Centrality and Content Creation in Networks - *

The Case of Economic Topics on 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

June 3, 2016

Abstract

We analyze the role of local and global network positions for content contributions to articles belonging to the category "Economy" on the German Wikipedia. Observing a sample of 7,635 articles over a period of 153 weeks we measure their centrality both within this category and in the network of over one million 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. Beyond the econometric analysis, our study sheds light on how the discipline of economics is represented on German Wikipedia. We find non-neoclassical themes to be highly prevalent among the top articles.

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 the editor of this journal and an anonymous referee for very useful suggestions. Moreover, we thank the Wikimedia Foundation for granting access to the Wikipedia data, Thorsten Doherr and Manfred Knobloch for support with the data processing, 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, Fabian Trottner and Lukas Trottner provided helpful research assistance. We acknowledge financial support from the WissenschaftsCampus Tübingen.

1 Introduction

User-generated content has proven 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.

As a use case for analyzing the role of hyperlinks, we consider a set of articles related to economic topics. Based on category labels assigned within German Wikipedia, we identify a main category 'Economy' and subcategories such as 'Economics', 'Economists' and 'Enterprise and trade.' We investigate whether links from articles that are semantically close (also in the category 'Economy') have a different impact than 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 thus exploit different dimensions of proximity that exist between articles, when analyzing the relation between centrality and content provision. To our knowledge, this is the first paper taking a systematic look at Wikipedia articles in the field of economics. Our analysis gives some idea about the kind of contributions that might be needed to cover this thematic area on Wikipedia more fully. Moreover, our sample of articles is of interest to economists in a more general way since it offers insights on how the discipline of economics is represented on one of the largest non-English Wikipedias.

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, the watchlist 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 against 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 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 also consider spillovers occurring in the production of knowledge. The channel of spillovers that we are analyzing consists in the hyperlinks pointing from one article to another. However, we are not looking for knowledge spillovers in the classical sense, but for spillovers in the level of production activity. On Wikipedia, this approach is based on the hypothesis that links placed on page A pointing to page B may attract the attention to page B. Consequently, the existence of an additional link may trigger the contribution of authors who might not have contributed in its absence. These

[1] http://de.wikipedia.org/wiki/Wikipedia:Verlinken, accessed on June 5, 2015.

spillovers affect the level of content provision on a page and also increase the knowledge contributed to the page. Note, however, that the dimension to which the notion of spillover applies in our context is not the knowledge itself but the attention and effort that authors direct to a particular page after they read another one pointing to it.

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 'Economy' 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.

Thematically we find a relatively strong bent towards heterodox economic schools of thought and anti-capitalism. This can be seen when analyzing the top 20 articles among those associated with the discipline of economics. By any of our five measures, Karl Marx turns out to be the most prominent economist on German Wikipedia during the period of observation.

2 Related Research

Our research is inspired by two strands of work on user-generated content: research on direct and indirect spillovers in networks of software and other peer production and research on motivations and patterns in collaboration among Wikipedia authors. In particular, we extend methodological aspects of earlier work on author networks to hyperlink networks.

For analyzing direct and indirect networks spillovers, we can build on the theory of social networks (cf. Jackson (2008) and Jackson and Zenou (2013)). Particularly relevant to our work are studies focusing on knowledge spillovers in production through social networks. Fershtman and Gandal (2011) analyze knowledge spillovers in the production of open source software and Claussen et al. (2012) in the electronic gaming industry. Both papers analyze the relationship between developers' network-position and the success of the project they are working on. Like Claussen et al. (2012) we use a panel design to account for unobserved and individual-specific heterogeneity, which in our case can stem from variations in the relevance of different articles or the expertise required for contributing to the subject matter. At the difference of these papers, we do not consider the social network of contributors but the hyperlink network of articles.

The mechanism by which articles would most likely benefit from more links is their

ability to channel users' attention and to attract new contributors to their target page. That links can channel users to other pages has been documented both for commercial retailing (Stephen and Toubia (2010), Carmi et al. (2012)) and on Wikipedia (Kummer (2013)). Such work typically builds on exogenous variation on networks around an individual node (ego-networks) and over short time horizons. Our paper complements their findings by analyzing how content contributions to a public information good interacts with an article's position in the link network over a long period of time and in a large and coherent set of articles.

Viewing hyperlinked articles as a citation network formed of directed links is well grounded in the literature. In fact, citation networks of scientific papers had been analyzed as early as the 1960s. Without using the more recently developed measures of network position it was still possible to evaluate citation data and to provide several interesting statistics on average references and citations in the network (cf. de Solla Price (1965)). More recently, Albert et al. (1999) have undertaken a similar endeavor for web pages. Like in social networks, phenomena like homophily and preferential attachment are important issues in hyperlink networks (Jackson (2008), Katona and Sarvary (2008)). Capocci et al. (2006) find for example, that, similar to the internet in general, preferential attachment is a highly prevalent phenomenon on Wikipedia. We can also borrow from the approach used by Halatchliyski et al. (2010) who analyzes authors' contributions in two related knowledge domains considering the article network. Our paper contributes to this literature by computing the network measures of an article both on the local network of articles (category) and the global network (entire German Wikipedia).

An important challenge for estimating spillovers in any hyperlink network is network formation, because on the web it is not costly to place links and the link network is formed over time. This leads to endogeneity bias, because the network structure itself might be the outcome of the utility maximization or other strategic considerations of the economic agents. For example, links may contain information, because they reflect a judgement of those who place them (cf. Page et al. (1999), Surowiecki (2005), Mayzlin and Yoganarasimhan (2012), Dellarocas et al. (2013), etc. for settings where links refer to more information or are placed strategically). Models which provide a micro foundation of such behavior in online settings have been provided by Mayzlin and Yoganarasimhan (2012) for the blogosphere and by Dellarocas et al. (2013) for commercial content provision. The problem of endogenous link formation when quantifying spillovers in networks is generally recognized in the literature on social networks. The existing literature can be grouped in two streams (cf. Dellarocas et al. (2013), Graham (2015)): the first focuses on strategic link formation of individuals in a network (Bala and Goyal (2000), Galeotti et al. (2010) and Goyal and Moraga-Gonzalez (2001)). The second stream analyzes the strategic choice of effort or input as a function of a given network structure (e.g. Bramoullé and Kranton

(2007)). Recent papers by Bramoullé et al. (2009) and De Giorgi et al. (2010) show that peer effects can be identified in network data if the graph is exogenously given and has many overlapping peer groups. Also, Lin (2010) exploits friendship information in a classroom setup (partially overlapping network) and applies a Spatial Autoregressive Model. While the first approach takes efforts as given, the second assumes the network to be exogenously formed. Approaches that tackle both aspects simultaneously are scarce. (cf. Dellarocas et al. (2013), Graham (2015) and Cabrales et al. (2011)).

