Connexions: a Social and Successful Anomaly among Learning Object Repositories

Xavier Ochoa
Centro de Tecnologías de Información, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
Email: xavier@cti.espol.edu.ec

Abstract— **Connexions is a Learning Object Repository that has gained notoriety as a successful example of collaborative creation of learning materials. In a previous quantitative study about the characteristics of learning material repositories, Connexions presented some anomalies that prevented it to be classified together with other Learning Object Repositories. This paper, working with updated data, finds that those anomalies are not errors of the previous study, but that the anomalies have increased and are more strongly expressed. Moreover, those anomalies, namely the exponential growth in the number of contributors and content, seem to be the reason behind the success of Connexions. It is concluded in this work that Connexions strategy of "release early, release often", together with the openness of its core community, that welcome and support newcomers and sporadic contributors, offer a reasonable explanation for the exponential growth trends found in Connection, that can be classified now as a Social Learning Object Repository.**

Index Terms— **Learning Object Repositories, Social Networks, Connexions, Social LOR**

I. Introduction

Connexions is a Learning Object Repository (LOR) created at Rice University in 1999 as an alternative way to create, share, maintain, use and reuse content [1], [2]. The primary goals of Connexions are to interconnect learning content across disciplines, courses and curricula and to create a collaborative environment where learning communities could share knowledge that would lead to the creation and improvement of materials [3]. One of the major differences between Connexions and other, more traditional LORs, is that the process of content development could be opened for collaboration between interested individuals worldwide.

In a previous work of the author of this paper [4], several systems for learning content publication were quantitatively examined in order to find common characteristics and behavioral patterns. The main conclusion of this previous paper is that there are different kinds of these systems where the characteristics are clearly different between Learning Object Repositories and Referatories (LOR), Open Courseware initiatives and Learning Management Systems (OCW and LMS) and Institutional Repositories for research papers (IR). While most of the systems, due to their shared characteristics, aligned quite clearly in one of those categories, Connexions presented some anomalies. Some of Connexions characteristics put it, as expected, inside the LOR group. However, one very important variable, the growth of the contributor base presented an exponential growth, while in all the other LORs, this growth was linear. The difference was so clear that it could not be assigned to biases in in the measurements or in the calculations. The only logical conclusion, implied in the discussion section, is that Connexions, due to its unique measured behavior, is a new type of repository. Given that the main difference of Connexions with more traditional LORs is the social interaction for the creation of materials, the name Social LOR is proposed for this new category.

The first objective of this work is to test if the anomalies detected in [4] are still measurable and if new anomalies could be found. To conduct this study an updated set of data is collected from Connexions and other representative LORs. This data is taken two years after the collection made for the original study. Several characteristics will be measured and analyzed using the same procedures used in the original work.

The second objective is to provide a in-depth quantitative analysis of the particular characteristics of Connexions and its community that could provide some explanation for the observed anomalies and could justify the creation of a new category of repository where Connexions would belong. First, the processes of publication and reuse will be analyzed and the obtained variables correlated. After, the social network that emerge from the collaboration to create materials will be analyzed and its characteristics also contrasted with the previous results.

This paper is structured as follows: Section II provides some background for this study, presenting related work made to understand Connexions and other Social LORs. Section III explains how the data was collected and prepared for the analyses. Section IV compares Connexions with more traditional LORs in order to establish if the anomalies are still present in the current version of the repository. Section V analyzes the particular characteristics of the production and consumption process in Connexions and provide reasonable explanation for two of the observed anomalies. Section VI analyzes the social network of Connexion contributors, the result of the analyze offer an explanation for the third observed anomaly. The paper finishes with a general discussion about the findings of the analyses and further research that should be done to obtain a better understanding of Social LORs.

doi:10.4304/jetwi.2.1.11-22

II. Related Work

While the qualitative and quantitive analysis of Learning Object Repositories is a not well researched area in the field of Technology Enhanced Learning, there are some works that explore the factors that contribute to Connexions and other Social LORs success. The most detailed description of the desirable characteristics of Connexions is presented by Dholakia et al. in [5]. They present Connexions as a exemplary case of a sustainable open educational program. The most important aspect of this work is the description of the characteristics that the authors considered are key factors behind the success of Connexions: 1) Increase the equity of the Connexions brand, 2) High quality, ample, modular, continually updated, personalized on assembly, published on demand content, 3) An engaged and involved user community and 4) Site usability. As will be presented in the following sections, the present work provides evidence to support factors number two and three.

In a more general approach, Monge et al. in [6], analyze what they considered are the characteristics of the so called "Repository 2.0". They perform a Qualitative study of repositories that have a community aspect as part of its constitution (Connexions among them). From the analysis of the characteristics of these repositories and inspired by Web 2.0 technologies, they recommend several strategies to create a community backed repository: 1) Clear authorship and use license attribution, 2) Rapid content creation, 3) Indexable content for search engines, 4) Social tagging, 5) Reputation systems for contents and 6) Social recommendation systems. The present work finds evidence to back strategy number two and five.

Petrides et al. in [7], present the first quantitive analysis of the production and reuse of content in Connexions. While it is not presented as conclusion of that work their results provide indication of exponential growth in the number of modules, as well as the inequality in the production of content. The main focus of their research is the analysis of the reuse behavior of the users based on the commentaries made when a new version of the module is published. The authors complement the quantitive study with a series of interviews made to selected contributors.

The present work will provide a more in-detail quantitative analysis of the characteristics of the Connexions repository than what have been provided before. This work will also compare those characteristics to those found in traditional LORs in order to detect anomalies and specific behavior that belongs only to Connexions.

III. Data Collection

In order to analyze Connexions characteristics, raw data was obtained for each one of the modules that were published in the repository until December 1st., 2009. Web scrapping was used to obtain the list of all the modules published from the search facility provided by the Connexions site. For each of the 15,504 published modules, two pages were retrieved: one containing the object itself, and another containing the detailed metadata about its authors and version history. Regular expressions were then used to extract the information form the downloaded HTML pages. Finally the extracted data was stored in text files for their processing in the analysis tools. The code used for the data collection, as well as the extracted data, can be downloaded from the author web page [8].

To redo the comparison between Connexions with other LORs, data was also collected from two well known repositories: Ariadne [9] and Merlot [10]. This two were selected because their are the most representative traditional Learning Object Repository (Ariadne) and Learning Object Referatory (Merlot). The information already collected for [4] was updated to December 1st. 2009, if possible. In the case of Ariadne, there has not been direct additions to the repository since January, 2008. Given the access that the author has to the Ariadne LOR, the data in this case was extracted directly from the core repository database. This collection only represents the objects that have been directly uploaded to Ariadne and not harvested from any other source [11]. From Ariadne, information about 5,112 learning objects was obtained. For Merlot, web scrapping and regular expressions were used, as in the case of Connexions. The information of a total of 21,520 objects were obtained from Merlot. The code used to extract the data from Ariadne and Merlot, as well as the data itself, are also available at [8].

