Wikipedia: A Complex Social Machine

Ramine Tinati
University of Southampton, UK
and
Markus Luczak-Roesch
University of Victoria, New Zealand

Wikipedia represents a successful peer-produced knowledge-resource constructed via the endeavours of millions of volunteers. We examine the activity of Wikipedia by analysing WikiProjects, an community-driven feature which allows communities of Wikipedians to coordinate their efforts in order to improve or produce Wikipedia articles. We harvested the content of over 600 active Wikipedia projects, which comprised of over 100 million edits and 15 million Talk entries, associated with over 1.5 million Wikipedia articles and Talk pages produced by 14 million unique users. Our analysis reveals findings related to the overall positive activity and growth of Wikipedia, as well as the connected community of Wikipedians within and between specific WikiProjects. We argue that the complexity of Wikipedia requires metrics which reflect the many aspects of the Wikipedia social machine, and by doing so, will offer insights into it's state of health.

1. INTRODUCTION

Wikipedia is one of the largest, global peer-produced knowledge resources on the Web. Consisting of over 12 million articles, available in over 270 languages, it is a clear example of how a social machine where the collaborative efforts of millions of volunteers can produce a highly useful resource for mankind [Hendler and Mulvehill 2016]. Wikipedia demonstrates many successful features of crowdsourced knowledge and expertise – socially, politically, and technically [Tinati et al. 2013]. Furthermore, the growth of Wikipedia has been organic in nature, with many of the newly introduced features being community-driven, with very little changes to the core platform technology.

Wikipedia has received much attention over recent years given its impact as a highly successful Web system, and its mainstream role in society [Kuznetsov 2006; Kamps and Koolen 2009; Kittur et al. 2007; Kittur and Kraut 2008; Liao 2008]. Common amongst such research is the general desire to better understand the activity, participation and retention of Wikipedians, and to better understand the community that engage with the platform. Most recently, research has begun to suggest that Wikipedia could be in a state of decline, or at the very minimum, no longer growing [Suh et al. 2009; Halfaker et al. 2013; Jullien et al. 2015]. This is particularly pressing issue, given the role that Wikipedia plays on the Web (e.g. Google's knowledge graph uses Wikipedia data).

Interested in understanding the stability of Wikipedia, in this article we examine Wikipedia beyond the traditionally used metrics. Our study focuses on the pages of 618 active WikiProjects which are associated with 1.5 million articles, with a related set of 1.5 million Talk pages. WikiProjects represents a community-driven feature, aimed to improve the breadth and quality of Wikipedia articles, for many domains and areas of interest. By using WikiProjects as a proxy of Wikipedia activity, we expose the activities of the sub-communities that have emerged as editors and maintainers of the Wikipedia corpora.

2. BACKGROUND AND RELATED WORK

The popularity *and* availability of Wikipedia research data [Stuckman and Purtilo 2009] has led to a rich set of interdisciplinary literature investigating how various parts of the Wikipedia ecosystem function. Studies have investigated of Wikipedia's structure and article connectivity [Buriol et al. 2006; Capocci et al. 2006; Stuckman and Purtilo 2009], finding power law characteristics [Lam and Riedl 2009], and reflecting structural properties of the Web graph [Barabasi and Albert 1999].

There is also a growing interest into the individuals who contribute to Wikipedia; studies have explored the motivations for participation, as well as their social interactions, such as collaborative editing behaviour[Kittur et al. 2007]. Examining how the community interacts (e.g. conflicts) has provided insight into virtual collaboration and coordination, offering recommendations for interface and platform functionality. Studies have been conducted on the collaborative process of article creation, examining the quality of articles based on the underlying social processes [Liu and Ram 2011], and developing metrics to measure article quality based on the conflicts within these processes [De la Calzada and Dekhtyar 2010]. Studies have examined barriers to adoption, and the social processes of an articles lifecycle [Hautasaari and Ishida 2012], as well as the effects of culture within the collaborative environment, identifying how Wikipedia is far from culturally neutral, which directly influence the collaborative efforts in article creation [Pfeil et al. 2006], and how external factors such as political, regional, or linguistic differences affect the policy and governance of Wikipedia in different countries [Liao 2008].

