Collaboration and Communication in Online Environments:
A Social Entropy Approach

Sorin Adam Matei
Kyoungae Oh
Robert Bruno
Department of Communication
Purdue University
2132 BRNG
100 N. University Drive
West Lafayette, IN 47907
Corresponding author:
smatei@purdue.edu
Tel: 765 494 7780
Abstract

Despite the fact that diverse and equal participation of members is considered to be the essence of true online communities, no compelling index has been proposed to measure the degree of diversity in terms of contributions of content to online social environments. The primary purpose of the paper is to reintroduce to the communication field a tool particularly suited for measuring diversity and equality of participation. This paper argues that the notion of social entropy, as defined in information theory, can be applied to measure participatory diversity and is especially suitable for characterizing online groups. Adopting social entropy researchers can better and more systematically (i) estimate participatory diversity of online communities or groups; (ii) compare different online groups in terms of participatory diversity; (iii) evaluate changes of contribution over time; (v) understand online interaction dynamics at small and large group scales.
Collaboration and Communication in Online Environments:
A Social Entropy Approach

Online collaboration has been historically embraced and promoted for its tremendous potential benefits: lower production costs, convenient access, and a theoretical ability to engage a larger array of minds and inputs compared to face-to-face communication (Berman & Weitzer, 1997; Braman, 1994; Kiesler, Siegel, & McGuire, 1984; Licklider & Taylor, 1968; Rheingold, 1993; Sclove, 1995; Sproull & Kiesler, 1991). The latter perceived benefit is probably the most enticing. This effect is premised on a number of characteristics considered to be intrinsic to online communication technologies. The most important is the functional equality of the nodes that constitute any computer network (Hauben, 1999). This led some first generation Internet researchers to believe that online group members would contribute to the collective effort equally (Hiltz, 1984; Hiltz & Turoff, 1978; Kiesler & Sproull, 1992). The built-in equality of CMC systems has been also believed to hinder or even preventing majorities from becoming tyrannical (Grossman, 1995; Hiltz & Turoff, 1978). Finally, it is also believed that the egalitarian nature of the medium will create new avenues for expressing non-mainstream views (Myers, 1987; Turkle, 1995). Within this set of interlocked expectations is hidden something akin to a Holy Grail of collaboration: a method for producing rich and diverse knowledge that grows out of the egalitarian efforts of the many, including the ignored, marginalized, and non-involved voices. This view was intensely and broadly popularized by books such as The Wisdom of the Crowds (Surowiecki, 2004), Out of Control (Kelly, 1995), The Wealth of Networks (Benkler, 2007), or Emergence (Johnson, 2001).

Despite the fact that diversity or equality of interaction and contribution has been a central assumption for many of users or observers of online environments, research that directly addresses these issues has been fragmentary and generally subordinated to other priorities. Most frequently, diversity and equality of collaboration are the implied desirable end-states of collaborative models to be studied, not the main characteristic to be studied. This means that the direct operationalization and integration of diversity in
predictive models related to communication are relatively rare (Hiltz & Turoff, 1978). One notable exception would be decades-old research on programming diversity in media economics and media agenda-setting (Chaffee & Wilson, 1977; Dominick & Pearce, 1976) or some of Osgood’s (Osgood & Wilson, 1955) or Schramm’s (Schramm, 1955) polymath considerations about entropy as a communicative issue.

What we are currently missing in the study of computer-mediated communication are the means to directly and objectively operationalize diversity and equality as system-level phenomena. This is particularly important if one considers the fact that equality and diversity of collaboration are viewed, as the long list of studies indicated at the very beginning of the paper suggests, as the most desirable and defining characteristics of computer-mediated communication.

The present paper proposes a central statistical measure and companion theoretical insight for characterizing computer-mediated communication collaboration. The goal is to demonstrate that concepts that can describe and explain diversity, such as “social entropy” can help us express the amount (quantity) of collaboration in online environments in such a manner that collaborative process can be compared across groups, settings, and time periods.

