Taming a Menagerie of Heavy Tails with Skew Path Analysis

Josh Introne
Department of Media & Information
Michigan State University
East Lansing, MI
jintrone@msu.edu

Sean Goggins
iSchool at
The University of Missouri
Columbia, MO
gogginss@missouri.edu

ABSTRACT

The discovery of stable, heavy-tailed distributions of activity on the web has inspired many researchers to search for simple mechanisms that can cut through the complexity of countless social interactions to yield powerful new theories about human behavior. A dominant mode of investigation involves fitting a mathematical model to an observed distribution, and then inferring the behaviors that generate the modeled distribution. Yet, distributions of activity are not always stable, and the process of fitting a mathematical model to empirical distributions can be highly uncertain, especially for smaller and highly variable datasets.

In this paper, we introduce an approach called *skew-path analysis*, which measures how concentrated information production is along different dimensions in community-generated data. The approach scales from small to large datasets, and is suitable for investigating the dynamics of online behavior. We offer a preliminary demonstration of the approach by using it to analyze six years of data from an online health community, and show that the technique offers interesting insights into the dynamics of information production. In particular, we find evidence for two distinct point attractors within a subset of the forums analyzed, demonstrating the utility of the approach.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing

G.3 [**Probability and statistics**]: Distribution Functions

General Terms

Measurement, Theory

Keywords

Power Law, Social Media, Diversity, Dynamics

1. INTRODUCTION

The term "heavy-tailed distribution" describes numerical distributions that have a longer tail of low-probability events than would be expected under an exponential distribution (a family that includes normal, Poisson, and binomial distributions). As with exponential distributions, there are a variety of heavy-tailed distributions, including log-normal and power law distributions.

Heavy-tailed distributions show up in many places, notably on the web and in social media, but also in natural and social settings [6,7]. Power law distributions, in which one quantity varies with the power of another, are particularly ubiquitous in social media, leading some to suggest that one should almost always apply log-transformations to data harvested from social spaces [17], and many others to seek general mechanisms that explain their abundance (e.g. [3,10]).

Determining which distribution best fits a given set of empirical data often involves maximum likelihood estimation. For example, Clauset et al. [6] recommend applying a bootstrapping procedure to estimate the maximum likelihood that a given distribution might fit a power-law. The procedure can be computationally intensive and does not provide conclusive determination; what may appear to be an acceptable power-law might still be a better fit for a log-normal or exponential distribution, and so if the class of distribution is important, the careful scientist should check her data against the likely alternatives.

For large datasets with stable distributions, this kind of model-based analysis is warranted and can yield deep insight. However, if one seeks to understand the dynamics of data with highly variable participation rates, or with relatively few data points, or to investigate what happens during the growth or decay of a given system, substantial uncertainty can make it difficult to distinguish between mathematical models [4].

One result of this is that research devoted to discovering patterns of human interaction that generate stable distributions tends to focus upon large systems with stable patterns of activity. This creates a bias in the research literature, and leaves a methodological challenge for the study of smaller, highly dynamic systems that populate a significant portion of the web ecosystem.

In this paper, we offer an analytical approach that can be used to characterize distributions regardless of the amount of data available, and this is especially useful for understanding the variations in the shape of these distributions over time. We define *skew*[1] as the basic measure underlying the approach, which is closely related to Shannon's entropy and based on measurements of diversity developed in ecology. A *skewed* distribution is the opposite of an *even* distribution, and can be loosely interpreted as the degree of concentration of a measured quantity such as posts per user, or links per node. When applied to dimensions that characterize aspects of user content creation, skew reflects the degree to which a small number of voices or ideas dominate. The

 ACM 978-1-4503-3672-7/15/06…$15.00 DOI: http://dx.doi.org/10.1145/2786451.2786484

[1] Not to be confused with *skewness*, which quantifies the asymmetry of a probability distribution. As it is used here, *skew* is more closely related to *kurtosis*.

intuitive accessibility of the measure thus makes it easier to interpret then more abstract properties such as entropy or kurtosis.

In the following, we first offer some background on various prior approaches to measuring the dynamic behaviors of online communities. We then describe the derivation of skew and provide supporting material to help develop the reader's intuition. Finally, we offer a preliminary case study of an online health forum using a visualization technique we refer to as *skew path analysis*, and discuss our findings.