2.1 WikiProjects

A WikiProject is a community driven feature introduced in Wikipedia which allows a collection of contributors to work together as a team to improve a set of Wikipedia articles, usually related to a specific area of interest or domain. Prominent examples of this are WikiProjects such as 'WikiProject Medicine' and 'WikiProject Military History', both containing thousands of contributors, working on many Wikipedia articles. From a social and governance perspective, there are no formal guidelines on how a WikiProject should be run, and there are no privileges for those contributing to a WikiProject. Furthermore, the Wikipedia articles which a WikiProject is contributing to is also part of the main corpus of Wikipedia articles, and an article can be part of more than one WikiProject.

Whilst WikiProjects are fairly unknown to non-Wikipedians, the study of the WikiProjects phenomenon has begun to provide insight into the smaller-scale team coordination and collaboration in Wikipedia [Morgan et al. 2013], and the motivations that driven these teams to contribute [Farič and Potts 2014]. Unlike large-scale studies of Wikipedia, study-

ing WikiProjects as individual or comparative set of communities offers a finer level of granularity in terms of why individuals contribute to, and the internal social processes that enable this to occur.

In this study we build on studies investigating the emergent socio-technical features of Wikipedia, with a particular interest in the role of WikiProjects for improving the general corpus of Wikipedia articles. We also take into consideration the existing studies suggesting that discussion and 'Talk' plays an important role in the production of content, and in the co-ordination of contributors.

3. EXPERIMENT DESIGN

The analysis uses data containing of over 1.6 million Wikipedia article pages and their corresponding Talk page (limited to the English language subset). In total we harvested over 3.2 million unique articles, which are associated with 618 active WikiProjects; dormant and retired projects were not include these in the dataset.

For each Wikipedia article, we harvested a historic log of all revisions (edits). Each revision was related to a user, a timestamp, and the changes made. If the user was not logged in when the edit was made, then the user's IP address is recorded, and 'anon' is appended to the entry. In addition to harvesting an article's edit log, we also extracted the corresponding 'Talk' page if available (not all articles have a Talk page with entries). Commonly, a Talk page will contain the discussions of Wikipedians, often related to the creation, modification, and conflicts of the Wikipedia article. For each Talk entry, the user, the timestamp, and the comment made is collected. In order to refine our harvesting strategy, we generated a list of all articles and Talk pages related for each WikiProject, identified duplicates (a page can be associated with more than one project), and then harvested the revisions and Talk entries.

In addition to the articles associated with the WikiProject, we also harvested the project's home (*root*) page. This page resembles a typical Wikipedia article structure, but contains specific content related to a project, including, contributors, articles associated with, and the 'project goal'.

3.1 Data

Table I shows a overview of the data collected for this study. This data was collected during November 2014. We have provided this data in a structured for, freely available via the Southampton Web Observatory [1].

4. RESULTS

We compared the 618 WikiProjects with respect to the set of articles, Talk contributions, and users, associated with a given WikiProject. We also computed the author co-relations between users who have made an edit entry on an article page as well as the corresponding

[1] WikiProject dataset available via the Southampton Web Observatory: **http://webobservatory.soton.ac.uk/**

Feature	Value
WikiProjects	618
Unique Wikipedia article Pages	1,673,826
Unique Wikipedia Talk Pages	1,051,062
Revisions Made	109,185,140
Talk Entries	15,310,649
Unique Wikipedians	14,993,913
Registered Users	2,544,224
Anonymous Users	12,449,689
First article Revision	2001-01-20
First Talk Entry	2001-02-06