It should be noted that the setting on Wikipedia is non-standard in a favorable way: While the difficulties of endogenous strategic link formation put a great challenge research on networks, links are in our context placed by "the Wikipedian," who has atypical and less strategic objectives. Consequently, the role of links in Wikipedia is different from the role of links on the web in two ways. First, Wikipedia articles do not compete for attention, second, links are in general not placed strategically, but to maximize the encyclopedic value of the entire website. Hence, while links certainly do contain some information and reflect relevance, the problems due to link formation in estimating spillovers are mitigated somewhat by this absence of strategic incentives. Links are neither a measure of popularity, nor is there any scope for strategic linking to maximize traffic. Instead, linking is driven by non-strategic motives and, most importantly, semantical connections.

We build on a second important strand of literature about collaboration between authors on Wikipedia. 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. Kriplean et al. (2008) focus on a non-monetary reward tool at Wikipedia, "Barnstars," which can be awarded to hard-working authors, and its contribution to content creation. Soto (2009) reviews further existing research based on Wikipedia data and (among other things) quantitatively analyzes the ten largest Wikipedias. 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. Algan et al. (2013) invited Wikipedians to participate in economic experiments and found that they can maintain higher levels of collaboration than an average student population.

Other related empirical analyses in this line of research focus on the determinants of the quality of articles on Wikipedia. Kittur and Kraut (2008) examine how the number of collaborating editors and their coordination methods affect article quality measured by peer evaluations in Wikipedia's quality assessment project. More editors to an article improve quality only when the editors use appropriate coordination techniques. Aaltonen and Seiler (forthcoming) provide evidence for a "rich get richer effect," that causes arti-

cles that were developed early to stay better than later cohorts of articles. Ransbotham et al. (2012) 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) show that the density of the authors' individual social networks can predict both norm violations and users' discouragement after deletions or 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 articles written by relatively "young" and relatively "old" teams face a longer time span until they are promoted than articles by teams with an average experience on Wikipedia. Halatchliyski et al. (2010) find that the most central authors also contribute to integrating the two fields of knowledge they consider in their paper. Greenstein and Zhu (2012a and 2012b) investigate the political language bias of articles and how it evolves over time. They find that an early bias of Wikipedia towards Democrat language has gradually disappeared over time. This erosion of the overall bias is driven by new articles, which use Republican vocabulary. Gorbatai (2011) shows that frequent editors of Wikipedia strongly react to (attempted) contributions of inexperienced users, as they are a sign of increased demand.

This paper contributes to the literature by analyzing the relationship of network centrality and content generation, and we highlight that local links (from the same content category) play a different role in content generation than links from more remote articles. Moreover, we provide a deep analysis of a single category of articles about economic topics, and investigate which type of articles receive most edits, and whether these are also the most central articles.

From a methodological point of view, we extend earlier work on Wikipedia, which 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 base our analysis on explicit direct links in the content network. We thus analyze the semantic network whereas earlier studies focused on the social network. We compute the network measures only for the articles inside the category (i.e., for roughly 10,000 nodes), but we use the links from all pages in the entire network (i.e., more than one million nodes) to compute them. This approach differs from previous work, where network measures are often computed only on subnetworks and abstracting from the existence of all the other articles. We 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.

3 Data and Network Measures

For the analysis in this paper, we downloaded a full-text dump from the Wikipedia toolserver and constructed the time-varying graph of the article network on a weekly basis. From this graph, we computed the measures of an article's network position that lie at the heart of our analysis, and augmented the database with data on readership. Moreover, we added information on article characteristics like the length of the page, the number of authors or the categories to which the article belongs. In the next subsection (3.1) we provide information about the definition of the category and the extraction of the dataset. Subsection 3.2 is devoted to a discussion of our network measures, and in Section 3.3 we describe the distribution of the main variables in our dataset. The online appendix describes the data extraction in more detail.[2]

3.1 Preparation of the Data and Selection of the Articles

In our analysis, we use data on 153 weeks between December 2007 and December 2010. At the beginning of the period of observation German Wikipedia, which exists since March 2001, covered already a large range of topics, more than 735,000 articles as of *Dec.* 2007. It continued to grow substantially to a volume of 1.2 *million* articles in *Dec.* 2010. Given the size of Wikipedia, we choose to focus on a particular category. Coming from the perspective of economic research, we identified all articles related to the categories and subcategories of the 'Economy' ('Wirtschaft').

While articles have been selected from one category, network measures account for links between these articles and the entire German Wikipedia. This dataset is too large to use only in-memory processing. Hence, we stored the data in a disk-based, relational database and queried the data using Database Supported Haskell (DSH) (Giorgidze et al. (2010) and (2011)). This is a novel high-level language which allowed us to formulate and efficiently execute queries on nested and ordered collections of data, and which was specifically developed for the application on similar datasets.[3]

The choice of articles sampled was based on Wikipedia's category tree. We sampled all the articles belonging to the categories and subcategories of 'Economy.' Next, we define the 'category network' as the set of nodes that remain within the category, and the 'global network' is composed of the entire German Wikipedia. Hence, the entire analysis is based on the *directed* network formed via incoming hyperlinks from the entire Wikipedia.

[2] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

[3] DSH queries are automatically translated into efficient lower-level query languages that the underlying database system understands. For this study, we utilized 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.

We also used the category tree to distinguish interesting subcategories. We create 11 non-overlapping content categories, which are constructed in a way to better characterize the most interesting aspects of the subcategory 'Economics' ('Volkswirtschaftslehre'; see section A1 of the online appendix .[4] Similarly, we create five categories of persons, with 'Economist' as the first category. This subclassification allows us to understand better how the academic discipline of economics is represented on German Wikipedia and what other contents related to the economy and economic matters are covered. We pursue a twofold approach for aggregating categories based on automatic keyword search and manual classification that allows to aggregate them to 11 conceptual categories of non-persons and 5 categories of persons. Our procedure is described in the online appendix .[5] In section 4, we discuss our findings on the representation of economic topics on German Wikipedia.

Note that we identified and excluded the revisions that were made by small programs, so-called 'bots,' before computing the values of the variables. Such programs automatically make small formal changes to ensure that a consistent style is maintained throughout Wikipedia. Since they are not made by humans, we did not consider the revisions that were carried out by bots and we also excluded bots from the author count.