IV. Connexions versus Traditional LORs

As mentioned in the Introduction, several measured characteristics of Connexions seemed to depart from the normal behavior of Learning Object Repositories and Referatories. Most prominently, while all the rest of repositories and referatories presented a linear growth in the number of authors, Connexions had a clearly exponential growth. Other characteristics, such as content growth and lifetime presented borderline distributions, but not clearly enough to conclude that Connexions was behaving differently.

This section will perform an updated comparison between Connexions and traditional LORs (Ariadne and Merlot), with data collected 2 years after the of the original data [4]. This comparison will include three key characteristics: the growth of the contributor base, the growth of the published content and the distribution of the lifetime between contributors. The objective of this updated comparison is to determine if the diverging behavior of Connexions was just an artifact from the measurement and analysis or if it has consolidated and increased with time.

A. Contributor Growth

The first step in the analysis of the growth of the contributor base was to determine the start date of each contributor. This was done searching for the earliest day in which the contributor became active in the repository. This is the day in which the contributor published its first object, not the day in which the contributor created his or her account. Afterwards, the list of contributor was sorted in descending order according to its star date and the cumulative sum is obtained for each day. The resulting growth function for the three repositories can be seen in Figure 1.

Once the empirical Contributor Growth function is obtained, it was fitted with several models: linear ($at+b$), bi-phase linear with breakpoint (a_1t for t <Breakpoint and $a_2t + b_2$ for $t \geq$ Breakpoint), bi-phase linear with smooth transition (ln(a*exp(bx)+c), exponential ($b*e^{at}$), logarithmic (b*ln(at)) and potential ($b * t^a$). This models were the same used during the first comparison in [4] and were selected because they could be similar to the shape of the empirical growth function. The fitting was performed using Generalized Linear Models with Least-Squares Estimation. The selection of the model was based on the Akaike information criterion (AIC) [12], that not only takes into account the estimation power of the model, but also its simplicity (less estimated parameters). The results of the fitting are presented in Table I. The code to perform these calculations over the data in R and MATLAB software can be downloaded from [8].

The results show that the exponential function is still the best fitting alternative for the growth of the contributor base of Connexions. Also, the fitted exponential rate of growth (λ) have not change significantly (less than 9%) from the original 1.2 x 10-3 to the current 1.1 x 10-3. Similarly, for Ariadne and Merlot the best fitting function is still the bi-phase linear growth. That means that their contributor base is in a mature phase, growing linearly each day. For Ariadne, their Mature Growth Rate (MGR) has remained in the vicinity to the what has been found in 2007 (changed less than 20%) . For Merlot, the MGR has increased significantly (from 0.54 to 0.95), but the growth is still visible linear over the last years. These results confirm what was found in [4]: Connexions attracts users in a fundamentally different way that traditional LORs.

B. Content Growth

To obtain the Content Growth empirical function, the date of the first publication for each object was obtained. In the case of Connexions, the date of creation of each module is mentioned in the metadata file associated with each module (version 1.1). The list of dates is then sorted in descending order and the cumulative sum of objects is obtained for each day. The resulting function is presented in Figure 2. The same six models and methodologies used to fit the Contributor Growth were fitted to the empirical data from the Connexions, Ariadne and Merlot repositories. Again, the AIC was used to select the best fitting model. The results are shown in Table II. The R and MATLAB code to perform these calculations can be downloaded from [8].

The most important result obtained from data is that now the exponential model seems to be a better fit for the Content Growth of Connexions. As it can be seen in Figure 2 (Connexions), the start of the curve can be fitted with a line (as it was done in the previous analysis), but

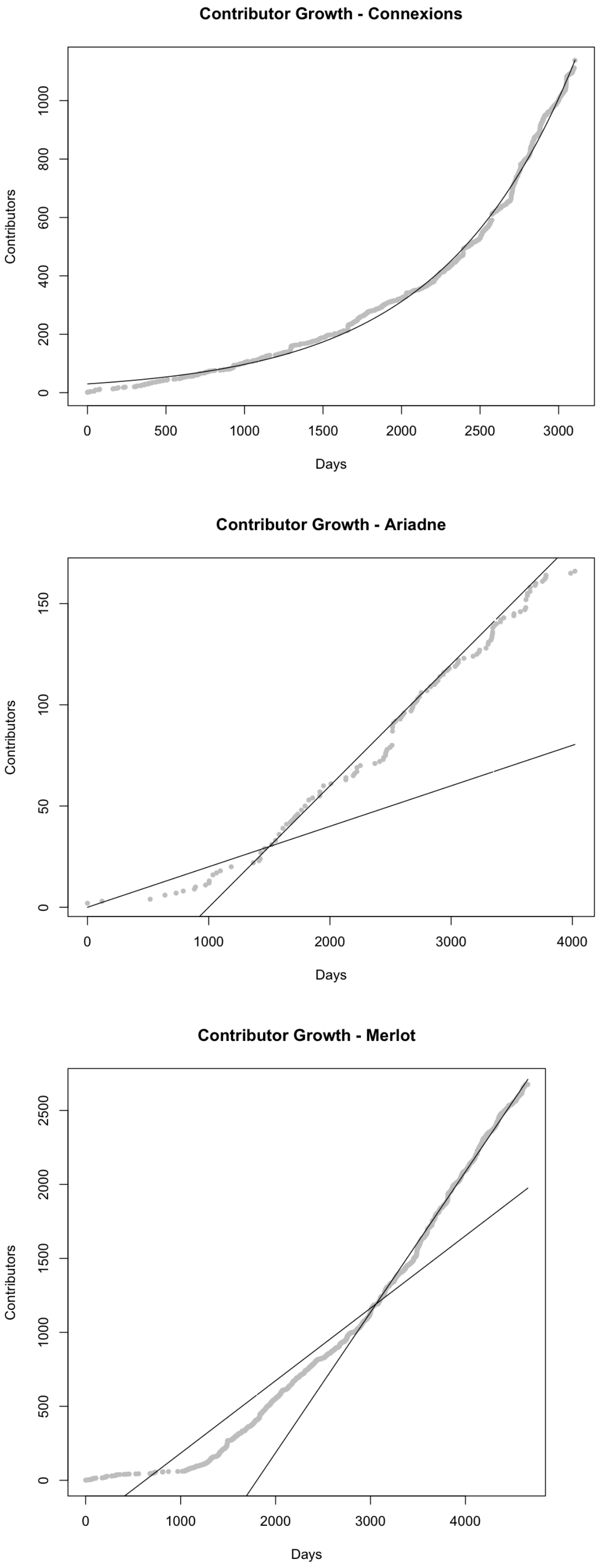

Figure 1. Contributor Growth Function for Connexions, Ariadne and Merlot

TABLE I.
RESULT OF THE ANALYSIS OF THE CONTRIBUTOR GROWTH. (IGR=INITIAL GROWTH RATE, MGR=MATURE GROWTH RATE)

