
Jinhyuk Yun (윤진혁), Sang Hoon Lee (이상훈), and Hawoong Jeong (정하웅)

1Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
2School of Physics, Korea Institute for Advanced Study, Seoul 02455, Korea
3Institute for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
4Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk 37673, Korea

(Dated: October 22, 2015)

Wikipedia is a free Internet encyclopedia with enormous amount of contents. This encyclopedia is written by volunteers with various backgrounds in a collective fashion; anyone can access and edit most of the articles. This open editing nature may give us prejudice that Wikipedia is unstable and unreliable sources; yet many studies suggest that Wikipedia is even more accurate and self-consistent than traditional encyclopedias. Scholars have attempted to understand such extraordinary credibility, but usually used the edit number without consideration of real-time. In this work, we probe the formation of such collective intelligence through the systematic analysis using the entire history of 34 534 110 English Wikipedia articles, between 2001 and 2014. From this massive data set, we observe the universality of both timewise and lengthwise editing scales, which suggests that it is essential to consider the real-time dynamics. By considering real-time, we find the existence of various growth patterns that are unobserved in terms of the number of edits as the time step. To account these results, we present a mechanistic model that adopts both the article editing dynamics based on editor-editor and editor-article interactions. The model successfully generates some key properties of the real Wikipedia articles such as distinct types of articles for the editing patterns characterized by the interrelationship between the numbers of edits and editors, and the article size. In addition, the model indicates that infrequently referred articles tend to grow faster than frequently referred one, and articles attracting high motivation of edit counterintuitively reduce the number of participants. We suggest that this decay of participants eventually brings inequality among the editors, which will be more severe with time.

I. INTRODUCTION

Humankind has accrued a priori knowledge since the onset of Homo sapiens. From ancient cave paintings to modern research papers, the species’ desire towards sedimentation has been displayed as documentary. Encyclopedia, a set of documents that contains a vast collection of information from the entire fields of human knowledge, has played a pivotal role to disseminate these legacies [1, 2]. Conventionally, a group of experts devote their expertise for these encyclopedias [3, 4]. Taking advantage of technological developments, media that publish encyclopedia keep abreast of the times: hand-writing, letterpress printing, and optical disks. Emergence of information technology has opened the new era of publishing traditional encyclopedia in World Wide Web [5], which offers a variety of references and up-to-date information. Although these new media can reduce publication price, encyclopedia editing is still costly.

Besides the improvement of traditional encyclopedias, new media enable fresh challengers to participate in the competition. Wikipedia [6], a representative player among the challengers, has proposed an entirely new manner: editing by volunteers with various backgrounds in a collective fashion. This new paradigm of sharing knowledge is one of the most famous examples of “collective intelligence.” However, due to the nature of open-edit policy, Wikipedia does not guarantee that contents are valid [7], thus it is regarded ambiguous and even inaccurate to utilize in scientific context [8, 9]. Despite such a long-standing bias against the credibility of Wikipedia, many studies suggest that Wikipedia is more reliable than our prejudice; Wikipedia itself tends to refer reliable scientific sources [10, 11], and only 13% of Wikipedia articles contain perceptible academic errors [12] and the quantity of factual errors, omissions, or ambiguous statements in scientific context of Wikipedia is comparable to traditional encyclopedias [13].

Gradually, prejudice against quality of Wikipedia’s articles has been eroded and the number of citations to Wikipedia in peer-reviewed scientific articles has increased over time [14]. A bizarre gap between such prejudice and the actual situation appeals to the scholars, who have analyzed Wikipedia’s external characters and internal dynamics. For example, researchers have investigated editors of Wikipedia and their editing patterns [15, 16], and occurring and resolving of conflicts in Wikipedia [17, 18, 19]. Despite the significant contributions of such endeavors, the previous studies mainly focus on the raw number of edits, and often neglect real-time and the different editing patterns for articles with different sizes and ages. In this paper, we examine an exhaustive set of English Wikipedia articles to understand how the article size and age displays external appearance in this open-editing encyclopedia. In particular, a simple time-rescaling method reveals articles belonging to various types, when we take account of the interrelation between observable parameters: the number of edits, the number of editors, and the article size.

