Global Reach Index in Online Courses

Ilia Rushkin

The global reach of an online course (MOOC or otherwise) could be characterized simply by the number of countries represented among the users. However, this measure is flawed: it counts countries even if there are very few users from it. Consider a course with 1,000 users: 99 users come from 99 other countries (just 1 user per country), and the remaining bulk of 901 users are all from the US. The number of countries represented is 100, yet this course clearly has a much smaller global reach than another course of 1000 users with 10 users per country.

In essence, we should have a participation cutoff and include a country in the count only if a large enough number of users comes from it. In the interest of robustness, the cutoff should be data-driven rather than set arbitrarily by hand.

A mathematically similar situation is described by the Hirsch citation index, where to characterize the impact of a researcher's publication record only the papers with a large enough number of citations are counted. Therefore, let us apply the same idea as in Hirsch citation index to countries to produce for any course a global reach index, or R-index for short.

Calculation

The definition of the R-index is as follows: *R-index is the greatest number R such that among the course users there are R countries with at least R users from each.*
To illustrate it graphically, numerate the represented countries in the order of decreasing number of users and plot the number of users vs. the country number. Because of the ordering, the plot will be a monotonically decreasing function, whose idealized continuous version is shown in the Figure:

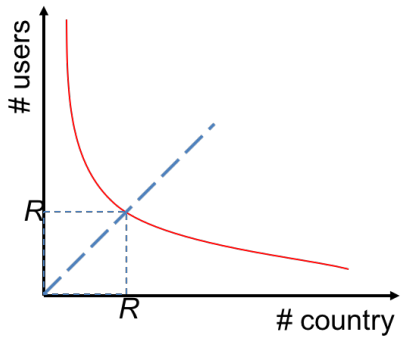

Figure 1. Illustration of the definition of the R-index.

Drawing a straight line with slope 1 from the origin, the x-coordinate of the intersection point is R (ignoring for simplicity the discreteness of the function plotted). In case of Hirsch citation index, the x-axis in this plot would represent a researcher's publications and the y-axis – the number of times a publication was cited. Prior to Hirsch, the same idea was known as the

Eddington number and was used by Arthur Eddington to measure the achievements of bicyclists: on the x-axis would be the number of occasions you rode a bicycle and on the y-axis – the number of miles you rode on each occasion.

The definition easily lends itself to generalizations. One might calculate the R-index on the basis not of countries but continents, or US states, etc. One may also introduce a scaling factor, i.e. use a slope $k \neq 1$.

A policy decision is who should be included into the calculation as a user: all registrants, or only those who had some activity, etc.

As defined, the R-index is sensitive to the total number of users in a course: a course with 10,000 users is likely to have a higher R-index then a course with 100 users. We might also be interested in a size-normalized version of the R-index, which could compare courses of different sizes on equal footing. To find such normalized R-index observe that if a course has N users, the maximum possible value of the R-index is the square-root of N (strictly achievable only if that is an integer, of course). Let us divide by that and introduce the normalized reach index as

$$\text{Rn-index} = \text{R-index} / N^{1/2}$$

For a course of any size, the Rn-index is a number between 0 and 1 (1 is achievable exactly only of the number of users in the course happens to be a perfect square), which can also be expressed as a percentage (0 to 100%). It is a percent of the theoretical highest achievable R-index for a course of this size.

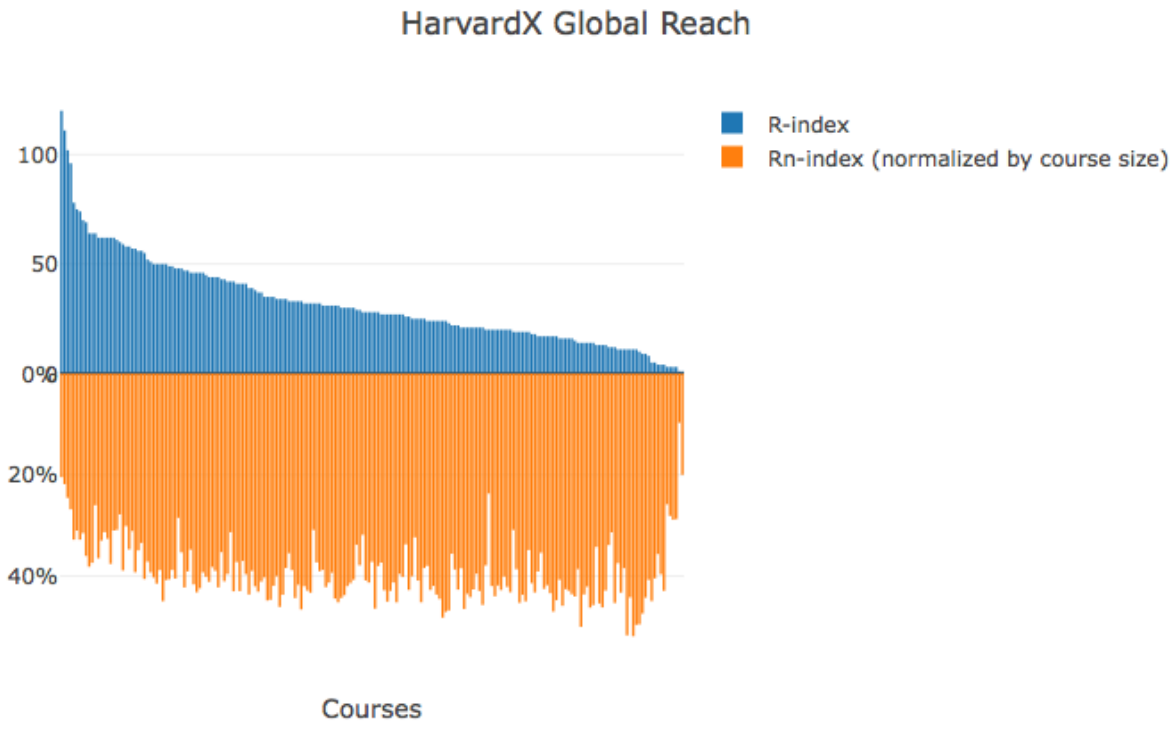

Figure 2. R-index and the normalized R-index for all HarvardX courses on record as of 09/2017. The calculation is done using users who visited at least one chapter ("viewed" the course) and have the country data. The courses are arranged in the order of decreasing R-index. Note the effect of normalization: the normalized index does not show a strong trend.