Modeling Dynamics of Wikipedia: An Empirical Analysis Using a Vector Error Correction Model

Feng-Jun LIU[1,a], * Jiang-Nan QIU[1,b] and Na ZHAO[2,c]

[1]*Faculty of Management and Economics, Dalian University of Technology, 116024 Dalian, Liaoning, China*
[2]*School of Management Science and Engineering, Shandong Technology and Business University, 264005 Yantai, Shandong, China*

[a]liufengjunmail@163.com, [b]qiujn@dlut.edu.cn, [c]zhaonawfxy@163.com

*Corresponding author: Feng-Jun LIU

Abstract. In this paper, we constructed a system dynamic model of Wikipedia based on the co-evolution theory, and investigated the interrelationships among topic popularity, group size, collaborative conflict, coordination mechanism, and information quality by using the vector error correction model (VECM). This study provides a useful framework for analyzing the dynamics of Wikipedia and presents a formal exposition of the VECM methodology in the information system research.

1 Introduction

Wikipedia has become one of the most striking emblems of mass collaboration. Its unprecedented success has posed challenges to traditional theories of public goods and collective-action, which has inspired many scholars from various fields to study it [1, 2]. Existing research highlights many factors that are crucial to the success of Wikipedia, including topic popularity, group size, collaborative conflict, coordination mechanism, and information quality, etc [3]. However, most of these studies examine the relationships among factors from a static perspective without considering the dynamic evolution of Wikipedia. Although some scholars have already explored the statistical properties in many aspects of Wikipedia by statistical methods and revealed the dynamic relations among factors by visual analysis tools [4-7]. there are lack of rigorous empirical studies. So far, the dynamic mechanism of Wikipedia is not completely known.

As an attempt complementary to the previous studies, this study constructs the PSCCQ model, a system model consisting of five microcosmic factors (i.e., topic popularity, group size, collaborative conflict, coordination mechanism, and information quality), with which we explore the fundamental dynamic mechanism behind Wikipedia.

Our study makes several key contributions. First, we build the PSCCQ conceptual model from a systemic perspective, and provide strong empirical evidence to support the validity and usefulness of this model. Second, the VECM methodology provides a statistically rigorous yet atheoretical approach for analyzing the dynamics of temporal relationships among variables without strong theoretical restrictions. In summary, this paper promotes the importance of taking a systematic view of the dynamics of Wikipedia by utilizing the VECM approach in concert with the PSCCQ model.

2 Research model

The process of collaborative knowledge building in Wikipedia demonstrates complex self-cleaning, self-regulating, and self-developing dynamics of the mass of participants that are akin to a kind of evolution [8]. The co-evolution theory provides an effective theoretical lens for analyzing the dynamics of Wikipedia. Based on the co-evolution theory, A few scholars have attempted to analyze the dynamic interactions among factors, modules, and subsystems of the Wikipedia system. For example, Cress and Kimmerle built a theoretical model for describing the co-evolution between the Wiki's social system and the individuals' cognitive systems [9]. Kimmerle et al. used the social network analysis to graphically visualize co-evolutionary processes of individual knowledge learning and collective knowledge building [7].

Based on the co-evolution theory, we build the PSCCQ model as a theoretical framework to analyze the dynamic interactions among topic popularity, group size, collaborative

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

conflict, coordination mechanism, and information quality, and reveal the dynamics of Wikipedia (cf. Fig. 1).

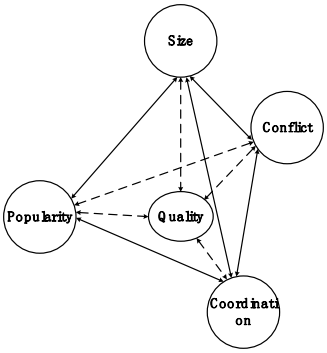


Figure 1. The PSCCQ model.