3.2 Measures of Centrality and Activity

Our main interest is the relationship of the centrality of pages to content generation activity. Hence, our main explanatory variables are measures of centrality in the network of incoming hyperlinks and the dependent variables are page length and the number of authors.

In social network theory, the notion of centrality a priori captures how well-connected a person is in a group of people (cf. Jackson (2008)). By well-connected we generally mean that a person is able to access a large number of people in the network through a small number of intermediaries. Such connectedness can be advantageous for that person, or it might simply reflect a node's importance. Network theory extends these notions to geographical entities or objects, such as road networks between cities or link networks between blogs or Wikipedia articles. Like humans, networked objects might benefit from their connectivity. Alternatively, connectivity might reflect importance or popularity, which was shown for blogs on the web (cf. Dellarocas et al. (2013)). While this fact introduces an important source of endogeneity which we will discuss throughout our analysis, it is worthwhile pointing out one important specificity of Wikipedia: Links on Wikipedia are motivated by the needs of the encyclopedia to connect the content and provide additional background. This is clearly different from other online contexts where

[4] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

[5] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

popularity is the main driver of the network. On Wikipedia there is no point to link an article even to the most popular pages, absent any real underlying relationship between the two topics.

Centrality measures can be based on undirected links, which are reciprocal, or on directed links, which are unidirectional (but might be reciprocated). In this study, we are working with *directed* links, namely links pointing from one article to another article. For any given article, e.g. 'Inflation,' these are not the hyperlinks marked within the text on inflation, which point for example to 'Price level stability' or 'Equation of exchange.' Instead, the centrality of 'Inflation' is defined by the links pointing *from* other articles *to* 'Inflation,' e.g. 'German Mark', 'Macroeconomics' and 'Price level stability.' Sometimes 'Inflation' links back to such a linking article, so that some of these links are in fact bidirectional. As it is standard in the analysis of directed networks, we do not treat such links differently from unidirectional links.

We define the network based on direct links in such a way that it includes those articles that lead to any given article, e.g. to 'Inflation,' by only one click. Our behavioral hypothesis is that these links attract some readers and authors that may otherwise have paid no or less attention to the linked article. The quantitative measure for centrality in terms of incoming links is called indegree centrality and is a simple count of the number of links to an article that were present at a given point in time each week. We compute two measures of indegree centrality: The first one is counting only links from within the category 'Economy' (indegree within category), which we construct as described in Section 3.1. The second measure counts links from the entire German Wikipedia (global indegree). The idea behind contrasting these measures is that articles from within this category are on average more closely related in content (e.g. the link from 'GDP' to 'Inflation') than articles from other categories (e.g. the article on the artist 'Marc Chagall' who lost wealth due to hyperinflation). We examine whether links within the category 'Economy' have a higher propensity to channel readership and content generation than links from outside.

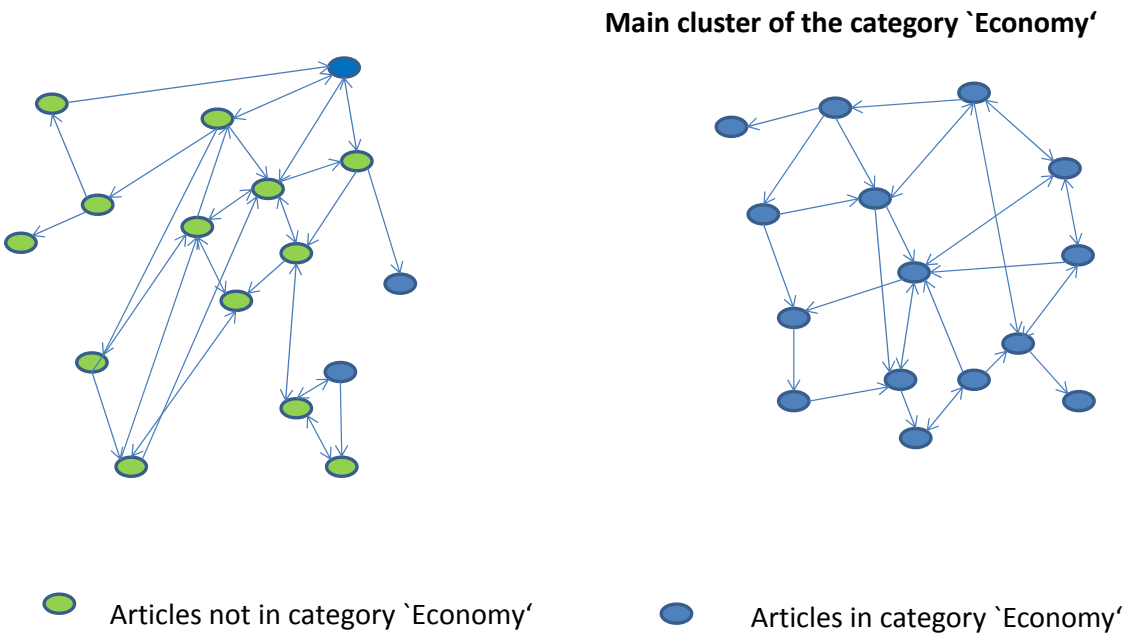
Social network theory uses a variety of more sophisticated network measures that take into account indirect links or the position of an article in the network. An indirect link in our application is a connection between two articles that is established only via another article. To account for the potential effect of these more complex linkages, we control for two more measures of closeness centrality in terms of incoming links. Specifically, we use the closeness rank in the network of the category and in the global network in our analysis. The closeness centrality is a standard measure in network theory. It is computed for every article in every period and then the articles are ranked according to their closeness. The details of these computations are somewhat technical and are presented in section A2 of the online appendix.[6] Other centrality measures may provide

[6] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

additional characterizations of the network. Betweenness can capture situations when nodes take a crucial role in connecting two otherwise disconnected groups. While we computed these centrality measures at least locally, we found that they add little to the econometric analysis. For betweenness, this might be explained by the network's density and the fact that there are usually hundreds of paths to get from one group of articles to another one. Centrality measures are computed using the igraph library by Csardi and Nepusz (2006). We account for the existence of redirect pages, by counting a link to a redirect page also as a link to its target page.

We point out an important special case of extremely low local (category) centrality that we observe in our data set. We frequently observed that single articles belonged to the category 'Economy' according to their labels but did not receive any link from another article within this large category. We call these articles 'disconnected' from the main cluster on 'Economy.' When such articles receive a link from the main category, their closeness centrality is suddenly drastically increased.

Figure 1: Articles belonging to the category 'Economy' that are disconnected from its main cluster.



