Supplementary Information 1

January 18, 2019

We have studied both real and fake corpora in this work. Only part of them are listed in Table I in the main text and their reference is listed in Tab. 1 here. Data of other corpora can be found in Supplementary Information 2. To clarify the origin of scaling structure, we generate three kinds of fake corpora:

1 Paper generator

We use SCIgen[1] to generate random Computer-Science research papers.

2 N-gram generator

Based on Sinica Corpus[2], we construct the data set of Chinese words. The N-gram generator[3] groups N random Chinese characters into one word. After building the word bank, the user can arrange for these words to obey Zipf's or other distribution such as Gaussian.

3 Mixed text

A text composed of work by various authors is also created. We generate two different mixed texts: Newspaper[4] and Mix (various authors). The former is the collection of China Times, Apple Daily, Liberty Times, etc. over the period of 2016∼2017, while the latter is the combination of three books[5, 6, 7].

Table 1: Summary of reference to Table I in the main text

No.	Sample	Reference
1	*Moby-Dick*	[8]
2	*The Hobbit*	[9]
3	Xu Zhimo 徐志摩	[10]
4	*Frog* 蛙	[5]
5	*Demi-Gods and Semi-Devils* 天龍八部	[11]
6	LIGO 2016	[12]
7	Chopstick	[13]
8	Empirical Test of Zipf	[14]
9	Newspaper	[4], Section 3
10	Mix (various authors)	Section 3
11	Paper generator	[1], Section 1
12	1-gram Fake	Section 2
13	2-gram Fake	Section 2
14	3-gram Fake	Section 2
15	4-gram Fake	Section 2
16	2-gram log-normal	Section 2
17	2-gram double power law	Section 2
18	2-gram exponential	Section 2
19	2-gram Gaussian	Section 2
20	Excerpts from Frog	[5]
21	Excerpts from Moby-Dick	[8]

References

[1] Stribling, J., Krohn, M. & Aguayo, D. SCIgen - An Automatic CS Paper Generator. Retrieved from `https://pdos.csail.mit.edu/archive/scigen/` (2017)

[2] Academia Sinica. Sinica Corpus. Retrieved from `http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm` (2017)

[3] Wu, Shan-Jyun. Fake Script Vietnamese. Retrieved from `http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm` (2017)

[4] Collected by Tsai, S. T. News collection from China Times, Apple Daily, Liberty Times, etc. (2016, 2017) This text file is included in Supplementary Information 2.

[5] Mo Yan 莫言. *Frog* 蛙. ISBN: 9787532136766 (Shanghai Wenyi, Shanghai, 2009)

[6] Yu Guangzhong 余光中. *Selected Poetry of Yu Guangzhong* 余光中詩選. ISBN：9576742730 (Hongfan Bookstore, Taipei, 2006)

[7] Chang Show-Foong 張曉風. *Zhang Xiaofeng Prose Collection* 張曉風散文集. ISBN: 9574441407 (Chiu Ko Publishing, Taipei, 2004)

[8] Herman Melville. *Moby-Dick*. Urbana, Illinois: Project Gutenberg. Retrieved from `https://www.gutenberg.org/ebooks/2701` (2008)

[9] Tolkien, J. R. R. *The Hobbit*. ISBN: 0618260307 (HMH Books, Boston, 2002)

[10] Xu Zhimo 徐志摩. Selected poems. Retrieved from `http://w3.loxa.com.tw/fxp6033/poet01.htm` (2017).

[11] Jin Yong 金庸. *Demi-Gods and Semi-Devils* 天龍八部. ISBN: 9789573256748 (Yuan-Liou Publishing, Taipei, 1996)

[12] Abbott, B. P.*et al.* Observation of Gravitational Waves from a Binary Black Hole Merger. Phys. Rev. Lett. **116**, 061102 (2016)

[13] Tsai, S. T. *et al.* Acoustic Emission from Breaking a Bamboo Chopstick. Phys. Rev. Lett. **116**, 035501 (2016)

[14] Maillart, T. , Sornette, D. Spaeth, S. & Von Krogh, G. Empirical Tests of Zipf's law Mechanism In Open Source Linux Distribution. Phys. Rev. Lett. **101**, 218701 (2008)

[15] Wu Jingzi 吳敬梓. *The Scholar* 儒林外史. Retrieved from `https://zh.wikisource.org/zh-hant/%E5%84%92%E6%9E%97%E5%A4%96%E5%8F%B2` (2017)

[16] Lung Yingtai 龍應台. 野火集. ISBN: 9576070589 (Eurasian Press, Taipei, 1985)

[17] Haruki Murakami 村上春樹. Kafka on the Shore (Chinese version) 海邊的卡夫卡. ISBN:2966622172 (China Times Publishing, Taipei, 2017)

[18] Jin Yong 金庸. *Mandarin Duck Blades* 鴛鴦刀. ISBN: 9789573256748 (Yuan-Liou Publishing, Taipei, 1996)

[19] Ni Kuang 倪匡. Wisely 衛斯理系列-第二種人. ISBN:9576455936 (Storm & Stress Publishing, Taipei, 1995)

[20] Zheng Chou-yu 鄭愁予. Selected poems. Retrieved from `http://w3.loxa.com.tw/fxp6033/poet03.htm` (2017).

[21] Rowling, J. K. *Harry Potter and the Sorcerer's Stone.* ISBN:0439554934 (Arthur A. Levine Books, New York, 1997)

[22] Rowling, J. K. *Harry Potter and the Chamber of Secrets.* ISBN:0439064864 (Arthur A. Levine Books, New York, 1999)

[23] Rowling, J. K. *Harry Potter and the Prisoner of Azkaban.* ISBN:043965548X (Scholastic, New York, 2004)

[24] Rowling, J. K. *Harry Potter and the Goblet of Fire.* ISBN:0439139600 (Scholastic, New York, 2002)

[25] Rowling, J. K. *Harry Potter and the Order of the Phoenix.* ISBN:0439358078 (Scholastic, New York, 2002)

[26] Rowling, J. K. *Harry Potter and the Half-Blood Prince.* ISBN:0439785960 (Scholastic, New York, 2006)

[27] Rowling, J. K. *Harry Potter and the Deathly Hallows.* ISBN:0545010225 (Scholastic, New York, 2007)