3 Research setting

3.1 Research context

We select a specific Wikipedia article—global warming as research object for several reasons. Firstly, global warming belongs to a featured article, which means that it is identified as one of the best articles. Secondly, it is also one of the most frequently viewed articles in Wikipedia. It has around 50 million page views, ranking 152 in English Wikipedia article traffic. Thirdly, global warming has always been one of the most intensely controversial topics in Wikipedia. In summary, this case represents one of the most typical, highly concerned and intensely controversial articles in Wikipedia, so it is ideal for deeply examining the dynamics of Wikipedia.

3.2 Data collection

Following Ransbotham, Kane, and Lurie [10], we use the relative search frequency in Google to reflect the topic popularity of global warming. We determine the number of times that users of Google search for keywords from the article title each month from Google Trends. The number of unique editors contributed to an article in Wikipedia is widely used to measure group size in previous studies [11], so we adopt it to measure group size during the monthly observation period. Rollback has been frequently used to model conflicts and identify edit wars in Wikipedia [12]. Therefore, we measure collaborative conflict by calculating the monthly number of rollbacks in the article. The most important coordination mechanism in Wikipedia is communication [13]. We operationalize coordination mechanism as the monthly accumulated number of discussions recorded in the article talk page during the study period. The number of edits provides a good indicator of a "high level of quality" for Wikipedia articles [3]. We quantify information quality as the number of edits in the given month.

Using monthly counts of each variable, we obtain five time series. With the data from February 2004 to November 2015, we have a total of 142 monthly observations in the form of time series, as shown in Fig 2.

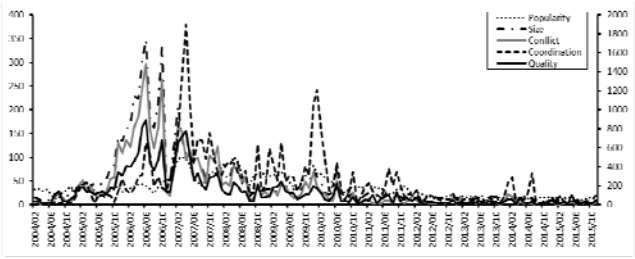


Figure 2. Time series plot.

4 Model specification and estimation

Selection of a VECM involves three basic decisions: (1) Unit root test, (2) lag length selection, and (3) co-integration test. Non-stationary data generally lead to spurious regression due to non-constant mean and variance [14]. Therefore, we first test the stability of variables. The results indicate that all variables are stationary at first differences implying that all variables are integrated of order one. Before co-integration test, we need to select an optimal lag length to ensure that the model is not mispecified [15]. The results show that the optimal lag length is identified to be lag 4. Finally, the Johansen co-integration test results show that there are co-integration relationships among the variables.

The estimates of VECM regression coefficients typically are not as informative as analyzing relationships among variables because of the complicated dynamics inherent in VECM models [15, 16]. Therefore, we report the general estimation results in Table 1 and then provide a detailed analysis using Granger causality tests, impulse response functions, and forecast error variance decomposition in the next section.

5 Empirical analysis

5.1 Granger causality tests

Granger causality tests help to determine whether the lagged values of one variable help to predict values of another variable [17]. Table 2 presents the results of Granger causality tests based on the VECM. Fig 3 provides a reduced form model of PSCCQ, as uncovered by the Granger causality tests.

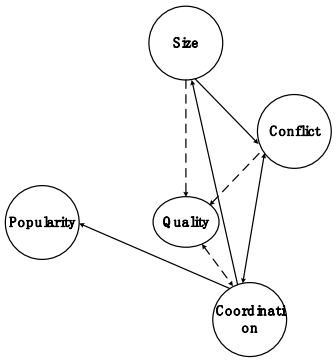
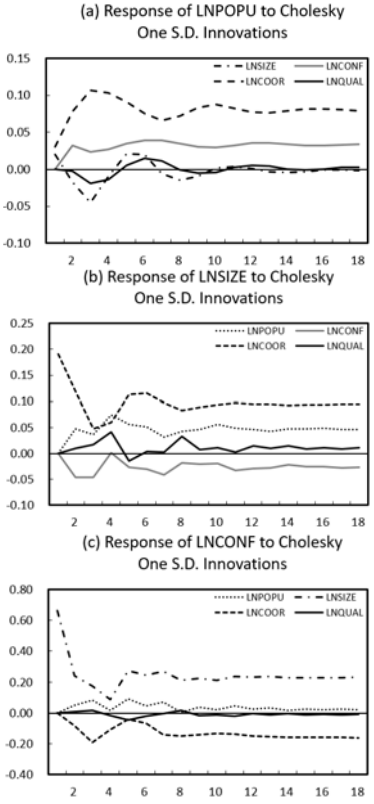


Figure 3. Reduced form model of the PSCCQ.

