Measuring wiki viability*

An empirical assessment of the social dynamics of a large sample of wikis

Camille Roth
CAMS
CNRS/EHESS
96, Bd Raspail
F-75006 Paris
France
camille.roth@polytechnique.edu

Dario Taraborelli
CRESS
Department of Sociology
University of Surrey
Guildford GU2 7XH
United Kingdom
d.taraborelli@surrey.ac.uk

Nigel Gilbert
CRESS
Department of Sociology
University of Surrey
Guildford GU2 7XH
United Kingdom
n.gilbert@surrey.ac.uk

ABSTRACT

This paper assesses the content- and population-dynamics of a large sample of wikis, over a timespan of several months, in order to identify basic features that may predict or induce different types of fate. We analyze and discuss, in particular, the correlation of various macroscopic indicators, structural features and governance policies with wiki growth patterns. While recent analyses of wiki dynamics have mostly focused on popular projects such as Wikipedia, we suggest research directions towards a more general theory of the dynamics of such communities.

Categories and Subject Descriptors

H.5.3 [**Information interfaces and presentation**]: Group and Organization Interfaces—*Web-based interaction, Collaborative computing, Evaluation/methodology*; K.4.3 [**Computers and society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Human Factors, Measurement

Keywords

wikis, web 2.0, online communities, governance, moderation, metrics, dynamics, viability.

1. INTRODUCTION

Mapping the wikisphere

Online communities have demonstrated their potential to leverage a vast amount of collaboratively contributed content. Famous examples include large open-source software development projects such as Mozilla Firefox or Linux [5] and wiki-based encyclopedias such as Wikipedia [4]. However not all projects achieve such successful outcomes. The destiny of an online community often relies on the capacity of its maintainers to attract new members, to develop policies to secure the commitment of contributors, and to promote high quality standards for its output.

The question of forging a sustainable community of active contributors is particularly crucial for individuals, companies and organizations willing to launch and develop wiki-based projects. This issue relates more broadly to the understanding of the general dynamics of content-based communities and, therefore, calls for research on a wide range of wikis at various stages of development. Previous studies have mainly focused on Wikipedia [10, 9]. While of great interest, it seems difficult to build on this knowledge to develop a more comprehensive theory of the social structure and content dynamics of wikis—Wikipedia being a special case in many respects (population, maturity, topical range, and media attention, among other factors). A comprehensive analysis of wiki dynamics is further hindered by the heterogeneity of wiki platforms and by the lack of tools and methods to collect data in a standardized way. Nevertheless, a first approach to generalization can be achieved by examining a set of wikis that share the same platform.

In the following sections, we present an analysis of the evolution of a large sample of wikis based on the MediaWiki engine. In particular, we aimed to identify factors that correlate with growth, as a preliminary step towards a full-fledged understanding of factors that determine different patterns of evolution in the wikisphere.

Wiki ecology: demographics and viability

The present study is, to our knowledge, the first longitudinal analysis of the content and population dynamics of a large set of wikis over time using data retrieved via an API-based service. As well as almost always focusing on Wikipedia, previous quantitative wiki research has mainly examined the topological structure of underlying interaction or hyperlink networks [3, 12] or article-level features [1, 11], with little interest in the specific dynamics of the demographic determinants themselves (with the exception of [7] which investigates Wikipedia's demographics of casual *vs*. committed contributors). In a previous study of a static dataset describing wikis we provided a glimpse of the demographic structure of a portion of the wikisphere in terms of both population and content sizes, but stopped short of investigating its evolution.[8]

As content-based online communities, wikis mostly vary in two dimensions: **(i) contributors**, who may or may not constitute an active community (as described e.g. in [2]); and **(ii) pages**, which may or may not amount to authoritative or useful content (as demonstrated for example by [4]). Users and pages are likely to obey a

*This work was supported by the EC-sponsored PATRES network (NEST-043268). We are grateful to *s23.org* for giving us the permission to harvest their MediaWiki statistics database and to four anonymous reviewers for valuable feedback on a previous version of this work.

WikiSym '08, September 8-10, Porto, Portugal.

