allow observing an article's network position in a panel design, we can effectively tackle the first two sources of endogeneity considered, which are constant heterogeneity specific to articles and time trends or time-dependent shocks that affect the entire network.

Tackling the third source of endogeneity, reverse causality from content to links, is more difficult in our data of connected articles as it cannot be dealt with by fixed effects alone. However, we can make use of a special type of pages in order to shed more light on the effect of network position on content provision. These are the articles that are initially disconnected from the large economics cluster. In order to understand why looking at these articles may be useful, note that authors in general do not observe whether an article is connected to a large component or not. Experienced users may look at the option that allows to display the direct links pointing to a page. Yet, users will not necessarily employ it when linking from another page and, more importantly, they will not see how the linking articles themselves are connected. Most authors will thus not consciously decide to link an article from a large cluster of several thousand articles from which it was previously

not accessible. The length of the page may influence the creation of links towards this page. But we expect that there is no systematic relation between page length and whether new links come from outside the category, from isolated pages within the category (which leaves the article disconnected from the cluster economics) or from the main cluster of the category. If we find an effect of getting connected to the large cluster of the category economics that is strong and lasting compared to the coefficients of the indegrees found in the sample of always connected articles, we consider that it plausibly results from the sudden sharp increase in connectedness. This sharp increase is reflected in a discontinuity in the closeness centrality.

4.2 Getting Connected to the Category of Economics

In order to analyze the effect of becoming part of the connected component in the category of economics, we put together a sample that includes articles that are at first not connected, but become connected from the category at some point during our period of observation. There are in total $1,237$ of these articles. Since the change in closeness centrality is very similar for all of them, we just consider a dummy for becoming connected. We do not consider additional changes in indegree, since we know that most articles change by one link at maximum in a given week and do not change in most weeks. Therefore accounting for getting connected and indegree simultaneously may result in overcontrolling. We analyze both the length and the rate of change of a page from five weeks before the page becomes connected until five weeks after. In a few cases we observe that a page was connected more than once. In those cases we consider only the last time when the page is connected in our sample.

For the eleven weeks in the sample, we regress page length on an indicator variable that takes the value of one if the page can be reached via the links from the main component of economics and zero otherwise. This means it takes the value zero in the five weeks before connection and the value one in the week when connection occurs as well as in the five weeks after. Furthermore we regress the first difference of page length over time on the same indicator variable. The two-way fixed effects regressions thus take the form:

$$(page\ length)_{it} = \alpha_i + \alpha_t + \beta * \iota(page\ connected)_{it} + \epsilon_{it} \tag{3}$$

$$\Delta(page\ length)_{it} = \alpha_i + \alpha_t + \beta * \iota(page\ connected)_{it} + \epsilon_{it} \tag{4}$$

with $t = 0$ at the period of the jump into the category and $t \in \{-5, ..., 5\}$.
In order to alleviate the concern that becoming connected is rather the effect than the cause of simultaneous editing of the target page and the pages pointing to it around week

0, we compare weeks -7 to -3 with weeks 3 to 7 in a further specification (reported in Table 9). While our approach reduces the vulnerability to simultaneity issues in important aspects, fully disentangling the factors that might drive simultaneity would require exogenous instruments or the ability to explicitly account for the identity of the linking articles and their properties, which we believe to be a fruitful avenue for further research.

5 Results

Table 5 shows the two-way fixed effects regressions corresponding to equation 1, where page length is regressed on several sets of network variables, article fixed effects and time fixed effects.[8] The table shows the result for $7,635$ articles from the category economics that belong to the large cluster in that category throughout the entire 153 weeks. The first column shows the coefficients for the number of links that the page receives from the entire Wikipedia and a squared term. Our estimates indicate that an additional link pointing to a page is associated with 13 more bytes of text. This corresponds to one or two words. The insignificant coefficient on the quadratic term indicates no curvature. A main question of our investigation is whether the effect of links from the category is different from the mean effect of all links. In the second column we add the number of links that the page receives from other pages of the category economics. These links represent a subset of the global links. The effect can be interpreted as the additional effect from a link being a category link. The coefficient for a category link is more than ten times higher than the coefficient obtained when not differentiating between the two groups of links. Moreover, the new variables render the coefficient for a link that comes from outside the category small and insignificant, suggesting that the explanatory power mostly stems from the category network. Since we run regressions with article fixed effects, the coefficients apply to deviations from the averages that are specific to the article. If the number of incoming links from the category exceeds this average by one, the target page is by 141 bytes longer (considering the sum of the two linear coefficients). For links from the category we estimate significant declining effects, with the coefficient for the quadratic term taking, however, a rather low value of -0.13.

