
Talking About Talk: Coordination in Large Online Communities

Jim Maddock

Northwestern University
Technology and Social Behavior
maddock@u.northwestern.edu

Aaron Shaw

Northwestern University
School of Communication
aaronshaw@northwestern.edu

Darren Gergle

Northwestern University
Technology and Social Behavior
dgergle@northwestern.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright is held by the owner/author(s).
CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
Copyright © ACM 978-1-4503-4656-6/17/05.
<http://dx.doi.org/10.1145/3027063.3053077>

Abstract

Social computing systems and online communities develop varying strategies for managing collaborative processes such as consensus building, task delegation, and conflict management. Although these factors impact both the ways in which communities produce content and the content they produce, little prior work has undertaken a large comparative analysis of coordination dynamics across linguistically diverse communities engaged in the same activity. We describe and model the coordination processes of Wikipedia editors across the 24 largest language editions. Our results indicate that language edition is associated with a difference in quantity of coordination activity, as measured by talk page posts, with increases as high as 60% when compared against pages in English.

Author Keywords

Social Computing; Online Collaboration; Peer Production

ACM Classification Keywords

H.5.3 [Information Interfaces & Presentation]: Group and Organization Interfaces—Computer-supported cooperative work, Evaluation/methodology

Introduction

The past two decades have seen an explosion of large peer production and social computing projects, such as

moving the section from "back ba

If that's the case, [canadian bacon](#) should have its own section. [MateoP](#) 00:28, 31 March 2006 (UTC)

I prefer to have all the information about bacon as one section. Excuse me if I don't reply promptly. [Tor](#)

Canadian bacon should be the exception because it's a page who want to find this information easily could

For the Nth time. Canadian bacon is back bacon. It explains the different cuts of meat known as back bacon, which is by far the most common sort of bacon consumed in Canada, as if it was something else entirely different. [April 2006 \(UTC\)](#)

Which is exactly why Canadian bacon needs its own section, just given a different name to sound exotic. [F](#) bacon article in any way.

WHAT IS YOUR ALTERNATIVE COMPROMISE?

Canadian bacon already has its own article, it's not stupid, that if the click on a link looking for [pavement](#) or found [eggplant](#) when I shouting. [Jooler](#) 22:22, 2 April 2006 (UTC)

Yes, since most Americans think Canadian bacon is no way. They would still type in "bacon" and be confusing to them. To make matters worse,

In thinking about this question I compared it with a Canadian bacon that needs an entire article? True only point is to disambiguate it from other bacon via a redirect. Canadian bacon should be a section within the back bacon article. This is perhaps the most hilarious debate I've seen

Third opinion

I'm not overly convinced at this point that Canadian bacon from each other, but there seems to be no explanation. Once all the variants have been listed, I would suggest

- Mini disambig at the top like: "Types of bacon":
 1. Type A, aka B (in the US), C (in the UK), D, etc.
 2. type B aka A in Canada, C, D, etc.
- and then have each of those items go to a different article. That, I believe, would reduce the confusion, as it w

Wikipedia, Zooniverse, and Linux. These projects are developed by increasingly diverse communities of contributors who employ varying collaborative processes, such as those involved in consensus building, task delegation, and conflict management [3, 13, 15, 17, 4, 7]. Previous work has (1) examined coordination dynamics within English Wikipedia [15, 11, 4, 18], (2) compared coordination practices in small samples across Wikipedia language editions [8, 6], and (3) analyzed content asymmetries across the Wikipedia platform [9, 2, 21, 5]. However, while these studies suggest that editors in different Wikipedia language editions may favor different coordination processes, existing research does not comprehensively examine coordination dynamics across many language editions.

We conduct a large, comparative study of Wikipedia's 24 largest language editions¹ in order to create a more general understanding of peer-production and online collaborative work across linguistic boundaries. Our results indicate that coordination activity—measured by posts to talk pages²—can vary dramatically across language editions, sometimes by as much as 60%.

While it may seem self evident that no two communities share an identical set of coordination practices, much of the prior work that ties productivity and quality metrics to group composition, network structures, and community norms focuses on single language speaking platforms [4, 7, 18, 14], implicitly suggesting that coordination practices are relatively universal. We show that the amount editors coordinate varies substantially across Wikipedia language editions.

¹For a list of language editions see https://en.wikipedia.org/wiki/List_of_Wikipedias

²We refer to edits to talk pages as "talk page posts" to avoid confusion with article edits.

Background and Related Work

Coordination in Online Communities

Coordination dynamics have been well documented across a range of communities and social computing systems, though primarily in studies that analyze communities where English is the dominant language. Previous studies on both English Wikipedia and Wikia (another wiki-style website) show that coordination and conflict management strategies impact the rate, quality, and the type of content editors generate [13, 14].