2. BACKGROUND

Research that has centered upon the appearance of heavy-tailed distributions on the web has sought to explain the local interactions that give rise to them. For example, Barabási and Albert [3] suggested that preferential attachment—in social networks, the tendency for popular individuals to be more attractive social partners for newcomers—is responsible for power-law distributions of network connectivity. Huberman and Adamic [10] found that this did not predict the structure of the Web, and proposed an alternative stochastic model. Within social information streams, Barabási demonstrated how auto-correlated individual behaviors might explain observed power-laws[2], but Malmgren et al. [15] offered an alternative, random model.

The drive to demonstrate a parsimonious mechanism underlying common distributions of activity across many complex, sociotechnical systems is understandable. Relatively less attention has been paid to the ways in which distributions vary within and across various sociotechnical systems, despite the fact that such variance exists. For instance, Guo et al. [9] show that a stretched exponential distribution is a better fit for user content creation in a number of different online social networks than a power-law distribution. Zhao et al. [22] identify different dynamic processes at various scales of analysis and over time within a large online social network. In particular, they find that preferential attachment decays over time for the social network they study.

Other research has sought to characterize the dynamics of sociotechnical systems without explicitly analyzing or fitting distributions. Both Kan et al. [13] and Viegas and Smith [21] develop new approaches to visualization that capture the variance in distributions over time. The work reported here is similar in spirit to these efforts, yet bears a clearer relationship to model-based approaches that explicitly seek to fit distributions to data.

2.1 Interpreting Skew

Within web-based research, distributions are generally obtained by comparing the number of *instances* performing at different *levels* of activity. For example, we may count the proportion of nodes (instances) with different degrees (levels) in a network, or the number of topics (instances) exhibiting different numbers of posts (levels). This counting procedure leads to an inference about a probability distribution that governs the appearance of instances at different levels of activity. A heavy tailed distribution differs from an exponential distribution because the appearance of instances at high-levels (further down the tail of the distribution) is more likely.

A similar form of analysis is of critical importance in ecology research; the distribution of individuals in each species in a bounded geographical area is the *diversity* of that area, and much thought has been put into developing measurements of diversity [12,20]. Note that there is a slight difference between the analysis of diversity, which examines the count of instances across *classes* (a categorical dimension) rather than *levels* (an ordered dimension). It is a simple matter to convert between the two framings; we adopt the ecological framing here.

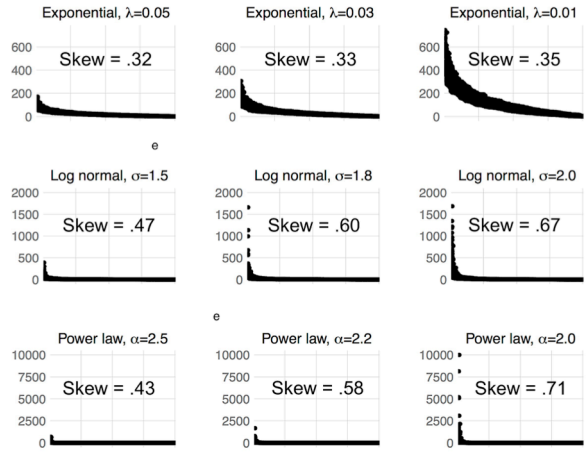

Figure 1: Average skew for different distribution families and parameters. Following the ecological framing of "members of a species", the x-axis is a categorical axis indicating class, and the y-axis is the number of instances of that class.

One widely used measure of diversity is the exponential of Shannon's entropy:

$$D = e^{-\sum_{i=1}^{S} p_i \ln p_i}$$

As discussed by Tuomisto [20:854–855], D is a composite of two distinct aspects of diversity—*species richness*, which is simply the number of distinct species $|S|$ found in the area being measured, and *evenness*, which is a measure of how evenly distributed individuals are across all species. D is maximized and equal to $|S|$ when each species contains the same number of individuals in an area, decreases both as the number of total species is reduced and as the distribution of individuals across species becomes less even.

By itself, evenness is one way of characterizing the "flatness" of a distribution. It varies from (0-1] can be obtained by dividing D by $|S|$. Isolating evenness achieves our original goal of finding a more flexible approach to characterizing changing distributions. While it may also be useful to consider diversity, we have found that when analyzing user content creation, the "noise" in diversity due to varying levels of traffic can hide more interesting variation in the shape distribution.