5.2 Impulse response functions

Impulse response functions (IRFs) plot the response of current and future values of the endogenous variables to a one-unit increase in the current value of a random disturbance term [18], which provide a more intuitive description of the dynamics of temporal relationships among variables. Fig 4 provides twenty possible impulse response functions for the estimated VECM.



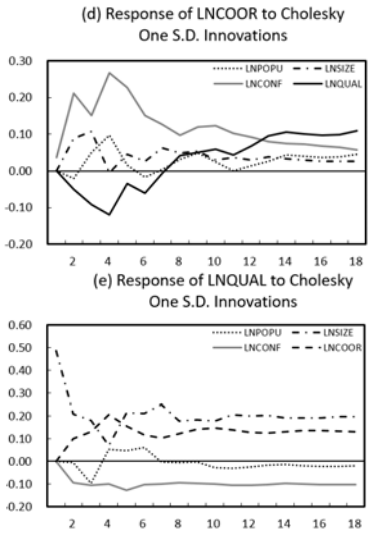


Figure 4. Impulse responses (Impulse→Response).

The results of the IRFs analysis basically corroborate the results of the Granger causality tests. Specifically, the feedback relationship between *LNCONF* and LNCOOR identified previously holds, which can be found in the significantly negative response of *LNCONF* at forecast horizons 3 and 8-18 in Fig 4c and in the significantly positive response of LNCOOR at forecast horizons 2-11 in Fig 4d. Similarly, we can also confirm the feedback relationship between *LNCOOR* and *LNQUAL*. Additionally, the unidirectional relationships among the variables identified previously also hold. The response of *LNPOPU* to a one standard deviation shock in *LNCOOR* has a maximum value of 0.107 at the 3th period, as shown in Fig 4a. Turning now to the response of *LNSIZE* to *LNCOOR*, we find that *LNSIZE* initially reaches the maximum value of about 0.2, then gradually decreases, and achieves stability until the 10th period (cf. Fig 4b). The significantly positive response of *LNSIZE* to *LNCONF* can be seen during the whole period, (cf. Fig 4c). Fig 4e reveals the *LNQUAL* responses to a shock in each variable. The response of *LNQUAL* to *LNSIZE* reflects a discernible decline, which declines from 0.48 to 0.19. In parallel, the magnitude of the effect of *LNCONF* on *LNQUAL* is about -0.1 which persists over the entire forecast period.

5.3 Forecast error variance decomposition

Forecast error variance decomposition (FEVD) analysis provides the relative importance of the variance of the error made in forecasting a variable because of specific shocks of all variables in the system at a specified time horizon [18]. Fig 5 provides a graphical representation of the FEVD, where each graph depicts the proportions of forecast error variance, up to 12 periods (one year) ahead, accounted for by shocks in each variable.