NOTES: The graph illustrates how certain articles can be classified as economic articles, and yet be unconnected to the cluster in terms of the link network. The blue nodes have a tag that categorizes them as economic articles, whereas green articles do not have such a tag (and are hence not in our sample). Especially in the beginning of the observed period, several articles were categorized as economic but had only links from non-economic articles.

Our dependent variables of interest are page length and the number of authors, which both measure content generation activity. We observe both on a weekly basis. Page length is measured as the number of bytes after the most recent revision, so that we can observe the weekly net difference in text length. Thus, the measure is robust to vandalism

or other attacks that are quickly undone. Moreover, we can observe the total number of authors that have contributed to an article up until this point.

Focusing on length and the number of authors is in line with Wikipedia's quality standards and the existing literature (cf. Kittur and Kraut (2008)), since these are the most reliable measures of *activity* for our data. As we will discuss in more detail, content changes infrequently for most articles (cf. Table 3). As a result, some good measures of quality, such as the number of references or images have in fact very low variability. Another good quality measure, the ratio of authors/content, has very high variability on young and short articles. Moreover, taken *together*, length and number of authors are a good indication of the quality of a Wikipedia article. In the early life of an article, its length is the crucial determinant, whereas later quality increases as more authors edit a page. Clearly, the quality also depends on the number of edits, which indicate more effort by the same individuals, but, especially in conflictive situations, more authors and the associated coordination and 'additional eyeballs' seem to be a stronger indicator of quality than merely additional edits.

3.3 The Anatomy of the Data Set

One large cluster within the category could be reached via the directed network of incoming links, and 7,635 pages are always part of this cluster. We refer to it as the 'connected component' in the category 'Economy' (or just 'connected' or 'reachable articles'). All other articles could not always be reached via the category network.[7] Hence, our main data set is a balanced panel observing the 7,635 articles that remain in the connected component during 153 weeks (1,168,155 observations). During the period of observation, 1,237 initially disconnected pages received an incoming link from the connected component in the category 'Economy,' and thus became part of that component. We use these articles in a second data set for our analysis. From these 1,237 pages we observe 203,031 weekly observations.

Table 1 provides summary statistics of our variables for the balanced panel of articles that are always reachable from the category 'Economy.'[8] 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

[7]We follow a widely used 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 'Economy.' Moreover, we find approximately 7,000 articles that were nonexistent at the beginning of our period of observation or ceased to exist before the end and are, hence, excluded from the analysis.

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

is 3630 bytes, and the median article was written by 16 authors. Our main centrality measures are indegree and closeness centrality. 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 within the category. Articles usually belong to more than one category, but we do not observe these additional categories.[9] The distributions of the centrality variables show that for many articles half or more of the links come from the category 'Economy.' 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.

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 (see section 3.2). 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. Age captures whether the article has been on the wiki for a long time or whether it is still 'under construction.' The indicator variable for redirect pages flags pages that were converted to a link page, which merely redirects the reader to the page of a synonym. The presence of a literature section, finally, indicates relatively long articles that draw extensively on scientific, literary or journalistic sources outside Wikipedia. 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 'Economy' during the period of observation. 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. 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, most articles do not change in any given period. For descriptive analysis and robustness checks in the regressions, we also measure the number of clicks in the 24 hours before the next due date in our weekly panel. In the online appendix, we provide further illustrations and descriptive tables of the data we used.[10] Figure D1 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. Finally, Table D1 displays the magnitude of changes for all observations with non-zero change.

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

[10] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

4 The Category 'Economy' on the German Wikipedia

In this section, we shed light on how the academic discipline of economics is represented on German Wikipedia and what other contents related to the economy and economic matters are covered. To do so, we create 11 non-overlapping categories of non-persons (see Table 5), which are constructed in a way that attributes most weight to the category 'Economy' ('Volkswirtschaftslehre'), which we are most interested in. We add other categories in declining order that are farther away from our field of interest. An article is only assigned to a category if it has not yet been assigned to a category of higher rank (e.g. if 'Monetary Theory' is already part of 'Economics,' it will not be assigned to 'Banking'). Similarly, we create five categories of persons, with 'Economists' as first category. Note that a different ordering of categories would lead to a different assignment of some articles. The details of the sampling and the construction of categories are explained in section A1 of the online appendix. [11]

4.1 Distribution of Articles Across Subcategories

Table 5 summarizes the distribution of articles in our main sample across subcategories and median values of article age, page length, the number of authors and the number of clicks. Moreover, it contains two measures of centrality, the number of articles within the master category 'Economy' hyperlinking to a specific article (indegree from category) and the number of articles from other categories hyperlinking to a specific article (indegree outside category). The category 'Economics' covers 13.4 percent of the articles on non-persons from the sample. While this is a substantial share, the majority of articles about the economy on Wikipedia have not been assigned any label that is associated with economics as an academic discipline. On the one hand, this is due to the importance of various institutions (banks, firms, government institutions, legal associations, etc.). On the other hand, economic issues are discussed by many other communities such as managers, worker representatives, politicians, researchers of other disciplines or people criticizing the prevailing economic systems. This is reflected in the entries we find in our sample. For example, nearly six percent of the articles fall into the category 'Labor, poverty' without falling in the categories 'Economics', 'Management' or 'Trade, enterprise.' Among the articles on persons, more than half are about economists. This includes persons trained as economists but not working as academic economists. Since we note that the movement criticizing globalization has relatively high prominence on German Wikipedia, we separately identify persons labeled as 'Globalization critiques' ('Globalisierungskritiker').

Looking at the article characteristics for non-persons, we observe that the articles on 'Economics' are oldest. The youngest articles on 'Trade, enterprise,' are created more

[11] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

than half a year later (evaluated at the median). The category with the longest articles is 'Politics, policy,' followed by 'Labor, poverty,' and 'Economics.' The number of authors is also highest in the category 'Politics, policy' with 19.7 authors in median, followed now by 'ICT' (18.8) and 'Trade, enterprise' (18.2). The most clicked articles are those on 'Management,' followed by 'Labor, poverty' and 'ICT' (both around 15 clicks). The median number of links from within the main category 'Economy' is highest for 'Economics.' This means that notions from economics are used in a large number of other articles. The number of links from the category is lowest, with values less than four, for 'ICT', 'Sociology, social matters' and 'Other economics topics.' From outside the main category, the largest number of links point to the category 'Politics, policy' with 12.4 links.