Diversity and Social Entropy: A Neglected Tradition

Characterizing (and quantifying) the state of a communicative system with respect to its level of diversity and organization is a relatively mature theoretical and methodological concern. It was formulated almost 60 years ago by Shannon and Weaver (Shannon & Weaver, 1998), who proposed “entropy” as a central measure of information system diversity and, through it, of information itself. Yet, with some limited use in theoretical discussions about communication as an “uncertainty reduction” process, for almost as many years the concept has been relatively ignored by communication scholars, especially in the context of describing social processes.

The relative absence of entropy in the study of communication should be, in view of its association with the birth of communication as a discipline, quite surprising. Entropy was first applied to human phenomena by Shannon and Weaver, the inventors of a profoundly innovative “theory of communication”
For them, communicative acts that carried meaning were characterized by two factors, redundancy and organization. The lower the level of randomness and the higher that of order, the more likely that a communication act carried meaning (i.e., the act can be characterized as “information” or as a “signal” that is distinguishable from “noise”). Shannon and Weaver borrowed the conceptual and mathematical tools needed to describe information load and by implication communicative system organization from physics where presence or absence of order in a system is designated as level of “entropy.” This measures how diverse a system is in terms of its constitutive elements. When the elements are present in an equal proportion, thus present a maximum level of diversity, the system is said to have a high level of entropy. When the elements present a level of imbalance (some are more prevalent than others), entropy and diversity are low.

The entropy concept, as we will discuss in the next section, has numerous implications. It reveals how organized or diverse a system is, and also if level of organization goes up or down. Furthermore, it can be also used for making some indirect inferences about the nature of the diversity that it describes.

Yet, these potential benefits and implications of employing the entropy concept have been for a long time ignored by communication scholars. In the interim, other sciences have further explored the issues of diversity and entropy, developing very sophisticated statistical methods for identifying their magnitude within and for determining their comparability between systems. Economics, sociology, geography and environmental sciences are some of the disciplines that have adapted and refined the concept of entropy (McDonald & Dimmick, 2003). For example, they have adapted the entropy measure to express the degree of organization, diversity, richness or equality of representations of firms, species, features, etc. in a given population, area, or society (Maignan, Ottaviano, Pinelli, & Rullani, 2003).

These methods and the theoretical implications they create, both of which will be discussed in the following sections, are now ripe to be incorporated in the communication disciplines, especially those focusing on the emergence and impact of new online communication environments. Communication technologies are especially in need of measuring and understanding diversity and entropy for two reasons.
The first reason, as already mentioned, is that online environments are considered to be intrinsically egalitarian and capable of fostering a sense of diversity (Hiltz & Turoff, 1978; Kiesler & Sproull, 1992). The other reason is the expectation that the egalitarianism of online collaboration and communication is not only morally good, but that it could be an alternative method for creating social order and social coordination (Johnson, 2001; Rheingold, 2002). Online environments are believed, especially by many practitioners, to have the ability to self-organize and to create “emergent” orders (Johnson, 2001; Ostrom, 1990; Raymond, 2001). The presupposition is that the more decentralized and egalitarian the interaction and the more massive the collaboration, the more likely the online systems would be able generate solutions that hierarchical, top-down methods of control and coordination could not (Raymond, 2001). Yet these claims cannot be verified and the nature of these phenomena, which include fascinating endeavors such as Wikipedia, cannot be understood until we get a better handle on measuring the over-time evolution of collaborative diversity and entropy.

The present low awareness (to say nothing of acknowledgment) of the importance of diversity and entropy of interaction online becomes more and more glaring and the need to redress this situation more urgent as collaborative communication environments have become mainstream tools. A profusion of many-to-many technologies, congregating millions of users, has emerged (Rheingold, 2002). These include auction sites, social networking or tagging software, wikis, blogs, Q&A sites, etc. (Blair & Thompson, 2005; Butler, 2001; boyd & Ellison, 2007) Boyd and Ellison that were launched and are consciously utilized precisely for their unique power to foster a diversity of opinions and increase equality of participation. These online systems have changed some fundamental aspects of everyday life, such as news gathering and consumption, social support, commerce, dating, etc (Rheingold, 2002).

The relative absence of research that addresses decentralization, the purported egalitarianism of CMC, and the underutilization of the tools and methodologies currently available in other scientific disciplines demand immediate action. In what follows, we will propose a number of methodological
strategies and theoretical principles that can be used for addressing the issues of operationalizing, measuring, and understanding entropy in online collaborative and communication environments.