Table I. General Dataset Overview

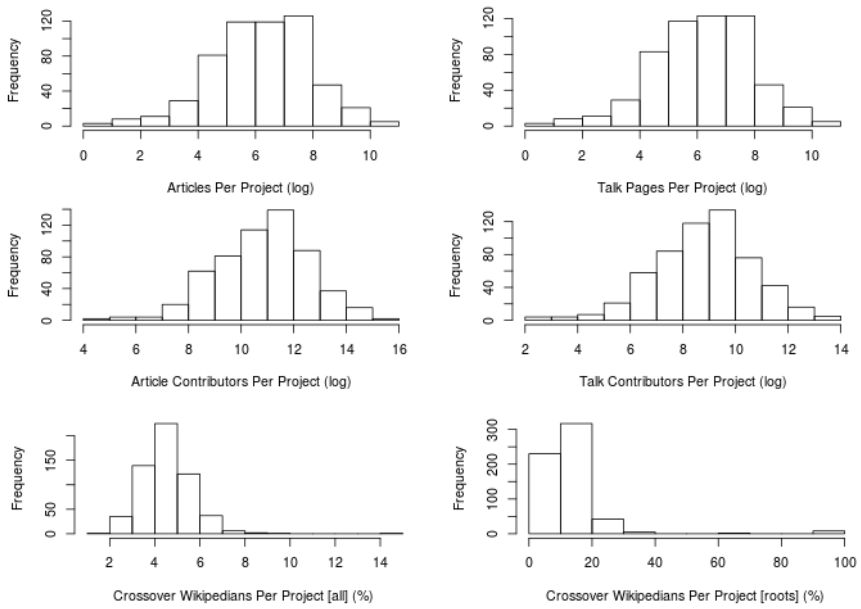

Fig. 1. Distribution of articles, Talk Entries, and Contributions for All WikiProjects. Crossover (Editors and Talkers) Wikipedians For All and Root Pages of a WikiProject

Talk page (we label this the *crossover* users). In addition to this, we also examine the co-relationship of users between projects.

To examine the role that WikiProjects played in the editing and Talk activity, we extracted the set of users who have made entries on the 'root' WikiProject set of pages and corresponding Talk page. The root pages represent the 'homepage' of a project. We also computed statistics for the total set of pages related to a project. Table II provides an overview of the statistics generated for this analysis.

Figure 1 presents the distribution of statistics generated in Table II, specifically, the article and Talk pages per project, the number of Wikipedians per project, and the proportion of Wikipedians in a project who edit as well as Talk (*crossover users*). Findings reveal a normal distribution between WikiProjects (number of article and Talk contributions), and

Measure	All Project Pages	Root Pages of Project)
Avg # of articles	1,718	1
Avg # of Talk Pages	1,700	1
Avg # of article Entries	177,533	453
Avg # of Talk Entries	25,028	234
Avg # of Wikipedians	43,036	117
Avg # of article editors	40,294	117
Avg # of Talkers	4,817	62
Avg # of Anon. editors	1,308	51
Avg # of Anon. Talkers	2,324	46
Avg # of Crossover Wikipedians (%)	1,868	12
Avg % are Crossovers Wikipedians	4.5	13

Table II. WikiProject Statistics for 618 active Wikipedia projects. Root pages represent the activity on the core WikiProject page (homepage)

Feature	Avg growth function	Std Dv. Growth Constant)
All article pages in Project	2.58x - 2.8	0.12
All Talk pages in Project	2.92x - 3.31	0.25
Root article pages in Project	3.68x - 4.32	0.74
Root Talk in Project	3.86x - 4.76	0.60

Table III. Regression Analysis of Editing and Talk Activity Growth for All and Root Pages

crossover users, however, for root pages of a WikiProject, we found a larger proportion of Wikipedians active on articles and Talk pages. As expected, we found a positive correlation between the size of a WikiProject (number of articles and Talk pages), and how many participants contributed to it. However, we noted that there are a number of mid-sized projects which contain many editors, but very few Talkers.

