ECONSTOR

WWW.ECONSTOR.EU

Der Open-Access-Publikationsserver der ZBW – Leibniz-Informationszentrum Wirtschaft
The Open Access Publication Server of the ZBW – Leibniz Information Centre for Economics

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

Working Paper

Centrality and content creation in networks: The case of German Wikipedia

ZEW Discussion Papers, No. 12-053

Provided in Cooperation with:
ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research

Suggested Citation: Kummer, Michael E.; Saam, Marianne; Halatchliyski, Iassen; Giorgidze, George (2012) : Centrality and content creation in networks: The case of German Wikipedia, ZEW Discussion Papers, No. 12-053, http://hdl.handle.net/10419/66112

Nutzungsbedingungen:
Die ZBW räumt Ihnen als Nutzerin/Nutzer das unentgeltliche, räumlich unbeschränkte und zeitlich auf die Dauer des Schutzrechts beschränkte einfache Recht ein, das ausgewählte Werk im Rahmen der unter
→ http://www.econstor.eu/dspace/Nutzungsbedingungen
nachzulesenden vollständigen Nutzungsbedingungen zu vervielfältigen, mit denen die Nutzerin/der Nutzer sich durch die erste Nutzung einverstanden erklärt.

Terms of use:
The ZBW grants you, the user, the non-exclusive right to use the selected work free of charge, territorially unrestricted and within the time limit of the term of the property rights according to the terms specified at
→ *http://www.econstor.eu/dspace/Nutzungsbedingungen*
By the first use of the selected work the user agrees and declares to comply with these terms of use.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der Leibniz-Gemeinschaft

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

Download this ZEW Discussion Paper from our ftp server:

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

Die Discussion Papers dienen einer möglichst schnellen Verbreitung von neueren Forschungsarbeiten des ZEW. Die Beiträge liegen in alleiniger Verantwortung der Autoren und stellen nicht notwendigerweise die Meinung des ZEW dar.

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.

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 ("Wirtschaft" - "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 “Wirtschaft” 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 “Wirtschaft”, 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

August 14, 2012

Abstract

When contributing content on large online platforms, producers of user-generated content have to 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. 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 German Wikipedia over 153 weeks and 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 nearly 0.5. Moreover we observe that the rate of content generation increases notably when previously unlinked articles get connected to the main cluster in the category.

JEL-Classification: L14, D83

Keywords: Wikipedia; network position; user-generated content; hyperlinks.

*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. 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 and the participants of the ICT Conference 2012 at ZEW in Mannheim. Burak Tuerkoglu, Sergiy Golovin, Benedikt Achatz and Fabian Trottner provided helpful research assistance. Hans-Martin von Gaudecker provided programming advice to ensure the reproducibility of this research. George Giorgidze, Michael Kummer and Marianne Saam 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 channel 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. By and large, administrators who wish to start or maintain a wiki faces three problems. First, they have to succeed in motivating potential users to give it a try. Second, users have to like what they find on the platform, so that they can connect, come back, and eventually contribute to it. Third, users have to contribute content that is useful to others, so that new users have something to connect and come back to. Particularly the third step is critical, and at the same time it can be very challenging to achieve. Not only must the content be good and trolls be discouraged (cf. Jian and MacKie-Mason (2012)), but contributing must also be fun and a credible leadership is needed to prevent the project from forking (cf. Lerner and Tirole (2002)). Non-voluntary organizations might be able to overcome this problem by mildly forcing their members (e.g. employees or students). By doing so, they can directly influence where and in which ways users participate and contribute. However, in voluntary organizations and on the open web, users are free like a flock of birds, and there does not seem to be a way of telling anybody *what* to do without the risk of scaring them away. This is even more true on big platforms, where the users are numerous, their contributions are often spontaneous and the content is vast.

In this paper, we study user-generated articles on German Wikipedia and the network that is formed by hyperlinks between them. We analyze where users decide to provide content on a platform characterized by the feature that many articles need to be written or improved. In particular, we analyze how the position of an article in the network of articles 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. This question is situated in the more general context of understanding how producers in peer production of information goods select their tasks.

