Medical Informatics and the Internet in Medicine
September 2007; 32(3): 215 – 223

informa
healthcare

Entry requirements and membership homogeneity in online patient groups

ROY RADA

Department of Information Systems, University of Maryland, Baltimore County, USA

(Received for publication 26 July 2006; accepted 6 December 2006)

Abstract
The objective was to explore a relationship between the economics of religion and the attributes of online patient groups by testing the hypotheses that (1) the harsher the entry requirements to an online patient group, the more active its members are; and (2) membership homogeneity in a given group is reflected in the educational level of group members. Online groups were randomly chosen from the 'Yahoo groups' category of 'Illnesses'. The hypothesis about entry requirements was narrowed by defining those requirements as either 'Open', 'Register', or 'Closed'. The number of messages over a 4-month period in each of 162 different groups was tallied. The hypothesis about membership homogeneity was refined by counting the citations in messages and by predicting the educational level of members (as reflected in the average word length of messages) based on these citation counts. Across 162 groups, the number of messages was significantly less in Open groups than in Register groups and less in Register groups than in Closed groups. Across 14 groups, the average word length of messages in a group positively correlated with the number of citations in that group. The hypothesis is supported that increased group entry barriers correspond to increased group message activity and members tend to be similar within a group. These attributes could be used to help design effective groups.

Keywords: *Group processes, online systems, consumer participation, cultural anthropology, information storage and retrieval*

1. Introduction

Research in the economics of religion shows that a strict membership requirement often goes hand in hand with strong growth of a religious group [1]. Religions which permit anyone to join and which make few demands on members may garner many members but paradoxically display little activity. The economics of religion draws on the theory of clubs. One challenge of a club is to minimize the number of non-participating or 'free-loading' members and to insure a certain homogeneity in the membership of the club [2]. One published research proposal [3] applied club theory to online groups but does not seem to have led to further publications. The work reported here applies results of the economics of religion and of club theory to online patient groups.

Correspondence: Roy Rada, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250, USA. E-mail: rada@umbc.edu

ISSN 1463-9238 print/ISSN 1464-5238 online © 2007 Informa UK Ltd.
DOI: 10.1080/14639230701447685

Many patients do not receive support-group interventions [4] because of practical problems, such as a lack of transportation or inconvenient meeting times. While patients prefer one-on-one dialog with the doctor or nurse over online information, online information is easier to obtain [5]. Online patient groups are an increasingly important source of information and support for patients [6]. Studies of online groups have shown that:

- Some patients in online discussions want information in the first phase when they learn about their disease but later want empathy [7].
- Some patients want an online system that requires a registration of participants, permits anonymity, and is backed by a recognized institution [8].
- Geographical proximity of members [9] affects group vitality but not group size [10].
- A group moderator can stimulate member participation [11] or interfere with it [12].
- A meta-analysis of Interactive Health Communication Applications concluded that patients who use computers benefit both socially and medically from this intervention [13].

Authors often conclude that the rigor of methods in studying online patient groups needs to increase [14].

2. Methodology

Patient online groups were sampled and studied in two, related experiments. One experiment focused on entry requirements, and the other, on membership homogeneity.

2.1. Entry requirement

Groups may be categorized by their entry requirements. A group might have a moderator determine whether someone can read the messages in the group (call this a 'Closed' group). In the 'Closed' type of group, a potential member must complete and submit an application before learning whether or not entry will be granted. Other types of group fall into at least two categories: requiring registration (the 'Register' group) and requiring nothing (the 'Open' group). To enter a 'Register' group, a person must have first registered with the group system (which anyone can do) and then enters the group with the system 'user name' and password. In the 'Open' group, an unidentified person can read group messages.

The vitality of a group can be measured via a content analysis of messages [15], visualizing patterns of interactions [16], or, as in this study, simply counting the number of messages that the group exchanges over a period of time [17]. The hypothesis is that 'Closed' groups will tend to exchange more messages than 'Register' or 'Open' groups. In this study, the number of members per group is also recorded, and the correlation between the number of messages and the number of group members is computed.