Column 3 and 4 add the relative closeness rank, which measures whether a page is located rather in the center of the network or rather in its periphery. Column 3 shows the specification of column 1 augmented with the relative rank in closeness on the entire Wikipedia. Given that we scaled the rank variable such that it ranges from 0 to 100, the coefficient indicates that a ten points improvement in the relative closeness position is associated with 150 additional bytes of content. In the descriptive statistics we saw that

[8] Time fixed effects were implemented manually by adding a dummy for each point in time in the regression.

the closeness of most articles changes by less than one point in any given week. From this point of view the effect looks small. Moreover, the size of the coefficient for indegree is barely affected and the added explanatory power of the new variable is rather low. Finally, Column 4 brings together all available network variables, including the measure of the closeness rank both on Wikipedia and inside the category. The coefficient of the closeness rank inside the category is insignificant and the coefficient of the closeness rank on the entire German Wikipedia is even smaller than in column 3. The coefficient of the number of links from the category remains very close to its value in column 2. The control dummies for redirects and a literature section have the expected signs. Older articles tend to be longer.

In Table 7, we report four robustness checks of the last result: In the first column, we replace the contemporaneous measures of centrality by the ones from the week before, which cannot be influenced by current editing behavior. This tests whether our result is mainly driven by a reverse effect of content generation on incoming links created in the same week. We find virtually no change in the results and thus consider this effect not to be important. In the second column, we eliminate outliers from the sample. We observe two kinds of outliers visible in Tables 1 and 4: articles that gain a lot of attention in the form of long contributions, many authors and many links (both from the entire Wikipedia and within the category), and articles that experience very high changes in these variables in at least some periods. We compute maxima of levels and changes per article. We eliminate articles that lie in the extreme two percent for any maximal change. Of the remaining articles, we eliminate those lying above the 95th percentile of the maximal levels of any variable. In total this eliminates 15 percent of the articles. The results show that both indegrees are estimated to have even larger coefficients, which sum to 222 bytes for a link from within the category. The quadratic specification now better captures a positive but declining influence of links from outside the category. In the third column, we perform the estimation for a different category, sociology, excluding those articles that overlap with economics. As in our main sample, links from within the category have a much stronger effect on page length (roughly three times as large as a link from outside the category). However, this coefficient is significant only at the ten percent level, which may be a consequence of the smaller sample size. The coefficient for links from the entire Wikipedia is now significant, which was not the case for economics. Column 4 finally reports how the results change when including a proxy for how often a page was clicked in the last week.[9] Not surprisingly, as articles are clicked more they get longer. However, the relationship of an article's network position and its length remains unaffected by the inclusion of this variable.[10]

[9]We measured the clicks in the 24 hours before the next due date in our weekly panel.

[10]We performed further robustness checks that did not affect the main conclusions. We excluded pages that merely redirected the reader to a different page and explanation pages. We also included a

Now we turn to the question whether the higher centrality is not only associated with more content but also with more authors. Table 6 shows the two-way fixed effects regressions corresponding to equation 2. It mirrors the specifications from Table 5, but the regressions have now the number of authors as the dependent variable. Columns 1 and 3 show the results when using the centrality measures from the entire Wikipedia. The results indicate that an additional link is associated with roughly 0.11 more authors, with a very weak curvature of the slope. Similarly to our results for page length, the effect is much stronger for links from the category: an additional link from the category corresponds to approximately 0.54 more authors (considering the sum of Wikipedia and category coefficients). The coefficient for links from outside the category is much smaller but remains significant in all specifications. The closeness rank has negligible effect in column 3, which turns insignificant in column 4.