Our analysis consists of both data analysis and modeling based on it. First, we use the entire edit history in Wikipedia to inspect Wikipedia’s growth, mainly focused on the number of edits, the number of editors, and the article size. In this process, we demonstrate that the consideration of real-time is essential to understand the underlying dynamics be-
hind the present Wikipedia. Second, to consider the forma-
tion of current Wikipedia in more detail, we develop an agent-
based model that imitates the interplay between an article and
the editors in a society. Our model shows inherent differences
of articles belonging to different types of growth pattern. The
results are consistent with real data, which suggests that a so-
ciety’s attitudes on Wikipedia articles determine the growth
pattern. We believe that this approach provides valuable in-
sights for the formation of collective knowledge. We focus on
the long-term formation of collective knowledge, which has
significant effects on progresses of humankind over a variety
of temporal scales. We hope that our work provides the in-
sights to solve some of the fundamental questions: why people
collaborate, how the collective memory is formed, and how knowledge is spread and descended.

The rest of the paper is organized as follows. In Sec. II we
introduce the Wikipedia data that we use in our investigation.
In Sec. III we propose a time-rescaling method and show that
the articles in Wikipedia can be classified into multiple types
based on their growth pattern. We present our model and re-
sults in Sec. IV including verification of our model with real-
data. Finally, we present conclusions and discussion in Sec. V.

II. DATA SET

For the analysis, we use the December 2014 dump of En-
glish Wikipedia. This dump contains the complete copy of
Wikipedia articles from the very beginning up to December
8, 2014, including the raw text source and metadata source
in the Extensible Markup Language (XML) format [20]. In
this data set, there are a total of 34,534,110 articles across all
categories with the full history of edits. Each article docu-
ments either Wikipedia account identification (ID) or Internet
protocol (IP) address of the editor for each edit, article size
and timestamp for each edit, etc. A single character in En-
glish takes one byte, so the article size is the direct measure
of article length [21]. There are 587,606,564 editing events
(“edits" from now on) for the entire Wikipedia articles in to-
tal, where individual articles’ edit numbers are ranged from 1
to 1,044,048.

Previous studies tend to sample data sets for various rea-
sons, and thus articles with small numbers of edits are nec-
essarily filtered out [8, 9, 12, 18]. However, Fig. 1 suggests
a fat-tailed distribution for the number of edits, so the major-
ity of articles are not edited as many times as the articles in
the tail part of the distribution and those articles should not be
neglected. Therefore, we consider all entries and use the en-
tire set for analysis. Additionally, we use ID and IP address,
for logged-in editors and unlogged-in editors, respectively,
to identify distinct editors. In total, 40,536,451 editors have
participated in the establishment of the current Wikipedia.
Among them, 83.9% (34,006,990) of editors are unlogged-in
and only 16.1% (6,529,461) of editors are logged-in. Interest-
ingly, the absolute share of logged-in editors are rather smaller
than that of unlogged-in editors, most of heavy editors tend to
be logged-in (Fig. 2).

There are possible biases for IP addresses when an IP ad-
dress is shared, e.g., editors who use library, public WiFi, vir-
tual private network (VPN), etc., or move the locations. In
those cases, there will be under- or over-estimation on the
number of distinct editors. Additionally, a small number of
edits does not specify the editor, yet we use other informations
even for such cases based on the article size and timestamp.

III. DATA ANALYSIS

Previous studies on Wikipedia data set did not use the infor-
mation about article size changes after the edits [18, 19] or real
timestamps of the edits [15, 17]. We combine such informa-
tion together with conventional measures, such as the number
FIG. 3. Distribution of time between two consecutive edits in Wikipedia. Each symbol corresponds to the range of ages to which articles belong. Specifically, ‘n-th year’ corresponds to the edit event occurring between first and last days of the ‘n-th year’ since the onset of the article. The time differences follow fat-tailed distribution that is a sign of the burstiness, with a daily periodic pattern (1 day = 86,400 seconds).