For our domain, we have also found it intuitively easier to think about *lack* of evenness, rather than evenness, and so we define the skew of a distribution to be 1-evenness, or:

$$Skew = \begin{cases} 1 - \dfrac{e^{-\sum_{i=1}^{S} p_i \ln p_i}}{|S|} & if\ |S| > 0 \\ 0 & if\ |S| = 0 \end{cases}$$

From the definition of entropy, and excluding the degenerate case (no instances to measure), skew is at a minimum for a given set of classes when all measured classes contain the same number of instances, and a maximum when only one class is populated. However, the existence of $|S|$ in the denominator means that the absolute minimum skew attainable varies with the number of classes measured. Furthermore, in many cases where we might assess skew (e.g. the number of posts per topic) it may not make

sense to infer the existence of a class unless it appears in our data, so the minimum inferred probability for any given class is $1/|S|$.

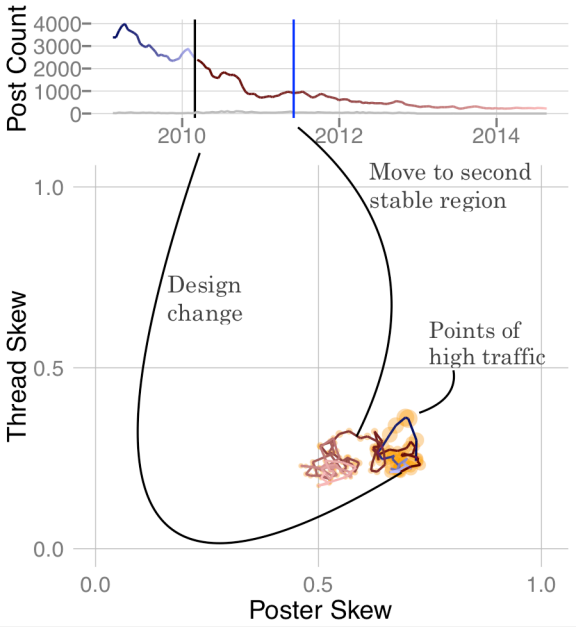

Figure 2: Traffic patterns and skew path visualization for the fibromyalgia forum. The skew path is slightly smoothed to ease interpretation. Two significant points are indicated in the diagram; the first point is a significant design change in the forum, and the second is the point at which the forum appears to move from one region of skew space to another.

To help develop an intuition about how skew behaves in context then, Figure 1 provides several examples drawn from several canonical distributions, and reports average skew values for each. For each graph, we generated 100 distributions of 100 different points drawn from a distribution with fixed parameters, and report the average skew for these distributions. In the case of the log-normal distributions, $\mu=0$. For exponential distributions, skew tends to remain low because there is a smoother transition from higher to lower probability events. Log-normal and power-law distributions have similarly high skew values for some parameter ranges, but power-laws produce the highest skew values as α decreases.

2.2 Skew Path Analysis

In this paper we use skew to analyze posting patterns in message forums hosted by WebMD, a large and popular online health information service. Our analysis focuses upon skew in two dimensions: *poster skew* is used to measure the distribution of posts across individual posters within a time window for a given forum, and *thread skew* is used to measure the distribution of posts across individual threads in a forum within a time window.

We visualize the change in skew over time as a path through this two dimensional space. Figure 2 provides an annotated sample of the visualization from a forum dedicated to the discussion of fibromyalgia. The visualization is tuned to make general features, such as the obvious "bunching" of the path, and the fact that high degrees of activity tend to occur at higher levels of poster skew, easy to assess. There are other ways to visualize skew and its correspondence with other values, but the approach taken here is intended to elucidate specific features at the expense of others.

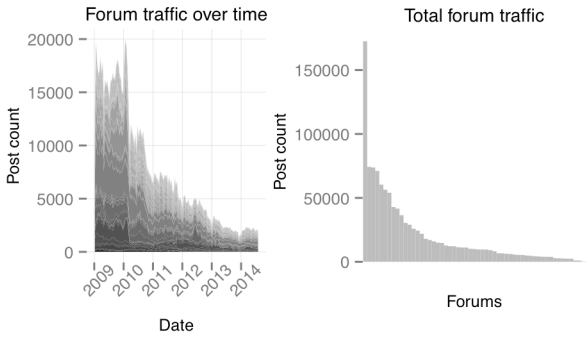

Figure 3: WebMD traffic patterns over time

3. Applying Skew Path Analysis to WebMD

To investigate the utility of skew path analysis, we applied it to study longitudinal data from an online health information service called WebMD. WebMD is a popular online health resource (as of January 2015, Alexa ranked WebMD.com as the 120th most popular site in the US, and second most popular health site, behind NIH.gov [23]). WebMD hosts a variety of forums, some of which have been in existence since 1998 [18], though they have gone through various changes since that time.