Repository	**Previous Results - 2007**		**Current Results - 2009**	
	Function	Parameters	Function	Parameters
Connexions	Exponential	λ=1.2x10-3	Exponential	λ=1.1x10-3
Ariadne	Bi-Phase Linear	IGR=0.02, MGR=0.06, BP=3.5 y.	Bi-Phase Linear	IGR=0.02, MGR=0.05, BP=3.5 y.
Merlot	Bi-Phase Linear	IGR=0.12, MGR=0.54, BP=1.1 y.	Bi-Phase Linear	IGR=0.43, MGR=0.95, BP=7.4 y.

the last part of the curve (after day 2000) is distinctively exponential. This finding is very relevant given that one of the most discouraging findings about traditional LORs is that they grow linearly with time [4]. Corroborating this conclusion, the model that best fit Ariadne and Merlot Content Growth is still the Bi-Phase linear.

The main implication of exponential growth is that Connexions has the potential, in the 2 or 3 years, to become the largest non-federated repository. This conclusion is obvious when the relative size are compared. For example, in the quantitative analysis done in 2007, Ariadne had around 4,900 objects; Merlot, 18,000; and Connexions, 5,000. After two years, Ariadne has 5,100; Merlot, 21,500; and Connexions, 15,500. If the trends continue, by 2011, Ariadne will have 5,300, Merlot, 25,000 and Connexions, 37,500. By 2015, Connexions will be bigger that the biggest referatory, Intute (www.intute.ac.uk), with more than 270,000 objects.

C. Lifetime Distribution

The final characteristic to compare is Lifetime Distribution. The lifetime of a contributor is defined as the time from its first to its last publication or edition in the repository. The lifetime can be considered as the period during which the contributor is active in the repository. The lifetime is calculated subtracting the date of the last publication, from the date of the first publication. However, while the start of the lifetime is always known, the end of the lifetime is not always accurate. A contributor could have published its first object two years ago and its last object one year ago. The measured lifetime will be 1 year. However, if the contributor published one object more the day after the data was collected, its actual lifetime will be 2 years. To cope with this limitation, the lifetime of a user is only considered finished if the time from the last object insertion is at least as long as the longest period without activity between two consecutive publications. If a lifetime is not ended, it will be assigned the time interval from the first object insertion until the date of data collection, biasing the lifetime to shorter values.

The lifetime of different contributors vary widely. If the distribution of the length of the lifetime is plotted, a L-shaped curve is obtained, tale-telling signature of the heavy-tailed distributions. For such distributions the concept of mean and standard deviation does not have the same meaning that for normal related distributions. For example, most of the contributors will have a very short lifetime (few days), while very few will have a very long lifetime (few years). To establish the nature of the lifetime distribution of the different repositories, the data is fitted with known heavy tailed distributions: Lotka (inverse power law), Lotka with exponential cut-off, Exponential, Log-Normal, Weibull and Yule. These distributions were selected because they have high skewness to the left and are commonly present in other Information Production Processes [13]. The Maximum Likelihood Estimation (MLE) method [14] was used to obtain the distribution parameters. To find the best-fitting distribution, the Vuong test [15] is applied on the competing distributions. When the Vuong test is not statistically significant between two distributions, the distribution with less parameters is selected. This methodology is recommended [16] to select among heavy tailed models instead of the more common Least-Squares Estimation and R^2 values used for Generalized Linear Models. The results of the model selection can be seen in Table III.

The updated results show a change in the best fitting distribution for the lifetime of the Connexions contributors. in the previous study [4], all LORs were best fitted by the exponential distribution, meaning contributors loss interest in publishing material in the repository with time. In other words, the novelty of the publication fade faster with each passing day. In the current study, the lifetime of Connexions contributors seems to be best fitted by Lotka distribution with exponential cut-off. This distribution is very similar to the Pareto distribution [17] that govern several unequal processes, for example the distribution of wealth among society. The majority have little, a very small minority have a lot. Indeed, in Connexions, more than 70% of the contributors that have at least 2 publications have a lifetime shorter than 10 days. However, a not insignificant 10% of the contributor population have a lifetime longer than a year. In other words, the majority of the contributors seems to be active for a short period of time (maybe an initial publishing followed by small edits) while a small but very influential majority seems to have a long term relationship with the repository. Further research is needed to understand the reasons behind the Lotka distribution of the lifetime of Connexions contributors.

D. Comparison Conclusions

The main conclusion from the previous analysis is that Connexions behaves differently than traditional LORs. Several key characteristics of the repository deviate, as hypothesized, from the ones found in several other repositories and referatories:

- In Connexions, the number of contributors grows exponentially, while in other LORs grows linearly;

TABLE II.
RESULT OF THE ANALYSIS OF THE CONTENT GROWTH. (IGR=INITIAL GROWTH RATE, MGR=MATURE GROWTH RATE)

Repository	Previous Results - 2007		Current Results - 2009	
	Function	Parameters	Function	Parameters
Connexions	Bi-Phase Linear	IGR=0.8, MGR=2.19, BP=2.8 y.	Exponential	λ=2.7x10-3
Ariadne	Bi-Phase Linear	IGR=2.9, MGR=0.66, BP=1.0 y.	Bi-Phase Linear	IGR=3.0, MGR=0.52, BP=0.9 y.
Merlot	Bi-Phase Linear	IGR=0.9, MGR=5.8, BP=2.8 y.	Bi-Phase Linear	IGR=3.6, MGR=6.7, BP=7.1 y.

TABLE III.
RESULT OF THE LIFETIME DISTRIBUTION.

Repository	Previous Results - 2007		Current Results - 2009	
	Function	Parameters	Function	Parameters
Connexions	Exponential	λ=1.2x10-3	Lotka with exp. cut-off	α=0.84, λ=0.7x10-3
Ariadne	Exponential	λ=1.0x10-3	Exponential	λ=1.2x10-3
Merlot	Exponential	λ=1.5x10-3	Exponential	λ=2.5x10-3

- In Connexions, the number of modules grows exponentially, while in other LORs the number of objects grows linearly;
- in Connexions, the time that the contributor stays active in the repository follows a Lotka distribution, while in other LORs it follows an exponential distribution.