To illustrate the practical implications of our argument we will discuss the collaborative diversity issues raised by one of the most intriguing technologies that have emerged in the field of online collaboration, Wikipedia. We will illustrate our main points about the utility and operationalization of entropy with a number of examples related to this online project. To illustrate our main points about using entropy in the study of online environments, we start by providing a brief overview of an emerging communication technology (Wikipedia) which can be used as a test bed for studying entropic processes. This will be followed by a discussion of some specific methodological and theoretical avenues related to diversity and entropy. We will conclude with a brief example of how these principles and methodologies can be used in practice for measuring collaborative diversity and entropy on Wikipedia and with the implications derived from this example for future research on online collaborative environments.

Wikipedia

One of the most significant technologies that have emerged in the last several years is Wikipedia, which is built around the very idea of equality and diversity of contribution. This is an online collaborative encyclopedia, created outside of traditional authorship, editorial and copyright constraints. An “open content” repository of encyclopedic knowledge, Wikipedia is designed from the ground up as a collective and distributed effort (Wikipedia, 2008a).

Akin with the open source software movement, described in the famous essay “The Cathedral and the Bazaar” (Raymond, 2001), Wikipedia relies, as its name suggests, on the wiki publishing paradigm (Leuf & Cunningham, 2001). Wikis are web-based collective and non-directed online publication systems shaped on the model of the Portland Pattern Repository.1) Created by Ward Cunningham in 1995, wiki systems use an open web-based editing interface, which allows any Internet visitor to the site to add, delete and/or publish content. The first wiki repository served as a historical record of computer programming ideas.
The wiki idea was transferred from computer programming to knowledge production in 2000 as an attempt to develop an encyclopedia (Nupedia) that would subsequently be submitted for peer review. Founded by Jim Wales, a commodity trader and pornography industry entrepreneur with a graduate degree in economics and philosophy, and by Larry Sanger, an academic philosopher, Wikipedia has known explosive growth over the past four years. As of October 2008 Wikipedia had about 2,600,000 articles and over 8,000,000 registered users for its English version alone (Wikipedia, 2008b).

Based on a simple web interface, Wikipedia allows any site visitor not only to read but, in case he or she disagrees or finds the content inaccurate or insufficient, to immediately alter or change the entries (Wikipedia, 2008c). The edits are subjected to only limited editorial gate-keeping. For most articles, a user needs to be registered for a number of days before her or his changes will be immediately accepted by the system. As for the editorial process itself, on each entry page there is an “edit” button; when pressed, this switches the page from a “display” to an “edit” mode. The reader can then make any changes immediately, and for most pages this means that alterations are recorded almost instantaneously. Although until recently all articles could be edited even without any registration, after a number of incidents, most notably that of John Seigenthaler, a former Robert Kennedy aide who was accused of participating in the conspiracy that killed his former employer (Seigenthaler, 2005), Wikipedia has enforced a “protection” policy (Hafner, 2006) for editing some high-profile pages (such as that dedicated to George W. Bush). This means that only administrators or selected users can edit “protected” articles. However, most articles are editable by most users after they register with the site and complete a minimum of editorial work.

The idea behind this process is to make any reader into a co-author of the Wikipedian project. Although Wikipedia utilizes a code of conduct, which specifies the manner in which the content should be changed (Wikipedia, 2008c), this is limited in scope and virtually unenforceable. Its main requirement is that the user abides by a “neutral point of view” writing policy (Wikipedia, 2008d). This means that the contributor should avoid making personal comments or being judgmental about the various perspectives that
might explain a specific topic. Disputations regarding the neutrality of a specific article are to be settled by arbitration, but this is a prolonged and generally avoided process.

More important than the code of conduct, however, is the expectation that even when partisan interests or inaccuracies seep into the read-write mechanism, the system itself provides for quick redress. It is assumed that by its very open nature, communal editing will ensure continuous vigilance and general objectivity. Wikipedia's editing process assumes that exposing an article to many users will result in accuracy. In theory, as soon as a biased contribution is posted, a thousand eyes will spot and correct it (Sanger, 2001).