4.1 WikiProject Growth

We computed the growth of each WikiProject in terms of the number of contributions, pages, and newly joined Wikipedians. Figure 2 provides an overview of the raw and normalised view of the growth in Edits and Talk entries, for all associated project pages (a) - (d). and for all root pages, (e) - (h). Each line in the plot corresponds to a WikiProjects, and the values represent the cumulative number of edits or Talk entries. As the Figures illustrate, the growth of WikiProjects appear to follow a linear function (article and Talk pages), and as the narrow upper and lower boundaries of the calculated standard deviation in Table III indicates, WikiProjects appear to grow at a similar rate. This is also true for just the root pages, although the rate to which WikiProjects have grown in Talk activity is slightly less unified, as (h) shows.

4.2 User Onboarding

We also computed the intake of new Wikipedians, by extracting their first entry on either an article or Talk page. Figure 3 provides an aggregate count of new article and Talk users, for both root pages and all associated WikiProject article pages. The plots for new article users show that there was an initial spike in activity, which then level off and remain steady. For Talk pages, examining only the root pages reveal a steady increase in activity, and for

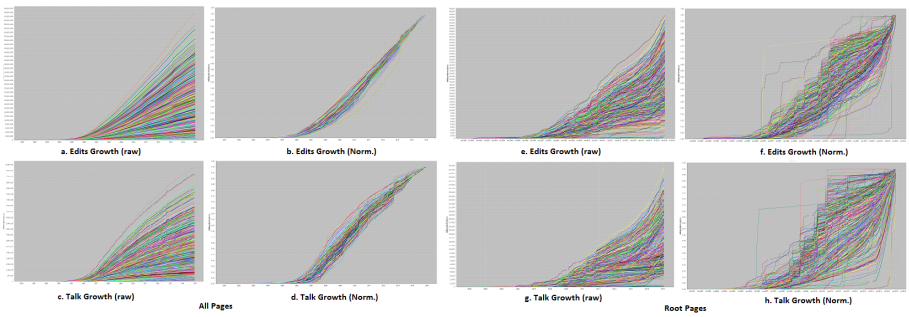

Fig. 2. Growth of Editing and Talk pages, aggregated by WikiProjects. (a) - (d) for All activity, (e) - (h) for root activity

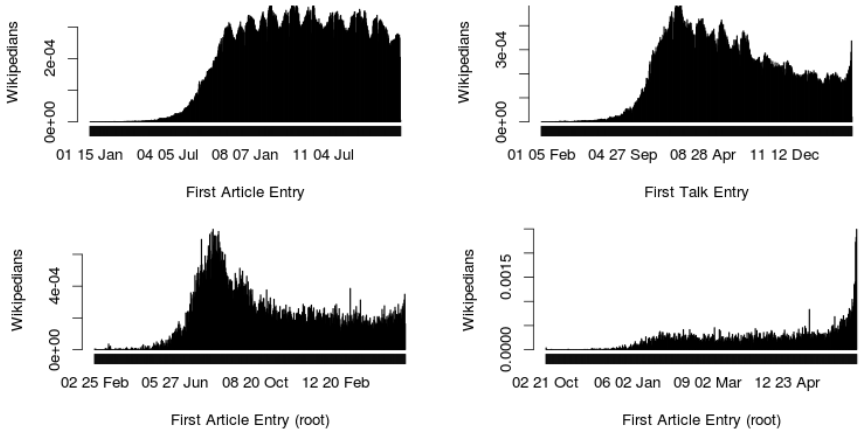

Fig. 3. New Editors and Talkers to Join Wikipedia. Identified by their First Entry on an article or Talk page

all WikiProject pages, we observe an initial spike and then a slight decline until it has reach a level of stability.

4.3 Cross-Project Contribution

We examined the co-occurrence network between root pages of WikiProjects with respects to Wikipedians who are active on more than one WikiProject. Figure 4 provides an illustration of the co-occurrence network between WikiProject root article pages, and root Talk pages. As Table IV shows, there is a large group of Wikipedians who contribute to one or more WikiProjects, and that both the root article and Talk pages are all part of one large component.