To find groups to study, Yahoo Groups (http://groups.yahoo.com/) was visited, and the Yahoo groups category of 'Illnesses' was chosen from the options at http://health.dir.groups.yahoo.com/dir/Health_Wellness/Support/Illnesses (call this the root node). The Yahoo 'Illnesses' category was the parent of a tree (in the graph theory sense) where inner nodes represent categories, and leaves represent patient discussion groups. This tree contained 13,741 leaves or patient discussion groups. By way of illustration, after the 'Illnesses' category was selected, the user was shown a page with 136 categories, such as 'Cancers' and 'Anemia', and after the 'Anemia' category was chosen, the user was presented 23 leaves or groups. A traversal of the 'Illnesses' tree identified the groups for the study.

The traversal algorithm was a modified depth-first search [18]. The search began at the root node and randomly chose a child of that node. Microsoft Excel's 'RandBetween' function was used to determine a random number in the range between 1 and the number of descendants of the root. The depth-first search proceeded from parent to child until a leaf was reached. The sibling leaves were randomly sampled until a maximum of 20 unique groups from that disease category had been collected, at which point the search algorithm backtracked and proceeded down another branch of the tree. In total, 162 groups were collected by manually implementing this algorithm.

For each of the 162 groups, values for various attributes were obtained from the Yahoo site. While the Closed groups required an application from the user before the user could read the group's messages, for all groups Yahoo provided a summary page that listed the number of group messages by month for the lifetime of the group. The attribute values collected were:

- group required registration (Yes or No);
- group required application (Yes or No);
- number of messages in September 2005;
- number of messages in October 2005;
- number of messages in November 2005; and
- number of messages in December 2005.

The collected information was recorded in an Excel spreadsheet.

2.2. Membership homogeneity

In addition to exploring membership entry requirements, this work explored homogeneity in the membership of a group. While most online group members are lurkers [19], messages in a group reflect a group culture. Some patients who go online for help are well educated [20] and may want to participate in a group with other well-educated people [21]. However, in Yahoo Groups, the educational level of participants is not directly given. Instead, the researcher can measure various attributes of messages that might relate to educational level. The two attributes that will be studied here are (1) the average number of citations in messages of a group and (2) the average word length of messages in a group. The amount of citing will be taken as the independent variable, while the dependent variable will be the average word length of messages. The hypothesis is that groups whose messages contain more citations are likely to have longer words in the messages. The average word length is clearly a surface feature, although it has been found to correlate well with deeper characteristics like discourse structure [22] and, in turn, educational level.

The sampled groups from the 'Entry Requirement' experiment were considered for this 'Membership Homogeneity' experiment. If an 'Open' or 'Register' group had more than 20 messages for each of September, October, November, and December 2005, then the group was eligible for inclusion in the 'Membership Homogeneity' study. Groups that required a membership application were not used. From a set of sibling groups (the leaves of one disease category), two groups were chosen. In this way, 14 groups from seven disease categories were selected. For instance, the 'Anemia' node has 23 leaves, and data from two 'Anemia' groups were used.

For each of the 14 groups, 40 messages in chronological order were selected. Information about each message was stored in a row of a Microsoft Excel spreadsheet. One column was used for the contents of a message. Other columns in the row were used to record derived attributes of the message.

One set of message attributes covered the extent to which a message cited other information. Pilot studies produced a categorization of citations. The categories of citation, with an example of each follow:

- Research Report: a journal article with the scientific details of the results of a clinical trial;
- Mass Media News: a newspaper article with a popularization of a new medical research result;
- Government Policy: a publication describing a policy, such as for government-sponsored coverage of a particular treatment;
- Medical Device: brochure about attributes of a wheel chair;
- Drug: a drug-company website where the details about a particular drug are documented; and
- Patient Support Source: a lecture to patients by a social worker.

