The Dynamics of Information Production in an Online Health Community

JOSHUA INTRONE, Michigan State University

SEAN GOGGINS, University of Missouri

1. INTRODUCTION

Power laws are ubiquitous in social media [2], but they exhibit great variety and do not appear for the same, or even agreed upon, reasons. For instance, individual behaviors that give rise to power-laws in social information streams can be explained via either random [6] or correlated behaviors [1]. And the well known "80/20" rule (80% of the work is done by 20% of the users) is often only a rough approximation when applied to user content creation [4].

Despite its existence, few studies have sought to exploit this variance for insight into the time varying nature of social media. Quantifying this variance offers two benefits. First, the "peakedness" of a power law can be used to measure the concentration of the measured phenomena; e.g. the concentration of posting activity in a forum, or the concentration of a community's topical focus. Thus, although it is a high-level aggregate measurement, it provides the analyst with insight into the dynamics of information production. Second, for the health community domain we describe here, observing these distributions over time reveals a process of punctuated equilibrium [3], suggesting the existence of stable point attractors. These stable points can be used to focus analysis and develop insights about how various contextual factors impact the dynamics of information production in an online community.

We introduce an approach to measuring distributions of user activity over time, and then apply it to ~6.5 years of data drawn from 55 discussion forums hosted by WebMD, a popular online health service. This data includes a sociotechnical design change, and we identify a significant shift in the dynamics of each forum following this change.

1.1 Indexing fat tails

Power laws are often found when we count the number of items within a class. For example, power law distributions may be encountered when we count the number of nodes (instances) with different degrees (classes) in a network, or the number of posts (instances) in each topic (classes).

Quantifying such distributions is of critical importance in ecological 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 [5,8]. One widely used measure of diversity is the exponential of Shannon's entropy, $D = e^{-\sum_{i=1}^{S} p_i \ln p_i}$.

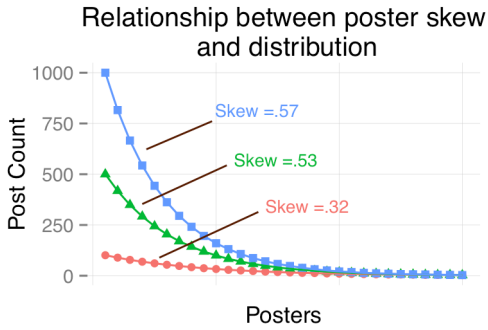

Figure 1: Sample skew levels from generated data from several logarithmic distributions; 30 posters shown

As discussed by Tuomisto [8:854–855], 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. D decreases as the distribution of individuals across species becomes less even. Dividing D by $|S|$ isolates evenness, which varies from (0-1] when there is at least a single individual present.

Evenness is thus one way of characterizing the "flatness" of a distribution. For our domain, we have found it intuitively easier to think about *lack* of evenness, rather than evenness, and so we define the *skew* of a distribution[1] to be *1*-evenness, or:

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

Here we will focus upon the skew of two distributions. *Poster skew* will be used to quantify the distribution of posts across individual posters within a time window for a given forum. *Thread skew* will be used to quantify the distribution of posts across individual threads within a time window. Figure 1 offers an example of how poster skew relates to several different distributions of activity.

To help visualize the relationship between thread skew and poster skew, we introduce a visualization we call a *skew path*. A skew path is a trajectory that plots the state of the system as measured by the two types of skew at successive points in time (at preset intervals). In the visualizations provided here, we annotate the skew path to indicate the point at which the redesigned WebMD site was launched and to provide a rough indication of the relative levels of traffic.

1.2 Analysis

We applied skew path analysis to roughly 6.5 years of forum posts from each of the 55 featured web-forums on WebMD. WebMD moderators and experts (with medical degrees) participate in these forums, but the participation has varied widely over time. In February 2010, there was a significant sociotechnical design change to the site; features that allowed members to rapidly determine who the participants in each thread were removed, and the number of experts on the site was increased by over 200%. Messaging on the site indicated that WebMD sought to both improve and highlight informational aspects of the site.

A representative sample of results is shown in Figure 2. Absolute traffic levels varied significantly among the forums, but all appeared to exhibit two distinct regions of stability within the skew path. The shift from one stable region to the other did not occur at precisely the same time as the launch of the redesigned site, but in all cases it did occur after the change.