Turning to the articles on persons, the category of 'Globalization critiques' stands out in several respects: Its articles are relatively old, with 7,565 bytes they are twice as long in median than the articles on other persons and they also have more than twice as many authors with 33.7, while the numbers for other articles about persons lie close to the median numbers for non-persons. Articles on persons tend to be clicked less often than articles on non-persons. Again articles on 'Globalization critiques' receive most interest. They are also nearly twice as hyperlinked from articles outside the main category 'Economy' as other articles on persons. Meanwhile, this category is not extremely central within the main category 'Economy.'

4.2 Top Articles on Economics

In the previous section, we have looked at the median characteristics of articles of different categories. At the top of the distribution of all variables, we observe values that exceed the median by far. Looking at the top articles according to different page characteristics gives a more precise idea of which articles are popular among contributors and readers of Wikipedia. Since we have a particular interest in the academic discipline of economics, we look at the top 20 articles only within the subcategory 'Economics' (Table 6). Within the subcategory, the article with the highest mean length during the period of observation is 'Marxian Economics.' Also ranks 2,7 and 12 are taken by articles related to Marxian economics. 'Neoliberalism' and 'Capitalism' are other terms frequently used in non-neoclassical economics showing up in the top 20. On the other hand, a number of general economic terms, mostly related to macroeconomics, are part of the list ('Unemployment statistics', 'Tax', 'Money', 'Property', 'Government debt', 'Inflation'). Some of the longest articles also rank within the top 20 with respect to the number of authors. But this is not the case of the articles on Marxian economics. Meanwhile, 'Criticism of capitalism' (rank 16) and 'Keynesian economics' (rank 19) are two articles on non-neoclassical topics included in this list. The highest numbers of clicks interestingly go to the two main target variables of macroeconomic policy: 'GDP' and 'Inflation.' Again we see some over-

lap with the two preceding rankings, but the readers of Wikipedia seem less focused on heterodox economic thinking and the criticism of capitalism than the contributors. Turning to centrality within the category 'Economy,' we observe a far stronger dominance of core notions of textbook economics than in the other rankings, including the articles on 'Economics' (which might include links from the labels), 'Production', 'Employee', 'Liquidity', 'Cost,' and 'Demand.' Interestingly the article that receives the most links from outside the category 'Economy' is 'Liberalism.' Further results for selected slices in the middle and at the bottom of the distributions are available upon request. They show a mix of topics from economic theory, institutions and policy with no particular dominance of specific themes. Typical articles with median page length are 'Endogenous Growth Theory', 'Manchester Capitalism,' and 'German Tax Reform in 2000.' At the bottom of the rankings, we tend to observe more specific topics: 'Degree of openness', 'Balance of payment deficit', 'Mass market,' and 'National Bureau of Economic Research.' But there is no obvious general rule, which topics rank in the middle and which topics rank in the bottom. More in-depth network and semantic analysis might yield additional insights as to which thematic areas are more developed on Wikipedia than others.

In Table D2of the online appendix,[12] we show the top 20 articles in terms of increase in content and centrality in German Wikipedia during the period of observation. We opt for absolute measures of increase, since relative measures would place articles that are initially extremely small at the top even if their absolute increase is small. The table shows the fastest growing 20 articles, assessed by comparing the average value during the first 10 weeks of the sample to the average value of the last 10 weeks. The fastest growing pages (in terms of length) are shown in column 2 alongside the growth (in bytes) in column 3. Column 4 shows the 20 articles which obtained most additional links (on top of the links they had during the first 10 weeks), and column 5 quantifies the increase in links. While the growth in indegree is strongest on topics related to main themes of economics, the growth in content covers again several topics with somewhat Marxist flavor: 'Capitalism', 'Profit', 'Ground Rent' or the 'Planned Economy.' The other dominant theme seems to cover issues which may be related to the European financial and economic crisis, such as the 'European Monetary Union', 'Balance of Trade' or 'Credit Default Swaps.' The strongest growth in links pointing them can be seen in articles which contain key definitions such as 'Product', 'Good,' the 'GNP' or 'Inflation.'

Since it might be of interest to readers who are themselves economists, we report the top 10 economists: Table 4 shows the most prominent economists on German Wikipedia in terms of our variables of analysis. The most striking result of this table is that Karl Marx is the most important economist on German Wikipedia by all five criteria. Friedrich Engels also appears among the top ten. Other well-known economists born before the

[12] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

20th century, such as Adam Smith, David Ricardo, David Hume, or Joseph Schumpeter, appear, but their importance varies by criterion. On ranks 2-10, interestingly, the ranking by indegree from category mentions most of the names that appear in the history of economic thought typically taught at university. The only living economists on the list are German: Horst Köhler, who is a former president of Germany, and Hans-Werner Sinn and Peter Bofinger, two economists who frequently appear in German media.

While views on this naturally differ, most economists would probably consider that certain topics are underrepresented on German Wikipedia. Moreover, it is worth noting that the ranking obtained from the link structure reflects the relevance in the scientific discourse of economic thought more closely than other performance or popularity measures, such as length or clicks. This fact highlights the link structure's ability to quantify the relevance of topics, but it also raises the question, to which extent hyperlinks can be a potential channel for attracting authors and contributions to articles. We investigate these issues in the following sections for the entire sample of articles.

5 Econometric Analysis of Centrality and Content Creation

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. In subsection 5.1 we analyze these questions for the main component of always connected articles. In subsection 5.2 we shed light on articles that get connected to the category 'Economy.'

5.1 Network Position and User-Generated Content

Our main explanatory variables are measures of centrality in the network of incoming hyperlinks. As described in section 3.2, we have four centrality measures: the number of incoming links within the category 'Economy' (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 variables enter our estimation only in a linear way. 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 (see Table 1).

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. The second source of endogeneity is the fact that Wikipedia is a collaborative site: The content matter of certain pages may be subject to unobserved exogenous shocks and seasonality. Sudden spikes of interest in certain issues might lead to more authors contributing to single pages or to the entire platform. Moreover, contributions to Wikipedia continuously grow and inevitably generate some hyperlinks, so that 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. Observationally equivalent problems are caused by temporal variations in other unobserved factors such as authors' idiosyncratic preferences, or article popularity in general, which influence both content creation and links. Note that measuring the position of articles based on a two-mode author-article network suffers from similar problems.