The first two differences have a clear positive effect in Connexions. Connexions is growing faster in number of contributor and objects that any other learning object repository or referatory. If the trends continue, Connexions will be by far the largest non-federated repository.

The effect of the third difference is more subtle. While it could be desirable that the majority of contributors stay active longer in the repository (for example in Weibull distributions), the majority of the contributors in Connexions stay active for a short period of time (10 days or less). However, the Lotka distribution also implies that a small but faithful group of contributors are ”always” active, contributing to the repository during long periods of time. The effect of these faithful contributors will be analyzed in the following sections.

While these analyses assert that Connexions is behaving better than traditional LORs, they do not explain why. The hypothesis of this work is that the social aspects of Connexions are what give it an advantage over other LORs. This hypothesis, however, cannot be proven or disproven by the previous analysis. The following sections will conduct analyses of intrinsic characteristics of Connexions behavior and the social network that is formed around the published objects in order to gain insight about the reasons behind Connexions success.

V. ANALYSIS OF CONNEXIONS CHARACTERISTICS

In order to understand the publication process in Connexions, this section will analyze several characteristics that are key features of Connections. For example, Connexions enable the collaboration between several authors to create and improve a module. Each new revision made to module is stored as a version. Connexions also publishes the number of times that objects are accessed through its web interface and also when the module has been included in a Connexions course. Finally, Connexions also allows its users to rate the content. These characteristics will be quantitatively analyzed and correlated in order to find explanations for the apparent successful behavior of Connexions as a LOR.

A. Content Creation

The main difference between Connexions and traditional LORs is that it provides a mechanism similar to a wiki to enable the collaborative creation of learning materials. Any registered user could create a module. That user could decide to share the creation/edition of the content with other users through a shared environment. Other users could ask the owner of content for authorization to access the shared environment to edit the content. Each published edit to the content generate a new version. The following subsections analyze three characteristics of the content creation process:

1) Contributors per Module: The Contributor per Module measure how collaborative is the process of content creation. Connexion list all the users that have made changes to each published module. Using this list, the number of contributors for each of the 15,504 modules was obtained. The analysis of these data shows that most of the modules (80%) only have one contributor. This means that the collaborative creation of content is not always exploited. The individual creation of material seems to be still the driving force behind Connexions. However, if the empirical distribution of the number of contributors per module is plotted with log-log axis, it is clear that the distribution is heavy tailed. Most of the objects have just one contributor, but few modules (3%) have 4 to 7 contributors (Figure 3). Using the same methodology that in subsection IV.C, the six mentioned heavy tailed distribution were fitted to the data. The best fitting distribution was Lotka with exponential cut-off (α=1.86, λ=0.49). The finding of a inverse power law is consistent with findings in other types of collaborative creation. For example, Voss [18] also found an inverse power law distribution in the number of authors of Wikipedia articles.

2) Versions per Module: The versions per module provide an idea of how “alive” is the content. That means

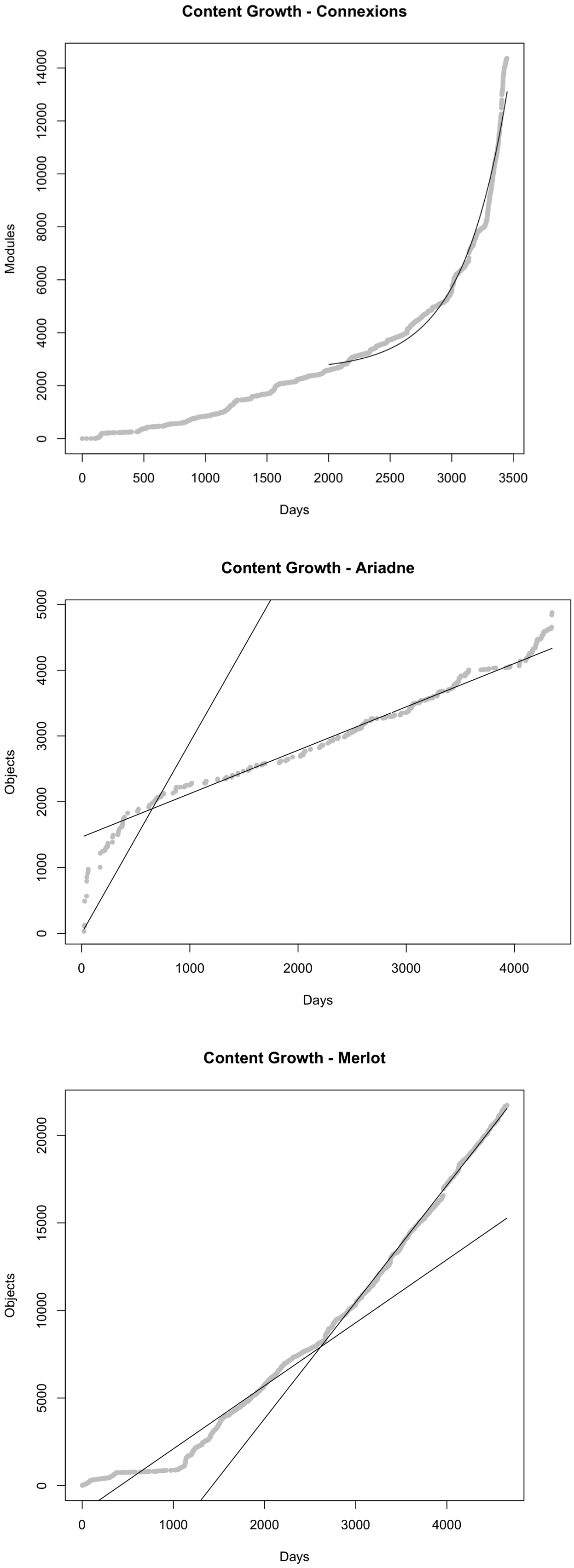

Figure 2. Content Growth Function for Connexions, Ariadne and Merlot

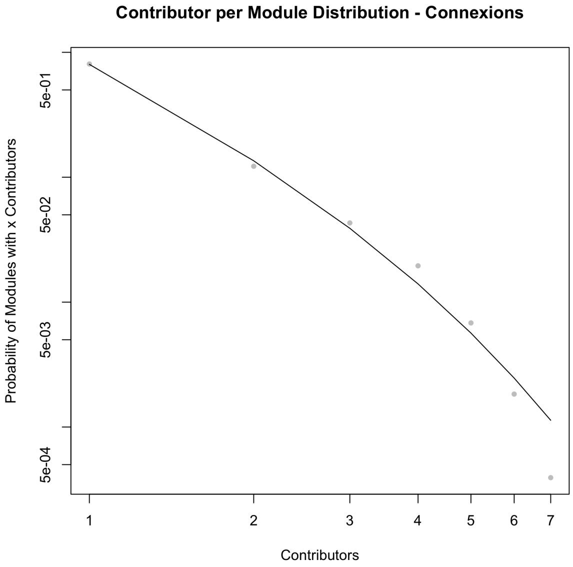

Figure 3. Contributor per Module Distribution. Fitted line: Lotka with exponential cut-off

the level of attention that the content receive from the contributor community to improve its quality. To analyze this characteristic in Connexions, the number of versions for each module is counted and its distribution plotted. Again a heavy tailed distribution is found (Figure 4). The best fitting distribution is also Lotka with exponential cut-off (α=1.53, λ=0.12). Most of the materials (60%) are never edited (only the original version is published). However, around 25% of the materials have been edited at least twice since its original publication. These number suggest that a considerable proportion of the material in Connexions is “alive”, meaning that is being altered after its publication. This differs from what happens in traditional LORs where the materials, in most cases, cannot be updated from its original publication.