The editorial mechanism described above relies to a significant degree on the expectation that the collaborative efforts will be diverse and egalitarian. The basic premise of a wiki system is that knowledge will be more abundant, reliable, and useful when it incorporates a large diversity of inputs and viewpoints from a large number of contributors. Trickling down, it is expected that the egalitarianism incorporated in the very method of publication, which does not require any kind of human editing and which allows any visitor to a wiki site to add to the common repository of information, will act as a motivator and will generate knowledge that will be at least equivalent, if not superior in quality and more reliable, to other types of knowledge.

Despite the importance of entropy and diversity to the Wikipedian model, no compelling explanatory models or even simple index measures have been proposed to measure degrees of participants’ diversity in terms of their contributions to Wikipedia. In what follows we will discuss and illustrate a number of theoretical and statistical tools for measuring the diversity and equality of participation in online collaborative environments in general and Wikipedia in particular.

Determinants of Diversity

Since the diversity measures we are introducing are relatively new to the discipline, at least in the way in which we propose their application, and one of the goals of the paper is to highlight their relevance for communication, we start by explaining their statistical characteristics, methodological advantages, and potential shortcomings. The discussion starts with the notion of diversity, which will then be related to that of entropy.
Diversity is the presence of variation in terms of qualities or attributes of interest within a given system (in our case online collaborative environment). The notion of diversity has been actively used to describe the structure of social or biological communities. For example, in the context of ecology or biology this concept is known as biodiversity, which is defined as a variety of life forms and is measured within a given ecological community (Maignan et al., 2003). In economics, various diversity measures are used to evaluate the structure of an industry or of a geographically situated industrial or business environment (Stigler, 1983). Applied to social contexts, the notion of diversity refers to the presence of a variety of opinions, cultures, ethnic groups, and socio-economic characteristics (Maignan et al., 2003; McDonald & Dimmick, 2003). Being so widely utilized, the operationalization of the concept tends to vary from discipline to discipline. For example, biodiversity in ecology includes diversity within species and among species, and comparative diversity among ecosystems. Furthermore, the definition of species is also dependent on research contexts.

To better understand the essence of the concept, we start with a broader, more abstract discussion of its characteristics.

Conceptual Underpinning

Suppose that we have online communication space \(O\), which has \(n\) number of opinions and \(m\) number of members.

\[ O = \{O_1, O_2, \ldots, O_n\} \]

Let \(C\) be a classification of \(O\). The opinions posted on the communication space can be classified by a certain criterion variable. For example, \(C\) might be a set of classification for each participant with number \(m\). Thus,

\[ C = \{C_1, C_2, \ldots, C_m\} \]

Assume \(C_i \cap C_j = 0\). So each opinion in \(O\) belongs to only one participant. Thus,
\[ \bigcup_{j=1}^{m} C_j = C = O. \]

\( S_i \) is the share (mathematical proportion) of the \( i^{th} \) individual in the opinion space \( O \).

\[ S_i = \frac{|C_i|}{\sum_{j=1}^{m} |C_j|}, \]

\[ \sum_{i=1}^{m} S_i = 1. \]

The question that we want to answer is how can we measure and quantify diversity of contribution/participation/involvement/presence in this situation? Suppose that in the online communication space \( O \) there is only one opinion, posted and one participant (say Tom).

\[ O = \{1\} \text{ and } P = \{\text{Tom}\} \]

, where \( P \) represents the set of participants in \( O \).

In this scenario, there is no uncertainty about who posted it. It is completely certain that only Tom posted his opinion in this communication space and the online communication space therefore has no diversity.

