Internet Interventions 1 (2014) 165–168

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.invent-journal.com/

Describing the distribution of engagement in an Internet support group by post frequency: A comparison of the 90-9-1 Principle and Zipf's Law

Bradley Carron-Arthur [a,*], John A. Cunningham [a,b], Kathleen M. Griffiths [a]

[a] National Institute for Mental Health Research, Australian National University, 63 Eggleston Road, Acton, Canberra, ACT 0200, Australia
[b] Centre for Addiction and Mental Health, 33 Russell Street, T526 Toronto, ON, Canada

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form 11 September 2014
Accepted 11 September 2014
Available online 28 September 2014

Keywords:
eHealth
Internet support group
Social network
Zipf's Law
90-9-1 principle
1% rule

ABSTRACT

Sustainable online peer-to-peer support groups require engaged members. A metric commonly used to identify these members is the number of posts they have made. The 90-9-1 principle has been proposed as a 'rule of thumb' for classifying members using this metric with a recent study demonstrating the applicability of the principal to digital health social networks.
Using data from a depression Internet support group, the current study sought to replicate this finding and to investigate in more detail the model of best fit for classifying participant contributions.
Our findings replicate previous results and also find the fit of a power curve (Zipf distribution) to account for 98.6% of the variance.
The Zipf distribution provides a more nuanced image of the data and may have practical application in assessing the 'coherence' of the sample.
© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

Online peer-to-peer support has many potential health benefits (Ziebland and Wyke, 2012). To date, systematic reviews have failed to find consistent evidence for the efficacy of online peer-to-peer support groups on health outcomes (Eysenbach et al., 2004; Griffiths et al., 2009). However, there is evidence that consumers value these groups (Horrigan et al., 2001) and there is increasing interest in identifying the key components of sustainable thriving online support groups (Young, 2013). It is generally agreed that one key component is highly engaged core members who contribute substantially to the community (Young, 2013). There is no consensus on what metrics should be employed to classify the contributions of members. Four studies have sought to identify highly engaged members in online peer-to-peer support groups using different combinations of metrics. These metrics include the number of posts made by members (Cobb et al., 2010; Jones et al., 2011; van Mierlo et al., 2012; van Mierlo, 2014), the number of threads initiated (Jones et al., 2011; van Mierlo et al., 2012), the number of different threads in which a member participates (Jones et al., 2011; van Mierlo et al., 2012), the level of connectedness to other members in the forum (Cobb et al., 2010) and time spent logged in (Jones et al., 2011). One metric common to them all was number of posts.

Recent research has used number of posts as a sole means of classifying members in Digital Health Social Networks (DHSN) with a peer-to-peer support group component (van Mierlo, 2014). The study investigated the 90-9-1 principle or the 1% rule. This rule describes a commonly reported phenomenon whereby the majority of content in an Internet community is produced by only 1% of the participants (referred to as 'superusers'), a minority of the content is produced by a further 9% of participants ('contributors') and 90% of people observe the content in the Internet community without actively participating ('lurkers') (Nielsen, 2014). The study sectioned the content attributed to these three groups and found that the sections contained 74.7%, 24.0% and 1.3% of the total posts in the DHSN respectively. It was concluded that the 90-9-1 principle applied to DHSN.

The DHSN study sought to verify the 90-9-1 principle rather than to determine the distribution which best fitted the data. Thus, the 90-9-1 principle may not provide the greatest accuracy in classifying participants in a DHSN. The aim of the current study is to further investigate the model of best fit for classifying participants in a DHSN, including but not limited to the 90-9-1 principle.

2. Method

This study used data from the peer-to-peer Internet support group — BlueBoard (blueboard.anu.edu.au). BlueBoard is predominantly used for peer-to-peer discussion about Depression (38.8% of content). It also includes forums on Bipolar Disorder (18.4%), Generalised Anxiety Disorder (5.0%), general discussion (22.1%) and other topics (15.7%).

* Corresponding author. Tel.: +61 02 6125 6825.
E-mail addresses: Bradley.Carron-Arthur@anu.edu.au (B. Carron-Arthur), John.Cunningham@anu.edu.au (J.A. Cunningham), Kathy.Griffiths@anu.edu.au (K.M. Griffiths).

http://dx.doi.org/10.1016/j.invent.2014.09.003
2214-7829/© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Fig. 1. BlueBoard homepage.