Since generating content on large platforms is highly complex, readers as well as authors take advantage of organizing mechanisms when identifying articles of interest. There are three main possibilities to find articles on Wikipedia: 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 meaning between the two subjects. The links need not to be reciprocal. The main guidelines on German Wikipedia say that an article must be readable without information from the linked pages. Within Wikipedia, links should point only to pages on technical terms or to pages that contain further information on topics that might be of particular interest to readers of the article.[1] It is not compatible with Wikipedia's rules to set links just to attract attention to an article without embedding its subject into the text pointing to it.

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 e.g. Medelyan et al. (2009)). Adafre and de Rikjke (2005) propose a procedure that automatically detects missing links between 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 (1) contributions to an article and (2) contributions by new authors.

In the context of economic research on production of information goods, we consider centrality in the network of articles as a possible channel of knowledge spillovers. Links may trigger the contribution of knowledge that might not have been contributed in their absence. In line with the vast literature on knowledge spillovers in different contexts, we investigate which dimensions of proximity affect the strength of the spillovers. We chose a sample of more than $7,000$ articles belonging to a particular category ("Wirtschaft" - "Economics"). 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 which are on average less close. Another dimension of proximity applied is the comparison of direct links, measured by the number of incoming links, to indirect links, measured by the closeness centrality.

Our main result is that an increase in the number of links from within the category is strongly associated with an increase in page length. In particular, we find that greater centrality of an article is associated with *new* authors contributing to it. However, evidence for a relation between links from outside the category and page length turns out to be rather weak. Social network analysis reveals that the category "Economics" is,

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

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.

2 Related Research

One of the first papers to investigate user-generated content, peer production and the economics of open source is that by Lerner and Tirole (2002). In their highly influential study, they analyze the production of open source software, which is, like Wikipedia, based on the provision of content by users. They point out that peer production[2] can succeed if three core conditions - inherent of Wikipedia - are fulfilled: (i) modularity - the overall project is divided into much smaller tasks, (ii) the existence of fun challenges to pursue, and (iii) a credible leadership, which keeps the project together and prevents forking or break-down. Points (ii) and (iii) are potentially conflicting goals, and administrators (especially of projects/platforms with limited support of contributors) have to strike a careful balance between what users are motivated to do and what needs to be done on the platform.

The 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.[3] 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 (e.g. 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. One of the most widely cited empirical works in this context is the study by Goyal et al. (2006), who 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

[2]With peer production we refer to goods produced by large groups of contributors, who produce tiny fragments without receiving monetary compensation, a production mode that is typical for open source software or Wikipedia.

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

analyzed as early as the 1960s.[4] More recently, Albert et al. (1999) have undertaken a similar endeavor for web pages.

Particularly relevant to our work are studies focussing on knowledge spillovers in production through social networks. Fershtman and Gandal (2011) analyse knowledge spillovers in the production of open source software and Claussen et al. (2012) in the electronic game industry looking at the network of developers. At the difference of these papers, we do not consider the social network of contributors but the hyperlink network of articles. In a broader sense, our work is related to knowledge spillovers that are measured via patent citations. However, hyperlinks in the main text of Wikipedia articles do not systematically point to knowledge that has been used but to items that are semantically related. Thus we do not expect that particular pieces of knowledge systematically spill over from one article to the other via hyperlinks. We rather expect that the quantity of producer effort and thus the quantity of knowledge that is contributed at all to a given article is influenced by the number and quality of hyperlinks.

Earlier work on Wikipedia has focussed on collaboration aspects, which have been the subject of several studies. Denning et al. (2005) discuss the collaboration of volunteers in Wikipedia. They point out some risks associated with the central idea of 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.

In addition, several 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

[4]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.

tend to reduce their contributions after the block. Ransbotham et al. (2012) empirically analyze the relation between the characteristics of the network of authors associated with the creation of collaborative user-generated content and the content value measured as article views. Their results based on social network analysis reveal a curvilinear relationship between the numbers of distinct contributors to user generated content and viewership. They conclude that network effects are stronger for newer user-generated content. 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. Greenstein and Zhu (2012a and 2012b) investigate the language bias of articles and how it evolves over time. Comparing Wikipedia articles to Democrat and Republican textbodies, 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. The study by Gorbatai (2011) uses 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.