FIG. 4. Size difference between two consecutive edits in Wikipedia, in the unit of kilobyte (KB). Each symbol corresponds to the range of articles’ size when the latter of two consecutive edits occurred.

of edits and the number of editors, to display the nature of Wikipedia. Our first analysis of time and size differences between two consecutive edits reveals regularity, regardless of articles’ age and size.

### A. Edit Scale of Wikipedia

The time between the consecutive edits follows a fat-tailed distribution with characteristic periodicity from the human circadian rhythm (Fig. 3), which suggests that the editing timescale of Wikipedia is intermittent or “bursty,” meaning that brief but intense activities are followed by much smaller activities for a long time [22][23]. These intense activities in Wikipedia are reported as “Wikipedia Edit War,” which refers to significantly rapid consecutive editing by various editors with conflicting opinions [24]. Our observation indicates that the “edit number” (or the number of edits), which many studies use as the proxy of the real time [17][19], is not an unbiased proxy of the time. Counterposed to the assumption that English Wikipedia has already become global media, we observe strong periodicity for the time between the consecutive edits in Fig. 3. The peaks are located at every 86,400 seconds or a single day, which implies that English native speakers (mostly people in the United States because of the relative population, we presume) still dominate English Wikipedia even though there is no barrier to the global access.

Besides the timescale, we observe that an article’s growth is mainly addition and subtraction with a characteristic size scale, which are rather independent of the current size (Fig. 4). This observation is counterposed to the recent report that growth of collaborated open-source software and mammalian body masses are proportional to their size [25], and implies that an influence of a single edit becomes smaller as article size is increased. Most previous researches [17][19] do not take account of the degree of the influence for a single edit, and thus consider all of the edits as affecting the article of Wikipedia equally. However, our observations propose the necessity of combining the time and size difference between the edits with the conventional measures.

### B. Time-rescaled Measures for Wikipedia

In Sec. III A, we have shown that the time between two consecutive edits is quite heterogeneous (Fig. 3). This global effect of various timescales itself makes it unfair to directly compare characteristic parameters of articles: the number of editors, edits, and the article size for different articles. To compensate such an effect, we employ rescaled measures for article $i$ as $N_e^*(i) = N_e(i)/T(i)$, $N_p^*(i) = N_p(i)/T(i)$, and $S^*(i) = S(i)/T(i)$, where $T(i)$, the age of article $i$, is measured as the time between the moment of onset and that of the latest edit of article $i$. $N_e(i)$, $N_p(i)$, and $S(i)$ are the number of edits, the number of editors, and the article size for article $i$, respectively. The rescaled measures are free from the temporal effects, making it possible to recruit myriad articles into the same ground for analysis in the sense of growth per unit time. For the number of edits, the number of editors, and the article size from the data, we hereafter use their rescaled values unless stated otherwise.

A natural step forward is to search for any possible interplay between $N_e^*$, $N_p^*$, and $S^*$ in the formation of current Wikipedia. One can suppose that the number of edits has varied gradually as a function of the number of editors, because both measures reflect the degree of interests in the article. Unexpectedly, we discover that the articles shows peculiar bimodality in their number of editors across entire value of the number of edits [see Fig. 5(a)]. The bimodality is characterized by the linear relation $N_e^* = k_1 N_p^*$ with two distinct proportionality constants, $k_1 = k_1^{\text{upper}} \approx 0.9$ and $k_1 = k_1^{\text{lower}} \approx 0.5$, respectively. In other words, there are two groups of articles, determined by the proportion between the number of editors involved in the

articles and the editors’ average activity; one group is dominated by relatively a small number of enthusiasts who edit articles frequently, and the other is composed of a relatively large number of editors who seldom edit articles. Besides the cases of edits and editors, Wikipedia shows a similar division of article size for given numbers of editors [see Fig. 5(b)]. There are two types of articles determined by the average article size produced by an editor per unit time. This relation is also described by the linear dependency \( S^* = k_2N_p \), where \( k_2 = k_{upper} \approx 900 \) for the upper mode and \( k_2 = k_{lower} \approx 25 \) for the lower mode. In other words, editors for some articles have generated about 900 bytes on average, meanwhile the editors of rest of the articles have generate only about 25 bytes on average.