BlueBoard is moderated by a team of paid personnel. Members are consumers and carers. BlueBoard's homepage is shown in Fig. 1. The data used in this study included all posts generated between 1st October 2008 and the 23rd May 2014 (n = 131,004 by 2932 members). Posts made by moderators (n = 352 by 10 moderators) were not included in the analysis. Data collection procedures were approved by the Australian National University Human Research Ethics Committee.

In order to replicate the analysis conducted by van Mierlo (2014), we separately calculated the total number of posts made by the 1% of registered members who contributed the most, the next 9% and the final 90%. To investigate alternative models of fit for the data we graphed on a log–log scatterplot the total number of posts of each member ranked in order of those who made most to least posts and fitted a power curve using Microsoft Excel.

3. Results

The percentages of posts made by participants in each of the three Sections 1, 9, and 90 were 85.8%, 11.2% and 3.0% of the total number of posts respectively. The corresponding number of members in each section and the range in the number of posts made by members in that section are shown in Table 1.

A log–log scatterplot showing the frequency of posts made by each member ranked in descending order is presented in Fig. 2. The best fitting curve was found to have the function $f(x) = 63935x^{-1.427}$ with correlation coefficient r = 0.993 and a coefficient of determination of 0.986. This indicates that the model accounts for 98.6% of the variance.

4. Discussion

The current analysis broadly replicated the findings of van Mierlo (2014), that the top 1% of registered members contribute the vast majority of posts, the next 9% a minority and the last 90% very few. Thus, the 90-9-1 principle appears to provide a reliable means of broadly categorising participant contributions in a DHSN. However, the graph in Fig. 2 and the associated best fitting power curve provide an alternate and more precise means of describing the distribution. In fact, the distribution in Fig. 2 adheres to Zipf's law — that the frequency of posts made by a member is inversely proportional to their rank in frequency. This is a widely observed phenomenon spanning areas such as linguistics, populations, income and internet traffic (Newman, 2006; Adamic and Huberman, 2002). This model gives a more nuanced image of the distribution. It shows a gradual reduction in contributions rather than a quantum leap at the boundary between superusers and contributors as the 90-9-1 principle implies. Researchers, developers and other stakeholders seeking to optimise the network effects associated with members who generate the highest levels of traffic in an Internet support group (van Mierlo, 2014) may benefit from the understanding that there is a predictable diminishing return associated with each individual member as opposed to categorical differences in types of users.

A range of explanations has been proposed to explain the occurrence of Zipfian distributions including, for example, the principle of least effort (Ferrer i Cancho and Sole, 2003), proportional growth processes (Gabaix, 1999) or a simple stochastic process (Miller et al., 1958). There is no consensus on which is correct and none allow a meaningful interpretation of the current data. However, a phenomenon associated with data which better fits the Zipfian distribution is that of greater 'coherence' in the sample (Cristelli et al., 2012). For example, ranking cities by population size in the USA fits the Zipfian distribution better than the European Union (EU). Furthermore, each individual country of the EU fits the distribution well in comparison to the EU as a whole, and conversely each individual state in the USA does not fit the distribution well in comparison to the USA as a whole. This is thought to reflect the time each has had to organically evolve as a collective unit (Cristelli et al., 2012). For Internet support groups, describing the distribution of engagement using the Zipfian distribution may allow researchers and developers to assess the coherence of the group versus the coherence of its subsets, such as the different forums within the group. In the current study, the best fit was found for the support group as a whole as opposed to any individual forum by topic.

Frequency of posts is one way of identifying highly engaged members in a network. It is not necessarily the most suitable method. Borgatti (2006) argues that key members in a network are most appropriately identified using the combination of metrics that identifies members whose engagement contributes the kind of value that reflects the reason they are being sought. In addition to the metrics which have been used in past research, future research may investigate other metrics such as the average word count of posts, time of day, regularity of posting or combinations of these. Since quantity does not necessarily reflect quality, content analysis of posts is required to determine if the highly engaged users are contributing informative and supportive content (Salem et al., 1997).

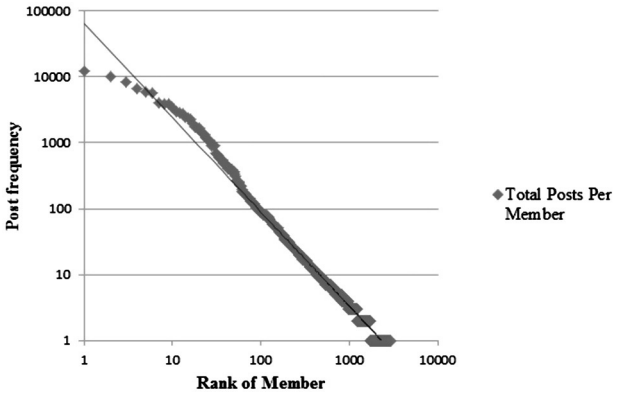

Fig. 2. Log–log scatterplot of the total posts made by each member ranked in descending order and a power curve which best fits the data.