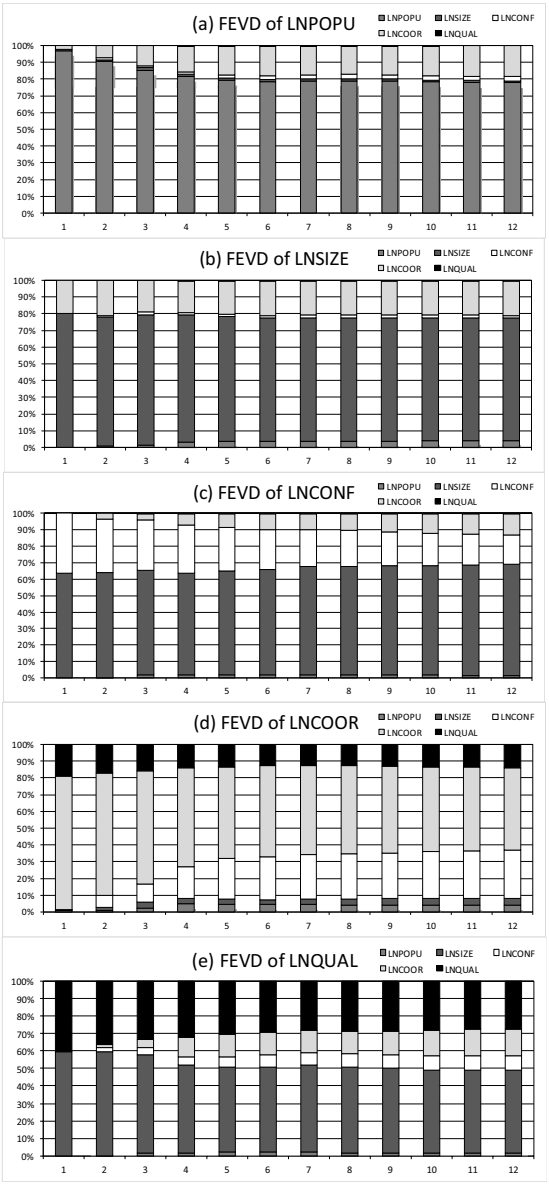


Figure 5. Forecast error variance decompositions.

As would be expected, the results of the FEVD further corroborate the results of the Granger causality tests and IRFs. Specifically, approximately 20% of the *LNPOPU* (or *LNSIZE*) error variance is accounted for by a shock in *LNCOOR* (cf. Fig 5a and Fig 5b). The *LNSIZE* accounts for the large majority of the *LNCONF* error variance, reaching nearly 67.73% at the end of the forecast period, while the explanation ability of *LNCOOR* is relatively low, only about 13% (cf. Fig 5c). Moreover, Fig 5d shows that *LNCONF* and *LNQUAL* together account for approximately 43% of the error variance in *LNCOOR*, about 29% and 14% respectively. With respect to the FEVD of *LNQUAL*, the result indicates that the *LNSIZE* has the greatest impact on *LNQUAL*, followed by *LNCOOR*, *LNCONF* and *LNPOPU* (cf. Fig 5e).

Discussion

We have systematically investigated the dynamic interrelationships among topic popularity, group size, collaborative conflict, coordination mechanism, and information quality in Wikipedia through a detailed empirical analysis. What's more, our study demonstrates the usefulness of the PSCCQ framework and VECM methodology can be applied to formally and practically analyze the dynamics of Wikipedia. Our first finding shows that the critical importance of coordination mechanism in effectively harnessing the "wisdom of the crowd" in the online collaborative environment. Our second finding shows that too many contributors involved in a particular project may be detrimental to group performance. Wikipedia managers should not necessarily pursue a more-is-better strategy towards the number of contributors.

This paper also has some limitations. First, a potential limitation of our study relates to sample data. A potential extension of the research is to apply panel vector autoregression (PVAR) to further explore the dynamics of Wikipedia with large sample. Second, we overlook the network characteristics of Wikipedia community. Future research should combine the network dynamics with the knowledge dynamics to give a fuller picture of the dynamics of Wikipedia.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 71573030) and the Doctoral Scientific Research Foundation of Shandong Technology and Business University (No. B5201705). The authors would like to thank Ye Zhang for excellent technical support and Professor Na Zhao for critically reviewing the manuscript.

References

[1] T. Yasseri and J. Kertész, "Value production in a collaborative environment," J Stat Phys, vol. **151**, pp. 414-439, (2013).

[2] M. Mesgari, C Okoli, M Mehdi, F. Arup Nielsen, and A. Lanamaki, "'The sum of all human knowledge': a systematic review of scholarly research on the content of Wikipedia," J Assoc Inf Sci Tech, vol. **66**, PP. 219-245, (2014).

[3] Y. Ren, J. Chen, and J Riedl, "The impact and evolution of group diversity in online open collaboration," Manage Sci, vol. **62**, pp. 1668-1686, (2015).