4.4 Editors-and-Talkers

In order to understand the role of Wikipedians active on the root WikiProject pages (articles and Talk pages) have within the wider set of WikiProject pages, we computed the

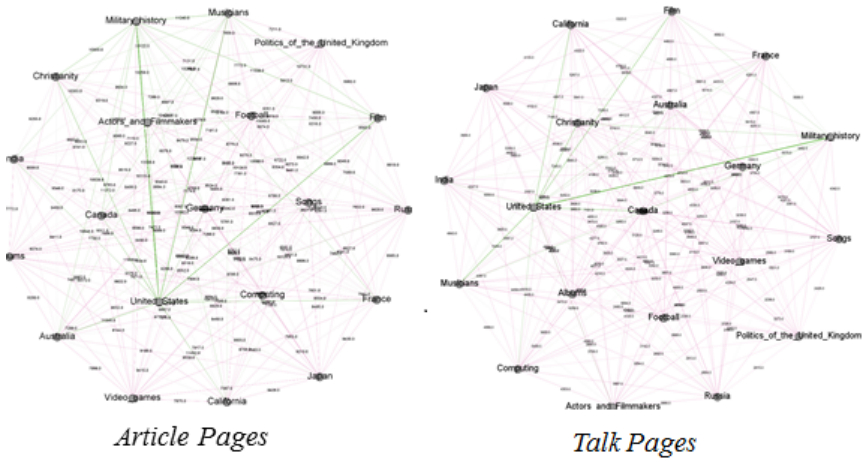

Fig. 4. Article and Talk Co-occurrence network between Wikipedians on root pages. Nodes represent WikiProjects

Graph Measure	Talk Pages	article Pages)
Nodes	603	608
Edges	34,1067	35,0852
Node Avg. In-degree	565	577
Node Avg. Out-degree	565	577
Connected Components	1	1
Diameter	4	4

Table IV. Co-occurrence network statistics of root article and Talk pages.

set of editors and Talkers that are active on both root pages and the WikiProject pages. Figure 5 shows the distribution of root editors and Talkers on WikiProject pages, as well as the relationship between the proportion of root editors and Talkers that are active on a WikiProject page. Each data point within the plot represents a WikiProject, and its position indicates the percentage of root users who edit or Talk on the associated set of WikiProject pages. As shown, there is a positive relationship between projects that have root users that edit and Talk on associated pages. There are also a number of projects where many of the root users do not edit but Talk a lot.

5. DISCUSSION

One of most important findings from our analysis is the ability to use Wikipedia's emergent features as a measure of the platforms activity. Unlike existing studies [Suh et al. 2009; Halfaker et al. 2013], our analysis, which uses WikiProjects as a proxy for Wikipedia activity, appears to show signs of growth. Our findings suggest that there is still significant activity in terms of new article edits, new Talk discussions, and a steady - and in Talk pages - growing number of new Wikipedians. It is as result of this that we consider emergent features such as WikiProjects an important area to study and understand, as these features may offer alternative ways to better monitor and observe a system's 'health'.

Moreover, whilst the primary task in Wikipedia is to create a comprehensive knowledge-