Similar to Kittur and Kraut (2008) and Ransbotham et al. (2012) we can tackle two of these three problems with a balanced panel structure. Specifically, 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. Several other variables come to mind that could be included as controls. However, fundamental differences between pages in the averages (of levels and growth rates) of other relevant variables, such as the number of references, the number of distinct authors that contributed in the past, will all be captured by the page fixed effects, which we include in every regression. We, therefore, opted for a succinct specification with only three controls. Since the data allow observing an article's network position in a panel design, we can effectively tackle the first

two sources of endogeneity, 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. The ideal experiment would exogenously add or remove randomly selected links between pages and compare the periods before and after such a treatment. Unfortunately, such an experiment would violate the guidelines of Wikipedia, and we are also not aware of any completely exogenous and quasi-experimental source of variation of articles' network position on the kind of large-scale observational data from Wikipedia we are analyzing.[13]

However, we are able to implement a research design, where we look at a large variation in network position that we consider as random and analyze the growth of that article before and after this event. To do so, we make use of a special type of pages. These are the articles that are initially disconnected from the main cluster of the category 'Economy' and that got connected in our period of observation. 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 displaying 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 'Economy' 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.

When looking at the articles that get connected to the category, we examine both the effects on the level of the page length and on the growth in page length. If we find a significant effect of getting connected on page growth, we consider it to be unlikely to rely on systematic correlations between connectedness and the error term, since this unobserved effect on the error term would have to coincide with connectedness not only

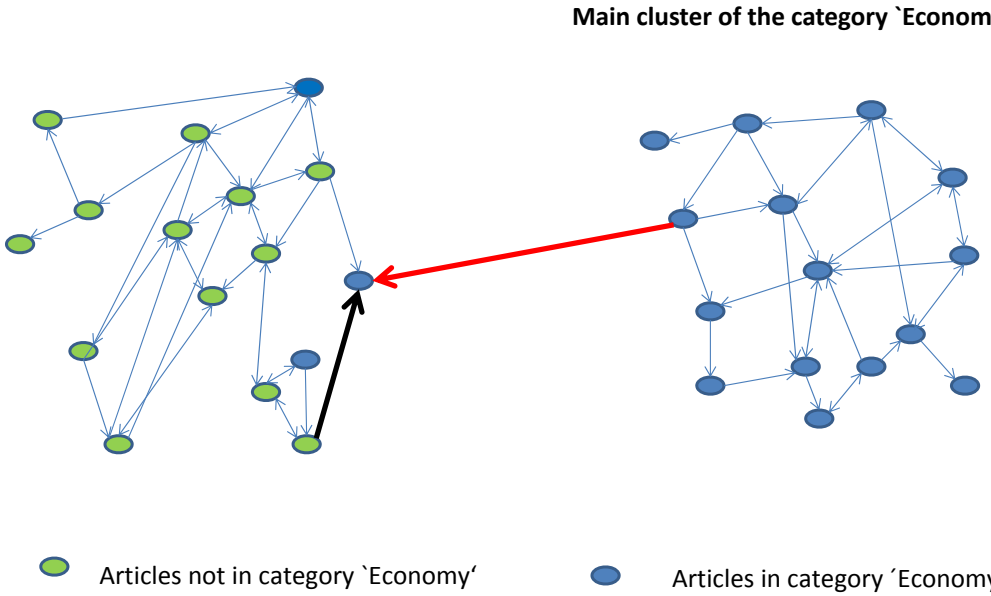
[13] Absent exogenous variation of the network position, Kummer (2013) takes an alternative approach that exploits the effect of quasi-experimental attention peaks on single pages to neighbors. This approach offers many useful insights, which are derived from restricting attention to specific clusters of articles, and shorter time spans. Our paper complements these insights by studying network position and content generation on a large set of connected articles and over a longer time horizon.

in the period of getting connected but also in future periods.

5.2 Getting Connected to the Category 'Economy'

In order to analyze the effect of becoming part of the connected component in the category 'Economy,' we put together a sample that includes articles that are at first not connected, but become connected to 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 and then consider only the last occurence in our sample.

Figure 2: An article connects to the main cluster of the category (illustration).



NOTES: The figure shows a schematic illustration of getting connected to the economics cluster. The black link is from outside the cluster, the red link establishes a connection to the cluster. Importantly, the link has to be embedded into the text on the other page, but leaves the linked article untouched. The corresponding analysis assumes that it is uncorrelated to the error term whether a new link connects to the cluster or not.

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 $t \in \{-5, ..., 5\}$.
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 10). 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.

6 Results of the Econometric Analysis

Table 7 shows the two-way fixed effects regressions corresponding to Equation 1. Page length is regressed on centrality variables, article fixed effects, and time fixed effects. The table shows the result for $7,635$ articles from the category 'Economy' 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.

One of our main questions 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 'Economy' and further augment the specification with the relative rank in closeness both inside the category and on the entire Wikipedia. The links from the entire Wikipedia represent a subset of the global links, and their estimated 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

139 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.

The relative closeness rank in Column 2 measures whether a page is located rather in the center of the network or rather in its periphery. Given that we scaled the rank variable such that it ranges from 0 to 100, the coefficient indicates that a ten point improvement in the relative closeness position is associated with 75 additional bytes of content. In the descriptive statistics, we saw that 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. The coefficient of the closeness rank inside the category is insignificant. The control dummies for redirects and a literature section have the expected signs. Older articles tend to be longer.

In Columns 3 and 4, we turn to the question whether the higher centrality is not only associated with more content but also with more authors. The columns show the regressions from Equation 2. They mirror the specifications from Columns 1 and 2, with now the number of authors as the dependent variable. Column 3 shows 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 the results for page length, the effect is much stronger for links from the category, as shown in Column 4: 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 a negligible and insignificant effect (Column 4).

In Table 8, we report four robustness checks of the main result for page length (from Table 1 Column 2): 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 D2:[14] 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

[14] Table D2 can be found in the online appendix, cf.:
https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrdW1tZXJzd29ya2luZ3BhcGVyc3xneDo0NzdjMjFiMzlmN2YwYTU5

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 'Economy.' As in our main sample, links from within the category have a much stronger effect on page length (in sum, three times larger than 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 the category 'Economy.' Column 4 finally reports how the results change when including a proxy for how often a page was clicked in the last week. Clicks and length are positively related, but the relationship of an article's network position and its length remains unaffected by the inclusion of this variable.[15]

Summing up our results for the connected component in the category 'Economy,' 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 9 and 10 use the information on the $1,237$ pages that get connected to the main cluster of economics during the period of observation. Table 9 shows the results when we consider 5 periods before and after the jump, also including 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 (Columns 2 and 4), ranging from 66

[15] 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 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 all effects are in the same direction and continue to matter (though sometimes less). The results are available from the authors upon request.

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 10 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.

7 Conclusions

The creation of user-generated content in a peer production setting requires mechanisms that help producers to identify where they want to contribute. We consider the network of hyperlinks between Wikipedia articles as a possible channel of spillovers in production activity 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 it receives, despite controlling for time-invariant unobserved heterogeneity, time effects and other variables.

