Proc. 13th Econophysics Colloquium (EC) and 9th Symposium of Physics in Economy and Social Sciences (FENS), 2017

The Heaping Effect in the Survey Data on the Number of Friends

M. Nawojczyk[a,*], M. Stojkow[a], D. Żuchowska-Skiba[a], M.J. Krawczyk[b] and K. Kułakowski[b]

[a]AGH University of Science and Technology, Faculty of Humanities, Al. Gramatyka 8a, 30-071 Kraków, Poland
[b]AGH University of Science and Technology, Faculty of Physics and Applied Computer Science, Al. Mickiewicza 30, 30-059 Kraków, Poland

The number of acquaintances is relevant for modeling social networks. Here we consider the data on the declared number of friends, as collected in 2000, 2007 and 2015 from Polish respondents above the age of 50. We demonstrate that the answers on the number of friends show sharp maxima at 10, 15, 20 and sometimes 30, which accompany a broader peak between 0 and 8. These results do not change qualitatively with sex and age of the respondents. The effect, known as data heaping, can be detected as a deviation from the Benford law.

DOI: 10.12693/APhysPolA.133.1417

PACS/topics: social systems, random walk, exit time, incomplete knowledge

1. Introduction

The number of acquaintances has been highlighted by Dunbar as a relevant measure of social functions of brain [1, 2]. This number is also relevant for modeling social networks [3–5]; some data refer to the number of friends, depending on an accepted definition of a social bond. Our aim here is to report a peculiar character of some survey data on the number of friends: some values — multiplies of five — are overrepresented in the data. This so-called heaping effect [6, 7] has been found for several self-reported data; examples can be found in [8–12] and references therein. The effect is absent in results of the related data mining; hence its origin lies in the respondents' minds.

The visibility of the heaping effect depends on details. The distribution of the number of social links, as measured in social media, has been reported as a scale-free function [13, 14]. Accordingly, the order of magnitude of the mean degree in social networks is often so large that the binning procedure hides some details of the distribution [15–18].

Here we are interested in the number of friends, as evaluated by Polish respondents of the social survey [19]. The effect is best visible for answers in the range 0–40 friends. Therefore we show the results for the cohort of persons above the age of 50. There are indications that, for younger persons, the effect could be overshadowed by the larger scale of results. For people above 50, the mean number of friends is not larger than 100 [20]. A series of indices have been proposed to evaluate the effect quantitatively [21]; below we use the Whipple index as an example [22].

*corresponding author

In Sect. 2 we demonstrate that the answers on the number of friends show sharp maxima at 10, 15, 20, and sometimes 30, which accompany a broader peak between 0 and 8. Also, the related values of the Whipple index are shown there. In Sect. 3 the results are interpreted as a demonstration of the size effect, which applies to the reported values as well as to their uncertainty [23]. The relation of the data to the Benford law [24] is also explored there.

2. The data

In Figs. 1, 2 we show the number of friends of Polish women and men above the age of 50, as collected in 2015 [19]. As we see, these results do not change qualitatively with sex and age of the respondents. We have checked that the results collected in 2000 and 2007 are qualitatively the same. To observe the age dependence of the effect quantatively, the values of the Whipple index $I_W(a)$ are calculated in these three years, as dependent on the age a of the respondents. The formula for the 10-range is [22]:

$$I_W(a) = 100\frac{P_a(0)+P_a(10)+P_a(20)+P_a(30)}{(1/10)\sum_{k=0}^{39}P_a(k)} = 1000\frac{\sum_{k=0}^{3}P_a(10k)}{\sum_{k=0}^{39}P_a(k)}, \quad (1)$$

where the numbers of friends $P_a(k)$ are calculated for respondents of age a. These results are shown in Figs. 3 and 4. Accordingly, for the 5-range

$$I_W(a) = 500\frac{\sum_{k=0}^{7}P_a(5k)}{\sum_{k=0}^{39}P_a(k)}. \quad (2)$$

These results are shown in Figs. 5 and 6.