Pointers to two categories of content were considered irrelevant and were excluded from further consideration; these two types were:

- Advertisement: a commercial for a product; and
- Identifier: group member information not related to the discussion topic, such as a signature line that points to the member's Web home page.

The spreadsheet had a column for each of the relevant citation types, and the number of citations of each kind was recorded in the appropriate cell of the column for each message. Although a citation might fit into multiple categories, it was placed into only the one category that seemed most appropriate. One column of the Excel spreadsheet held for each message the sum across relevant message types of the number of citations.

In addition to content type, a citation might vary by medium type. For example, a citation might include a Web address (such as www.sciencemag.org), a journal title (such as *Journal of Radiobiology*), or a pointer to a lecture event. Separate tallies were not kept of citations by medium type.

One column in the spreadsheet stored for each message the average word length of each message. To determine average word length of a message, the Excel functions:

- TRIM to remove extra spaces and
- LEN to count the number of characters in a message

were used. For a message in cell D2, the formula LEN(TRIM(D2))/IF(LEN (TRIM(D2)) = 0,0,LEN(TRIM(D2))-LEN(SUBSTITUTE(D2,″ ″,″″)) + 1)] computed average word length. This formula essentially determined the number of strings without blanks and divided that number into the number of non-blank characters in the message.

3. Results

The results are presented in the two categories of 'entry requirement' and 'membership homogeneity'.

3.1. Entry requirement

Each of the 'Open', 'Register', and 'Closed' types had more than 35 groups within it (Table I). For each group, the total number of messages (call this value 'TOT') over the

4 months was computed, and the average TOT across the groups of a type was computed (Table I). The distribution of TOT across the groups of a type was highly skewed.

To visualize the distribution of TOT for a group type, histograms are presented. To draw the histogram, the following 12 bin sizes were defined: [0 – 1), [1 – 8), [8 – 16), [16 – 32), [32 – 64), [64 – 128), [128 – 256), [256 – 512), [512 – 1024), [1024 – 2048), [2048 – 4096), [4096 – 8192) where '[a-b)' indicates that the bin includes any number greater than or equal to 'a' but less than 'b'. Twenty-one Open groups belonged to the '[0 – 1)' bin (Figure 1). In other words, 21 groups had no messages in the 4 months from September 2005 through December 2005. The group might have never shown activity after it was created or might have been once very active but died. By contrast, the three bins from 1024 to 8192 contain together only one group.

The group categories of 'Register' and 'Closed' show less skewness in TOT than the 'Open' group but remain highly skewed. The histogram of the 'Closed' group (Figure 2) shows that 11 of the 38 groups had zero members, but the distribution suggests a higher median than for the 'Open' groups.

Given the distribution of messages across groups, a non-parametric statistical test was chosen to evaluate the significance of differences in the distribution of messages by group types. The hypothesis is that at the 0.05 confidence level the median TOT of:

- Open groups is lower than the median TOT of Register groups; and
- Register groups is lower than the median TOT of Closed groups.

The non-parametric test chosen was the Wilcoxon Rank-Sum Test for two independent samples with correction for ties [23]. Both hypotheses are supported by the statistical test.

Table I. Number of groups and average number of messages for the three group types.

Group type	Number of groups	Average number of messages
Open	49	178
Register	75	169
Close	38	364

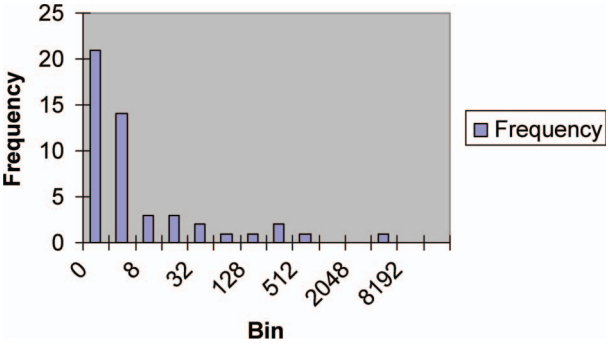

Figure 1. Histogram showing the distribution of messages across groups in the category 'Open', namely groups that required no registration. The bin sizes indicate the total number of messages in a group over 4 months and are bounded by 0, 1, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, and 8192, as indicated on the *x*-axis. The number of groups in a bin is indicated on the *y*-axis.