In our sample of articles on economic topics, one more link is associated with 13 bytes more text, which corresponds to one or two words. When differentiating between links within the category 'Economy' and links from other Wikipedia pages, we find a large discrepancy in the effects. One more link from an article from inside the category is related to an 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 become 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 contributions 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 all 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. For the two categories we were able to scrutinize, 'Economy' and 'Sociology,' we find that semantical relatedness matters more than the mere presence of direct links between pages in generating spillovers in content provision. Note, however, that our categories are closely related to academic disciplines. Whether our findings generalize to other categories, such as sports, fashion or non-academic topics is beyond the scope of this study, and remains to be shown in future research.

Beyond econometric evidence, our analysis is to our knowledge the first to show how economics as a discipline is represented on Wikipedia. Among the longest articles and those attracting the most contributors, we find a high share associated with Marxian economics and other heterodox views. This tendency is less observed among the most clicked articles and absent from the most central articles within the category 'Economics.' Extrapolating these observations to hypotheses about whether hyperlinks may also affect the dominating views to which authors contribute is not possible based on the top 20 articles only. Future research could explore the potential of hyperlinks to channel different thematic contributions using semantic analysis.

An important limitation is the degree of freedom for authors to affect the link structure. While strategic linking itself is not allowed on Wikipedia, the way authors write the content of article A (e.g., 'Marc Chagall') affects whether this article will link to article B ('Inflation') or not. Much of this linking is driven by the content of article A and not the content of article B, which represents our main dependent variable. However, we cannot completely exclude endogeneity at this point, at least not our main data set. In recent research, one of the authors uses more experimental evidence and finds that a node can expect to receive a spillover in attention of approx. 25 percent of the average number of clicks on its neighbors (Kummer (2013)). We see this as complementary to the present work, which uses a sample that is much more representative of an entire category of Wikipedia. Moreover, we are able to consider a subset of article getting connected to the category 'Economy,' which we consider a large and exogenous variation in centrality.

Two further properties of our study require caution when generalizing its result: 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

Aaltonen, Aleksi and Stephan Seiler, "Cumulative Growth in User Generated Content Production: Evidence from Wikipedia," *Management Science*, forthcoming.

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.

Algan, Yann, Yochai Benkler, Mayo Fuster Morell, and Jérôme Hergueux, "Cooperation in a Peer Production Economy Experimental Evidence from Wikipedia," *Workshop on Information Systems and Economics, Milan, Italy*, 2013, pp. 1–31.

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.

Bala, Venkatesh and Sanjeev Goyal, "A noncooperative model of network formation," *Econometrica*, 2000, *68* (5), 1181–1229.

Bramoullé, Yann and Rachel Kranton, "Public goods in networks," *Journal of Economic Theory*, 2007, *135* (1), 478–494.

_ , Habiba Djebbari, and Bernard Fortin, "Identification of Peer Effects through Social Networks," *Journal of Econometrics*, 2009, *150* (1), 41–55.

Cabrales, Antonio, Antoni Calvó-Armengol, and Yves Zenou, "Social interactions and spillovers," *Games and Economic Behavior*, 2011, *72* (2), 339–360.

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.

Carmi, Eyal, Gal Oestreicher-Singer, and Arun Sundararajan, "Is Oprah Contagious? Identifying Demand Spillovers in Online Networks," *NET Institute Working Paper.*, 2012, *No. 10-18.*, Available at SSRN: http://ssrn.com/abstract=1694308 or http://dx.doi.org/10.2139/ssrn.1694308 (August 3, 2012).

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.

Dellarocas, Chrysanthos, Zsolt Katona, and William Rand, "Media, Aggregators, and the Link Economy: Strategic Hyperlink Formation in Content Networks," *Management Science*, 2013, *59* (10), 2360–2379.

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).

Galeotti, Andrea, Sanjeev Goyal, Matthew O Jackson, Fernando Vega-Redondo, and Leeat Yariv, "Network games," *The review of economic studies*, 2010, *77* (1), 218–244.

Giorgi, Giacomo De, Michele Pellizzari, and Silvia Redaelli, "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, 2010, *2* (2), 241–75.

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, Sanjeev and Jose Luis Moraga-Gonzalez, "R&d networks," *Rand Journal of Economics*, 2001, pp. 686–707.

Graham, Brian S., "Methods of Identification in Social Networks," *Annu. Rev. Econ.*, 2015, *7*, Submitted. Doi: 10.1146/annurev-economics-080614-115611.

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, Matthew O. and Yves Zenou, *Economic Analyses of Social Networks, The International Library of Critical Writings in Economics*, London: Edward Elgar Publishing, 2013.

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.

Katona, Zsolt and Miklos Sarvary, "Network formation and the structure of the commercial World Wide Web," *Marketing Science*, 2008, *27* (5), 764–778.

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.

Kummer, Michael E, "Spillovers in Networks of User Generated Content," *ZEW-Centre for European Economic Research Discussion Paper*, 2013, (13-098).

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

Lin, Xu, "Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables," *Journal of Labor Economics*, 2010, *28* (4), 825–860.

Mayzlin, Dina and Hema Yoganarasimhan, "Link to success: How blogs build an audience by promoting rivals," *Management Science*, 2012, *58* (9), 1651–1668.

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.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The PageRank citation ranking: Bringing order to the web.," 1999.

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.

Stephen, Andrew T and Olivier Toubia, "Deriving value from social commerce networks," *Journal of marketing research*, 2010, *47* (2), 215–228.

Surowiecki, James, *The wisdom of crowds*, Anchor, 2005.

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.

8 Tables

8.1 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 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

8.2 Characteristics of the Category 'Economy'

Table 4: 'Top economists' on German Wikipedia, by centrality, clicks and editing.

	Page length	No. of authors	Indegree categ.	Global indeg.	Clicks
1.	K. Marx	K. Marx	K. Marx	K. Marx	K. Marx
2.	N. I. Bucharin	H. Köhler	A. Smith	F. Engels	M. Weber
3.	S. Gesell	A. Smith	J. M. Keynes	M. Weber	J. M. Keynes
4.	D. Hume	M. Weber	D. Ricardo	H. Köhler	F. Engels
5.	J. S. Mill	J. M. Keynes	M. Weber	D. Hume	A. Smith
6.	F. Oppenheimer	M. Friedman	M. Friedman	F. Lassalle	M. E. Porter
7.	A. Smith	F. A. v. Hayek	J. Schumpeter	J. S. Mill	P. Bofinger
8.	F. A. v. Hayek	F. Lassalle	F. Engels	A. Smith	F. A. v. Hayek
9.	F. Lassalle	S. Gesell	F. A. v. Hayek	N. I. Bucharin	H. W. Sinn
10.	M. Weber	H.W. Sinn	W. Eucken	J. M. Keynes	J. Bentham

NOTES: The table shows the top 10 economists by page length, number of authors, indegree from category, indegree from Wikipedia and clicks. Economists are identified via the Wikipedia label 'Economist' (Ökonom) in German Wikipedia.