Variable	*Quantile*				
	1	**2**	**3**	**4**	**5**
edits per user	$[0.14, 3.67[$	$[3.67, 9.80[$	$[9.80, 24.8[$	$[24.8, 61.2[$	$[61.2, 1903]$
admins per user	$[.00169, .00192[$	$[.00192, .00347[$	$[.00347, .00576[$	$[.00576, .01[$	$[.01, 1.45]$
edits per page	$[0.12, 9.2[$	$[9.2, 14.3[$	$[14.3, 21.8[$	$[21.8, 35.1[$	$[35.1, 47\,245]$
admins per page	$[4.16 \cdot 10^{-6}, .00103[$	$[.00103, .00309[$	$[.00309, .00926[$	$[.00926, .0299[$	$[.0299, 2.65]$
users per page	$[.00119, 0.22[$	$[.220, .728[$	$[.728, 2.07[$	$[2.07, 7.05[$	$[7.05, 363]$
	phase diagram, boundaries: {0.00119497, .116, .307, .619, 1.24, 2.44, 4.99, 12.1, 363}				

population quantiles		**1**	**2**	**3**	**4**
	mean quantiles:	$[400, 615[$	$[615, 1075[$	$[1075, 2407[$	$[2407, 19909]$
	phase diagram, mean boundaries over all 'users/page' quantiles: {400, 504, 676, 949, 1313, 2162, 3787, 19 909}				

Table 1: Quantile boundaries and sets in the *clean* wiki dataset.

dual dynamic: while more users may contribute to more pages, content proliferation in turn seems to require more attention from users. As a first approximation, it may thus seem judicious to assess the healthiness of a wiki through these variables, taken as demographic proxies for its actual growth and activity.[1]

2. EMPIRICAL PROTOCOL

2.1 Dataset

We constructed a dataset made of simple statistics gathered for a large number of MediaWiki-based wikis,[2] which enabled us to consider the same set of variables across all wikis and make sure these variables were generally available. The data was collected over the period August 2007–April 2008 from a publicly-available database[3] totaling 11 500+ wikis.

2.2 Variables

We considered a set of four quantitative variables:

- *population size* (U), measured by the number of *registered* users;
- *content size* (P), measured by the number of so-called "good" pages (i.e. actual *content* pages excluding default pages created by the wiki engine), hereafter indifferently called "pages", "good pages" or "articles" ;
- *administrator population* (A), or the number of users who are granted "administrator status", i.e. special rights to modify sensitive content and perform maintenance activity;
- *editing activity* (E), measured by the total number of edits;

as well as one qualitative variable indicating the presence of an access control mechanism:[4]

- *editing permission* (R), or editing access control, i.e. the possibility of creating a page for unregistered/anonymous users. $R \in \{0, 1\}$, where $1 \equiv$ "anonymous editing allowed".

However simplistic these variables may be, they provide key indicators of the global dynamics of a wiki, and shed light on diverse aspects of its structure and evolution. We collected the values of these variables for each wiki on a daily basis and over a period of 250 days, i.e. approximatively 8 months.

[1] It should be noted that raw content growth per se may not be a good indicator of a sustainable wiki, as studies on wiki proliferation seem to suggest. [6]

[2] This initial dataset includes among others a large set of Wikipedias.

[3] Available from *http://s23.org/wikistats/largest_html.php*. The database is maintained by a user called "Mutante" who graciously granted us the permission to automatically harvest this data.

[4] This indicator was not part of the original dataset. To obtain this specific information we crawled each wiki in the dataset, at the beginning of the study, using a robot that attempted to determine whether page creation was possible without prior user registration.

2.3 Scope restrictions

Platform exclusions

A large number of wikis in the database are based on *wikifarms*, i.e. platforms hosting several wikis and providing services easing wiki creation and management. Some of these platforms return system-wide rather than wiki-specific figures when queried for wiki statistics, resulting in spurious data. To avoid this, all wikis explicitly hosted on platforms exhibiting this behavior (such as wikis hosted on *wikia.com*) were excluded from the dataset.

Population range

In [8] we observed, on a similar but static dataset, that a large majority of wikis are both thinly populated (i.e. often less than 10 users) and/or do not have a significant number of pages (i.e. generally less than 10 pages). In the present study, in order to focus on a relatively homogeneous wiki population, we excluded both wikis with very large populations and those with very few users, so as to avoid basing the analysis on data spanning too many orders of magnitude in terms of population. Included in the dataset, therefore, were only wikis whose user population was within the range $[400, 20\,000]$ on the first day of data collection (note that in this region, content sizes are widely spread, from a few to hundreds of thousands of pages).