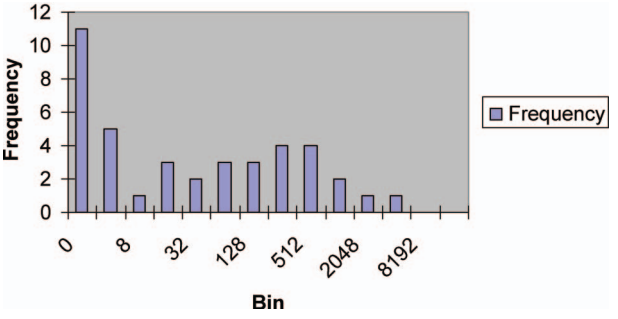

Figure 2. Histogram showing the distribution of messages across groups in the category 'Closed', namely groups that required members to apply to join. The bin sizes on the *x*-axis indicate the total number of messages in a group over 4 months. The number of groups in a bin is indicated on the *y*-axis.

Outliers are evident. The group with the most messages among the Open, Register, and Closed groups was an Open group called the 'colon cancer support' group (http://health.groups.yahoo.com/group/colon_cancer_support/) which had 6,830 messages over the 4 months studied. Some of the Closed groups never had a single message; requiring an application before entry does not guarantee that any one will want to submit an application. Nevertheless, the statistical results are consistent with the predictions from the economics of religion: the groups with harsher membership requirements tend to be more active.

To test the relationship between activity and membership size, a correlation coefficient between the total number of messages over 4 months for a group and the number of members in that group was computed. The correlation coefficient was only 0.370. This low correlation is partly because of the way messages and members were tallied: namely, messages were tallied for the last 4 months of the year 2005, but members who joined at any times in the group's life were tallied. Yahoo Groups provides message counts by month but only provides a single figure for the number of group members.

3.2. Membership homogeneity

The data from the 14 groups in the 'membership homogeneity' study show a wide variation in the number of citations made in groups (Table II). The question of whether a higher-citing group used longer words is addressed with two tests:

- First, the citations and average word length are ranked, and a Spearman coefficient of rank correlation is computed. At the 0.05 confidence level, one can affirm a positive relationship between (a) the total number of citations in the 40 messages of a group and (b) the average word length of those messages.
- Second, for each of the seven pairs of groups (one pair per disease category), the group with the greater number of citations also had a larger average word length than the group with the lower number of citations. This result can be compared with tossing a coin seven times and getting seven 'heads'; in other words, by chance this would rarely happen.

This result needs to be tempered with the observation that a citation with a Web address might be counted as a very long word by the Excel functions used in this experiment. The correspondence between the Web address being both a long word and part of a citation might

Table II. Number of relevant citations over 4 months and the average word length of messages in those 4 months for each of the 14 groups[a].

Group label	Cites	Average word length
A1	4	6.77
A2	0	5.76
B1	131	6.41
B2	0	5.29
C1	7	6.01
C2	3	5.82
D1	5	5.52
D2	16	5.65
E1	1	5.34
E2	0	5.28
F1	18	7.49
F2	13	6.35
G1	2	5.77
G2	8	6.1

[a]Two groups with the same 'disease category' parent have the same first letter in their group label.