These factors, however, have yet to be analyzed in other language editions, and previous cross-cultural work suggests that community dynamics and their effects on content production may not generalize beyond English Wikipedia [6]. For example, Hara et al. performed a mixed methods study that analyzed four language editions of Wikipedia and found differences in the ways that editors coordinate to produce content within East/West and large/small communities [8].

We advance this line of research and diverge from the approach of Hara et al. in two important ways. First, we expand our analysis beyond a small subset of 120 talk pages to include the entirety of 24 Wikipedia language editions, which after processing amounts to 11,067,512 talk and article pages. Second, our analysis does not rely on Hofstede's dimensions of cultural diversity [10], which have drawn criticism in the past decade for taking a limited, uni-dimensional approach to culture [12].

Cultural and Content Variation

Though little research compares *online coordination dynamics* in peer production across language boundaries, previous work indicates that the *content* contained within different language editions of Wikipedia varies greatly. En-

Figure 1: An example Wikipedia talk page where editors discuss future changes to the article. Captured from <https://en.wikipedia.org/wiki/Talk:Bacon>

| variable | median |
|---------------------|--------|
| $Posts_{talk}$ | 3.00 |
| $Edits_{article}$ | 38.00 |
| Age_{talk} | 2.78 |
| $Age_{article}$ | 7.23 |
| $Editors_{talk}$ | 3.00 |
| $Editors_{article}$ | 19.00 |

Table 1: Population median of each variable

| variable | var |
|---------------------|--------|
| $Posts_{talk}$ | 10506 |
| $Edits_{article}$ | 109183 |
| Age_{talk} | 8 |
| $Age_{article}$ | 10 |
| $Editors_{talk}$ | 446 |
| $Editors_{article}$ | 18529 |

Table 2: Population variance of each variable

cyclopedic content on Wikipedia does not cover topics universally across all language editions [9, 2, 20]. For example, the French Wikipedia article for “Conspiracy theory” at one point contained information about the Algerian War, while the English version did not [2]. Furthermore, content variation manifests in distinct patterns of gender bias. Women and men are portrayed differently across language editions in terms of coverage, lexical quality, article structure, and article visibility [21].

While substantial differences exist between encyclopedic content produced in different language editions, this prior work focuses on outputs instead of coordination and production processes. Given (1) differences in content across different language editions, (2) variations in coordination processes revealed through qualitative research, and (3) ties between coordination processes and content generation in English Wikipedia, we expect that coordination practices will also differ between language editions.

Our Study

We compare coordination practices across 24 different Wikipedia language editions, filling a gap between the existing literature on online coordination—which focuses on English-speaking communities or uses small sample sizes—and prior multi-language research, which generally targets content. While we might expect differences in coordination processes to exist between these communities, previous studies that investigate online coordination implicitly assume that these processes are somewhat universal, if only because they analyze single language editions. Therefore, identifying and quantifying these differences against our null hypothesis—that coordination patterns are somewhat universal and therefore that language edition does not matter—represents an important step towards creating a more contextualized view of coordination, which we capture

in the following research question:

RQ: To what extent is Wikipedia language edition associated with variations in coordination behavior?

Though we quantify coordination differences between Wikipedia language editions, our research does not directly measure the effects of language or culture on coordination. For example, we do not test whether speaking Hebrew causes editors to coordinate more than they would in English; instead we examine whether editors who contribute to the Hebrew language edition of Wikipedia on average coordinate more than editors in the English language edition. While our research design does not allow us to identify a causal effect or mechanism, we control for several confounding variables and offer potential explanations in our discussion.

Following previous work, we operationalize coordination through post counts to Wikipedia talk pages, which provide a measure of coordination between editors [15, 14]. In order to test the association between language edition and coordination, we build two models. We choose posts to talk pages as our dependent variable, because each post represents a contribution to a discussion about content and artifact production, similar to a post on a forum.