Earlier work on Wikipedia used a two-mode author-article network where a link between articles was established by the fact an author contributed to two articles (Ransbotham et al. (2012), Kittur and Kraut (2008)). By contrast, we exploit the information on the hyperlinks between articles and hence base our analysis on explicit direct links in the content network. We analyze the semantic network whereas earlier studies focused on the social network.

3 Data

3.1 Preparation of the Data and Definition of the Economics Category

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 content of articles including the hyperlink network for the entire encyclopedia. From the resulting

tables, we constructed the time varying graph of the article network and computed our weekly measures of an article's network position, which lies 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, the number of revisions, etc. 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 (i.e., terabytes of data), 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 query 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. 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 to efficient bundles of SQL queries. For comparison, we have formulated several DSH queries used for the Wikipedia data analysis directly in SQL as well and found that the equivalent DSH queries were much more concise, easier to write and maintain (mostly due to DSH's support for order, nesting, abstractions for query reuse and concise comprehension notation).

Equipped 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 to be 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 (e.g. islands in the north sea). 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 in 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 widely used in papers dealing with large content networks like Wikipedia. We do not rely exclusively on

the subset of articles that we sampled. While we compute the social network measures only for the articles *in* the sample, we compute them using links to the articles of the entire network. In previous work, network measures are often computed only on subnetworks, i.e. 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 only 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. Beyond answering the question which links channel spillovers, this allows us to shed light on possible distortions of an analysis that remains entirely limited to a category of a network. As we are interested not only in the network position within a category of articles, but in the position on the entire Wikipedia, we have to handle the large mass of more than a million articles. Using the igraph-library by Csardi and Nepusz (2006), we compute the number of incoming links and the closeness centrality for each article at every week, both in the category and globally. 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 generally not observed.

In order to ensure internal consistency and ease of use at the same time, Wikipedia collects all the 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[5]. Thus, Wikipedia can make sure that every user finds the entire information about the item researched and that all contributions are made in one single place. Since a link that points to a redirect is just as good as a link to the page itself, these redirects have to be taken into account when generating the graph that represents the network. Thus, 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 the target of the redirect 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. Using network analysis we identify one large cluster within the category that is connected via the directed network of incoming links. We observe $7,635$ pages that are always part of this cluster, which we call strongly connected component in the category "Economics". The other pages could not always be reached via the categorial network. During the period of observation, $1,237$

[5]To give an example, a user who searches for "Schumpeter" rather than "Joseph Schumpeter" will be redirected to the latter almost silently.

of these pages received an incoming link from the strongly connected component of the economics category, becoming than strongly connected.

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 strongly connected component during 153 weeks. It contains in total 1,168,155 observations.[6] The second data set consists only of articles that got connected to the economics category during the period of observation. In total we observe 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 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 "strongly connected" articles.[7] 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. It should be kept in mind that the links underlying the centrality measures on the directed network are not created by editing the page itself but by editing the pages pointing to it. We observe the length of a page in bytes, and also when a page was created. One byte corresponds roughly to one letter. The median page length is 3630 bytes and the median article was written by 16 authors. Our main centrality measures are the number of direct links pointing to a page (termed in-degree centrality in social network analysis) from the entire German Wikipedia ("Links from Wikipedia") and from the category the article was drawn from ("Economics"). Since articles from the category are also contained in the entire Wikipedia, we report the difference of the two in-degrees. By sample construction, every page has a link from the category. The median page has 11 links from Wikipedia, 4 of which are from the category. Articles usually belong to more than one category, but we do not observe these additional categories. 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 in-degree centrality lie far above of the 99% percentile. The closeness centrality measures represent the inverse average distance