According to the discussion in [22], the values of the Whipple index for the 10-year range between 100 and 150 mean that the effect is low, between 150 and 250 — that the effect is moderate, and high above 250. The

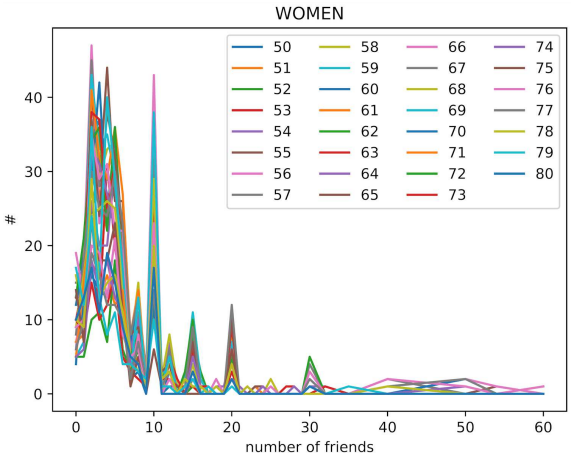

Fig. 1. The number of friends, as reported in 2015 by Polish women above the age of 50 [19]. Different colors mark the respondents' age.

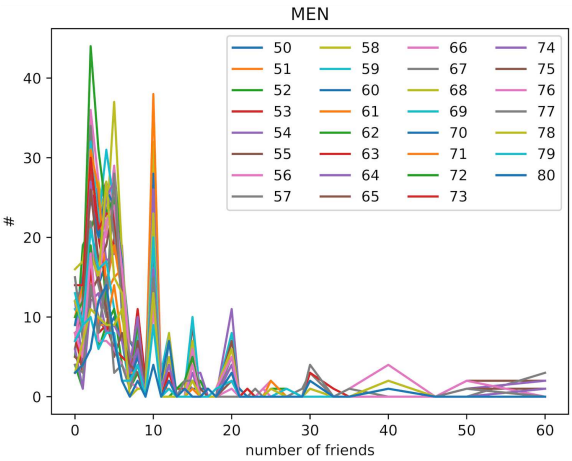

Fig. 2. The number of friends, as reported in 2015 by Polish men above the age of 50 [19]. Different colors mark the respondents' age.

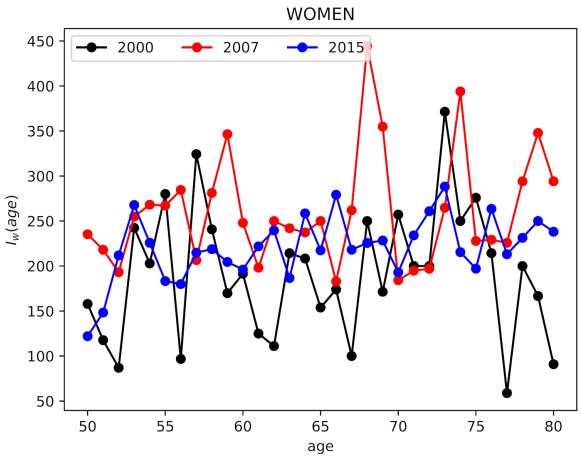

Fig. 3. The Whipple index $I_W(age)$ for the 10-range as dependent on the respondents' age, for Polish women above the age of 50 [19]. Different colors mark the data collected in 2000, 2007 and 2015.

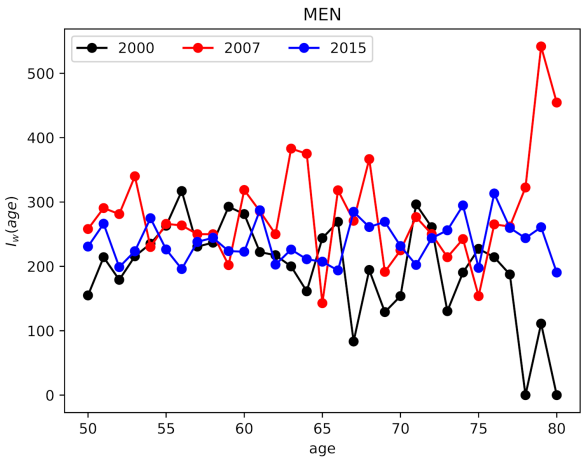

Fig. 4. The Whipple index $I_W(age)$ for the 10-range as dependent on the respondents' age, for Polish men above the age of 50 [19]. Different colors mark the data collected in 2000, 2007 and 2015.