Growth discontinuities

Some wikis experience abrupt changes of one or even several orders of magnitude in population or content size. There are many possible causes of such changes, including spam attacks, and administrative decisions to transfer, create, merge or suppress pages in bulk or admit or ban classes of user.[5] To exclude wikis exhibiting such a "suspicious" behavior, on the assumption that no recruitment of users or creation of pages of a significant magnitude could "naturally" happen within a period as short as 24 hours, a threshold percentage of change α between two successive days, for both population and content sizes, was set. Wikis whose daily growth in either content or users was above the threshold were excluded from the dataset. α was set to 0.05, i.e. wikis were excluded if they had ever experienced more than a 5% increase or decrease in users or pages over a period of 24 hours.

Clean dataset

To sum up, the final, "clean" dataset is thus made up of about 360 wikis, all of which have an initial population between 400 and 20 000 users, are not hosted by 'wiki farms' that do not report useful data, and which have no major discontinuity in the daily change

[5] We acknowledge that cases of spam attack are evidence that wiki sustainability is already highly damaged. However, without a method for systematically distinguishing these cases we prefer to leave aside this portion of the dataset for the sake of the present analysis.

of their population or content.

3. WIKI DYNAMICS

We assessed wiki dynamics by comparing their diverse paths with respect to a set of independent variables. 'Growth' is defined in terms of population and content size change: user growth G_U (resp. page growth G_P) is the ratio between final and initial populations (resp. content sizes): $G_U = U_{\text{last}}/U_{\text{first}}$ (resp. $G_P = P_{\text{last}}/P_{\text{first}}$). Wiki dynamics were studied as a function of the variables listed in Section 2.2:

(I) DESCRIPTIVE INDICATORS, i.e. variables on which wiki administrators have no direct control: **(a)** *user activity*, i.e. the proportion of edits per user (E/U), **(b)** *user density*, i.e. the proportion of users per page (U/P), and **(c)** *edit density*, i.e. the proportion of edits per page (E/P).

(II) GOVERNANCE FACTORS, variables that wiki administrators can directly control: **(a)** *administrator ratio*, i.e. the proportion of users who are granted administrator status (A/U), **(b)** *administrator density*, i.e. the proportion of administrators per page (A/P), **(c)** *editing permission* (R).

For each continuous variable, instead of carrying out a delicate analysis by dealing with clouds of points, we adopted a more insightful approach by dividing wikis into five quantiles, each including exactly 20% of all wikis in the clean dataset (see Table 1). We then computed and compared growth ratio means over all wikis for each quantile. Additionally, we distinguished population quantiles in order to control for user size-related effects. To this end, we plotted a growth landscape that consists of a two-dimensional representation of the various growth ratios. This representation was applied to all the above-mentioned variables, except for R where there are only two "quantiles" (0 or 1). For each variable except R, the upper graphs indicate the mean values and confidence intervals ($p < 0.05$) of each quantile on the variable considered, while the lower graphs show contour plots for the same variable with brighter areas corresponding to higher growth ratios.

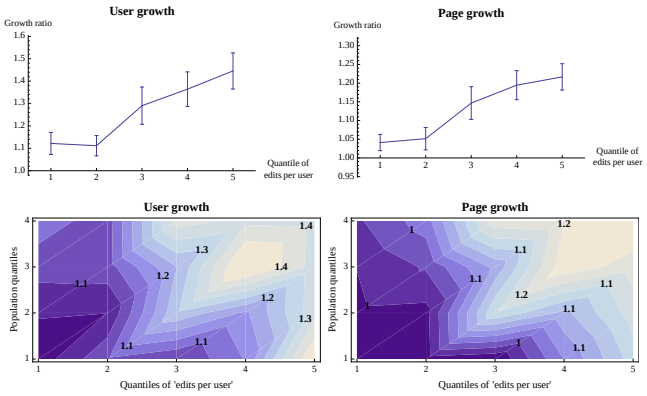

Figure 1: Growth landscape with respect to the proportion of *edits per user*.

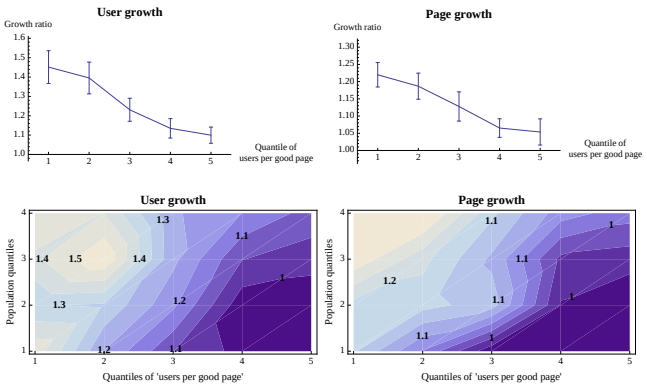

Figure 2: Growth landscape with respect to the proportion of *users per good page*.