Summing up our results for the connected component in the category of economics, we find that a higher number of links from articles in the same category is associated with more content generation and additional authors. The increase in page length related to an additional link from the category may look small since it corresponds to a short sentence. From the descriptive statistics we saw, however, that small changes are an essential ingredient of the development of Wikipedia. Consequently we consider the effect as non-negligible. The effect of links from outside the category is insignificant in our main specification and significant but about three times smaller in some robustness checks. The effect of closeness centrality is negligible.

The regressions in Tables 8 and 9 use the information on the $1,237$ pages that get connected to the main cluster of economics during the period of observation. This is associated with a discontinuous jump in closeness centrality at the time of connection, which can be identified and used to contrast the level (and the growth) of the content before and after this event. Table 8 shows the results when we consider 5 periods before and after the jump including also the period of the jump itself. The first two columns show the results from a simple pooled OLS regression, whereas columns 3 and 4 show the two-way fixed effects results when including both time and article dummies. The coefficients affecting the level of the page length (column 1 and 3) indicate that getting connected is associated with an increase in page length by approximately 400 bytes. This effect is both significant and sizeable compared to the effect of one additional link in the previous sample. The explanatory power of the regression is, however, very low. The cumulative effect over five weeks is even stronger for the first differences of page length

measure of how often pages linking to the page under consideration were viewed. Next we included several other (potentially endogenous) measures that better describe the pages (number of revisions, number of references). We repeated everything for authors, where some effects are somewhat reduced, but they are always in the same direction and continue to matter. The results are available from the authors upon request.

(columns 2 and 4), ranging from 66 bytes per week in the pooled regression to 195 bytes per week when including time and article fixed effects. These are sizeable effects which cannot be expected to last forever. It might be that a share of the additional content is provided in the same week as the article gets connected.

In Table 9 we account for that possibility, by excluding the week of the "jump" into the connected component and the two weeks before and after. Instead we consider two five-week intervals that are separated by the interval two weeks before and after the jump (i.e., week -7 to -3 vs. week 3 to 7). As expected, the coefficients get smaller, which indicates that a substantial fraction of the newly generated content is provided within the weeks after the new connection was established. However, the effects remain by and large positive and indicate that an article grows by 9 (pooled) to 21 bytes per week (fixed effects) faster during weeks 3 to 7 after being connected. We still observe not only a level but also a growth effect.

6 Conclusion

The creation of user-generated content in a peer production setting requires mechanisms that help producers to identify content they want to contribute to. We consider the network of hyperlinks between Wikipedia articles as a possible channel of spillovers that attracts more producer effort to more central articles. We find that the page length of an article is positively associated with the number of links pointing to it after controlling for time-invariant unobserved heterogeneity, time effects and several other variables.

On average, one more link is associated with a page length that is 13 bytes higher, which corresponds roughly to one or two words. When differentiating between links within the category economics, which we selected as sample, and links from other Wikipedia pages, we find a large discrepancy in effects. One more link from an article from inside the category is related to a increase in page length of around 140 bytes. This is a sizeable effect given that the median weekly change in page length excluding observations without any change is only 18 bytes. At the same time, the coefficients for links from outside the category becomes insignificant. The importance of links from the same category is corroborated in several robustness checks which persistently confirm that the effect of links from outside the category is much smaller. Moreover, links from the category are strongly related to new authors' contributions. On average every second additional link from the category is associated with a new author contributing to the page. These results are all obtained in a balanced sample of articles that are always connected to the large cluster of the category. Articles that are initially not connected increase by more than 300 bytes in length during the five weeks after connection.