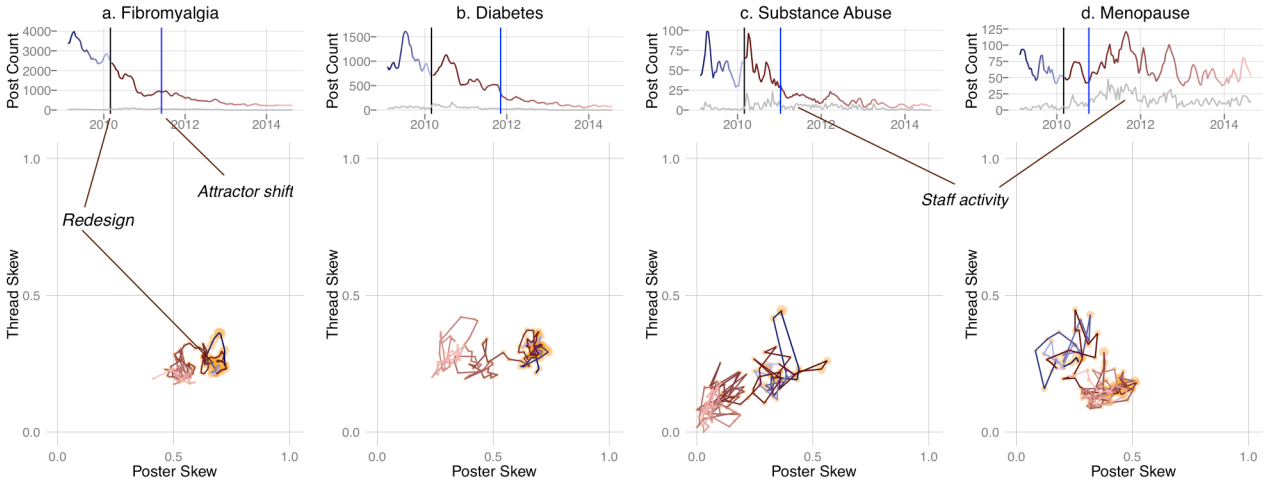

Figure 2: Posting traffic and skew paths for four representative forums. All data is binned in two-week intervals, and smoothed using a lagged moving average window (size = 2 periods). Top charts indicate post count for each two-week period; the hue change (from blue to red) corresponds to the design change. First vertical line is the point of the design change, second is the point of the shift to the second stable region of the path. Top charts also include staff activity (lower, gray line). Bottom graphs are skew paths; line color corresponds to color in the top chart. Highlighted areas are regions of relatively high traffic for that forum.

[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*.

In the first stable region for most forums (e.g. Figure 2a-c) poster skew and posting rate were relatively high, and thread skew was relatively low. In the second stable region, poster skew decreased, traffic was on average lower, and thread skew either remained roughly the same or decreased. However, in a few cases (e.g. Figure 2d) poster skew actually increased, and traffic rates remained stable or did not decrease significantly.

We split the data into the two stable regions for each forum as follows. We split each skew path into two clusters using a modified *k-means* clustering to find the best break point, and evaluated the resultant clustering using silhouette analysis [7]. This produces a measure varies from [1…-1], where positive values indicates that points on average are clustered correctly. Clustering fitness based on this analysis was normally distributed around a mean of .34 (*median=.35, s.d.*=.15), and was positive in all cases. This indicates that clustering was in all cases effective; however, the compactness and relative distance of clusters varied.

Figure 3: Comparison of the relationship between poster skew and posts per period from the two stable regions across all forums

We then analyzed the two stable regions across the dataset using quantitative and qualitative methods. Among our findings, we discovered a significant difference in the relationship between traffic and poster skew between the two clusters (Figure 3); the slope of the best fit line (*poster skew* x *log(posts)*) was 5.44 in the first stable region and 4.11 in the second. Extrapolating from this relationship, the expected posting rate at maximum poster skew is ~2469 posts per two week period for the first stable region and ~927 posts per period for the second. Thus the presence of dominant, high volume posters leads to higher levels of traffic in the first stable region of the skew path than in the second—there is more "bang for the buck" for having high-frequency posters in the first stable region of the skew path than the second.

In those forums with anomalous skew paths (such as shown in Figure 2d), the increase in poster skew and concomitant reduction in thread skew was due to the presence of highly active experts, who were very aggressive in fielding questions from the community. Across the dataset, the presence of highly active experts tended to reduce thread length overall, transforming the community dynamic from an emotional support group to a more stylized Q & A interaction.

This analysis suggests two stable modes of production in the WebMD health forums. In one mode, information production is concentrated in a handful of members who are frequently engaged in highly interactive, social conversations. These conditions may seem ripe for the spread of misinformation, but member activity is maintained. In the second mode, staff members play a significant role in information production and attend to a broader range of topics, but member activity wanes or remains relatively low. Moreover, our analysis suggests that the shift in dynamics is caused by the design change, highlighting the influence of sociotechnical design on forum-based communities.

Our results have a variety of implications for the design of knowledge-based communities in general, demonstrate the value of skew path analysis, and illustrate one approach to characterizing the complex dynamics of information production in online communities.

REFERENCES

1. Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature 435*, 7039 (2005), 207–211.
2. Clauset, A., Shalizi, C., and Newman, M. Power-Law Distributions in Empirical Data. *SIAM Review 51*, 4 (2009), 661–703.
3. Gersick, C.J.G. Revolutionary Change Theories: A Multilevel Exploration of the Punctuated Equilibrium Paradigm. *Academy of Management Review 16*, 1 (1991), 10–36.
4. Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. (Eric). 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 (2009), 369–378.
5. Jost, L. Entropy and diversity. *Oikos 113*, 2 (2006), 363–375.
6. Malmgren, R.D., Stouffer, D.B., Motter, A.E., and Amaral, L.A.N. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences 105*, 47 (2008), 18153–18158.
7. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, (1987), 53–65.
8. Tuomisto, H. A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia 164*, 4 (2010), 853–860.