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

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

of one article to all other articles in the relevant network.[8] We observe in our data that the original closeness measure is mainly driven by the variations in the share of disconnected articles and in the network size over time (results not reported). In order to abstract from these effects, we compute the relative closeness rank within our balanced panel (though considering links from the entire Wikipedia and "Economics" network, not only from the sample). This procedure may be useful in work on dynamic networks in general. Again, the measure is computed both on the local network made up by pages in the category and on the entire German Wikipedia. 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 literature section usually points to an article that draws extensively on scientific, literary of journalistic sources outside Wikipedia and therefore tends to be longer. The median age of articles 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 got connected to the category of economics during the period of observation. We consider the sample over the entire 153 weeks, which means that all articles of the sample are part of the time disconnected and later on strongly connected to the category "Economics". 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 these articles of $3,044$ bytes is about 600 bytes shorter than the median page length of articles which are always strongly connected. The number of links within the category is by definition smaller, since the articles are part of the time of observation disconnected from the main component. This means that they do not have any links from other articles in the category except maybe from a small number of articles which are also disconnected from the main cluster. The number of links from outside the category is similar in median in both samples but considerably smaller in the higher percentiles of the sample of articles that are initially disconnected. This means that an article that has a high centrality in Wikipedia and belongs to the category of economics is unlikely to be disconnected from the category in terms on incoming links. 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.

[8] 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:

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

To see how often the variables for individual pages typically change, 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 frequency of changes in variables. The changes are reported for our main sample of articles that are always strongly connected. Less than 25% of the observed pages never experience any change in their number of incoming links and less than ten percent have never been edited or never received an additional author during the three years period of observation. 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 not only on the activity at the page but rather 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 distributions of changes more visible, which are dominated by zeros otherwise. The median change in page length is 18 bytes in a week, which corresponds to about two words. This makes obvious that minor changes play an important role in the work many authors contribute to Wikipedia in order to improve the quality of the 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. Eighty percent of the changes amount to far less than one rank 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 greater centrality in the article network is associated with (i) more content being generated (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" (in-degree centrality within category) and from the entire German Wikipedia (global in-

degree centrality) 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 in-degree centralities maybe 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. Hence, like with almost all network data, several 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 of 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. Finally, articles might be affected by 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, in particular when taking into account that the number of authors constantly grows over time.

Consequently, it is important to go further than a simple analysis in the cross section, since comparing different articles about very different topics would not be appropriate, whenever article A is more relevant and more frequently edited than article B. Instead, like Kittur and Kraut (2008) and Ransbotham et al. (2012) we intend to 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. Taken both fixed effects together, we estimate two-way fixed effects panel regressions based on the following equations:

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

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

where $centrality_{it}$ is a vector of the four centrality measures mentioned above and where 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. But we can make use of a special type of pages in our data in order to shed more light on the relationship of network position and user-generated content. 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 an author does not observe on a page the list of other articles that point to it. He may have an idea about this if he is familiar with a set of related articles, but he may not observe the network status exactly. Hence he will not know whether an additional link to an article will connect this article to 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 (which leaves the article disconnected from "Economics") or inside 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 in-degree centralities 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 strongly connected, the sample includes articles that are at first disconnected and become strongly connected 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 articles that become connected, we just consider a dummy for becoming connected. We do not consider additional changes in in-degree

centrality, since we know that most articles change by one link at maximum in a given week and do not change in most weeks. So accounting for getting connected and in-degree centrality 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 strongly 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 11 weeks in the sample, we regress page length on an indicator variable that takes the value of one if the page could be reached via the links from an economics page and zero otherwise. This means it take 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. Still, 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 regression equation 1, where page length is regressed on several sets of network variables, article fixed effects and time fixed effects.[9] 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 received 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

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

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 received from other pages of the category "Economics". These two sets of links are not mutually exclusive. The effect can rather 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 the 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 $-.13$.

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 the closeness of most articles changes by less than 1 in any given week. We verified that the changes are not much stronger for observations with a non-constant in-degree centrality. From this point of view the effect looks small. Moreover, the size of the coefficient for in-degree centrality is barely affected and the added explanatory power of the new variable is rather low. Finally, Column 4 brings together all the 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 within 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.

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 regression 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. As for page length, the effect is much stronger for links from the category: an additional link from the category corresponds to approximatively 0.54 more authors (considering the sum of Wikipedia and category coefficients). On average every second additional link from the category is associated with a new author contributing to the page. The coefficient for outside links is much smaller but remains significant in all specifications. The closeness rank has negligible effect in column 3, which turns insignificant in column 4.

In sum, 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 effects of the other centrality measures are small.