Taken together the results suggest that adding missing hyperlinks to Wikipedia or

extending the content of articles in a way that it connects better to other articles may not only improve the quality of the information but also foster further contribution by authors that have not yet contributed to the newly linked articles. While the size of the additional contributions that may be expected is not very high, these changes of a few words or one sentence constitute a large part of contributions to Wikipedia. This strategy is expected to work best within a cluster of thematically related articles. Links from articles that do not share a central category with the target article seem to enhance content generation much less. Thus we find new evidence that semantical relatedness may matter more than the mere presence of direct links between pages in generating spillovers in content provision.

From a researcher's perspective, our results suggest that it may be an acceptable strategy in the context of content networks to use only a smaller group of articles (nodes) for network computations, which share a common category, as long as one does not extrapolate the result to the unobserved nodes. This should not be said without adding a word of caution: First, our results are not based on a two-mode author-article network considered in several other studies but on the link network of Wikipedia articles. Whether they extend to two-mode contexts remains to be tested. Second, our conclusions are obtained based on data from relatively mature articles and should be reexamined for newly created articles.

References

Adafre, S. F and M. de Rikjke, "Discovering missing links in Wikipedia," in "Proceedings of the 3rd International Workshop on Link Discovery" 2005, pp. 90–97.

Albert, R., H. Jeong, and A.L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, 1999, *401* (6749), 130–131.

Audretsch, D.B. and M.P. Feldman, "R&D spillovers and the geography of innovation and production," *The American Economic Review*, 1996, *86* (3), 630–640.

Capocci, A., V.D.P. Servedio, F. Colaiori, L.S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia," *Physical Review E*, 2006, *74* (3), 036116.

Claussen, J., O. Falck, and T. Grohsjean, "The strength of direct ties: Evidence from the electronic game industry," *International Journal of Industrial Organization*, 2012, *30* (2), 223–230.

Csardi, G. and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems*, 2006, *1695.*

de Solla Price, D.J., "Networks of scientific papers," *Science*, 1965, *149* (3683), 510.

Denning, P., J. Horning, D. Parnas, and L. Weinstein, "Wikipedia risks," *Communications of the ACM*, 2005, *48* (12).

Fershtman, C. and N. Gandal, "Direct and Indirect Knowledge Spillovers: The 'Social Network' of Open Source Software," *RAND Journal of Economics*, 2011, *42* (1).

Giorgidze, G., T. Grust, N. Schweinsberg, and J. Weijers, "Bringing Back Monad Comprehensions," in "Proceedings of the 4th ACM SIGPLAN Haskell Symposium, Tokyo, Japan" ACM ACM 2011, pp. 13–22.

_ , _ , T. Schreiber, and J. Weijers, "Haskell Boards the Ferry: Database-Supported Program Execution for Haskell," in "Revised selected papers of the 22nd international symposium on Implementation and Application of Functional Languages, Alphen aan den Rijn, Netherlands," Vol. 6647 of *Lecture Notes in Computer Science* Springer 2010. Peter Landin Prize for the best paper at IFL 2010.

Gorbatai, A., "Aligning Collective Production with Demand: Evidence from Wikipedia," *Working Paper*, 2011.

Gorbatai, A.D. and M. Piskorski, "Social Structure of Contributions to Wikipedia," *Working Paper*, 2012, *downloaded from http://www.wjh.harvard.edu/ hos/papers/AndreeaGorbattai/AndreeaGorbattai.pdf.*

Goyal, S., M.J. Van Der Leij, and J.L. Moraga-González, "Economics: An emerging small world," *Journal of Political Economy*, 2006, *114* (2), 403–412.

Greenstein, S. and F. Zhu, "Collective Intelligence and Neutral Point of View: The Case of Wikipedia," *Working Paper*, 2012.

_ **and** _ , "Is Wikipedia biased," in "American Economic Review, Papers and Proceedings" 2012.

_ **and M. Devereux**, "Wikipedia in the Spotlight," Technical Report 5-306-507, Kellogg School of Management 2009.

Griliches, Z., "The Search for R&D Spillovers," *Scand. J. of Economics*, 1992, *94*, 29–47.

Halatchliyski, I., J. Moskaliuk, J. Kimmerle, and U. Cress, "Who integrates the networks of knowledge in Wikipedia?," in "Proceedings of the 6th International Symposium on Wikis and Open Collaboration" ACM 2010, p. 1.