We chose WebMD for a variety of reasons. At least 5 years of forum postings were available for 55 "featured" forums, and WebMD's terms of service explicitly allow use of this information for non-commercial purposes[2]. The featured forums are maintained and do not contain spam. The data also spans several sociotechnical design changes, involving a significant interface redesign in March of 2010, and various changes to policies for staffing of forums with moderators and medical experts. Others have shown that technology changes tend not to be received well in online health communities [16], that the involvement of designated experts can disrupt social communication [18:109], and that heavy handed moderation can dampen overall conversation[11]. We were very interested to know if skew path analysis would provide additional insight.

Most importantly for our analysis, there is a great deal of variance in the participation levels both within and between forums (Figure 3). The total number of posts in each forum varied widely, from ~175K posts (fibromyalgia) to only 276 posts (raising fit kids), and the vast majority of forums have experienced significant declines in participation since the redesign in 2010. This high degree of variance makes it difficult to use model-based techniques, for reasons discussed above.

We scraped all publicly available data from the 55 featured forums hosted on WebMD in August of 2014, yielding a total of 1.1M posts spanning seven years. The earliest post in any forum was on 11-30-2006, and latest first post in any forum was on 11-09-2011. However, it became apparent that significant portions of data preceding 01-01-2009 were missing for many of the forums,

[2] From the terms of service: "WebMD authorizes you to view or download a single copy of the material on the WebMD Site solely for your personal, noncommercial use if you include the copyright notice located at the end of the material, for example: '©2013, WebMD, LLC. All rights reserved'"; http://www.webmd.com/about-webmd-policies/about-terms-and-conditions-of-use?ss=ftr#part3

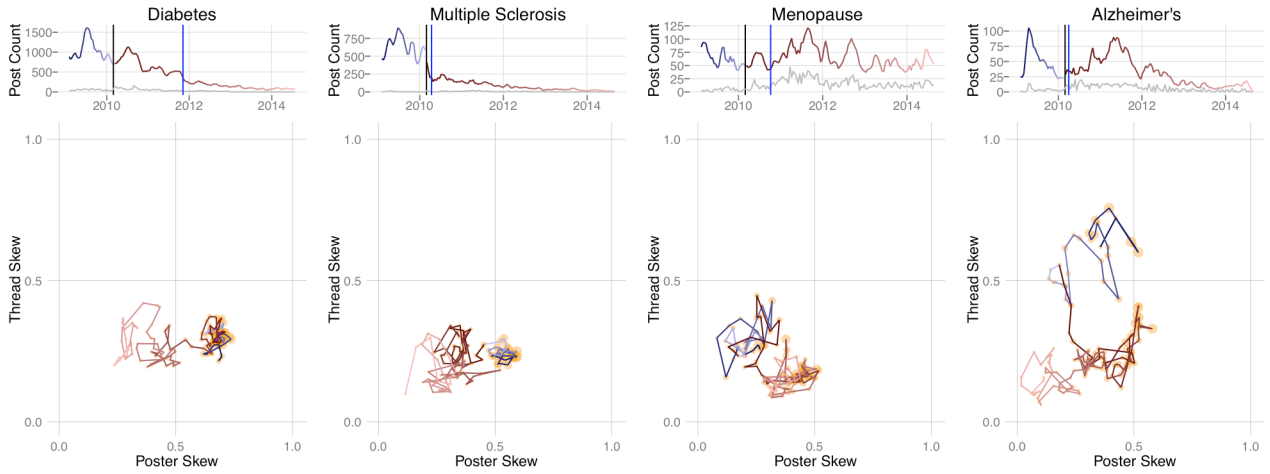

Figure 4: Traffic patterns and skew path analysis for a sampling of forums. The gray lines in the top graphs capture moderator involvement. Note the distinct appearance of (at least) two clusters in each forum, but the position of the clusters varies. Vertical lines in the traffic graphs capture the point of the design change and the point at which the system appear to move to a second stable region of skew space.

and so we restrict our analysis here to all data collected after that point.