Suppose that there is another participant (say Sara) in \( O \). In other words,

\[ O = \{1\} \text{ and } P = \{\text{Tom, Sara}\}. \]

In this situation the contributions to the online site can be made either by Tom or Sara. In this scenario, a degree of uncertainty about contributions necessarily occurs. From the perspective of information theory (Shannon & Weaver, 1949), it is said that this sort of a question having two possible answers (Tom or Sara) carries 1 bit of information. If we had \( m \) participants in a communication space, the question on who posted would have \( m \) possible outcomes, thus carries \( \log_2 m \) bit of information (Cover & Thomas, 2006). To simplify, this measure tells us a rather trivial fact: as more people participate in an online community, the social diversity tends to increase.
The value of measuring diversity with a mathematical formula becomes clear with the situation in which there are many members and many opinions in a communication space. For example:

\[ O = \{ *_{1}, *_{2}, *_{3}, \#_{4} \} \quad \text{and} \quad P = \{ \text{Tom, Sara} \}. \]  

(1)

\[ C_{\text{Tom}} = \{ *_{1}, *_{2}, *_{3} \}, \]

\[ C_{\text{Sara}} = \{ \#_{4} \}. \]

In this example (1), where star (*) and sharp (#) notations represent opinions, Tom posted three opinions and Sara one. Tom’s contribution accounts for 75% and Sara’s 25% of the total opinions. Extending the example, we can consider more diverse communication spaces with equal contributions by members.

\[ O = \{ \Delta_{1}, \Phi_{2}, \Omega_{3}, \Psi_{4} \} \quad \text{and} \]

\[ P = \{ \text{Tom, Sara, Kati, John} \}. \]  

(2)

\[ C_{\text{Tom}} = \{ \Delta_{1} \}, \quad C_{\text{Sara}} = \{ \Phi_{2} \}, \quad C_{\text{Kati}} = \{ \Omega_{3} \}, \quad C_{\text{John}} = \{ \Psi_{4} \}. \]

The equal amount of contribution by participants, \( \frac{1}{n} \times 100\% = 25\% \), strongly implies that the level of diversity of the communication space is higher than the previous communication space in the example (1). A uniform distribution of contribution by members leads to the highest diversity of a communication space.

Social entropy as measure of diversity

In sum, the diversity of opinion in communication spaces is a function of the number of participants \( m \) and the shares of participants \( S_{i} \). More participants means more diverse participation – and the more uniformly distributed the contributions by members of a community imply more diverse participation. In this respect we are dealing with a higher level of uncertainty and “disorder” which can be translated conceptually as a higher level of “entropy.” How can we translate this into a synthetic indicator?

We can do it, as Shannon and Weaver suggested, by measuring the relative degree of disorganization found in any system. Disorganization can be thought of as the random mixing of various elements, whose
relative presence should thus be equal. In this situation we can also say that the diversity of the system is at a maximum, since all elements are equally (randomly) present.

Shannon’s entropy index takes a value of 0 when there is absolute order in the system (one element is prevalent at the expense of all others) and a maximum value (which varies from system to system), when there is perfect disorder and diversity (all elements are equally present). Entropy is a synthetic measure that tells us at a glance how well represented are the different components of a social or communicative space.

Mathematically, the entropy of a random variable $X$ (in this case, the level of contribution) has a probability mass function $p(x)$ and is defined as follows:

$$H(X) = -\sum_{i=1}^{m} p(x) \log_2 p(x).$$

The entropy varies from zero to $\log_2 m$, as previously explained.

How do we apply this measure to online collaboration environments? Consider an online communication space in which there is a uniform distribution of contributions by four members, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The entropy of this communication space is

$$H(X) = -\sum_{i=1}^{4} S_i \log_2 S_i = -\sum_{i=1}^{4} \frac{1}{4} \log_2 \frac{1}{4} = \log_2 4 = 2.$$

Now, consider another communication space with four members. Assume that the shares of contribution by these members are unequally distributed, $(\frac{3}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$. The entropy of this communication space is

$$H(X) = -\sum_{i=1}^{4} S_i \log_2 S_i = -\frac{1}{2} \log_2 \frac{3}{10} - \frac{3}{10} \log_2 \frac{3}{10} - 2 \cdot \frac{1}{10} \log_2 \frac{1}{10} = 1.69.$$

The entropy of the former communication space with a uniform distribution of contributions is higher than the latter one with unequally distributed contributions.

Normalized social entropy as a diversity/evenness measure
Although entropy is an elegant modality to measure diversity in a system, there are some potential limitations. Entropy reflects not just one, but two system dimensions: richness and evenness. When we collapse them into one index score, there is a loss of information (Balch, 2000). Moreover, the two dimensions can contribute in different ways to entropy scores that are very similar, which can lead to all sorts of confusion.