Jackson, M.O., *Social and economic networks*, Princeton Univ Pr, 2008.

Jian, L. and J. MacKie-Mason, "Incentive-Centered Design for User-Contributed Content," in M. Peitz and J. Waldfogel, eds., *The Oxford Handbook of the Digital Economy*, Oxford University Press Oxford 2012, pp. 399–433.

Kittur, Aniket and Robert E. Kraut, "Harnessing the wisdom of crowds in wikipedia: quality through coordination," in "Proceedings of the 2008 ACM conference on Computer supported cooperative work" CSCW '08 ACM New York, NY, USA 2008, pp. 37–46.

Kriplean, T., I. Beschastnikh, and D.W. McDonald, "Articulations of wikiwork: uncovering valued work in wikipedia through barnstars," in "Proceedings of the ACM 2008 conference on Computer supported cooperative work" 2008.

Lerner, J. and J. Tirole, "Some Simple Economics of Open Source," *Journal of Industrial Economics*, 2002, pp. 197–234.

Medelyan, O., D. Milne, C. Legg, and I. H. Witten, "Mining meaning from Wikipedia," *International Journal of Human-Computer Studies*, 2009, *67* (9), 716–754.

Piskorski, M.J. and A. Gorbatai, "Testing Coleman's Social-norm Enforcement Mechanism: Evidence from Wikipedia," *Working Paper*, 2010.

Priedhorsky, R., J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, Destroying and Restoring Value in Wikipedia," *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, 2007, pp. 259–268.

Ransbotham, S. and G. Kane, "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia," *MIS Quarterly*, 2011, *35* (3), 613–627.

_ , G.C. Kane, and N. Lurie, "Network Characteristics and the Value of Collaborative User-Generated Content," *Marketing Science*, 2012, *31*, 387–405.

Romer, P.M., "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98, Number 5 (2)* (5), 71–102.

Soto, J., "Wikipedia: A quantitative analysis." PhD dissertation 2009.

Zhang, X. and F. Zhu, "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *The American Economic Review*, 2011, *101*, 1601–1615.

7 Tables

7.1 Summary Statistics

Table 1: Summary statistics of main variables. Connected articles.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	20	1049	1872	3630	7470	14089	229379
Number of authors	1	6	9	16	30	56	821
Links from Wikipedia	1	2	5	11	28	76	7981
Links from Wikipedia excl. categ.	0	0	2	6	17	53	7750
Links from category	1	1	2	4	10	23	667
Rel. closeness rank (Wikipedia)	.013	10	25	50	75	90	100
Rel. closeness rank (category)	.013	10	25	50	75	90	100
Dummy: literature section	0	0	0	0	0	1	1
Dummy: page is redirect	0	0	0	0	0	0	1
Age (in months)	1	113	162	217	271	316	492

Articles that were always connected to econ. main component. Number of observations: 1168155

Table 2: Summary statistics of main variables. Articles that get connected to category.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	19	915	1653	3044	5207	9231	67988
Number of authors	1	5	8	12	20	33	267
Links from Wikipedia	1	2	4	7	13	24	3914
Links from Wikipedia excl. categ.	0	1	2	5	10	21	3910
Links from category	0	0	1	1	2	4	122
Dummy: literature section	0	0	0	0	0	1	1
Dummy: page is redirect	0	0	0	0	0	0	1
Age (in months)	1	84	129	181	236	283	451

Number of observations included: 203031.

Table 3: Summary statistics of the frequency of changes of main variables.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	0	3	5	11	22	36	136
Number of authors	0	2	4	7	14	24	123
Links from Wikiped (excl. categ.)	0	0	1	4	12	34	152
Links from categ.	0	0	1	3	7	15	121
Rel. closeness rank (Wikipedia)	152	152	152	152	152	152	152
Rel. closeness rank (categ.)	149	151	152	152	152	152	152

The unit of observation is a page over entire period. Number of pages included: 7635

Table 4: Weekly changes of main variables.