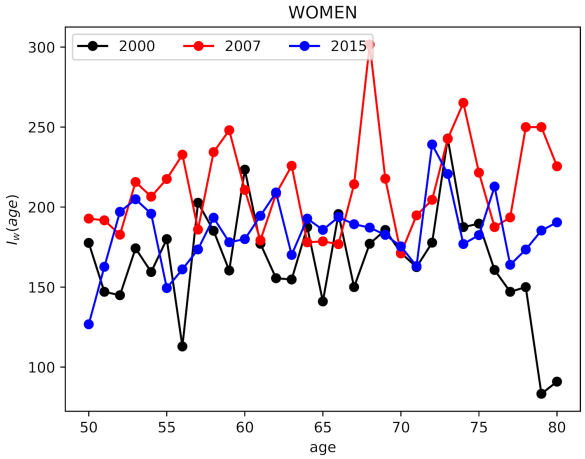

Fig. 5. The Whipple index $I_W(age)$ for the 5-range as dependent on the respondents' age, for Polish women above the age of 50 [19]. Different colors mark the data collected in 2000, 2007 and 2015.

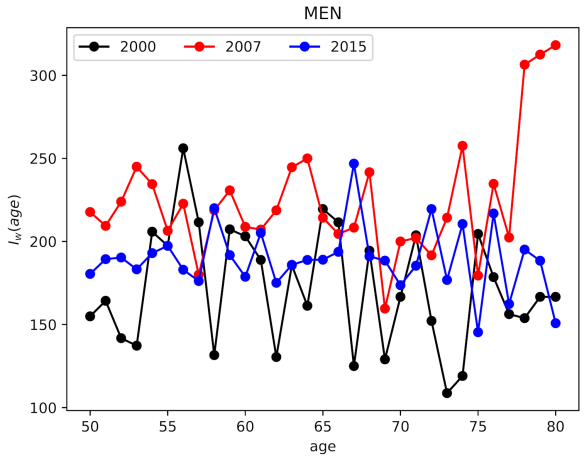

Fig. 6. The Whipple index $I_W(age)$ for the 5-range as dependent on the respondents' age, for Polish men above the age of 50 [19]. Different colors mark the data collected in 2000, 2007 and 2015.

mean value $I_W = 231 \pm 70$ (mean $\pm$ standard deviation), observed here, means "moderate". For the 5-year range, the data give the mean value $I_W = 192 \pm 36$. As it exceeds 175, an inference on the age distribution should be evaluated as "very rough" [25]. Yet, a well-defined procedure how to eliminate the bias is missing.

3. Discussion

In demographic data, age heaping has been ascribed to respondents' illiteracy [26]; in our case this factor can be safely excluded, yet this association refers to some primitive method of numbers evaluation. For a physicist, the Taylor law [27] seems to provide a natural context. The law states that the variance of a fluctuating variable increases with its mean as a power function. Consequently, the uncertainty of evaluation of any quantity increases with its value. For a psychologist, the Weber–Fechner law is a more appropriate starting point for the research of heaping. The law states that the change of a stimulus that will be just noticeable is a constant ratio of the original stimulus [28]. This law can be treated as a general hint when we ask how numbers are represented in our memory, while the Taylor law is a characteristics of numerous real data.

In Ref. [7], we read that multiples of ten receive most heaping, next are multiples of five and, finally, multiples of two. This rule, supported by data on respondents' age and seen also in Figs. 1 and 2, is told to be related to the number system "used by the estimator" [7]. The same idea has been mentioned when discussing an evaluation of elapsed time [29]; in the rounding procedure described there, the decimal system is accompanied by multiples of 30 days, which stand for months.