Table 1
Posts and members in each section.

Percentile	Members (N)	Percentage of posts (N)	Range in the number of posts (N)
1 (1%)	1–74 (74)	85.8% (112,373)	11,994–142 (11,852)
2–10 (9%)	75–743 (669)	11.2% (14,669)	141–5 (136)
11–100 (90%)	744–7434 (6691)	3.0% (3,962)	5–0 (5)

5. Conclusion

The 90-9-1 principle and Zipf's Law both provide a means of describing the distribution in engagement of members by post frequency in the internet support group but Zipf's law provides a more precise description of the data.

Acknowledgements

The authors would like to thank Anthony Bennett, Julia Reynolds and Kylie Bennett for their contributions to establishing and maintaining BlueBoard. We thank Anthony Bennett and Kylie Bennett for assistance in downloading the data. We also thank Professor David Hawking for his expert input regarding Zipf's law. BlueBoard is funded by the Australian Government Department of Health. B. Carron-Arthur is supported by an Australian Postgraduate Award. K.M. Griffiths is supported by the Australian National Health and Medical Research Council (NHMRC) Research Fellowship 1059620.

References

Adamic, L.A., Huberman, B.A., 2002. Zipf's Law and the Internet. Glottometrics 3, 143–150.

Borgatti, S.P., 2006. Identifying sets of key players in a social network. Comput. Math. Organ. Theory 12, 21–34.

Cobb, K.N., Graham, A.L., Abrams, D.B., 2010. Social network structure of a large online community for smoking cessation. Am. J. Public Health 100, 1282–1289.

Cristelli, M., Batty, M., Pietronero, L., 2012. There is more than a power law in Zipf. Sci. Rep. 2.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., Stern, A., 2004. Health related virtual communities and electronic support groups systematic review of the effects of online peer to peer interactions. BMJ 328.

Ferrer I Cancho, R., Sole, R.V., 2003. Least effort and the origins of scaling in human language. Proc. Natl. Acad. Sci. U. S. A. 100, 788–791.

Gabaix, X., 1999. Zipf's law for cities an explanation. Q. J. Econ. 114, 739–767.

Griffiths, K.M., Calear, A.L., Banfield, M., 2009. Systematic review on Internet Support Groups (ISGs) and depression (1): do ISGs reduce depressive symptoms? J. Med. Internet Res. 11, e40.

Horrigan, J.B., Rainie, L., Fox, S., 2001. Online Communities: Networks That Nurture Long-Distance Relationships and Local Ties. Pew Internet & American Life Project, Washington, DC.

Jones, R., Sharkey, S., Smithson, J., Ford, T., Emmens, T., Hewis, E., Sheaves, B., Owens, C., 2011. Using metrics to describe the participative stances of members within discussion forums. J. Med. Internet Res. 13, e3.

Miller, G.A., Newman, E.B., Friedman, E.A., 1958. Length-frequency statistics for written English. Inf. Control. 1, 370–389.

Newman, M.E.J., 2006. Power laws, Pareto distributions and Zipf's law. Contemp. Phys. 46, 323–351.

Nielsen, J., 2014. Participation Inequality: Lurkers vs Contributors in Internet Communities, (URL: http://www.nngroup.com/articles/participation-inequality/. Accessed: 2014-06-05. (Archived by WebCite® at http://www.webcitation.org/6Q7EwEncA).

Salem, D., Bogat, A., Reid, C., 1997. Mutual help goes on-line. J. Community Psychol. 25, 189–207.

Van Mierlo, T., 2014. The 1% rule in four digital health social networks: an observational study. J. Med. Internet Res. 16, e33.

Van Mierlo, T., Voci, S., Lee, S., Fournier, R., Selby, P., 2012. Superusers in social networks for smoking cessation: analysis of demographic characteristics and posting behavior from the Canadian Cancer Society's smokers' helpline online and StopSmokingCenter.net. J. Med. Internet Res. 14, e66.

Young, C., 2013. Community management that works: how to build and sustain a thriving online health community. J. Med. Internet Res. 15, e119.

Ziebland, S., Wyke, S., 2012. Health and illness in a connected world how might sharing experiences on the Internet affect people's health. Milbank Q. 90, 219–249.