	Min	p1	p10	p25	p50	p75	p90	p99	Max	Obs.
Length of page (in bytes)	-95,222	-868	-42	-1	18	70	309	2,739	83,235	124,771
Number of authors	1	1	1	1	1	1	2	3	76	82,260
Links from Wikipedia	-439	-2	-1	1	1	1	2	8	1,455	121,589
Links from Wikiped (excl. categ.)	-439	-2	-1	1	1	1	2	9	1,455	90,214
Links from category	-130	-2	-1	1	1	1	1	4	80	46,304
Rel. closeness rank (Wikipedia)	-92	-1.6	-0.58	-0.23	-0.02	0.18	0.51	1.8	91	1,137,528
Rel. closeness rank (category)	-99	-0.98	-0.20	-0.11	-0.04	0.03	0.13	2	85	1,090,973

Articles that were always connected to econ. main component. Only observations with non-zero changes.

7.2 Regression Results

Table 5: Relationship of page length and centrality.

	(1) Wiki degree	(2) Wiki & cat.	(3) add closeness	(4) all vars
Links from Wikipedia	13.333***	2.958	12.934***	2.931
	(3.18)	(1.22)	(3.14)	(1.22)
(Links from Wikipedia)^2	-0.000	0.001**	-0.000	0.001**
	(-0.54)	(2.04)	(-0.47)	(2.07)
Links from category		138.129***		135.871***
		(8.80)		(8.47)
(Links from category)^2		-0.130***		-0.127***
		(-5.24)		(-5.02)
Rel. closeness rank (Wikipedia)			15.216***	7.505***
			(6.17)	(3.08)
Rel. closeness rank (category)				-1.230
				(-0.67)
Dummy: literature section	1295.963***	1249.985***	1287.521***	1248.055***
	(6.11)	(5.95)	(6.07)	(5.94)
Age (in months)	10.648***	8.361***	10.692***	8.416***
	(21.55)	(22.76)	(21.85)	(22.46)
Dummy: page is redirect	-546.408	-742.157	-590.851	-767.075
	(-0.57)	(-0.77)	(-0.59)	(-0.77)
Constant	3336.571***	2803.789***	2582.005***	2501.686***
	(30.10)	(22.18)	(16.28)	(15.73)
Time dummies	Yes	Yes	Yes	Yes
Observations	1168155	1168155	1168155	1168155
Groups	7635	7635	7635	7635
Adj. R^2	0.107	0.130	0.109	0.131

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included. Dependent variable: page length.

* p<0.10, ** p<0.05, *** p<0.01

Table 6: Relationship of number of authors and centrality.

	(1) Wiki degree	(2) Wiki & cat.	(3) add closeness	(4) all vars
Links from Wikipedia	0.112***	0.073***	0.111***	0.072***
	(4.25)	(3.24)	(4.23)	(3.23)
(Links from Wikipedia)^2	-0.000**	-0.000**	-0.000**	-0.000**
	(-2.51)	(-2.05)	(-2.50)	(-2.04)
Links from category		0.468***		0.476***
		(6.39)		(6.38)
(Links from category)^2		-0.000***		-0.000***
		(-3.06)		(-3.18)
Rel. closeness rank (Wikipedia)			0.017**	-0.007
			(2.29)	(-1.22)
Rel. closeness rank (category)				-0.009*
				(-1.65)
Dummy: literature section	1.552***	1.393***	1.543***	1.406***
	(4.78)	(4.53)	(4.76)	(4.57)
Age (in months)	0.072***	0.064***	0.072***	0.064***
	(26.10)	(44.22)	(26.07)	(43.66)
Dummy: page is redirect	0.269	-0.399	0.220	-0.434
	(0.13)	(-0.19)	(0.10)	(-0.20)
Constant	6.127***	4.376***	5.291***	5.140***
	(13.05)	(11.30)	(13.73)	(13.00)
Time dummies	Yes	Yes	Yes	Yes
Observations	1168155	1168155	1168155	1168155
Groups	7635	7635	7635	7635
Adj. R^2	0.463	0.495	0.463	0.495

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included. Dependent variable: number of authors.