Roughly .1% of all posts were associated with a date that conflicted with the post ordering on the site. We examined these conversations and found post ordering to me more reliable than the timestamp, and so adjusted those posts with apparently incorrect timestamps to be equidistant from correctly sequenced, bracketing posts. This did introduce some error into a small proportion of the data, but as we are not concerned with precise timing of posts in this investigation, the introduced error should have little to no impact on the following analysis.

3.1 Initial Findings

To use skew path analysis to understand the dynamics of each forum, we binned data in two-week periods starting with the first post in each forum. Prior work has binned similar data from one-week to one-month intervals [e.g. 18], and we found that two-week intervals captured dominant trends without obscuring important local variation. Figure 4 presents a representative sample of our results. Data is smoothed in the figure to make patterns more distinct, but the following analysis use unsmoothed data.

Figure 4 captures two interesting patterns observed in the visualizations. Most of the forums exhibited a two-phase evolution in time, but this evolution took two different forms. For the more highly trafficked forums (e.g. the Diabetes and Multiple Sclerosis forums in Figure 4), the first phase was characterized by high posting activity, high poster skew and relatively low thread skew. In the second phase, these forums moved to a somewhat more diffuse pattern, characterized by lower poster skew, slightly lower thread skew, and lower posting traffic.

However, a handful of the less heavily trafficked forums (e.g. the Menopause and Alzheimer's forums in Figure 4) exhibited a different pattern, in which poster skew started out relatively low and then stayed the same or increased somewhat in the second phase. However, thread skew in these forums dropped significantly between the two phases. In these forums, we also note a relatively high degree of staff postings.

For all forums, the design change occurred while the system was operating within the first stable region. Across communities, the immediate response to the design change varied, but (with the exception of anomalous forums such as the Alzheimer's forum) traffic dropped off in the later parts of 2010. At some point during this decay, most forums moved to a second stable region in skew space. The distribution of traffic over threads did not change significantly, but the distribution of poster activity became more even at this point.

To help quantify the observation of two apparently stable regions in the skew path across all forums, we used a modified *k-means* procedure to identify the point in time which minimized the within cluster sum of squared errors. As a rough indicator of clustering quality, we applied silhouette analysis [19], which evaluates the cluster assignment for every point in a dataset. A point is assigned a positive value (maximum of 1) if it is closer on average to other points in it's assigned cluster, and a negative value (minimum of -1) if it is closer to points in the other cluster. The average of all points varies from [-1,1] and is an indicator of how distinct the clusters are.

Five of the forums did not have any traffic prior to the design change in 2010, and these forums did not have distinct clusters. Among the remaining forums, clustering fitness was normally distributed around a mean of .34 (*median=.35, s.d.*=.15), and was positive in all cases. This indicates that points were in general clustered correctly; however, the compactness and relative distance of clusters varied.

4. ANALYSIS AND CONCLUSION

Although the preceding findings indicate a dominant pattern in the evolution of the dynamics of the WebMD forums, whereby the forums jump from one apparently stable set of dynamics to another. We hypothesize that these regions of stability are point attractors operating within a significant proportion of the forums examined.

The term *attractor* is used within the literature on complex systems to characterize stable dynamics of a complex system [1]. Once a dynamic system has attained an attractor it will tend to stay there unless there is some significant perturbation to the

system. Gersick's theory of punctuated equilibrium [8] and Kuhn's theory of scientific revolutions [14] are well-known examples of attractors operating within sociotechnical systems. More directly related to web science, but without explicitly using the language of dynamic systems, Butler [5] offers compelling arguments about how the interplay between resource availability and benefit provision might lead to the existence of a stable attractor in online social media.

What causes the observed regions of stability in our data is not yet clear, and whether or not these stable regions can be generalized are questions for future work. Our early findings suggest that, for more highly trafficked forums, the shift often co-occurs with the loss of or significant decline in the activity of high degree-centrality users. These users also have tight couplings with one another within forums. In the less highly trafficked forums, the reduction in thread skew seems to result from increased activity due to moderators answering questions from new users. Further investigation is necessary to verify these observations, but if they are accurate, then the shift in skew may be due to a differential impact of the design change on a particular class of users, leading to a change in the overall dynamics.

The identification of apparent stability amidst a multitude of complex interactions is exciting because it suggests that there may be a simple, possibly general explanation for this stability. This is why the surprising ubiquity of power-laws and the attempts to infer the mechanism responsible for them have attracted so much attention within the scientific community. In this paper, we have introduced a new tool that can be used to hunt for stability that may appear in corners that model-based approaches cannot reach. Our initial results suggest that skew path analysis holds some promise for the endeavor.