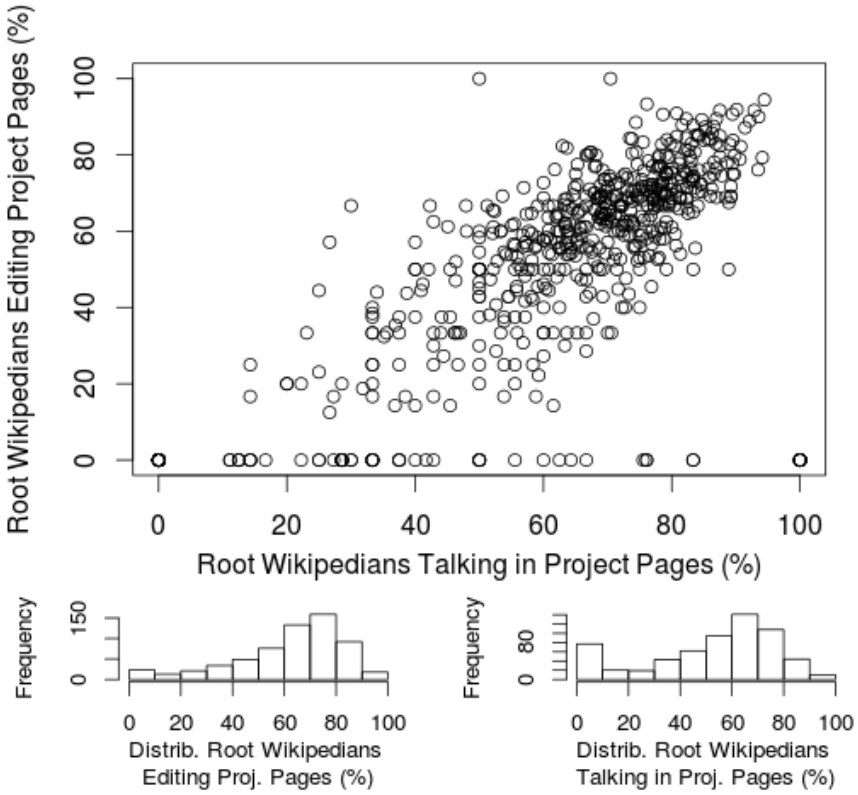

Fig. 5. Root WikiProject Page and All Editors and Talkers. Plot compares WikiProject Root Wikipedians participation in associated project pages

base using crowdsourcing techniques, features such as Talk pages have emerged to become important in the process of article production [Ferschke et al. 2012]. Our analysis of the activity on article and Talk pages revealed a fairly similar outcome, however the proportion of WikiProject members who contributed to both edits and Talk pages revealed a fairly low percentage. As Table II shows, on average only 4.5% of Wikipedians participated in editing and Talking on the associated pages of a WikiProject. Yet, in comparison to this, 13% of root page Wikipedians contributed to both Talk and edits. A possible reason for this might be that the community that forms around root pages is significantly smaller, as shown in Table II, and the type of content on a root page may promote more discussion, given that the content is concerned with WikiProject activity, goals, and achievements.

We also found that they were interconnected by a network of Wikipedians. As Figure 4 and Table IV show, there are a core network of Wikipedians which are active on more than one WikiProject. As the co-occurrence network metrics in Table IV show, nearly all projects shared common editors or Talkers, even though projects are extremely diverse in topic (e.g. from 'Academic Journals' to 'Indian Music'). Taking this into consideration, these Wikipedians may represent some of the core community responsible for sustaining

and improving the current state of Wikipedia [Kittur et al. 2007].

6. CONCLUDING REMARKS

Our analysis of WikiProjects has revealed features about how this emergent feature contributes to the main corpus of Wikipedia's articles, and importantly, how these communities reflect healthy statistics of new users, edits, and discussions. Whilst further insight is required to understand the context and relationship between Talk and article pages, and the role of WikiProjects, our findings suggest that Talk has integrated itself to become part of the core Wikipedia workflow, and that WikiProjects has reconfigured the article production and improvement process.

As Wikipedia has evolved over the course of 1.5 decades, so has its complexities, features, and role in modern society. Features such as WikiProjects represent community-led, reconfigurations of a technology. By developing analytical approaches which can take advantage of these organic, community-inspired features, it is possible to gain insight which may not possible by using traditional, system-level metrics.

Taking things forward, future work in this area needs to explore the extent of how WikiProjects facilitates the lifecycle of article production, and how talk pages plays a role in this process. Another interesting area of investigation would be to explore the role of WikiProjects in connecting articles and Wikipedians across linguistic boarders, given that Wikipedia is a multi-language resource.

Acknowledgements

This work is supported under SOCIAM: The Theory and Practice of Social Machines, funded by the UK EPSRC under grant EP/J017728/2.

REFERENCES

BARABASI, A.-L. AND ALBERT, R. 1999. Emergence of Scaling in Random Networks. *Science 286,* October, 509–512.