The regressions in Tables 7 and 8 use the information of pages getting connected to the main cluster of the category. 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 7 shows the results, when we consider 5 periods before and after the jump including also the period of the jump itself.[10] 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 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 effect is even stronger for the first differences of page length (columns 2 and 4), ranging from 66 bytes *per period* in the pooled regression to 195 bytes per period when including time and article fixed effects. These are sizeable effects which cannot be expected to last forever. They are more likely to occur only for a few periods and it might be that a share of the additional content is be provided in the same week when the article was connected.

In Table 8 we account for that possibility, by excluding the week of the "jump" into the category and the two weeks before and after the article was connected. Instead 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

[10]The number of observations is much smaller in Tables 7 and 8 for two reasons: first the number of articles that got connected is much smaller (1237 vs. 7635), second we only consider 11 or 10 periods per observed page (not 153 as before).

within weeks after the new connection was established. However, the effects remain by and large positive and our results indicate, that an article grows by 9 (pooled) to 21 byte per week (fixed effects) faster three to seven weeks after being connected to the new category. 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 identifying content they want to contribute to. We consider the network of hyperlinks between Wikipedia articles as a possible channel of spillovers in attracting 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 difference in effects. One more link from another article from the category is related to a increase in page length by 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. One additional link from the category "Economics" is related to an increase in the number of authors by 0.5. These results are all obtained in a balanced sample of articles that are always connected the to the large cluster of the category "Economics". Articles that are initially not connected increase by more than 300 bytes in length during the five weeks after connection.

Taken together the evidence suggests 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. The size of the additional contributions that may be expected is not very high. These small changes of a few words or one sentence constitute, however, a large part of contributions to Wikipedia. According to our evidence, this strategy will only work within a cluster of thematically related articles. Links from articles that do not share a central category with the target article do not seem to enhance content generation. Moreover, we conclude that administrators should prevent the formation of disconnected islands of content. Since the number of links pointing *to* an article is not directly visible to authors, metadata could be used by advanced contributors or administrators to detect and set these missing links.

From a researcher's perspective, our results suggest that it is an acceptable strategy in

the context of content networks to use only a smaller group of articles/nodes for network computations, as long as one does not extrapolate the result to the unobserved nodes. However, 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. Whether they also hold for newly created articles cannot be answered at this point and is studied in ongoing research. Finally further research in how to adequately specify exogenous variation in the network position of articles would be fruitful.

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.

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, 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.

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.

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. Strongly connected articles.

	Min	p10	p25	p50	p75	p90	Max
Length of page	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: 1,168,155

Table 2: Summary statistics of main variables. Articles that get connected to category during the period of observation.

	Min	p10	p25	p50	p75	p90	Max
Length of page	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: 203,031.

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

	Min	p10	p25	p50	p75	p90	Max
Length of page	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. Strongly connected articles.

	Min	p1	p10	p25	p50	p75	p90	p99	Max	Obs.
Length of page (in bytes)	-95,222	-868	-42	-1	18	70	309	2739	83235	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	1455	121,589
Links from Wikipedia excl. categ.	-439	-2	-1	1	1	1	2	9	1455	90,214
Links from category	-130	-2	-1	1	1	1	1	2	80	46,304
Rel. closeness rank (Wikipedia)	-91.74	-1.63	-0.58	-0.23	-0.02	0.18	0.51	1.76	90.84	1,137,528
Rel. closeness rank (category)	-99.15	-0.98	-0.20	-0.11	-0.04	0.03	0.13	2.04	84.94	1,090,973

Articles that were always connected to econ. main component.

7.2 Regression Results

Table 5: Relationship of page length and centrality.

	(1) Wiki links	(2) Wiki & categ.	(3) Links & 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	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 strongly connected articles were included. Dependent variable: page length.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Relationship of number of authors and centrality.

	(1) Wiki links	(2) Wiki & categ.	(3) Links & 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	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 strongly connected articles were included. Dependent variable: number of authors.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Relationship of the growth of page length and the page becoming strongly 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 articles and time fixed effects.

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

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Relationship of the growth of page length and the page becoming strongly connected, excluding the period of the jump itself and the 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 Pooled OLS-Regressions, Columns 3 and 4 include articles and time fixed effects.

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

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$