Data

Our dataset covers the 24 largest language editions of Wikipedia through October 2, 2015. Our measures focus on editing activity to article and talk pages (for example, Figure 1). *Article Pages* are the default view of Wikipedia and contain encyclopedic content. *Talk Pages* contain back-channel conversations. Each article-talk pair has a unique title, which we used to match talk pages with corresponding articles. Not all article pages possess a talk page (talk pages are not automatically created for each article and

| | Model 1 |
|-----------------------|------------------------|
| $Authors_{talk}$ | 0.0754*** (198.09) |
| Age_{talk} | 0.0621*** (175.73) |
| $Edits_{article}$ | 0.000297*** (51.72) |
| $Age_{article}$ | 0.0149*** (83.60) |
| Constant | 0.954*** (799.89) |
| Inalpha | |
| Constant | -1.114*** (-322.59) |
| Observations | 5533756 |
| Wald chi ² | 1443392.55 |
| Pseudo R ² | 0.2234 |
| BIC (df=6) | 28600000 |

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

*** $p < 0.001$

Table 3: Negative binomial regression on number of posts to talk pages ($Posts_{talk}$). Raw coefficients reported with z statistics in parentheses.

frequently small articles do not require coordination), so after matching we discarded all 19,234,735 unpaired pages. We also dropped 25,502,980 pages which contained only a single post, edit, or editor because they do not constitute collaborative activity.

“Bots,” which are scripts written to automatically perform routine cleanup tasks, account for a large number of edits and posts on Wikipedia. We collected an automatically generated list of registered Wikipedia bots³ and dropped 191,301,248 edits and posts made by these accounts from our dataset. Additionally, editors and bots frequently “archive” old talk page content in order to streamline coordination. In order to capture the entire post history of a talk page, we matched archived talk pages to their non-archived counterparts and report the aggregate post count. After processing, this dataset consists of 11,067,512 individual articles and talk pages, or 5,533,756 article-talk pairs.

Analytical Approach

Our unit of analysis is the population of talk-article page pairs within the 24 largest language versions of Wikipedia. Following prior work [14], we aggregate the edits and posts made to each talk-article page pair within our dataset. We then report two models of coordination activity, which we measure with posts to talk pages. Model 1 regresses coordination activity on control measures including article edits, article age, talk page age, and number of talk page editors, which we describe in Table 1 and Table 2.⁴ Model 2 includes the language version measure and allows us to estimate the additional variation in coordination explained by language edition.

³collected from

https://en.wikipedia.org/wiki/Category:All_Wikipedia_bots

⁴We do not include a control for the number of article editors because this measure is almost perfectly collinear with article edits (Pearson correlation coefficient = .95).

The first model illustrates the amount of coordination activity for a given article-talk pair based on the number of editors who have contributed to that talk page, the article-talk pair’s ages in years, and the number of edits to the article page. More formally:

$$Posts_{talk} = \beta_1 + \beta_2 Editors_{talk} + \beta_3 Age_{talk} + \beta_4 Edits_{article} + \beta_5 Age_{article} + \epsilon \quad (1)$$

This model demonstrates the relationship between the number of posts to a talk page and the number of editors who contribute to the talk page, the number of edits made to the article, and the page pair’s age. It assumes that language edition is not a significant factor.

Comparison between the first model and the second reveals the association between language edition and talk page posts. In addition to the controls listed above, we added 23 dummy variables, one for each language:

$$Posts_{talk} = \beta_1 + \beta_2 Editors_{talk} + \beta_3 Age_{talk} + \beta_4 Edits_{article} + \beta_5 Age_{article} + \beta_6 LanguageEdition + \epsilon \quad (2)$$

In the second model we use English Wikipedia as the baseline category. Therefore, all language version coefficients represent the estimated variation in our outcome for the given language version in comparison to the corresponding level of coordination on English Wikipedia.

Talk page posts are an over-dispersed count measure, so we use negative binomial regression to fit both models [16]. We report cluster and heteroskedasticity robust standard errors [1]. We also dropped one outlier observation because it prevented model convergence: the English Wikipedia’s “Talk: Main Page” receives an abnormally high volume of posts—126,202 in our dataset—which was nearly three times second highest talk page post count.

| | Model 2 |
|--------------------------------|------------------------|
| <i>Authors_{talk}</i> | 0.0755*** (197.92) |
| <i>Age_{talk}</i> | 0.0626*** (178.55) |
| <i>Edits_{article}</i> | 0.000293*** (50.91) |
| <i>Age_{article}</i> | 0.0147*** (82.75) |
| <i>Language</i> | 0.00275*** (27.73) |
| Constant | 0.933*** (646.84) |
| Inalpha | |
| Constant | -1.114*** (-322.50) |
| Observations | 5533756 |
| Wald chi ² | 1448440.64 |
| Pseudo R ² | 0.2234 |
| BIC (df=29) | 28300000 |

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

*** $p < 0.001$

Table 4: Negative binomial regression on number of posts to talk pages ($Posts_{talk}$). Raw coefficients reported with z statistics in parentheses.