3) Time between Versions: The time between versions provide a measurement of how easy is to edit content and how responsive is the contributing community. To analyze this characteristic, the dates of the different versions is subtracted for one another in each of the modules in Connexions. As expected, a considerable amount of edits (22%) occur the same day that the previous edit is published. This changes most probably obey to errors detected after the publish button has been pressed. The remaining 78% of edits occurs in a wide range of days. If the data is plotted in a log-log scale, a heavy-tailed distribution is found (Figure 5). The best fitting distribution is again Lotka with exponential cut-off (α=0.90, λ=1.2x10-3). This distribution implies that most of the edits occurs in the very first days after the previous version of the content is published (50% of the edits occur in the 10 days following the publication of the previous version). The distribution also determine that 10% of the edits occur after a period of time longer than a year. From this results it can be concluded that the contributing com-

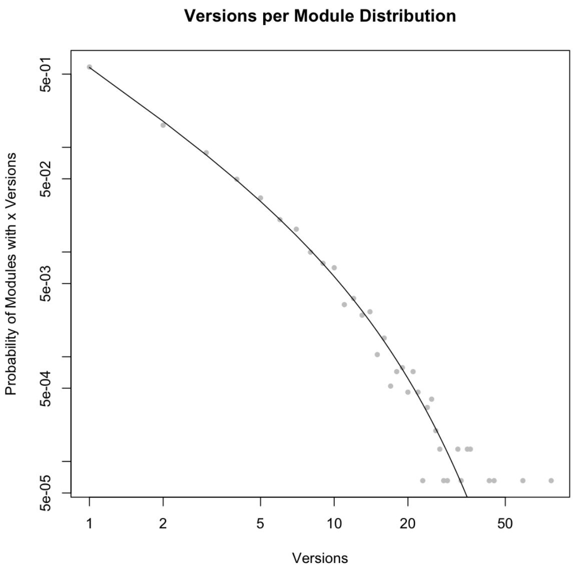

Figure 4. Versions per Module Distribution. Fitted line: Lotka with exponential cut-off

munity is very responsive. It is fast improving material recently published, but is also producing new versions of older materials continuously. If we compare this situation with the almost static nature of traditional LORs, this responsiveness can be seen as a distinctive advantage of Connexions.

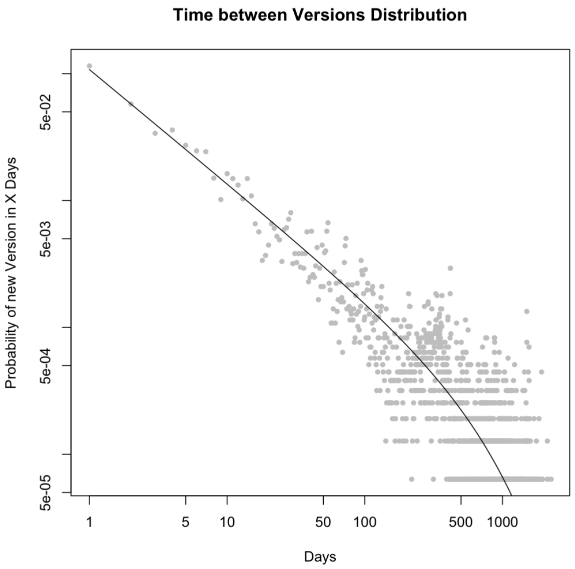

Figure 5. Time between Version Distribution. Fitted line: Lotka with exponential cut-off

B. Content Consumption

As important as the production process is how the content is accessed, rated and reused. Connexions stores statistics of all of these three indicators of the consumption of the content. The main difference between Connexions and traditional LORs is that Connexions allows the remix of materials in what is called courses. For example, a published module can be included as part of one or more courses. This characteristic provide a way to measure the reuse of the module. The following subsections analyze the Connexions characteristics related to consumption:

1) Popularity per Module: The popularity of a module indicates how many visits or hits the module receive per day. It can be seen as a proxy measurement of the attention than a given content generates among the user community. As most popularity distributions [19], the visits per day data is best fitted by a log-normal distribution (μlog=0.05, σlog=0.01) (Figure 6). Most of the content is rarely accessed, while very few but very popular modules, receive hundreds of visits per day. In this respect, the access to the content in Connexions is similar to the access of normal web pages.

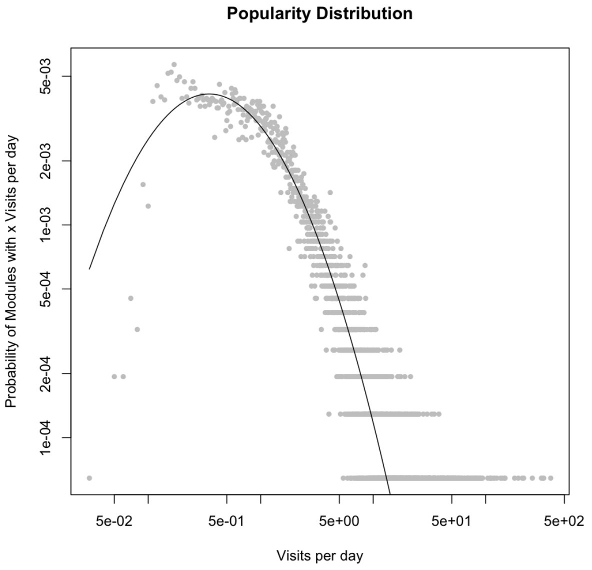

Figure 6. Popularity Distribution. Fitted line: Log-Normal

2) Ratings per Module: Connexions provide to its users the capability to rate the modules. This is a basic community feature implemented in most content repositories. This feature could help to find quality material, using social filtering [20]. For each module, the rate obtained and the number of users that provide those ratings were obtained. The main result obtained from the analysis of the data is that only 0.1% of the modules in connexions have received any kind of rate. This amount of data could not be used to draw any useful conclusion, apart that the rating facility of Connexions is basically ignored by the users. In contrast, Merlot rate and review covers more than 25% of its materials [21].