BURIOL, L. S., CASTILLO, C., DONATO, D., LEONARDI, S., AND MILLOZZI, S. 2006. Temporal Evolution of the Wikigraph. In *Proceedings of Web Intelligence*. IEEE CS Press., 45–51.

CAPOCCI, A., SERVEDIO, V. D. P., COLAIORI, F., BURIOL, L. S., DONATO, D., LEONARDI, S., AND CALDARELLI, G. 2006. Preferential attachment in the growth of social networks: the case of Wikipedia. *Physical Review E 74,* 3, 4.

DE LA CALZADA, G. AND DEKHTYAR, A. 2010. On measuring the quality of wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility*. WICOW '10. ACM, New York, NY, USA, 11–18.

FARIČ, N. AND POTTS, W. H. 2014. Motivations for contributing to health-related articles on wikipedia: An interview study. *J Med Internet Res 16,* 12 (Dec), e260.

FERSCHKE, O., GUREVYCH, I., AND CHEBOTAR, Y. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, 777–786.

HALFAKER, A., GEIGER, R. S., MORGAN, J., AND RIEDL, J. 2013. The rise and decline of an open collaboration system: How wikipedia's reaction to sudden popularity is causing its decline. *American Behavioral Scientist 57,* 5 (May), 664–688.

HAUTASAARI, A. AND ISHIDA, T. 2012. Analysis of discussion contributions in translated wikipedia articles. 57–66.

HENDLER, J. AND MULVEHILL, A. 2016. *Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity*. Apress.

JULLIEN, N., CROWSTON, K., AND ORTEGA, F. 2015. The rise and fall of an online project: is bureaucracy killing efficiency in open knowledge production? In *Proceedings of the 11th International Symposium on Open Collaboration*. ACM, 13.

KAMPS, J. AND KOOLEN, M. 2009. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09. ACM, New York, NY, USA, 232–241.

KITTUR, A., CHI, E., PENDLETON, B., SUH, B., AND MYTKOWICZ, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web 1,* 2, 19.

KITTUR, A. AND KRAUT, R. E. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. ACM, 37–46.

KITTUR, A., SUH, B., PENDLETON, B. A., AND CHI, E. H. 2007. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. ACM, New York, NY, USA, 453–462.

KUZNETSOV, S. 2006. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc. 36,* 2 (June).

LAM, S. T. K. AND RIEDL, J. 2009. Is wikipedia growing a longer tail? In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*. GROUP '09. ACM, New York, NY, USA, 105–114.

LIAO, H.-T. 2008. Conflictual consensus in the chinese version of wikipedia. 1–10.

LIU, J. AND RAM, S. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst. 2,* 2 (July), 11:1–11:23.

MORGAN, J. T., GILBERT, M., ZACHRY, M., AND MCDONALD, D. 2013. A content analysis of wikiproject discussions: Toward a typology of coordination language used by virtual teams. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*. CSCW '13. ACM, New York, NY, USA, 231–234.

PFEIL, U., ZAPHIRIS, P., AND ANG, C. S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication 12,* 1, 88–113.

STUCKMAN, J. AND PURTILO, J. 2009. Measuring the wikisphere. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*.

SUH, B., CONVERTINO, G., CHI, E. H., AND PIROLLI, P. 2009. The singularity is not near: Slowing growth of wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. WikiSym '09. ACM, New York, NY, USA, 8:1–8:10.

TINATI, R., CARR, L., HALFORD, S., AND POPE, C. J. 2013. The htp model: Understanding the development of social machines. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*. WWW '13 Companion. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 921–926.

Ramine Tinati is a New Frontier's Fellow at the University of Southampton, UK. Previously, he was a Senior Research fellow on the EPSRC's SOCIAM project. His research focuses on the space between machine and human interaction, with a specific focus on understanding human behaviour at the scale of the Web.

Markus Luczak-Roesch is a Senior Lecturer in Information Systems at the School for Information Management, Victoria Business School, Victoria University of Wellington (NZ). , Markus asks research questions on the fundamental properties of information in socio-technical systems as well as humans in the information age.