

Measuring the Creative Commons

Giorgos Cheliotis[†], Ankit Guglani[†] and Giri Kumar Tayi[‡]

[†]School of Information Systems
Singapore Management University

[‡]Department of ITM/MSIS
School of Business
SUNY at Albany

28 February 2007

Submitted to the 3rd Symposium on Statistical Challenges in E-commerce
Research

Contact: giorgos@smu.edu.sg

Extended Abstract

The Creative Commons (CC) is a set of licenses used often when content is published online, and with the aim of complementing existing copyright law by providing more options to authors and their audience. Authors licensing their work under the Creative Commons framework may choose to grant their audience the freedom to create derivatives of that work, such as in making a remix of a music piece. They may also impose constraints on this freedom: they may require that the derivative work is licensed under the exact same Creative Commons license as the original work, or they may only allow non-commercial uses of the work.

We believe it is important to measure the spread of CC as well as understand the motivations for (and implications of) the use of CC licenses for the following reasons:

1. Given that there is evidence that CC is spreading fast in online communities it could become the licensing scheme of choice for online marketplaces. Pricing of online content may increasingly depend on the associated license.
2. By studying CC usage patterns we hope to understand the incentives of individuals when confronted with the option of releasing their work under more liberal or more restrictive terms. This will give a measure of their valuation of these terms.
3. CC is specifically designed with re-use of content in mind, allowing authors to grant or deny others the right to create derivative works. As such activities are becoming increasingly popular, understanding CC usage can help shed more light into the nature of these activities.

Evidence suggests that a large number of online content is already licensed under CC. Some data has been made available on the web, however, to the best of our knowledge, there exists no comprehensive analysis of the use of CC licenses. We use primarily two methods for data collection:

- We conduct a backlink search with Google and with Yahoo to find all pages linking to the description of a Creative Commons license on the CC homepage, across all jurisdictions
- We use the CC-Search functions of Google and Yahoo to find all pages containing CC-specific metadata and filter these results by top-level domain

We call the first two data sets GBL and YBL respectively, while the datasets based on CC-Search we refer to as GCC and YCC, depending on whether we use Google or Yahoo for the data collection. The challenge we face is to use the collected data and other data sources (discussed in detail in the paper) in order to make statistically sound statements about the use of CC.

Sampling error is in theory not relevant because each method should return the total sum of all licenses. We are conducting in this sense a census of the CC population. The problem is that in practice the targets of our census are not CC users, they are Google’s and Yahoo’s indexes and search algorithms, whose inner workings are not visible to us. Also, we are not counting CC-licensed ‘works’, we are counting web pages, so our data is only a proxy for the actual number of works licensed under CC. Nevertheless, by combining more than one method of measurement we can increase our confidence in our findings, even if in most cases we cannot statistically characterize this confidence.

Method	Backlinks	CC-Search
Yahoo	YBL	YCC
Google	GBL	GCC

Although we continue to improve our methodology and add more data points to our analyses, we share some early findings:

1. There are significant differences in licensing across jurisdictions
2. The majority of authors choose the most restrictive licenses
3. A large minority of authors prefers the most liberal licenses
4. From the above it follows that authors tend to flock to the two extremes of either very liberal or very restrictive licensing
5. The more moderate license types are *consistently* underutilized across all jurisdictions (in contrast to the first finding)
6. Licenses which allow for commercial use of the underlying content are less favoured, compared to those that do not

These findings and the descriptive statistics of the distributions of licenses across jurisdictions appear to be surprisingly robust with respect to the choice of measurement method, in spite of the fact that the absolute numbers returned by each method vary considerably. We discuss our findings and measurement methodology in more length in the paper.