C. Reuse per Module

To gain more insight in the reuse process, the distribution of reuse among different module was analyzed. The first step in this analysis was to obtain the total number of reuses for each module inside a course. The histogram of the data was plotted to obtain a first indication of the type of statistical distribution that could fit to the data. The resulting histogram indicated a heavy tail distribution (Figure 7). The data was fitted with the six distribution used previously. The best fitting distribution was log-normal (μlog=0.09, σlog=0.62). A third of the material (34%) is never reused inside a course. 44% is only used once. The 22% remaining is reused between 2 ant 8 times. The amount of reuse per module can also be considered a type of popularity measurement.

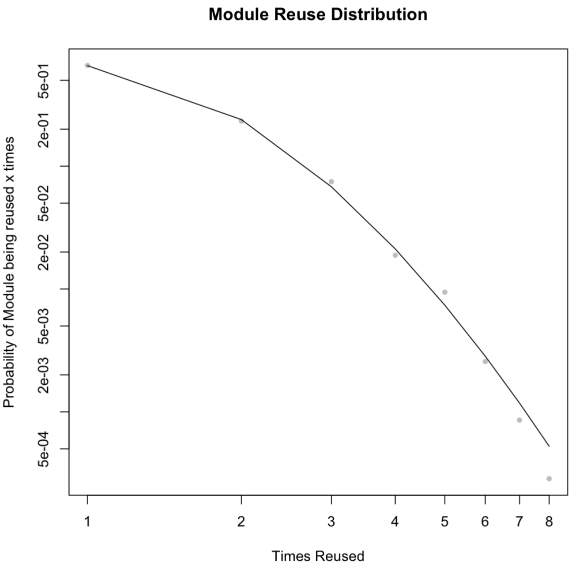

Figure 7. Reuse per Module Distribution. Fitted line: Log-Normal

D. Correlation between Characteristics

There are several interesting correlations (or lack of correlation) between several of the intrinsic characteristics measured before. This subsection present the correlation analysis and a brief discussion of the results. The analysis consists in obtaining the Kendall's tau correlation coefficient between the ranks of the different characteristics. The most common Pearson's coefficient is not used because there is no guaranty that the values come from a bi-variate normal distribution.

1) Number of Contributors versus Number of Versions: A larger contributor group should mean a more active edition of the module. To test this hypothesis, the number of contributors is correlated with the number of versions per module. The analysis shows that there is indeed some correlation (kendall τ=0.58, p-value $<$ 0.01) between these two quantities. Larger collaboration groups seems to indicate more active content development.

2) Number of Contributors versus Popularity: One expected outcome of collaborative creation of material is that the quality of the resulting content would be higher than for material that is created individually. It could be expected that materials with a larger number of contributors would be of higher quality and therefore be more popular inside the repository. To test this hypothesis the number of contributor of a module was correlated with the visits per day that that module receives. The result of the analysis is that is very little correlation (kendall τ=0.29, p-value $<$ 0.01) between those characteristics. The materials created collaboratively are not necessarily more popular that materials created individually.

3) Number of Versions versus Popularity: It could be hypothesized that the modules with larger number of edits are of higher quality and, therefore, should receive more visits. The correlation analysis indicate that there is some correlation (kendall τ=0.43, p-value $<$ 0.01) between these two quantities. An alternative explanation for this result is that more trafficked objects received a higher level of scrutiny by other users and, therefore, they changed more.

4) Popularity versus Reuse: It would be expected that popular content get reused more frequently and also that object that is present in several courses to be accessed more frequently. However, the correlation analysis shows that there is no correlation (kendall τ=-0.02, p-value $<$ 0.05) between the popularity and the reuse of a module. The "visits" popularity is not an indication of the reusability of the module.

E. Analysis Conclusions

From the analysis presented in this section, the most important result was to find that, in Connexions, most modules are updated several times and very frequently. This level of activity in the repository is a consequence of the wiki-like capabilities of Connexions that are uncommon in traditional LORs. This easiness to edit the content of the repository could encourage users to become contributors, not only of small edits, but also of new content. This could be an explanation for the exponential grow in the number of contributors and the resulting exponential growth in the number of modules measured in section IV.

Two surprising, although discouraging, results were also found. The first discouraging result was the lack of correlation between Popularity and Reusability. Connexions, as most LORs, equates the popularity of the resource with its quality and potential for reuse. Popularity is usually the base of the ranking algorithm that order the search result list. Better automatic metrics should be developed to measure the quality of the content [21]. The second discouraging result was the absence of correlation between the number of contributor to a module and the popularity of that module. Models that are the result of extensive collaboration have the same chance of being popular as modules created individually. This conclusion is alleviated by the previous mentioned result, popularity does not necessarily means quality.

These analyses, however, do not provide any explanation for the contributor lifetime distribution observed in Connexion. In order to gain a better insight on the forces that keep the Connexion community growing and contributing, the next section will analyze the social network resulting from the collaborative creation of learning materials.

VI. Analysis of the Social Network

In Connexions, the collaborative creation of content requires extensive communication between the interested contributors. These communications create links between the contributors, that aggregated can be considered the social network behind Connexions. This section presents several analysis of the characteristics of this network and the relation that it has with the characteristics analyzed in the previous sections.

To build this network, the authors of each module were extracted. A link was created between two contributors if they collaborated in one module. If the link between those contributors already existed, the strength of the link was increased by one. The final network was stored in a semi-colon separated text file. All the network analysis was performed with Cytoscape software [22]. The network and the code used to analyzed can be downloaded from [8].

A. Network Characteristics

The first step to analyze the network was to visualize it to understand its structure. The network contains 281 contributors (around 25% of the total contributors) and 691 links between them. The network presents a giant connected component that absorbs the majority of the contributors (202 contributors and 638 links). This giant connected component can be considered the main Connexion community. This community could be seen in Figure 8.

The main characteristics of the Giant Connected Component are:

- The average number of neighbors is 6.3. That means that, in average, each contributor in the main Connexion community have worked with six other contributors. This average, however, is misleading as the distribution of neighbors is fitted by an inverse power law with α=1.1. Most of the contributors only have one or two neighbors. On the other hand, there is a few well connected individuals that have from 20 to 70 neighbors.
- The average shortest path between nodes is almost 3. That means that the Connexions community is a "small world" [23] as there is only 3 degrees of separation between any two contributors, in average. The longest separation (network diameter) between two contributors in the main community is 8 links. The distribution of the length of the shortest paths follow a Normal distribution.