Results

Our models reveal the estimated association between language edition and median posts to a talk page. Model 1 (Table 3) indicates that for a prototypical talk page with a mean number of posts and editors, a one editor increase corresponds to an 7.8% increase in posts, and each year of talk page age corresponds to a 6.4% increase in posts. We also account for additional variance that could result from the associated article. For an average talk page, a one year increase in the corresponding article's age is associated with a 1.5% increase in talk posts, though a single article edit increase corresponds to only a .03% increase in talk posts. These results indicate that older articles do tend to have more active talk pages, but that in general, only large increases in editing activity to an article are associated with more posts to a talk page.

Model 2 (Table 4) introduces the language edition variables. Comparison of the results between Model 1 and Model 2 reveals that the addition of language variables improves overall model fit as the BIC measure decreases (reported in Tables 3 and 4). The coefficient for each of the language variables represents the expected change in number of talk posts (with respect to English) associated solely with the language edition of the talk page, which we report as a percentage in Figure 2. In other words, because we control for variance due to age, number of edits, and number of editors, the 23 language percentage estimates show expected change in posts irrespective of our control variables. Table 3 and Table 4 provide the coefficient estimates for Model 1 and Model 2, z-statistics, and goodness of fit measures.

The results of Model 2 reveal that coordination practices, as operationalized through talk page posts, vary substantially by language version. Generally speaking, editors coordi-

nate more for every article edit in some languages editions than in others.

We see that while certain language editions are associated with relatively similar numbers of posts to a given talk page, others differ dramatically. Though Ukrainian falls within 2% of our reference category and is therefore relatively similar to English, others show substantial differences. Our results suggest that a given talk page in Hebrew or Farsi would have 69% or 62% more posts than an English page, while a similar Portuguese page would have 23% fewer posts. Put another way, consider our median talk-article pair: a talk-article pair with 38 article edits over 7.23 years, and a talk page that is 2.78 years old and has 3 editors. In English we would expect this talk page to have 4 posts. However, in Hebrew or Farsi the same talk page would have 6 posts, while in Portuguese it would have 3 posts.

Due to our variables' long-tailed distributions these median values are relatively low, and the association between coordination activity and language edition becomes substantially more pronounced for older pairs that attract more activity. Our prototypical talk-article pair at the 99th percentile has 1251 article edits over 12.9 years, and a talk page that is 10.5 years old with 41 editors. In English this page would have 165 posts, in Hebrew or Farsi it would have 280 or 271 posts, and in Portuguese it would have only 127 posts.

Discussion

Our results show that language edition is associated with the amount of coordination that takes place surrounding a specific article, regardless of age and number of editors. In those language editions with large predicted percent changes with respect to English, individual editors generally coordinate more while producing content, while in language editions with lower predicted percent changes editors are

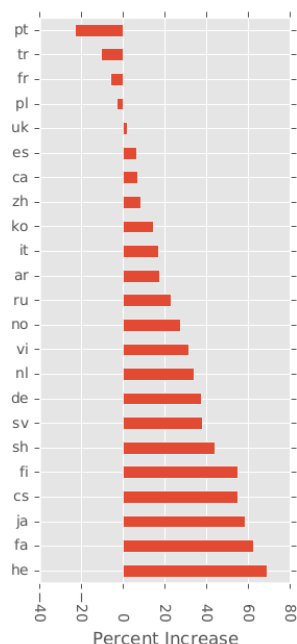


Figure 2: Language coefficients as percents

more likely to coordinate less. In the Portuguese or Turkish Wikipedia, editors make content related edits with substantially less coordination through Wikipedia talk pages, while in Hebrew or Farsi editors “talk” or post to talk pages more about each content related edit. Generally, this means that editors in some language editions coordinate their efforts more while creating or editing content, while in other language editions coordination is less common. Differences in these practices could be due to both population wide differences in coordination behavior and variances within the population.

Potential Explanatory Mechanisms

Several different mechanisms might explain these differences. We describe these briefly below with the caveat that our analysis cannot determine whether some have greater validity than others.

One potential explanation lies in organizational culture [19]. Organizational cultures consist of shared values and norms, which determine how individuals within a community act and feel. These norms influence diverse aspects of group behavior, from problem solving and innovation to new member socialization. Though the empirical foundation for this theory comes from research conducted on firms, the multiple coordination strategies employed by and within different Wikipedia language editions could result from each community’s unique organizational culture.