Table 5: The most important subcategories in the category 'Economy.'

	Share of articles	*Median values of*					
		Age Dec 2007	Page length	No. authors	Clicks	Indegree from category	Indegree outside category
Categories of non-persons							
Economics	.1346	162	4209	15.9	10.6	6.6	4.9
Management	.2210	143	3737	16.3	16.3	4.7	5.4
Trade, enterprise	.0664	131	4357	18.2	9.9	4.9	7.5
Finance	.0840	138	3196	13.8	11.5	5.8	3.6
Banking, insurance	.1059	139	4142	14.6	8.0	4.6	4.0
Labor, poverty	.0517	153	4514	17.6	15.4	4.7	9.6
ICT	.0412	152	3860	18.8	15.2	3.3	5.2
Politics, policy	.0319	158	4767	19.7	9.4	4.6	12.2
Sociology, social matters	.0340	145	4169	16.4	9.1	3.5	7.3
Other economic topics	.0242	139	2961	14.5	6.6	3.5	5.8
Other	.0826	158	3629	17.5	12.1	4.0	8.2
Categories of persons							
Economists	.0650	134	3766	15.3	3.5	3.9	6.9
Managers or entrepreneurs	.0307	124	3399	15.3	6.3	2.0	8.3
Globalisation critiques and non of the above	.0063	153	7565	33.7	9.2	3.6	16.2
Other professors	.0061	99	3706	13.7	3.9	3.1	4.8
Other	.0144	135	3915	14.1	2.8	2.5	7.4

NOTES: The table shows the distribution of articles in our sample over subcategories and provides the median values of article age, page length, the number of authors and the number of clicks. It also shows two measures of centrality, the local indegree (number of articles within the master category 'Economy' pointing to a specific article (indegree from category) and the global indegree (all links including outside category).

Table 6: 'Top pages' in the subcategory 'Economics'.

	Page length	No. authors	Clicks	Indegree from categ.	Indegree outside categ.
1.	Marxian Economics	Money	Gross domestic product	Economics	Liberalism
2.	Primitive accumulation of capital	Poverty	Inflation	National Economy	Infrastructure
3.	Poverty	Neoliberalism	Business cycle	Inflation	Per capita income
4.	Inter-state fiscal adjustment	G8	Liberalism	Competition (econ.)	Economics
5.	Unemployment statistics	Liberalism	Depreciation	Production	Inflation
6.	Tax	Tax	Human Developm. Index	Employee	Capitalism
7.	Labor theory of value	Capitalism	Money	Capital	Trade
8.	Money	Inflation	Market economy	Tax	Poverty
9.	Capitalism	Imperialism	Social market economy	Money	Tax
10.	Mercantilism	Market economy	Government debt	Liquidity	Gross domestic product
11.	Property	Gross domestic product	Poverty	Gross domestic product	Household
12.	Underconsumption	Business cycle	Tax	Product (business)	Monopoly
13.	Harrod-Domar model	Government debt	Capitalism	Economic policy	Property
14.	Prisoner's dilemma	Property	Innovation	Cost	National economy
15.	Journal of Economic Literature	Economics	Balance sheet	Demand	Imperialism
16.	Neoliberalism	Criticism of capitalism	Imperialism	Price	Production
17.	profit rate's tendency to fall	Game theory	Deflation	Monopoly	Institution
18.	Liberalism	Prisoner's dilemma	Social insurance	Social insurance	Revenue
19.	Government debt	Keynesian economics	G8	Profit (economics)	Merger
20.	Inflation	Human Developm. Index	Circular flow of income	Income	Privatization

NOTES: The table shows the top subcategories by page length, no. of authors, indegree from category, indegree from Wikipedia and clicks. Categories are constructed in non-overlapping fashion, using a list of key labels. An article is only assigned to a category if it has not yet been assigned to a category of higher rank on the list (e.g. if the article on 'monetary theory' is already part of the category 'Economics,' it will not be assigned to 'Banking.'

8.3 Regression Results

Table 7: Relationship of page length and centrality.

	Page Length		Number of Authors	
	(1) Wiki degree	(2) all vars	(3) Wiki degree	(4) all vars
Links from Wikipedia	13.333***	2.931	0.112***	0.072***
	(3.18)	(1.22)	(4.25)	(3.23)
(Links from Wikipedia)^2	-0.000	0.001**	-0.000**	-0.000**
	(-0.54)	(2.07)	(-2.51)	(-2.04)
Links from category		135.871***		0.476***
		(8.47)		(6.38)
(Links from category)^2		-0.127***		-0.000***
		(-5.02)		(-3.18)
Rel. closeness rank (Wikipedia)		7.505***		-0.007
		(3.08)		(-1.22)
Rel. closeness rank (category)		-1.230		-0.009*
		(-0.67)		(-1.65)
Dummy: literature section	1295.963***	1248.055***	1.552***	1.406***
	(6.11)	(5.94)	(4.78)	(4.57)
Age (in months)	10.648***	8.416***	0.072***	0.064***
	(21.55)	(22.46)	(26.10)	(43.66)
Dummy: page is redirect	-546.408	-767.075	0.269	-0.434
	(-0.57)	(-0.77)	(0.13)	(-0.20)
Constant	3336.571***	2501.686***	6.127***	5.140***
	(30.10)	(15.73)	(13.05)	(13.00)
Time dummies	Yes	Yes	Yes	Yes
Observations	1168155	1168155	1168155	1168155
Groups	7635	7635	7635	7635
Adj. R^2	0.107	0.131	0.463	0.495

NOTES: 2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors). Only articles connected over entire period were included. Dependent variables: page length (col. 1 & 2) and the number of authors (col. 3 & 4). t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01.

Table 8: 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

NOTES: 2-way fixed effects OLS regressions with both time and article dummies (robust SE). Only articles connected over entire period were included. t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01.

Table 9: 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

NOTES: 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; t statistics in parentheses.

Table 10: Effect when page becomes connected, excluding the periods (+/- 2) around the jump.

	(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

NOTES: 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. 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; t statistics in parentheses.