As it can be seen in Figure 8, the main community of Connexions have a backbone formed by committed, very productive members (large and blueish nodes). The official user of Connexions ("cnx.org") is part of this backbone. Attached directly to this backbone, there are several small, short lived contributors. The most probable explanation for this feature is that the group of committed users (backbone) is always searching to help with the development of others materials. They get to contribute in a lot of modules (therefore their large production) and help novice contributors. These group of committed contributors could be comparable to the top-Wikipedians responsible for most of the work in Wikipedia [24].

Another salient feature of the main community is the existence of few short lived but highly productive members, that can be seen in the periphery of the graph (large, redish nodes). A closer inspection of this nodes, reveal that they are mainly institutional accounts that dump a large number of objects into Connexions (usually an existing catalog of materials) and then disappear. For example, the series of three large red nodes at the top of the graph are called: "vocw", "vocw_2" and "vocw_3". These users belong to the Vietnamese Open Courseware initiative.

The most important feature of the network is the mixture of long and short lived members in this main community. New contributors are free to contact or are contacted by older contributors to collaborate in the development of material. This openness to new members provide a hint to why the contributor base is growing exponentially and why the mixture of a lot of short lived contributors with a small but highly committed group of contributors seems to provide good results in Connexions.

B. Relation between Network and Author Characteristics

Similarly to the analysis performed in section V.D, several characteristics of the nodes in the social networks will be correlated with characteristics of the contributors. This correlations will provide information about the structure of the network and the implication of this structure in the unique characteristics of Connexions.

The Kendall τ rank correlation is used instead of the Pearson correlation coefficient p, because some of the characteristics have heavily skewed distributions.

1) Modules Contributed versus Number of Neighbors: It is expected that the more modules a user has published or contributed, the largest the possibility to collaborate with other users. The analysis of the data confirms this supposition, because there exist indeed a considerable correlation (kendall τ=0.67, p-value < 0.01) between these two quantities inside the main Connexions community. This could be the reason of the existence of the highly productive individuals in the backbone in the community. Those committed contributors obtain value and visibility from their social position inside the community.

2) Lifetime versus Number of Neighbors: Another valid supposition is that the longer the user keeps active in the repository, the higher the probability to connect with other contributors. The analysis shows that the level of correlation between this values is lower than expected (kendall τ=0.44, p-value < 0.01). Users with short lifetimes are also able to connect with a considerable amount of other contributors (between 3 and 10). This

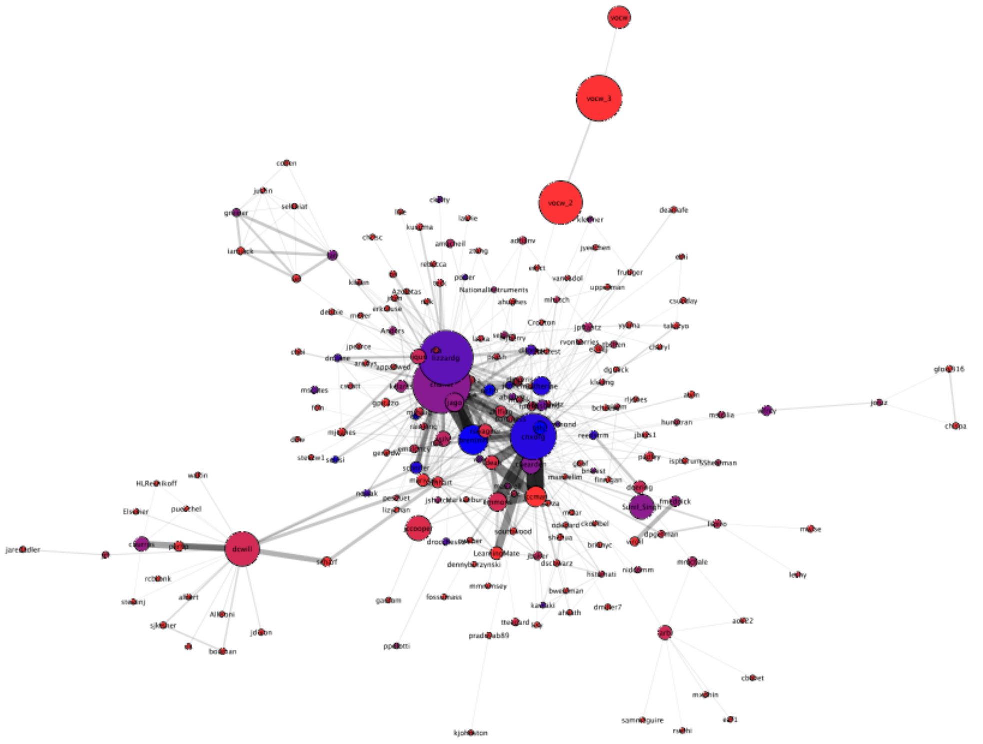

Figure 8. Giant Connected Component (Main Connexion Community). Node Size represents the number of modules published. Node Color represents the lifetime (red means few days, blue means several years). Link Width and Opacity represent the number of modules contributed together.

result implies that Connexions community is open also to very short-lived contributors. This confirms the conclusion reached through visual inspection of the graph in the previous subsection.

3) Date of Arrival versus Number of Neighbors: This correlation will establish if it is easier for old members, even if they are not active, to collect more relationships in the community. The analysis shows that there is low correlation (kendall τ=0.23, p-value < 0.01) between the date of arrival of the contributor and the number of links in the network. The conclusion that could be obtained from this result is that newcomers are welcomed inside the community to collaborate with existing groups or are helped by several contributors to improve their material. This confirms again the openness of the Connexion community.

C. Conclusions of the Analysis of the Social Network

The main finding obtained from the network analysis is that the community of users in Connexions is integrated by a backbone of committed individuals ready to help and accept short-lived and new contributors. This openness paired with continued provided by this backbone could be the drive behind the exponential growth observed in Connexions. However, this conclusion should be tempered, given the fact that only 20% of the contributors are part of this community. The large portion of the contributors (75%) work alone as in traditional LORs and it is not clear the impact that the existence of this community has in their work.

VII. Conclusions and Further Work

This work was inspired by slight anomalies found in the behavior of Connexions compared to traditional Learning Object Repositories [4]. A re-analysis of the anomalies with updated data show that they have not disappeared, but increased with time and expanded to previously normally behaving characteristics. Given that this characteristics are deeply related with the kind of repository, this work hypothesizes that Connexions is the first in a new class of LOR, the Social LOR.

To provide initial support for this hypothesis, the intrinsic characteristics of Connexions were analyzed to find possible explanations for these new behaviors, namely exponential growth and very unequal engagement with the repository. In the analysis of these intrinsic characteristics, It was found that the materials in Connexions are edited or updated several times and at a rapid rate. This goes in line with the open source community dictum: "release early, release often". This strategy is recommended in order to attract other programmers to experiment with the code and improve it. This strategy seems to be working in Connexions, where the number of contributors that edit or publish material is growing exponentially and is causing an exponential growth in the number of available material.