* p<0.10, ** p<0.05, *** p<0.01

Table 7: Robustness checks for the relationship of page length and centrality.

	(1) Lagged cent.	(2) Excl. outliers	(3) Sociology	(4) Add clicks
Links from Wikipedia	2.946	62.605***	31.015***	2.940
	(1.22)	(8.01)	(4.19)	(1.22)
(Links from Wikipedia)^2	0.001**	-0.264***	-0.011***	0.001**
	(2.04)	(-5.07)	(-7.50)	(2.07)
Links from category	134.937***	159.688***	59.778*	135.463***
	(8.42)	(7.07)	(1.71)	(8.46)
(Links from category)^2	-0.125***	-1.230**	-0.104***	-0.126***
	(-4.95)	(-2.14)	(-2.74)	(-5.03)
Rel. closeness rank (Wikipedia)	7.505***	1.818	10.421	7.473***
	(3.10)	(1.23)	(1.28)	(3.07)
Rel. closeness rank (category)	-1.182	-4.608***	-8.406*	-1.205
	(-0.65)	(-3.86)	(-1.96)	(-0.66)
Dummy: literature section	1248.652***	1002.577***	338.438	1247.455***
	(5.92)	(9.60)	(0.77)	(5.93)
Age (in months)	8.396***	3.576***	11.154***	8.502***
	(22.30)	(17.61)	(9.62)	(22.56)
Dummy: page is redirect	-718.429	110.313	0.853	-771.635
	(-0.74)	(0.39)	(0.00)	(-0.77)
Clicks				0.233**
				(2.38)
Constant	2516.272***	1933.826***	3633.877***	2481.048***
	(15.89)	(21.14)	(6.95)	(15.47)
Observations	1160520	994041	195381	1168155
Groups	7635	6497	1277	7635
Adj. R^2	0.130	0.227	0.095	0.131

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included.

* p<0.10, ** p<0.05, *** p<0.01

Table 8: Relationship of the growth of page length and the page becoming connected.

	(1) OLS Levels	(2) OLS Differences	(3) 2-Way FE Levels	(4) 2-Way FE Differences
Dummy: page is connected to cat.	439.133***	66.343***	317.699***	194.809***
	(5.49)	(6.06)	(5.72)	(5.48)
Constant	4059.235***	10.458***	2584.101***	-2056.589***
	(70.60)	(4.10)	(6.22)	(-4.76)
Time dummies	No	No	Yes	Yes
Observations	14376	14324	14376	14324
Groups			1327	1327
Adj. R^2	0.002	0.002	0.037	0.007

t statistics in parentheses

Columns 1 and 2 show pooled OLS-Regressions, Columns 3 and 4 include article and time fixed effects.

All Regressions use heteroscedasticity-robust standard errors. Dependent Variable: page length.

* p<0.10, ** p<0.05, *** p<0.01

Table 9: Relationship of the growth of page length and the page becoming connected, excluding the period of the jump itself and the 2 periods before and after.

	(1) OLS Levels	(2) OLS Differences	(3) 2-Way FE Levels	(4) 2-Way FE Differences
Dummy: page is connected to cat.	369.197***	8.650**	255.683***	21.334**
	(4.38)	(2.17)	(3.61)	(2.02)
Constant	4049.740***	7.293***	3654.610***	-116.975
	(69.20)	(4.50)	(12.47)	(-1.30)
Time dummies	No	No	Yes	Yes
Observations	12283	12237	12283	12237
Groups			1268	1268
Adj. R^2	0.001	0.000	0.042	0.002

t statistics in parentheses

Columns 1 and 2 show ooled OLS-Regressions, Columns 3 and 4 include article and time fixed effects.

All Regressions use heteroscedasticity-robust standard errors. Dependent Variable: page length.

* p<0.10, ** p<0.05, *** p<0.01

7.3 Figures

7.3.1 Development of variables over time

Figure 1: The figure shows the development of the median of the outcomes and the indegree in both groups, the always connected sample and the smaller jumping sample.