5. REFERENCES

1. Holly Arrow, Joseph E. McGrath, and Jennifer L. Berdahl. 2000. *Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation*. SAGE.
2. Albert-László Barabási. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039, 207–211. http://doi.org/10.1038/nature03459
3. Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439, 509–512. http://doi.org/10.1126/science.286.5439.509
4. Maurice C. Bryson. 1974. Heavy-Tailed Distributions: Properties and Tests. *Technometrics* 16, 1, 61–68. http://doi.org/10.1080/00401706.1974.10489150
5. Brian S. Butler. 2001. Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Information Systems Research* 12, 4, 346–362. http://doi.org/10.1287/isre.12.4.346.9703
6. A. Clauset, C. Shalizi, and M. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review* 51, 4, 661–703. http://doi.org/10.1137/070710111
7. Mark E. Crovella, Murad S. Taqqu, and Azer Bestavros. 1998. A Practical Guide to Heavy Tails. In Robert J. Adler, Raisa E. Feldman and Murad S. Taqqu (eds.). Birkhauser Boston Inc., Cambridge, MA, USA, 3–25. Retrieved March 18, 2015 from http://dl.acm.org/citation.cfm?id=292595.292596
8. Connie J. G. Gersick. 1991. Revolutionary Change Theories: A Multilevel Exploration of the Punctuated Equilibrium Paradigm. *Academy of Management Review* 16, 1, 10–36. http://doi.org/10.5465/AMR.1991.4278988
9. Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. 2009. Analyzing Patterns of User Content Generation in Online Social Networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 369–378. http://doi.org/10.1145/1557019.1557064
10. Bernardo A. Huberman and Lada A. Adamic. 1999. Internet: Growth dynamics of the World-Wide Web. *Nature* 401, 6749, 131–131. http://doi.org/10.1038/43604
11. Jina Huh. forthcoming. Clinical Questions in Online Health Communities: The Case of "See your doctor" Threads. *Proceeding of the 2015 conference on Computer Supported Cooperative Work*, ACM Press.
12. Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2, 363–375. http://doi.org/10.1111/j.2006.0030-1299.14714.x
13. Andrey Kan, Jeffrey Chan, Conor Hayes, Bernie Hogan, James Bailey, and Christopher Leckie. 2013. A time decoupling approach for studying forum dynamics. *World Wide Web* 16, 5-6, 595–620. http://doi.org/10.1007/s11280-012-0169-1
14. Thomas S. Kuhn. 2012. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press.
15. R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luís A. N. Amaral. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 47, 18153–18158. http://doi.org/10.1073/pnas.0800332105
16. Diane Maloney-Krichmar and Jenny Preece. 2005. A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community. *ACM Trans. Comput.-Hum. Interact.* 12, 2, 201–232. http://doi.org/10.1145/1067860.1067864
17. Daphne Ruth Raban and Eyal Rabin. 2007. *Statistical Inference from Power Law Distributed Web-Based Social Interactions*. Social Science Research Network, Rochester, NY. Retrieved March 18, 2015 from http://papers.ssrn.com/abstract=999286
18. Catherine Ridings and Molly McLure Wasko. 2010. Online discussion group sustainability: Investigating the interplay between structural dynamics and social dynamics over time. *Journal of the Association for Information Systems* 11, 2. Retrieved from http://aisel.aisnet.org/jais/vol11/iss2/1
19. Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. http://doi.org/10.1016/0377-0427(87)90125-7
20. Hanna Tuomisto. 2010. A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* 164, 4, 853–860. http://doi.org/10.1007/s00442-010-1812-0
21. F.B. Viegas and M. Smith. 2004. Newsgroup Crowds and AuthorLines: visualizing the activity of individuals in conversational cyberspaces. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004*, 10 pp.–. http://doi.org/10.1109/HICSS.2004.1265288
22. Xiaohan Zhao, Alessandra Sala, Christo Wilson, et al. 2012. Multi-scale Dynamics in a Massive Online Social Network. *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, ACM, 171–184. http://doi.org/10.1145/2398776.2398795
23. Alexa - Top Sites by Category: Health. Retrieved January 24, 2015 from http://www.alexa.com/topsites/category/Top/Health