The analysis of the social network that emerge from the collaborative creation of materials help us to understand the unequal engagement of contributors with the repository and why it is not affecting negatively to the growth of Connexions. The main feature found in the community formed by Connexion contributors is its openness to accept sporadic and new contributors. A backbone of committed members seems to provide help and support to non-expert contributors and in exchange receiving recognition and community status. The unequal distribution of the lifetime is a natural effect of the unequal distribution of interest in the repository. The negative effect of the inequality is mitigated by the fluid social interactions between experts and novices. It seems that the same forces that contribute to the success of Wikipedia and other wikis are also pushing the success of Connexions, and differentiating it from traditional LORs.

While the proposed interpretations of the results provide a reasonable explanation for the differences found in Connexions, the exploratory and quantitive nature of this work opens more questions than it tries to solve. Further qualitative research is needed to prove the proposed interpretations are correct. Also further quantitative research is needed to solve some of the questions raised by the results of the analyses. A list of what can be considered the most relevant questions to answer after this work are:

- What is the actual nature of the interactions between the network backbone and the novice users? (Qualitative)
- How the community grows over time? (Quantitative)
- What are the main factors that encourage contributors to share their publishing process with others? (Qualitative)
- What is the actual meaning of popularity and how it is related to the quality of the modules? (Quantitative and Qualitative)

Understanding how Connexions and the hypothesized new class of Social LORs work could help the Technology Enhanced Learning community to design better, more interactive and exponentially growing repositories of learning materials, that could finally fulfill the promise of providing relevant, high quality content to anyone, anywhere.

References

[1] R. Baraniuk, C. Burns, D. Williams, B. Hendricks, G. Henry, A. Hero, D. Johnson, P. Schniter, D. Jones, and J. Kusuma, "Connexions: DSP education for a networked world," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 4, 2002, pp. 4144–4147.

[2] R. Baraniuk, C. Burrus, D. Johnson, and D. Jones, "Sharing Knowledge and Building Communities in Signal Processing," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 10–16, 2004.

[3] R. Baraniuk, *Opening Up Education: The Collective Advancement of Education through Open Technology, Open Content, and Open Knowledge*. MIT Press, 2008, ch. Challenges and opportunities for the open education movement: A Connexions case study, pp. 229–246.

[4] X. Ochoa and E. Duval, "Quantitative analysis of learning object repositories," *IEEE Transactions on Learning Technologies*, vol. 2, no. 3, pp. 226–238, 2009. [Online]. Available: http://ariadne.cti.espol.edu.ec/xavier/papers/Ochoa-TLT2009b.pdf

[5] U. Dholakia, W. King, and R. Baraniuk, "What makes an open education program sustainable? The case of connexions," OECD, Tech. Rep., 2006.

[6] S. Monge, R. Ovelar, and I. Azpeitia, "Repository 2. 0: Social Dynamics to Support Community Building in Learning Object Repositories," *Interdisciplinary Journal of Knowledge and Learning Objects*, vol. 4, pp. 191–204, 2008.

[7] L. Petrides, L. Nguyen, A. Kargliani, and C. Jimes, "Open Educational Resources: Inquiring into Author Reuse Behaviors," in *Proceedings of the 3rd. European Conference on Technology Enhanced Learning: Times of Convergence: Technologies Across Learning Contexts*. Springer, 2008, p. 353.

[8] X. Ochoa, "Data files and calculations code," Online, 2009. [Online]. Available: http://ariadne.cti.espol.edu.ec/xavier/papers/connexions

[9] E. Duval, K. Warkentyne, F. Haenni, E. Forte, K. Cardinaels, B. Verhoeven, R. Van Durm, K. Hendrikx, M. Forte, N. Ebel, *et al.*, "The ariadne knowledge pool system," *Communications of the ACM*, vol. 44, no. 5, pp. 72–78, 2001.

[10] T. Malloy and G. Hanley, "MERLOT: A faculty-focused Web site of educational resources," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 2, pp. 274–276, 2001.

[11] S. Ternier, K. Verbert, G. Parra, B. Vandeputte, J. Klerkx, E. Duval, V. Ordez, and X. Ochoa, "The ariadne infrastructure for managing and storing metadata," *IEEE Internet Computing*, vol. 13, no. 4, pp. 18–25, 2009.

[12] H. Akaike, "An information criterion (AIC)," *Math Sci*, vol. 14, no. 153, pp. 5–9, 1976.

[13] L. Egghe and R. Rousseau, "Systems without low-productive sources," *Information Processing and Management*, vol. 42, no. 6, pp. 1428–1441, 2006.

[14] M. Goldstein, S. Morris, and G. Yen, "Problems with fitting to the power-law distribution," *The European Physical Journal B-Condensed Matter*, vol. 41, no. 2, pp. 255–258, 2004.

[15] Q. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, no. 2, pp. 307–333, 1989.

[16] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," 2007, 26 pages. Arxiv preprint arXiv:0706.1062.

[17] J. Hosking and J. Wallis, "Parameter and quantile estimation for the generalized Pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987.

[18] J. Voss, "Measuring Wikipedia," in *Proceedings of the 10th. Internationa Conference on Scientometrics and Informetrics*, 2005.

[19] S. Sinha and R. Pan, "How a hit is born: The emergence of popularity from the dynamics of collective choice," *Econophysics and Sociophysics: Trends and Perspectives*, 2006.

[20] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating word of mouth," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1995, pp. 210–217.

[21] X. Ochoa and E. Duval, "Relevance ranking metrics for learning objects," *IEEE Transactions on Learning Technologies*, vol. 1, no. 1, pp. 34–48, 2008.

[22] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, p. 2498, 2003.

[23] D. Watts and S. Strogatz, "Collective dynamics of small-

world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[24] F. Ortega, G. Barahona, and M. Jesus, "Quantitative analysis of thewikipedia community of users," in *Proceedings of the 2007 International Symposium on Wikis*. ACM, 2007, p. 86.

Xavier Ochoa is a professor at the Faculty of Electrical and Computer Engineering at Escuela Superior Politécnica del Litoral (ESPOL) in Guayaquil, Ecuador. He coordinates the research group on Teaching and Learning Technologies at the Information Technology Center (CTI) at ESPOL. He is also involved in the coordination of the Latin American Community on Learning Objects (LACLO) and several regional projects. His main research interests revolve around measuring the Learning Object economy and its impact in learning. More information at http://ariadne.cti.espol.edu.ec/xavier.