[4] J. Yun, S. H. Lee, and H. Jeong, "Intellectual interchanges in the history of the massive online open-editing encyclopedia, Wikipedia," Phys Rev E, vol. **93**, pp. 012307, (2016).

[5] T. Yasseri, R. Sumi, A. Rung, K. András, and K. János, "Dynamics of conflicts in Wikipedia," PloS One, vol. **7**, pp. e38869, (2012).

[6] F. Flöck, D. Laniado, F. Stadthaus, and A. Maribel, "Towards better visual tools for exploring Wikipedia article development–the use case of "gamergate controversy," ICWSM. pp. 1-8, (2015).
[7] J. Kimmerle, J. Moskaliuk, A. Harrer, and U. Cress, "Visualizing co-evolution of individual and collective knowledge," Inform Comm Soc, vol. **13**, pp. 1099-1121, (2010).
[8] U. Cress, I. Feinkohl, J. Jirschitzka, and J. Kimmerle, Mass Collaboration and Education, vol. **16**, Switzerland: Springer, pp. 85-104, (2016).
[9] U. Cress and J. Kimmerle, "A systemic and cognitive view on collaborative knowledge building with Wikis," Int J Comp.-Support Collab Learn, vol.**3**, (2008).
[10] S. Ransbotham, G. C. Kane, and N. H. Lurie, "Network characteristics and the value of collaborative user-generated content." Mark Sci, vol. **31**, pp. 387-405, (2012).
[11] G. C. Kane and S. Ransbotham, "Research note—content and collaboration: an affiliation network approach to information quality in online peer production communities," Inf Syst Res, vol. **27**, pp. 424-439, (2016).
[12] M. Jankowski-Lorek, S. Jaroszewicz, L. Ostrowski, and A. Wierzbicki, "Verifying social network models of Wikipedia knowledge community," Inf Sci, vol. **339**, pp. 158-174, (2016).
[13] A. Kittur and R. E. Kraut, "Beyond Wikipedia: coordination and conflict in online production groups," CSCW, pp. 215-224, (2010).
[14] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: representation, estimation, and testing," Econometrica, vol. **55**, pp. 251-276, (1987).
[15] W. Enders, Applied Econometric Time-Series, New York: John Wiley & Sons Limited, 1995, pp. 127-134.
[16] H. Lütkepohl, New Introduction to Multiple Time Series Analysis, Berlin: Springer Heidelberg, pp. 269-324, (2005).
[17] C. W. J. Granger, "Investigating causal relationships by econometric models and cross-spectral methods," Econometrica, vol. **37**, pp. 424-438, (1969).
[18] J. H. Stock and M. W. Watson, "Vector autoregressions," J Econ Perspect, vol. **15**, pp. 101-115, (2001).

Table 1. General VECM estimation results

	***D*(*LNQUAL*)**	***D*(*LNPOPU*)**	***D*(*LNSIZE*)**	***D*(*LNCONF*)**	***D*(*LNCOOR*)**
R-squared	0.417172	0.288289	0.30871	0.437351	0.363726
Adj. R-squared	0.333911	0.186617	0.209954	0.356973	0.27283
Mean dependent	-0.00396	-0.005373	-0.000913	-0.008516	0.015874
S.D. dependent	0.773045	0.219413	0.482391	1.037013	0.955626
F-statistic	5.010411	2.83546	3.125993	5.441155	4.001551

Table 2. Granger causality tests based on the VECM.

Variables	***D*(*LNQUAL*)**	***D*(*LNPOPU*)**	***D*(*LNSIZE*)**	***D*(*LNCONF*)**	***D*(*LNCOOR*)**
D(*LNQUAL*)		2.5455	5.0556*	7.8581**	6.8042**
D(*LNPOPU*)	1.0086		2.2433	4.5436	8.4576**
D(*LNSIZE*)	4.3569	1.5993		1.1003	7.9319**
D(*LNCONF*)	0.8781	1.5987	7.0872**		7.5498**
D(*LNCOOR*)	6.5173*	4.8347	2.9189	8.6165**	

Note: *, ** and *** indicate significance at 10%, 5% and 1% level, respectively.