Our finding of bimodality in the two relations \((N_p^* \text{ versus } N_p \text{ and } S^* \text{ versus } S^*)\) triggers an interesting question: does each of the modes in one relation correspond to each mode in the other relation [Figs. 5(a) and (b)]? It seems natural to speculate that such modes have the counterparts in the other relation, or at least one is subordinative of the other. Contrary to this speculation, our observation suggests that there is no visible relationship between the two different types of bimodality [see Figs. 5(c)–(f)]. The points in the figures are colored according to the modes to which the corresponding articles belong in the criteria based on \( N_p^*/N_p^* \text{ or } S^*/N_p^* \). We simply tear off the upper and lower modes by drawing a line [the green lines in Figs. 5(c) and (f)] between the two modes and assign purple and blue colors for the points in the upper and the lower modes, respectively. Those purple and blue points are totally mixed when the criterion is based on the other parameter relation. Taken together, we conclude that there are at least four different groups of articles, which can be categorized by its growth per unit time. The possible mechanism behind the division is suggested based on our modeling study in Sec. IV.

**IV. MECHANISTIC MODEL OF WIKIPEDIA**

To understand the underlying dynamics of the observed patterns, we develop a mechanistic model of editing dynamics by identifying two key factors that drive the evolution of Wikipedia articles. We assume that there are fundamental and inherent properties of an article, in particular, reflecting the society’s viewpoint on the article’s topic. First, Wikipedia is a more reliable source for certain topics than other conventional media or friends. Because of the nature of open-edit policy, there are long-lasting arguments of credibility especially for the scientific contexts. As the result, people avoid referring Wikipedia to reinforce their contention for scientific topics when they debate. Nevertheless, several topics are almost free from the trust issue and Wikipedia can be considered as trustworthy source of knowledge. The subcultures such as animations, movies, and computer games are good examples, because the editors are not only the fan of the topic but also the creators of such cultures. In those cases, therefore, members of a society do not hesitate to utilize Wikipedia as their grounds for the arguments.

Second, there are different levels of psychological barriers in editing, depending on the topic. People tend to edit the article about which they have enough knowledge. Thus, average “editability” of articles, for members of a society, is diverse by its nature from the casual ones which are easily editable to the formal ones. This editability also depends on collective motives, which describe the significance of the topic as the common goal of social movements. Therefore, the intrinsic rate of edit should be taken into account. Besides these two key factors, editors are also engaged in articles when they have already given more efforts to the articles by editing them, representing the feeling of attachment. Additionally, it is hard to edit an article that already has massive amount of information, so the motivation to edit will be reduced as the article size is increased.

**A. Model Description**

By incorporating the aforementioned factors, we create a mechanistic model of the article growth. The model comprises \( N \) agents where the individual agents represent members of a society connected to each other randomly, and all of the agents are also connected to a single Wikipedia article. Note that we take a single Wikipedia article in our model, as we assume that different model parameters yield different types of articles in real Wikipedia. Such connections between agents stand for various relationships in society: friends, coworkers, even enemies. Every agent \( i \) has its own opinion \( O_i(t) \in [0,1] \) (real numbers between 0 and 1) at time \( t \), where 0 and 1 are the two extremes of conflicting opinions. Initially, \( O_i(t = 0) \) is assigned as randomly generated number from the uniform distribution in the interval \([0,1] \). The Wikipedia article also has its own opinion \( W(t) \) at time \( t \), which is the overall stance of Wikipedia on the topic. We set \( W(t = 0) = 1 \), to get the insights of the situation that agents and the Wikipedia article adjust their opinions to the most radical one. We also introduce the length of the article \( L_o(t) \) with the initial value \( L_o(t = 0) = 1 \), where 1 is the minimum length that agent can reduce the article, so \( L_o(t) \geq 1 \) always. For each time step \( t \), the agent-agent interaction described in Sec. IV A1 and the agent-Wikipedia dynamics described in Sec. IV A2 occur in turn.