A more detailed mechanism of data heaping remains elusive. In Ref. [23], experiments have been reported on comparison of magnitudes, where the respondents were asked which number out of two is larger. There, an attempt is made to separate the "distance effect" and the "size effect". The former term means that "close numbers are more difficult to compare that numbers further apart". The size effect appears when "for a given distance, comparison difficulty increases with increasing size". The authors claim that they are able to separate both effects. As their analysis has been limited to one-digit numbers, it seems that the size effect could hardly be observed. In our case, the issue is not as the comparison of magnitudes, but just an evaluation of a considered quantity. Yet, to speak on the size effect seems appropriate here. The reason is that the distance between subsequent "prototypic" [29] values, i.e. the values promoted by heaping, increases with the values themselves. This is seen in Figs. 1 and 2, where the marked peaks appear at 10, 12, 15, 20, and 30, with differences between their positions clearly increasing with the positions themselves.

The overrepresentation of some numbers in data files, reported in the preceding sections, can be detected as a deviation from the Benford law [24]. The law states that the probability distribution of numbers' first digits d decreases like $\log_{10}(1+1/d)$. In Fig. 7 we show the data on the frequency of first digits in the data [19], collected in the years 2000, 2007, and 2015. Clearly, the digit 5 is overrepresented when compared with the Benford law. We have checked that the same deviation appears also for all data collected between 2000 and 2015 (2003, 2005, 2009, 2011, and 2013). Then, a comparison of data with the prediction of the Benford law can be useful to detect the heaping.

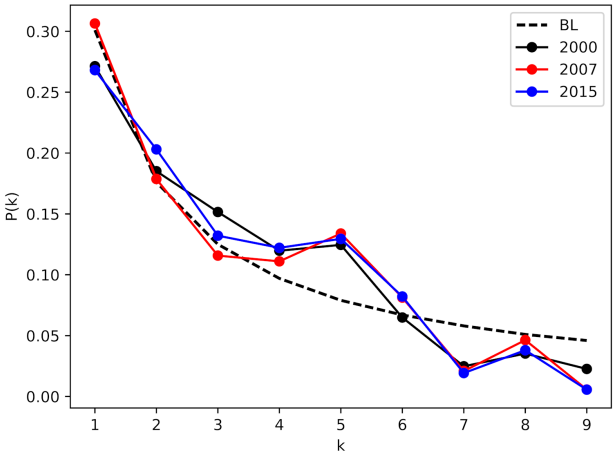

Fig. 7. The frequency of first digits in the data for Polish respondents above the age of 50 [19]. Continuous lines of different colors mark the data collected in 2000, 2007 and 2015. The black dotted line shows the predictions of the Benford law.

We note that in [24], the law has been applied to six social networks, as Facebook, Live Journal, Twitter etc. A good accordance of the distribution of first significant digits has been found there, except the network Pinterest Followers, where 5 appears about four times more often than the value suggested by the Benford law. However, this discrepancy is explained in [24] as a consequence of specific demands of Pinterest Followers: each new user must declare at least five initial relationships. In general, the bias reported in the present work applies to declared data, and not to real data: it applies not to what is, but what it seems to be.

Our analysis could be placed in new science of network (NSN) paradigm [30]. The heaping effect on the size of social networks is found to be large enough to seriously disturb the size evaluation. Fortunately, the effect is absent in the data collected directly in Internet. It can play a role, however, when we ask respondents about their beliefs and these beliefs influence their behavior [31, 32].

4. Conclusions

To conclude, we reiterate a few points:

• the goal of this paper is to identify a deviation of the Benford law in self-reported data on the degree distribution of social networks. In this way we provide a warning

against using parameters from declared data in modeling social networks. As such modeling is of common interest in sociophysics, our warning is justified there;
• the degree distribution of numerous social networks has been found to fit the power-law distribution [3, 16, 33]. As it is known that the Benford law is satisfied for data with a power-law distribution, it makes sense to look for deviations from the Benford law in the data related to the degree distribution in social networks.

We state that the origin of the above deviation is the heaping effect. We demonstrate that this effect is present in the declared data on social networks in Poland. We expect that this effect is present also in data collected for other nations (see [34–36] for examples); in this sense it can be considered as universal.

Acknowledgments

This work was partially financed supported by the Faculty of Physics and Applied Computer Science (11.11.220.01/2) and by the Faculty of Humanities (11.11.430.158) AGH UST statutory tasks within subsidy of Ministry of Science and Higher Education.