3.1 Significant descriptive indicators

We found significant correlations between a number of descriptive indicators of wiki structure and their content and population growth rates.

Figure 1 shows the effect of **user activity** (measured as the proportion of *edits per user*) on growth rates. The results suggest that user activity correlates very strongly with wiki growth, not only in terms of content production (which is to a certain extent unsurprising) but also new member recruitment. The effect becomes stronger with initially more populated wikis: the more users are actively editing, the more a wiki grows in content *and* population.

Figure 2 shows the impact of **user density** on growth. The results suggest that a higher number of contributors per page does not necessarily indicate mushrooming wikis: for an identical content size, we found a significant correlation between a lower number of users and higher growth ratios, both in content and new members.

To better visualize the effect of user density on growth, we represented the dependent variables G_U and G_P, independent vari-

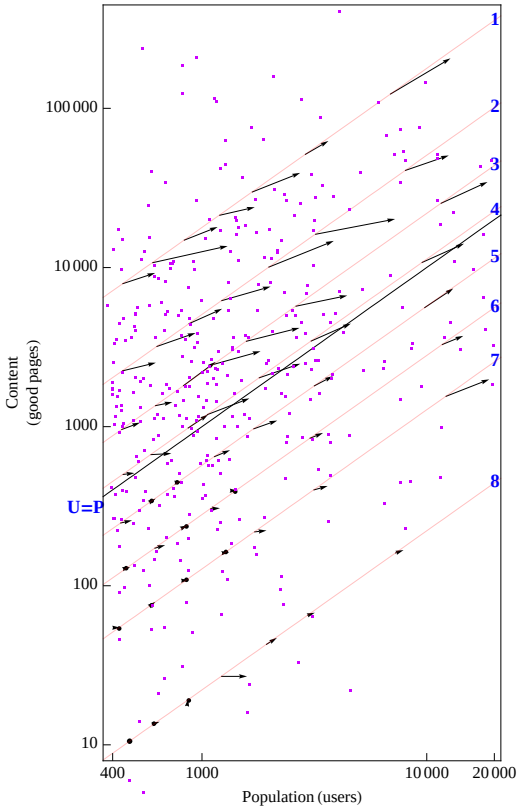

Figure 3: Phase diagram in the content/population space of wikis belonging to the clean dataset.

able U/P, and initial wiki positions U_{first} and P_{first} altogether on the same graph, yielding a phase diagram as plotted on Fig. 3.[6] In this diagram, each dot (light color) corresponds to a wiki in the database. Each arrow corresponds to a pair of quantiles "users per page, population". Widths and heights are proportional to user and content growth ratios, respectively. The size of the arrow represents the strength of the observed growth in content and population for wikis in a given region of the wikisphere.

This graph should be regarded as a *map* of a portion of the wikisphere, showing the expected destiny of a wiki in terms of content and population growth as a function of its initial position in the same space. This diagram broadly suggests that a wiki's position is correlated with its subsequent fate. More precisely, it illustrates that wikis in the upper/upper-right portion of the diagram are growing faster, and more interestingly it provides an overview of demographic dynamics in this region of the wikisphere.

3.2 Significant governance factors

Turning to governance features, we first analyzed the effects of the **administrator density** on wiki dynamics by looking at the overall proportion of administrators per page.

Figure 4 shows that having a relatively high number of administrators for a given content size is likely to reduce growth. There is a strong effect of the proportion of admins per page both on user and page growth. For instance, while the last quantile of admins/page ratio enjoys near-zero growth rates over 8 months, the first quantile tops overall rates (∼+50% for users, ∼+25% for pages). This effect may be interpreted as the impact of strong governance activity on the proliferation of content and users.

[6] For this diagram, an increased level of detail called for a larger grid, here of 8×7 quantiles; U/P quantile means are represented by diagonal straight lines labelled "1–8".