1. **Agent-Agent Dynamics**

Our model colligates resolving of conflicts between agents with the contribution of agents to modify Wikipedia. In our model, all members of society are open-minded and they can change their mind. For each timestep \( t \), a pair of agents are chosen and they try to convince each other for the topic of the Wikipedia article. We assume that agents rely on references to reinforce their opinion. There are two types of references: Wikipedia and general media. General media represent the ordinary viewpoint of the society toward the topic. As we described above, Wikipedia is more reliable sources for certain topics. Hence, we set a probability \( p \) corresponding to the re-
FIG. 5. Interplay between time-rescaled measures for Wikipedia in December 2014. Each point corresponds to an individual Wikipedia article. For panels (a)–(f), we randomly sampled one article per every 100 articles to display clearly, yet results does not change by sampling. Note that bimodality is observed only when we check these time-rescaled measures. (a) The number of edits $N_e$ versus the number of editors $N_p$. (b) The number of editors $N_p$ and the article size $S^*$. The black dashed guidelines in panels (a) and (b) have the slope $1 (N_p = 0.7N_e$ and $S^* = 100N_p$), implying the linearity. The bimodal structure in (a) is highlighted in (c) by manually dividing the upper (purple) and lower (blue) modes via the black dashed guideline in (a), but this division of modes does not divide the modes of (b), as shown in (d) where we show the same plot as (b) but with the individual points’ color of (c). Likewise, the bimodal structure in (b) is highlighted in (f) by manually dividing the upper (purple) and lower (blue) via the black dashed guideline in (b), but this division of modes does not divide the modes of (a), as shown in (e) where we show the same plot as (a) but with the individual points’ color of (b). The axes of (c) and (e) are identical to (a) and those of (d) and (f) to (b), so we omit the axes in (c)–(f) for better visualization.

liability of Wikipedia, and agents choose Wikipedia as their reference with the probability $p$ [see Fig. 6(a)]. Otherwise, agents decide to follow standards of society by following general media’s opinion, which is set as the average opinion of entire agents in the system. In other words, the reference opinion

$$ R(t) = \begin{cases} W(t) & \text{with probability } p \\ G(t) = \frac{1}{N} \sum_i O_i(t) & \text{with probability } 1 - p. \end{cases} $$

Then, as the interaction between agents, we select an agent $i$ and its random neighbor $k$ in the network. Without loss of generality, suppose that agent $i$ is the one whose opinion is closer to the reference, i.e., $|O_i(t) - R(t)| \leq |O_k(t) - R(t)|$ (if it is not the case, we just switch the update rules for $i$ and $k$) [see Fig. 6(a)]. The opinions of agents $i$ and $k$ at the next step $t + 1$ are given by $O_i(t + 1) = O_i(t)$ and

$$ O_k(t + 1) = \begin{cases} (1 - \epsilon_m)O_k(t) + \epsilon_mO_i(t) & \text{if } |O_i(t) - O_k(t)| \leq |O_k(t) - R(t)| \\ (1 - \epsilon_m)O_k(t) + \epsilon_mR(t) & \text{otherwise,} \end{cases} $$

where $\epsilon_m \in [0, 1]$ represents the tolerance parameter of agents, which indicates the psychological limit to change the opinion.

2. Agent-Wikipedia Dynamics

A distinct character of our model, compared to opinion spreading models with external field [34], is that agents can also modify the media, which corresponds to Wikipedia in our model. Suppose that agent $e$ is chosen, then it modifies the Wikipedia article with the probability

$$ r_e(t) = q[W(t) - O_e(t)]A_e(t) + 1 \sum_{s \neq e} A_s(t) + 1 L(t), $$

where $q$ is base activity, and $A_e(t)$ is attachment of agent $e$ for the Wikipedia article [30], which is assigned as 0 at the beginning and increased by unity every time an agent edits the Wikipedia article [see Fig. 5(b)]. The term $1/L(t)$ in Eq. 6 represents the reduced motivation as the article size is increased, due to the amount of information [27]. A recent report that growth of Wikipedia is slowed down supports this assumption [35]. If an agent decides to edit an article, Wikipedia’s opinion changes as