For example, two very different online communication groups in terms of composition and contributions can have entropy characteristics that seem to be very similar (Balch, 2000). For example, communication space \((C_1)\) has four opinions, expressed by two participants. The shares of two the participants are equal, \((\frac{1}{2}, \frac{1}{2})\). The second communication space \((C_2)\) has 64 opinions, with seven participants. The shares of the seven participants are unequally distributed, \((\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})\). However, the calculation of entropy provides a counterintuitive result: while the entropy of the first communication space \((C_1)\) is 1, the entropy of the second communication space \((C_2)\) is 2, despite the fact that in \((C_2)\) opinions are less evenly distributed compared to \((C_1)\). This is because \((C_2)\) has more participants. The fact that their contributions are unequally distributed is hidden.

This problem can be solved by normalizing the entropy values. This enables us to compare the evenness of two communication spaces, including over time, by controlling for the number of elements that compose each of them. Normalization can be obtained by dividing the raw entropy score by its maximum \(\log(m)\), which limits its range from 0 to 1.

\[
H_o = \frac{H}{H_{\text{max}}}, \quad 0 \leq H_o \leq 1, \text{ where } H_{\text{max}} = \log_2 m.
\]

Normalized entropy is particularly useful for handling the “lurker” problem in studying diversity in online environments. Lurkers are users who do not make any contributions to an environment; he or she is
just an observer. Lurkers can make an environment potentially richer, but they can also impact diversity. How can we capture both of these aspects of the lurker behavior?

Suppose that there are two communication spaces. In both, the contributing members make equal contributions:

\[ C_1 = \{\Delta, \Omega\}, P_1 = \{Tom, Jane\}, \text{ and the share distribution is } \left(\frac{1}{2}, \frac{1}{2}\right). \]

\[ C_2 = \{\Delta, \Omega\}, P_2 = \{Tom, Jane, Sara\}, \text{ and the share distribution is } \left(\frac{1}{2}, \frac{1}{2}, 0\right). \]

In the second communication space, however, is lurker Sara, who did not contribute to the interaction. Despite this important difference, the non-normalized entropy of the two communication environments is the same, 1.

\[
H(X_{c_2}) = -\sum_{i=1}^{3} S_i \log_2 S_i = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} - 0 \log_2 0 = 1.
\]

Normalizing the entropy values highlights the presence of the lurker in one of the spaces. For example, the max \( \log(m) \) entropy value for \( C_1 \) is

\[
H_{\text{max}}(X_{c_1}) = \log_2 2 = 1.
\]

while its normalized value will be

\[
H_o(X_{c_1}) = \frac{H}{H_{\text{max}}} = \frac{1}{1} = 1.
\]

For \( C_2 \), the maximum entropy value will be: \( H_{\text{max}}(X_{c_2}) = \log_2 3 = 1.58 \)

Thus, the normalized social entropy of \( C_2 \) is \( H_o(X_{c_2}) = \frac{H}{H_{\text{max}}} = \frac{1}{1.58} \approx 0.63 \)

Comparing the normalized entropy of two communication spaces shows that the first communication space is more diverse than the second one, because the second formula takes into account the presence of the lurker.

Social entropy can be used to measure changes in terms of ecological diversity of contributions to Wikipedia articles (entries). The question that can be asked and positively answered is: “Are contributions to Wikipedia equally distributed among the users or not?” This allows making inferences about the diversity of content, equality of contributions, and, when tracked over time, potential system-level diversity changes. The example below illustrates how entropy can be used to determine the evolution of collaborative diversity in specific Wikipedia articles.