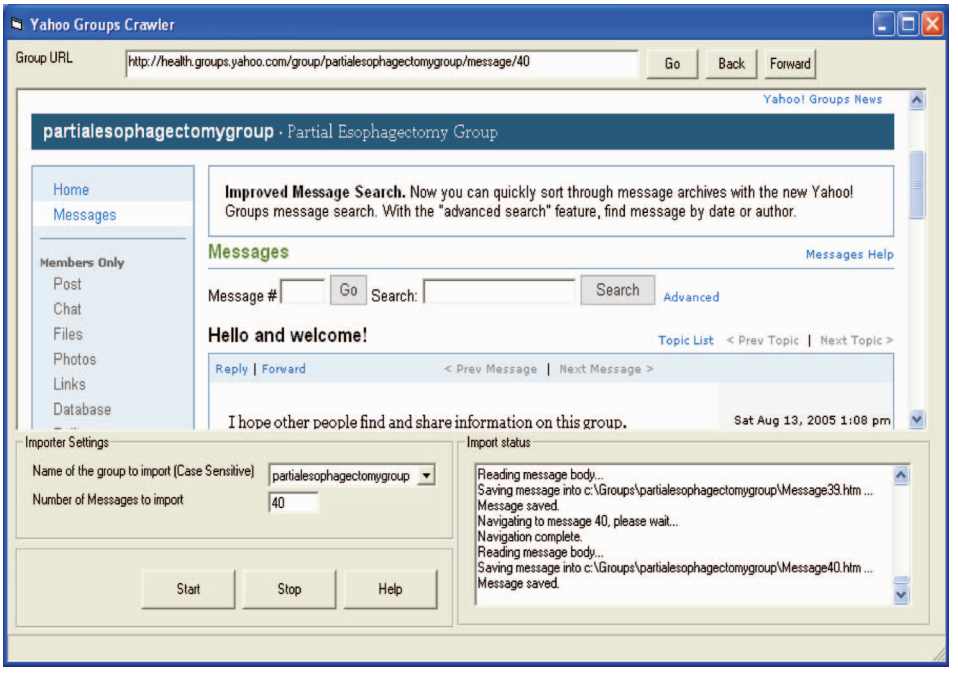

Figure 3. Screenshot showing software that the researchers developed to semi-automate data collection and analysis. It was developed in Visual Basic and allows the user to specify a number of messages to import (in the lower-left corner) from a Yahoo health group chosen from a pull-down menu (also in the lower-left corner). The actual Yahoo screen is enclosed in the upper half of the window. In this screenshot, the view is onto the Yahoo group for 'partial esophagectomy'.

bias the results. However, the overall impression is that groups that cite external sources more often have members who are better educated and that groups are self-selecting for homogeneous member attributes.

4. Conclusion

The economics of religion and club theory suggest unorthodox examinations of entry requirements and homogeneity in patient online groups. Yahoo health groups were studied for the impact of entry requirements on the activity level in the group. Homogeneity in groups was addressed by studying the extent to which a group members made citations and wrote messages with long words.

Yahoo health groups have three different kinds of entry requirements: (1) open to anyone, (2) only requiring to register with a Yahoo user name and password, or (3) requiring both a Yahoo registration and submission of an application to a moderator of the group who decides whether or not the applicant deserves admission to the group. These three groups were called, respectively, Open, Register, and Closed. Consistent with results from the economics of religion, the 'Closed' groups were the most active, and the 'Open' groups were the least active, as measured by the average number of messages exchanged over 4 months.

Fourteen groups were studied for the number of citations in the messages and the average word length of messages in the group. Statistical analysis shows a positive correlation between these two attributes. Group systems, such as Yahoo Groups, might calculate the average word length of messages in a group and make this value known to people who might join the group. A patient who writes messages with a certain average word length might want to join a group whose members write similarly.

This work might be extended in its method and its theory. From the methodological perspective, the 'Fleisch Kincaid Readability Rating' might be a better measure of complexity of messages than is average word length. Natural language processing programs could further dissect patterns in messages. Another way to extend the methodology and facilitate data collection would be to implement software to traverse groups and collect messages; the author's team has developed a software tool for this purpose that is illustrated in Figure 3.

The theories of clubs and religions could be further applied to the study of online patient groups. For instance, the birth or death of a religion shows fascinating adaptations to socio-economic circumstances [24]. Longitudinal studies of the birth and death of patient online groups might be very revealing.