$$ W(t + 1) = W(t) + \epsilon_e[O_e(t) - W(t)]/L(t), $$

where $\epsilon_e$ represents the physical and psychological limit for editing [25]. Figure 4 indicates that the impact of a single
FIG. 6. Schematic diagrams describing how our model works. (a) Flow diagrams of the interaction between agents $i$ and $k$ described in Sec. IV A 1) and IV A 2) are crucial to generate the splits of modes into different groups: $q$ is essential to reproduce a separation of $N_p^*(i)/N_p^*(i)$ [see Fig. 7(a)] and $p$ is indispensable for the division of $S^*(i)/N_p^*(i)$ [see Fig. 7(b)]. In the early stage, $N_p^*(i)/N_p^*(i)$ is almost unity across the systems with the entire

edit event should be decreased as the article size is increased, because the absolute amount of changing is preserved.

After the update, the article length is updated as

$$L(t+1) = L(t) + C_i(t),$$

where the random variable $C_i(t)$ is chosen according to the following rule. If the agent has modified the article toward an extreme position (0 or 1), we suppose that the agent tends to append new contents to the article. In contrast, agents are likely to replace the contents to neutralize the article’s opinion. If $O_i(t) \leq W(t) \leq 1/2$ or $O_i(t) \geq W(t) \geq 1/2$, the article size is increased by $C_i(t)$ drawn from the interval $[0, \epsilon_i]$ uniformly at random, to reinforce the argument. Otherwise, we consider that the article is neutralized and $C_i(t)$ is drawn from the interval $[-\epsilon_i/2, \epsilon_i/2]$ uniformly at random. The fixed parameter $\epsilon_i$ is related to the physical limit in Fig. 4. See Fig. 6(b) for the illustration on the $C_i(t)$ criterion. In Sec. IV B we discuss how the modes in Fig. 5 are formulated during the evolution of Wikipedia in our model.

B. Model Results

1. Conditions for Bimodality

For both $N_p^*(i)$ versus $N_p^*(i)$ and $N_p^*(i)$ versus $S^*(i)$ relations, our mechanistic model captures essential features of the observed empirical relations reported in Sec. III B with proper parameter values. As we shown in Sec. III, the proportionality coefficients between characteristic parameters are classified into two modes:

$$N_p^*(i)/N_p^*(i) \approx 0.5 \text{ or } 0.9$$

$$S^*(i)/N_p^*(i) \approx 25 \text{ or } 900$$

In particular, both $p$ for the agent-agent interaction (in Sec. IV A 1) and $q$ for the agent-Wikipedia interaction (in Sec. IV A 2) are crucial to generate the splits of modes into different groups: $q$ is essential to reproduce a separation of $N_p^*(i)/N_p^*(i)$ [see Fig. 7(a)] and $p$ is indispensable for the division of $S^*(i)/N_p^*(i)$ [see Fig. 7(b)]. In the early stage, $N_p^*(i)/N_p^*(i)$ is almost unity across the systems with the entire
parameters space composed of \( p \) and \( q \), which corresponds to the single (or unimodal, in contrast to the bimodal pattern shown in real data) linear relation. While this single linear relation is characterized at the early stage, as time goes by, we observe the decreasing trend of \( q \). Despite the fact that the \( N_*^q(i)/N_*^p(i) \) decrement over time occurs for the entire parameter space, the pace of decreasing is determined by \( q \), the base rate for editing an article. \( N_*^q(i)/N_*^p(i) \) drops much slower for smaller \( q \) values, which leads systems to fall into two different regimes: \( q \geq 0.5 \) and \( q \leq 0.5 \) (Fig. 7[a]). Interestingly, this divarication solely depends on the value of \( q \). On the other hand, \( S^*(i)/N_*^q(i) \) also shows unimodality in the early stage, but it is suddenly increased with time only for \( p \leq 0.5 \) (Fig. 7[b]). Analogous to \( N_*^q(i)/N_*^p(i), S^*(i)/N_*^p(i) \) is also almost solely driven by \( p \), but there also exists a small amount of influence by \( q \); small values of \( q \) do not guarantee the large article size across the entire \( p \) values, but low \( q \) yields large article size at similar values of \( p \).