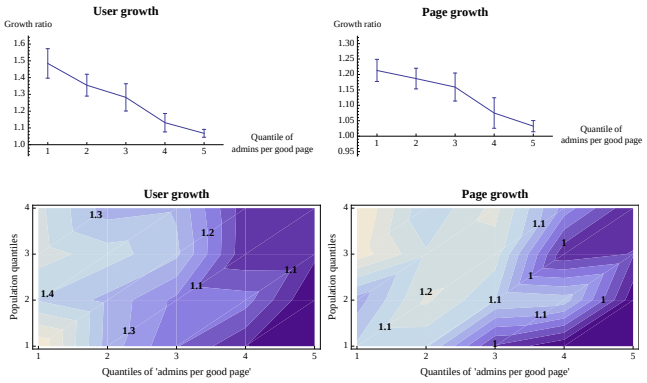

Figure 4: Growth landscape with respect to the proportion of *admins per good page*.

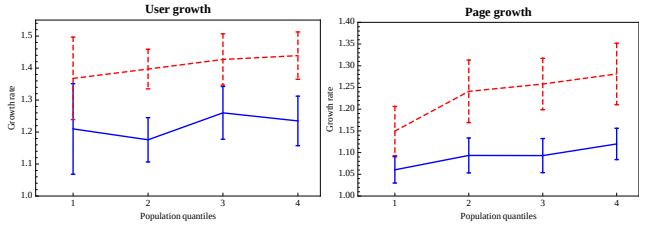

Figure 5: Growth landscape with respect to *editing permission*: *red dashed* refers to anonymously editable wikis, while *blue solid* to wikis editable by registered users only.

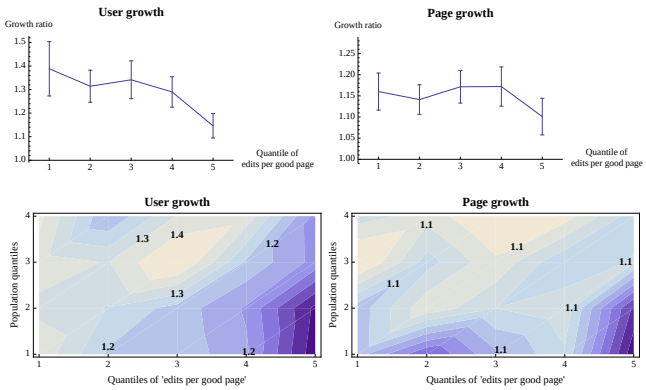

Figure 6: Growth landscape with respect to the proportion of *edits per good page*.

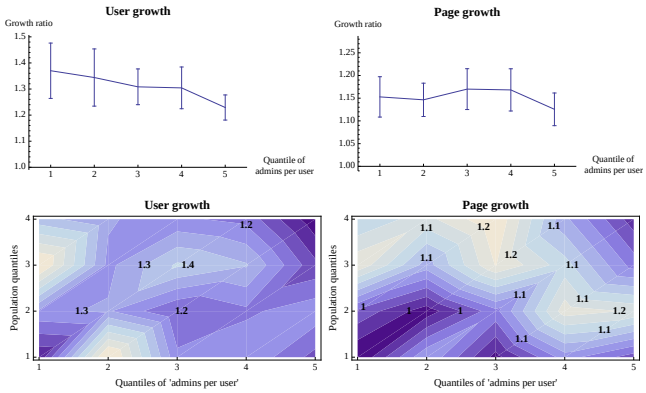

Figure 7: Growth landscape with respect to the proportion of *admins per user*.

We identified another significant effect when we considered **editing permission**. As a binary variable, the editing permission variable generates only two groups of wikis (wikis that allow anonymous editing *versus* wikis that restrict editing to registered users only). The growth landscape is consequently limited to a one-dimensional comparison over population quantiles. The results in Figure 5 show that for both dimensions—population and content—having no access control is likely to favor growth. While a stronger page growth is quite unsurprising in wikis where no registration is required, the fact that this factor also fuels user registration is more puzzling. One might expect that if users can participate without the need of registration, few would be inclined to register. Our results suggest on the contrary that wikis with unrestricted registration trigger participation more easily than wikis that restrict access.

3.3 Neutral indicators

Finally, we consider two indicators that showed a markedly milder correlation with wiki dynamics.

On the one hand, we found that **edit density** (i.e. edits/page) correlates in a moderately negative way with user growth—with a relatively stronger effect depending on initial population size—while there is surprisingly no significant correlation with page growth (Figure 6).