References

1. Iannaccone LR. Introduction to the economics of religion. Journal of Economic Literature 1998;36:1465–1496.
2. Sandler T, Tschirhart J. Club theory: Thirty years later. Public Choice 1997;93(3–4):335–355.
3. Fickas S, Arrow H, Orbell J. Join the Club: Enabling Self-Organizing Groups on the Net. In: Finin T, ed. American Association for Artificial Intelligence 2000 Workshop on Knowledge-Based Electronic Markets. Austin, TX: AAAI; 2000. pp 1–10.
4. Taylor SE, Falke RL, Shoptaw SJ, Lichtman RR. Social support, support groups, and the cancer patient. Journal of Consulting and Clinical Psychology 1986;54:608–615.
5. Lock K, Willson B. Information needs of cancer patients receiving chemotherapy in an ambulatory-care setting. Canadian Journal of Nursing Research 2002;34(4):83–93.
6. Rimer KB, Lyons JE, Ribisl MK, *et al.* How New subscribers use cancer-related online mailing lists. Journal of Medical Internet Research 2005;7(3):e32.
7. Arnold Y, Leimeister JM, Krcmar H. CoPEP: A development process model for a community platform for cancer patients. In: XIth European Conference on Information Systems (ECIS), Naples; 2003.
8. Ebner W, Leimeister JM, Krcmar H. Trust in virtual healthcare communities: Design and implementation of trust-enabling functionalities. In: Sprague R, ed. 37th Annual Hawaii International Conference on System Sciences. Hawaii: IEEE; 2004. pp 182–192.
9. Lutters W, Ackerman M. Joining the backstage: locality and centrality in an online community. Information Technology & People 2003;16(2):157–182.

10. Schoberth T, Preece J, Heinzl A. Online communities: a longitudinal analysis of communication activities. 36th Annual Hawaii International Conference on Systems Sciences. Hawaii: IEEE; 2003:10.
11. Preece J, Nonnecke B, Andrews D. The top five reasons for lurking: Improving community experiences for everyone. Computers in Human Behavior 2004;20(2):201–223.
12. Chim H, Liu BJ, Deng X. A group decision approach for information assessment. IASTED International Conference on Internet and Multimedia Systems and Applications, EuroIMSA. Grindelwald, Switzerland: IASTED; 2005. pp 7–12.
13. Murray E, Burns J, See T, Lai R, Nazareth I. Interactive Health Communication Applications for people with chronic disease. The Cochrane Database of Systematic Reviews 2005(4):CD004274.
14. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. British Medical Journal 2004;328(7449):1166–1172.
15. De Wever B, Schellens T, Valcke M, Van Keer H. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. Computers & Education 2006;46(1):6–28.
16. Viegas F, Smith M. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In: Hawaii International Conference on System Sciences, Hawaii; 2004. p 40109b.
17. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Transactions on Computer–Human Interaction 2005;12(2):210–232.
18. Rada R. Searching and gradualness. BioSystems 1981;14:219–226.
19. Preece J. Online communities: Designing usability, supporting sociability. Chichester, UK: Wiley; 2000.
20. Gitlow S. The Online Community as a Healthcare Resource. In: Nash D, Manfredi M, Bozarth B, Howell S, eds. Connecting with the new healthcare consumer: Defining your strategy. Gaithersburg, MD: Aspen; 2001. pp 113–134.
21. Fahy P. Two methods for assessing critical thinking in computer-mediated communications (CMC) transcripts. International Journal of Instructional Technology and Distance Education 2005;2(3):13–28.
22. Brown J, Eskenazi M. Student, text and curriculum modeling for reader-specific document retrieval. In: The IASTED International Conference on Human-Computer Interaction. Phoenix, AZ: International Association of Science and Technology for Development; 2005. pp 44–50.
23. Ferguson G. Statistical Analysis in Psychology and Education. 2nd ed, New York: McGraw-Hill; 1966.
24. Wilson DS. Testing major evolutionary hypotheses about religion with a random sample. Human Nature 2005;16(4):419–446.

Copyright of Medical Informatics & the Internet in Medicine is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.