2. Model Verification with Real Data

Based on our model results, we suggest a possible mechanism that yields the bimodality in Fig. 5, which encourages us to verify the model results compared to the real data: either parameter \( q \) or \( p \) should be a decreasing function of \( N_*^q/N_*^p \) and \( S^*/N_*^p \), respectively. However, we cannot extract simulation parameters \( p \) and \( q \) directly from the data. We therefore use a bypass to estimate \( q \) and \( p \). Fortunately, Wikipedia offers page view statistics of articles that can be used for estimate such parameters. We assume that this page view in a certain period reflects the degree of interest of Wikipedia users in the articles, and the number of edits in the same period naturally displays the editing frequency. Thus, the ratio of the number of edits to this page view for a certain period can be related to the base edit rate \( q \). Analogous to our presumption, this ratio is a decreasing function of \( N_*^q/N_*^p \) (see Fig. 8).

To treat the other parameter \( p \), we should employ the proxy that can reflect general interest of the entire society in the topic. We suggest that the Google books n-gram, a vast digitized collection of documents produced in the world is a suitable choice. Google books n-gram is a database containing about 6% of English books ever published. This data set offers yearly number of occurrences for any phrase less than 6 words from 1800 to 2008, and this number of occurrences can be considered as the proxy of interests in society for certain phrase. In our model, \( p \) is the proportion of degree of interest in Wikipedia versus that of the entire society. In other words, Wikipedia page view on a certain topic versus its n-gram frequency can be the estimator of \( p \). For fair comparison, we only take the Wikipedia articles that satisfy the following conditions. First, the title of article should exist in Google 1-gram data set in 2008, the latest year of the data set. Second, the article should be visited at least once in 2014. To avoid the effect of inflectional variation of words, we use

FIG. 7. Snapshots of (a) \( \langle N_*^q(i)/N_*^p(i) \rangle \) and (b) \( \langle S^*(i)/N_*^p(i) \rangle \), where \( \langle \cdot \cdot \cdot \rangle \) is the averaged quantity over all of the agents (= \( \sum_i \cdots /N \)) at five time points from \( t = 3000 \) to 185000. \( N = 10^5 \) agents are connected by the Erdős-Rényi random network with connecting probability 0.1. Although we present the results from the random network case, our results hold for other networks, e.g., the static scale-free networks [31], as well as the fully connected network case. Each point is the average of 1000 independent runs of simulation with \( \epsilon_e = \epsilon_w = \epsilon_i = 0.1 \). (a) \( \langle N_*^q(i)/N_*^p(i) \rangle \), which corresponds to the proportionality constant in Fig. 5(a). We observe the ratio gradually falls into different regimes that are determined by \( q \) as time passes. The plots in panel (b) represent \( \langle S^*(i)/N_*^p(i) \rangle \) corresponding to Fig. 5(b).
The heterogeneity for the ratio of the number of editors to that of edits, \( \frac{N^*_p}{N^*_e} \), leads us to the eventual question: is this heterogeneity from structural inequality? In other words, does the existence of dictatorship or monopoly of small group editors, or super editors \([17]\), make it difficult for others to participate in editing processes? To find the answer, we use the Gini coefficient, which is a common measure for inequality in economics \([40]\) ranged from 0 for the minimal inequality (or the maximal equality) to 1 as the maximal inequality. We consider the number of edits for individual editors as the wealth variable in Gini coefficient. The trend of the Gini coefficient as a decreasing function of \( \frac{N^*_p}{N^*_e} \) shown in Fig. 9(a) suggests the modes with slope \( \approx 0.9 \) and \( \approx 0.5 \) in Fig. 5(a) are in equilibrium and non-equilibrium states, respectively.

Additional analysis of the Gini coefficient in terms of the \( q \) estimator (the ratio of the number of edits in 2014 to the page view statistics in 2014) also indicates that the larger \( q \) induces more severe inequality for editing [see Fig. 9(b)]. This is counterintuitive because it actually means that articles inducing larger motivation to edit eventually set a larger barrier to participate in editing. It is doubtful that the phenomenon is caused by the amount of information \([27]\), since the Gini coefficient does not vary much according to its amount of information [see Fig. 9(c)].