On the other hand, higher **administrator ratios** (i.e. admins/user) have no significant effect on content or population growth, as evidenced by the contour plot on Figure 7.

	Variable	*Growth rate*	
		Population	Content
STRUCTURAL INDICATORS	User activity (E/U)	+ +	+ +
	Edit density (E/P)	-	—
	User density (U/P)	- -	- -
GOVERNANCE FACTORS	Editing permission (R)	+ +	+ +
	Admin ratio (A/U)	—	—
	Admin density (A/P)	- -	- -

Table 2: Effect of different factors on wiki growth rates.

3.4 Summary of findings

The results of this study suggest that different structural and governance-related factors have significant effects on the content and population dynamics of a wiki. Table 2 and Figure 8 summarize the correlations found between growth ratios and each of the variables we considered, by comparing the gain in the population and content sizes between the last and the first quantile for each variable (variables in Figure 8 are ranked from the most positively to the most negatively correlated).

If we focus on *structural aspects* of wikis, we note that the higher the ratio of *edits per user* the faster the wiki grows, both in terms of content and population. Wikis with very active user communities are not only likely to grow in content, but also to attract a large number of new contributors. This result contrasts with the opposite effect produced by high user density per page.

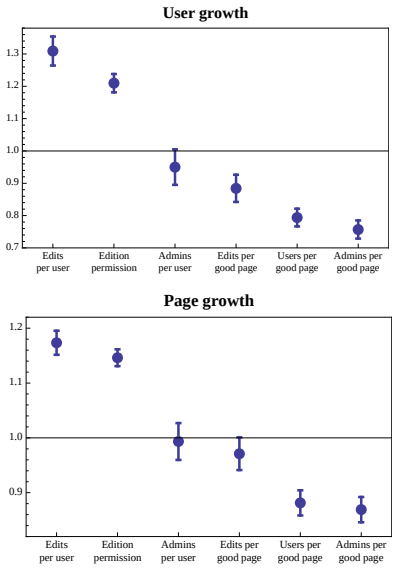

Figure 8: Comparison of growth ratios between the last and first quantiles, for each variable considered.

As far as *governance factors* are concerned, we observed the singular fact that population growth is in average more than 20% faster for anonymously editable wikis. This seems to support the intuition that less barriers favor population growth. Furthermore we observed that, while too many administrators per page may hinder the growth of a wiki (in terms of content size), the proportion of administrators per user does not appear to show a significant influence on growth. In all the above cases, we observed a striking correlation between content and population growth.

4. CONCLUDING REMARKS

The main outcome of this study is an account of the factors that wiki communities should take into account in order to control their demographics. In this respect, we showed the remarkable dynamical intertwinement of population and content growth, which suggests that models of wiki dynamics will probably need to focus on the strong interrelations between these two variables.

Representing via phase diagrams the impact of specific variables on wiki dynamics can be a valuable solution for wiki administrators for monitoring purposes and for social scientists as a first step towards modeling. However, in order to develop accurate models of wiki dynamics, further empirical evidence is needed. To make the data tractable for this study, we restricted the dataset in several ways. A more comprehensive study, beyond the scope of the present paper, should endeavor to investigate a larger spectrum of wiki-based communities.

5. REFERENCES

[1] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 179–186, 2007.

[2] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Group'05, Sanibel Island, FL, USA*, Nov 6-9 2005.

[3] A. Capocci, V. Servedio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia. *PRE*, 74(3):036116, 2006.

[4] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[5] M. W. Godfrey and Q. Tu. Evolution in open source software: A case study. *ICSM*, 00:131, 2000.

[6] H.-J. Happel and M. Treitz. Proliferation in enterprise wikis. In *Proceedings of the 8th International Conference on the Design of Cooperative Systems (COOP 08)*, Carry-le-Rouet, France, May 2008.

[7] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *ALT.CHI*, 2007.

[8] C. Roth. Viable wikis: struggle for life in the wikisphere. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 119–124, New York, NY, USA, 2007. ACM.

[9] B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1037–1040, New York, NY, USA, 2008. ACM.

[10] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with *history flow* visualizations. In *CHI '04: Proceedings of the 2004 conference on Human factors in computing systems*, pages 575–582. ACM Press, 2004.

[11] D. Wilkinson and B. Huberman. Assessing the value of cooperation in Wikipedia. *First Monday*, 12(4), 2007.

[12] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *PRE*, 74(1):016115, 2006.