References

[1] R.I.M. Dunbar, *Evolution. Anthropol.* **6**, 178 (1998).
[2] R.I.M. Dunbar, *The Social Brain Hypothesis and Human Evolution*, Oxford Research Encyclopedia of Psychology, Mar. 2016.
[3] M.E.J. Newman, *SIAM Rev.* **45**, 167 (2003).
[4] R. Toivonen, L. Kovanen, M. Kivelä, J.-P. Onnela, J. Saramäki, K. Kaski, *Social Networks* **31**, 240 (2009).
[5] B. Ball, M.E.J. Newman, *Network Sci.* **1**, 16 (2013).
[6] R.J. Myers, *Trans. Actuarial Soc. Am.* **41**, 395 (1940).
[7] S.H. Turner, *Proc. Soc. Statist. Sect.*, Am. Stat. Assoc., Washington D.C., 248, 1958.
[8] J.S. Siegel, D.A. Swanson, *The Methods and Materials of Demography*, Elsevier AP, Amsterdam 2004.
[9] A. Crockett, R. Crockett, *Historical Methods* **39**, 24 (2006).
[10] Hao Wang, D.F. Heitjan, *Statist. Med.* **27**, 3789 (2008).
[11] S. Zinn, A. Würbach, *J. Appl. Statist.* **43**, 682 (2016).
[12] L. Bermúdez, D. Karlis, M. Santolino, *Computat. Statist. Data Anal.* **112**, 14 (2017).
[13] T. Hogg, G. Szabo, *EPL* **86**, 38003 (2009).
[14] J. Kunegis, *KONECT: The Koblenz Network Collection*, in: *Proc. 22nd Int. Conf. on World Wide Web Companion*, Int. World Wide Web Conf. Steering Committee, Switzerland, 2013, p. 1343.
[15] S. Strogatz, *Nature* **410**, 268 (2001).
[16] L. Muchnik, S. Pei, L.C. Parra, S.D.S. Reis, J.S. Andrade Jr., S. Havlin, H.A. Makse, *Sci. Rep.* **3**, 1783 (2013).
[17] T.H. McCormick, M.J. Salganik, T. Zheng, *J. Am. Stat. Assoc.* **105**, 59 (2010).
[18] What Japan Thinks (access 21.03.2017).
[19] Diagnoza społeczna: zintegrowana baza danych (in Polish) (access 2.03.2017).
[20] S. Wolfram, Stephen Wolfram Blog (access 21.03.2017).
[21] E.G. Stockwell, J.W. Wicks, *Soc. Biol.* **21**, 163 (1974).
[22] K. Borkotoky, S. Unisa, *PLOS One* **9**, e90113 (2014).
[23] T. Verguts, F. van Opstal, *Psychonom. Bull. Rev.* **12**, 925 (2005).
[24] J. Golbeck, *PLoS ONE* **10**, e0135169 (2015).
[25] G.S. Pardeshi, *Indian J. Commun. Med.* **35**, 391 (2010).
[26] *Population and Health in Developing Countries*, International Development Research Centre, Vol. 1, Ottawa 2002.
[27] L.R. Taylor, *Nature* **189**, 732 (1961).
[28] Encyclopaedia Britannica, Weber's law (access 1.04.2017).
[29] J. Huttenlocher, L.V. Hedges, N.M. Bradburn, *J. Exp. Psychol. Learning Memory Cognit.* **16**, 196 (1990).
[30] D. Watts, *Everything is Obvious: Once You Know the Answers*, Crown Business, New York 2011.
[31] D.A. Prentice, D.T. Miller, *J. Personal. Soc. Psych.* **64**, 243 (1993).
[32] R.K. Merton, *Soc. Forces* **74**, 379 (1995).
[33] M.E.J. Newman, *Networks. An Introduction*, Oxford UP, New York 2010.
[34] V. Vehovar, K.L. Manfreda, G. Koren, V. Hlebec, *Soc. Networks* **30**, 213 (2008).
[35] R.C. Cherry, D.L. Poston, Jr., in: *2011 Annual Meeting of the Population Association of America*, Washington, DC, 2011 (access 2018.01.07).
[36] F.W. Crawford, R.E. Weiss, M.A. Suchard, *Ann. Appl. Statist.* **9**, 572 (2015).