Similar to the real Wikipedia, our model also supports the observed inequality. Although we use a simplified estimator of \( q \) in our real data, the ratio of the number of edits in 2014 to the page view statistics, the Gini coefficient is a increasing function of \( q \) in the model as in the real data [see Fig. 9(d)]. Additionally, since \( q \) has a limited effect on the article size (see Fig. 5), the model observation of the Gini coefficient is compatible to our observation that article size does not have large effect on the Gini coefficients. Such logical elimination suggests that few engaged and dominating editors make it indeed hard for laypersons to participate in editing processes. There are “democratic” articles (with slope of \( \approx 0.9 \) in Fig. 5) and “dictatorial” articles (with slope of \( \approx 0.5 \) in Fig. 5). In short, inequality exists indeed.

V. CONCLUSION

Traditionally, collaboration used to be mainly regional and face-to-face interactions were demanded, which had prevented the world-wide formation of collective intelligence. Nowadays, improvements of modern information technology bring us a whole new stage of online collaboration. In this study, we have examined such a new paradigm of collective intelligence through long-term data from Wikipedia \([6]\). People believe that such a new paradigm will eventually yield democratization of knowledge \([41]\). As a representative medium, Wikipedia is also considered as a spearhead of such pro-democracy movements \([42]\).

However, our observation suggests that current status of Wikipedia is still apart from the perfect world-wide democracy. The observed periodicity for the time between edits al-
FIG. 9. The Gini coefficient values for Wikipedia articles classified by various measures. Each sample in the data points corresponds to an individual Wikipedia article. The average coefficient is plotted as functions of (a) $N_p^*/N_e^*$, (b) the estimator of $q$: ratio of the number of edits in 2014 to the page view statistics, and (c) the article size at the end of 2014. Panels (d) and (e) display the observation from our model, as functions of (d) $q$ and (e) the timestep. For panels (a)–(d), the shaded areas correspond to the standard deviation of Gini coefficient for given values on the horizontal axis. Panels (a)–(c) with the Wikipedia data are drawn from the same sampled set of 678,255 articles used in Fig. 8. The model results in (d) and (e) are also done by the same condition as in Fig. 7, but with the fixed $p = 0.5$.

Incluces that the English Wikipedia is still regional for English natives (see Fig. 3). Bimodality and its inequality index suggests that there are articles dominated by small number of super editors (Figs. 5–9). Notwithstanding the fact that there is no explicit ownership for Wikipedia articles, some kind of privatization by dedicated editors for given topics are happening in reality. The value of such dedicated editors should not be devalued, of course. Their dedication has indeed played the main role to keep the current state-of-the-art accuracy in the current Wikipedia [12, 13]. However, in the long run, knowledge cannot survive without collaboration between experts and society [38]. Although most of advanced knowledges are invented by experts, such experts occupy a rather small proportion in a society; thus, knowledge without supports from other members of the society will lose its dynamic force to sustain. Additionally, despite our findings that the amount of contents created by an editor ($S^*/N^*$) mainly depends on the degree of referring Wikipedia (namely $p$), an equitable opportunity for participation also increases such individual productivity (see Fig. 7).

Our study not only gives a significant insight on the formation and current state of Wikipedia, but also offers the future direction of Wikipedia. Our simulation results suggest that such inequality is increased with time, which may result in less productivity and less accuracy as by-product in the future than now [see Figs. 7(a) and 9(e)]. It is indeed already reported that growth of Wikipedia slowing down [35] and our observation suggests that it will become more slower if we do not take any active action. To sustain collaborating environments, it is worth giving more motivation and incentives to the newbies to reduce monopolized structure in Wikipedia. We hope that extending our approach to various collaboration environments such as open-source movement [29, 33] might give us the insight for the future investment that brings us a new level of collaborating environments. Finally, we would like to emphasize that the result and implication of our study are not restricted to the Wikipedia or online collaboration systems, but have much wider applications in human or non-human interactions in the world.
ACKNOWLEDGMENTS

We appreciate Beom Jun Kim (金범준), Pan-Jun Kim (김관준), and Hyunggyu Park (박형규) for insightful comments. This work was supported by the National Research Foundation of Korea through Grant No. 2011-0028908 (J. Y. and H. J.).