Sociocultural psychology provides a second potential explanation for variations in coordination activity between language editions [12]. Sociocultural psychology posits that an individual’s psychology and the sociocultural contexts in which that individual exists are mutually constituted. The sociocultural contexts of Wikipedia editors may explain the varying editing behaviors we observe between different language editions. For example, prior work shows that in

English Wikipedia, editors use different collaborative strategies depending on the context of the article—in this case articles either about breaking news or historical events [11]. Sociocultural psychology predicts that these strategies also vary between language editions based on editors’ cultural contexts external to Wikipedia, such as geographic location or language. Therefore, Spanish Wikipedia editors contributing to breaking news stories would employ different strategies than English Wikipedia editors contributing to the same type of articles.

Implications for Future Studies

An important implication of this work is that future online coordination studies should not assume that platforms consist of a single, homogeneous community. Researchers should account for this diversity by studying multiple communities and operationalizing these groups within their models when making generalizable claims that relate productivity and quality metrics to group composition, network structures, and community norms.

Acknowledgements

We would like to thank the BYOR workshop and our lab-mates at Northwestern University and the University of Washington for their valuable feedback. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak super-computer system at the University of Washington.

REFERENCES

1. Joshua D. Angrist and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, Princeton, NJ.
2. Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In

- Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1075–1084. <http://dl.acm.org/citation.cfm?id=2208553>
3. Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer Production: A Form of Collective Intelligence. In *The Handbook of Collective Intelligence*, Michael Bernstein and Thomas Malone (Eds.). MIT Press, Cambridge, MA.
 4. Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't Look Now, but We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1101–1110. DOI : <http://dx.doi.org/10.1145/1357054.1357227>
 5. Ewa S. Callahan and Susan C. Herring. 2011. Cultural Bias in Wikipedia Content on Famous Persons. *Journal of the American Society for Information Science and Technology* 62, 10 (Oct. 2011), 1899–1915. DOI : <http://dx.doi.org/10.1002/asi.21577>
 6. Brenda Danet and Susan C. Herring (Eds.). 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press, Oxford ; New York.
 7. Samer Faraj, Sirkka L. Jarvenpaa, and Ann Majchrzak. 2011. Knowledge Collaboration in Online Communities. *Organization Science* 22, 5 (Feb. 2011), 1224–1239. DOI : <http://dx.doi.org/10.1287/orsc.1100.0614>
 8. Noriko Hara, Pnina Shachaf, and Khe Foon Hew. 2010. Cross-Cultural Analysis of the Wikipedia Community. *Journal of the American Society for Information Science and Technology* 61, 10 (Oct. 2010), 2097–2108. DOI : <http://dx.doi.org/10.1002/asi.21373>
 9. Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and its Applications in a Multilingual Context. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 291–300. <http://dl.acm.org/citation.cfm?id=1753370>
 10. Geert H. Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, Calif.
 11. Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Do Editors or Articles Drive Collaboration?: Multilevel Statistical Network Analysis of Wikipedia Coauthorship. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 427–436. <http://dl.acm.org/citation.cfm?id=2145271>
 12. Shinobu Kitayama. 2007. *Handbook of Cultural Psychology*. Guilford Publications, New York.
 13. Aniket Kittur and Robert E. Kraut. 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. <http://dl.acm.org/citation.cfm?id=1460572>
 14. Aniket Kittur and Robert E. Kraut. 2010. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 215–224. <http://dl.acm.org/citation.cfm?id=1718959>
 15. Aniket Kittur, Bryant Lee, and Robert E. Kraut. 2009. Coordination in Collective Intelligence: the Role of Team Structure and Task Interdependence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1495–1504. <http://dl.acm.org/citation.cfm?id=1518928>

16. J. Scott Long. 2014. *Regression Models for Categorical Dependent Variables Using Stata* (third edition. ed.). Stata Press Publication, StataCorp LP, College Station, Texas.
17. Siobhan O'Mahony and Fabrizio Ferraro. 2007. The Emergence of Governance in an Open Source Community. *Academy of Management Journal* 50, 5 (2007), 1079–1106.
18. Sam Ransbotham and Gerald C. Kane. 2011. Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia. *MIS Quarterly-Management Information Systems* 35, 3 (2011), 613. <http://aisel.aisnet.org/cgi/viewcontent.cgi?article=2973&context=misq>
19. Edgar H. Schein. 1984. Coming to a New Awareness of Organizational Culture. *Sloan Management Review* 25, 2 (1984), 3–16. <http://search.proquest.com/docview/1302988646/citation/53454F08575C42F7PQ/1>
20. Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. WikiBrain: Democratizing Computation on Wikipedia. In *Proceedings of The International Symposium on Open Collaboration (OpenSym '14)*. ACM, New York, NY, USA, 27:1–27:10. DOI : <http://dx.doi.org/10.1145/2641580.2641615>
21. Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv preprint arXiv:1501.06307* (2015). <http://arxiv.org/abs/1501.06307>