We selected the entry on the Indian film Naina (Wikipedia, 2008e), a relatively short and simple article. Written in May, 2005, by February, 2006 it had been edited five times. In May 16, 2005, Hemanshu created the article on the film Naina using 16 words: “Naina is a Hindi movie to be released in India in 2005. It stars Urmila Matondkar.” In July 4, 2005, an anonymous user (A1) added 15 words at the end of Hemanshu’s original entry. “It’s genre is horror. It is having great similarities over the English film The Eye.” In December 24, 2005, a second anonymous user (A2) contributed 24 more words to the article. In December 30, a third anonymous user (A3) proofread and edited the article, deleting 21 words and adding 10 of his or her own (in bold).

\[
\text{Naina is a Hindi movie released in India in 2005. It stars Urmila Matondkar. It’s genre is horror. It has many similarities with the English film The Eye. It’s release created controversy in India because of the fact that the lady had eye transplant before experiencing extra sensory perceptions in the film, and that discouraged many people from receiving eye transplants.}
\]

In January and February 2006, NilsB and DomLachosicz added external links, deleted 13 words and added 12 new words. Figure 1 visualizes the changing pattern of relative contributions by participant and Table 1 provides the descriptive and entropy statistics associated with the changes. As shown in Table 1, the length of the article has increased overtime, from 16 words to 69 words. As a consequence of the increase in textual contributions and number of contributors, the social entropy of the article has also been increasing over time. Figure 2 shows that social entropy increased from 0 (no diversity) at the initial point in time to 2.53 in the
most recent version, which demonstrates the increasing trend of diversity of participation for the article. Yet, significantly, normalized entropy has fluctuated over time (see Figure 3). To review, normalized entropy represents equality or evenness of participation in the collaborating environment when the number of members remains constant. The second version of the article has .998 of normalized entropy, which is very close to the maximum value, 1. This means that Hermanshu and the first anonymous user (A1) contributed to almost the same degree. In the third version of the article, the normalized entropy drops slightly because of the fact that the second anonymous user (A2) added a relatively large amount of textual information (44% of total words). Finally, the normalized entropy tends to increase as other new contributors (A3, NilsB, and DomLach) participated in this collaborative writing process. Even after controlling the number of contributors, we found that the relative share of textual contribution tends to be balanced.

What we notice from this simple example is that we can synthetically characterize the over-time evolution of the article, measuring its diversity both in absolute terms and normalized. Comparing the two versions of entropy (raw and normalized) we noticed that while the first one captures only the constant increase, the second was capable of indicating a subtle fluctuation. Moreover, if we compare figure 2 with figure 3 we notice that the slope for the increase in raw entropy is far more abrupt than the one for normalized entropy. The fluctuation and the slope difference can even help us formulate a tentative hypothesis. We can speculate that the effect size on diversity of newly added information will diminish as the total amount of textual information increases. Consequently, the normalized social entropy may tend to level off as the article evolves over time.

Discussion and Conclusions

Social entropy is a very promising way of measuring group dynamics, especially in terms of measuring contributions to online communication. This is relevant not only to wiki articles, but also to other types of virtual communities: social networks, newsgroups, email lists, blogs, etc. A main impact of this measure would be to rekindle interest in system-level processes. Social entropy enables us to characterize the status of a system with respect to its level of diversity and equality in a very parsimonious manner. Questions such as: does this communication system contribute to an equal or diverse communication process
will in the future be directly addressable – and measurable. Of course, the assumption that open, networked systems would automatically engender egalitarian environments has likely been relatively naïve. As online communication systems have matured, we already suspect that these environments are often dominated by the few at the expense of the many.

However, detecting diversity and equality of contributions (or the lack of same) should not become an end in itself but a means toward better understanding social and communicative systems. In fact, although seeming directly intuitive, equality and diversity of system-level processes are not always one-and-the-same as their normative counterparts. The claim that open communication systems should be diverse or tend toward diversity and equality, that is that their level of entropy should increase, should be carefully investigated using theoretically-grounded hypotheses.

Such hypotheses should take into account the fact that diversity and equality, although desirable societal and moral goals, might not always be present or even functionally desirable in the structure of a communicative group. Processes related to group structure and functional differentiation of the members, creation and maintenance of group communicative standards, preservation of a given meaning over time and prevention of ambiguity might lead to cycles in the life of a group. As groups gel and functional hierarchies are formed, entropy will not be maximized, but rather reduced or maintained at an optimal level. From this perspective, one of the most important theoretical contributions of research conducted from a social entropy perspective would be to find what entropy levels are optimal for any given online communicative and collaborative system. Furthermore, which entropy level corresponds to what level of functional differentiation and hierarchical organization?

What makes the social entropy index so enticing is that it gives researchers a gauge to directly measure and assess these levels of collaboration and their significance. But the theoretical and practical ramifications of such a measure spread much farther and wider. For example, teachers and managers alike could simply utilize the entropy index as a tool in manipulating individual contributions to group assignments and team projects to facilitate optimal collaboration levels and, therefore, favorable outcomes once desirable levels of diversity of contribution have been determined.
We hope that these research avenues, along with their broad implications and promise, will offer the community of communication scholars a number of topics for consideration which will lead to important future projects and studies.
References


Footnotes

1 Wiki is the Hawaiian word for quick, implying the speed with which changes can be operated (Wikipedia, 2006).

2 Since this present study employs logarithms to base 2, the entropy will then be measured in bits. When we use 2 for the base of the logarithm, there is an advantage since the measured entropy can be said to be in bits. Other values such as e or 10 can be used for the base. The choice of the base is relatively arbitrary (Lemay, 1999)

3 The number of characters can be also used as a measure of information contributed by participants of Wikipedia.

4 In Wikipedia, registered users have their own screen name and it appears in the history page for every article. Peripheral users with no membership have no screen name. However, it is identifiable who wrote what because of the fact that the wiki makes anonymous users’ IP address visible.

5 Measuring text is straightforward. In contrast, it is challenging how to count and measure visual content such as pictures, photographs, hyperlinks. It calls for further scholarly attention and discussion.
Table 1.
Social Entropy of the Wikipedia Article on Naina

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Words (%)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>16 (100)</td>
<td>31 (100)</td>
<td>55 (100)</td>
<td>61 (100)</td>
<td>69 (100)</td>
</tr>
<tr>
<td>Hemanshu</td>
<td>16 (100)</td>
<td>16 (52)</td>
<td>16 (29)</td>
<td>14 (23)</td>
<td>14 (20)</td>
</tr>
<tr>
<td>A1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>__</td>
<td>15 (48)</td>
<td>15 (27)</td>
<td>11 (18)</td>
<td>8 (12)</td>
</tr>
<tr>
<td>A2</td>
<td>__</td>
<td>__</td>
<td>24 (44)</td>
<td>16 (26)</td>
<td>14 (20)</td>
</tr>
<tr>
<td>A3</td>
<td>__</td>
<td>__</td>
<td>__</td>
<td>20 (33)</td>
<td>15 (22)</td>
</tr>
<tr>
<td>NilsB</td>
<td>__</td>
<td>__</td>
<td>__</td>
<td>__</td>
<td>7 (10)</td>
</tr>
<tr>
<td>DomLach</td>
<td>__</td>
<td>__</td>
<td>__</td>
<td>__</td>
<td>11 (16)</td>
</tr>
<tr>
<td>Social Entropy&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0</td>
<td>.99</td>
<td>1.55</td>
<td>1.96</td>
<td>2.53</td>
</tr>
<tr>
<td># Participants&lt;sup&gt;m&lt;/sup&gt;</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Max. Entropy&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0</td>
<td>1.00</td>
<td>1.58</td>
<td>2.00</td>
<td>2.58</td>
</tr>
<tr>
<td>Normalized H&lt;sup&gt;e&lt;/sup&gt;</td>
<td>__</td>
<td>.998</td>
<td>.980</td>
<td>.982</td>
<td>.986</td>
</tr>
</tbody>
</table>

a. The percentile represents relative shares (Si) of each contributor in textual content.

b. A1, A2 and A3 represent anonymous contributors respectively with no screen name.

c. Social entropy (H) can be calculated with the formula as follows: \(-\sum_{i=1}^{m} S_i \log_2 S_i\)

d. Maximum entropy (H<sub>max</sub>) can be calculated with the formula as follows: \(\log_2 m\).

e. Normalized social entropy can be calculated with the formula as follows: \(H_{\max} / H\).
Figure 1.

Trend of Contributions over Time

$^a$ A1, A2 and A3 represent anonymous contributors respectively with no screen name.
Figure 2.

Trend of Social Entropy

\[ a \] $V_i$ represents each version of the article on Naina at the point of time, $i$. 
Figure 3.

Trend of Normalized Social Entropy

$V_i$ represents each version of the article on Naina at the point of time, $i$. 