Local Internet Content

Emmanouil Tranos
and
Christoph Stich

August 6, 2019

1 Measuring Local Internet Content

Each entry of the JISC UK Web Domain Dataset (Jackson, 2017), which is a subset of the Internet Archive and curated by the British Library and includes all the archived webpages under the .uk top level domain[1], contains a timestamp, the URL of the archived website as well as the British postcode found on each site. We can use this dataset to derive a measure for how much local internet content (hereafter LIC, Tranos & Stich 2019) there exists across the UK.

However, the dataset includes websites with differing geographic reach; some websites may refer to single postcode, while others may refer to several postcodes all over the UK. As described in Tranos and Stich, 2019, this poses a problem when trying to ascertain whether localised internet content is a driver of online behaviour. We thus need a way to discount websites that have less of a local focus.

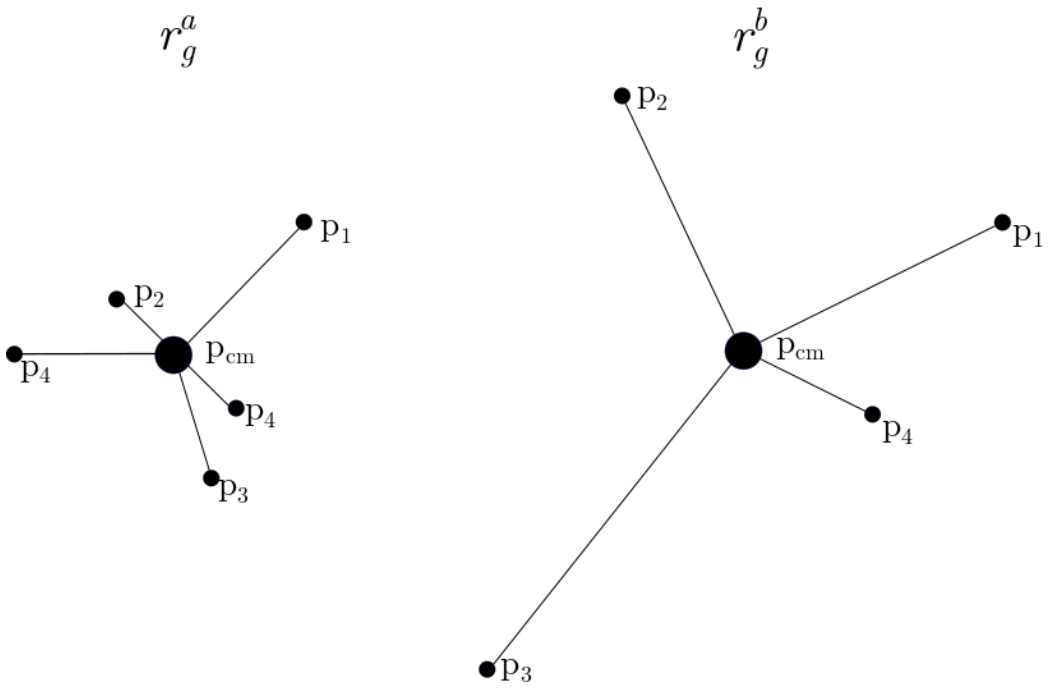
The underlying idea is that websites that have a high geographic dispersion are less "local". To compute the geographic dispersion of a websites' set of postcodes p we calculate the Radius of Gyration r_g of p in kilometres. The Radius of Gyration is defined as follows (Gonzalez, Hidalgo, & Barabasi, 2008):

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{cm})^2},$$

[1] http://data.webarchive.org.uk/opendata/ukwa.ds.2/

where p_i represents the $i = 1, ..., n$ spatial coordinates of each postcode recorded for each domain and $p_{cm} = 1/n \sum_{i=1}^{n} p_i$ is the geographic centre of mass of said domain. For an example of two r_g see figure 1.

Figure 1: An Example of Two Different r_g



$r_g^a < r_g^b$ as the average squared distance from the centre of mass is much smaller for r_g^a than for r_g^b.

A website with a high r_g will be of national interest, while a website with a low r_g will have a very local geographic presence. As local geographical units we utilise the Middle Layer Super Output Areas (MSOA) for England and the Intermediate Zones (IZ) for Scotland[2].

For each MSOA/IZ with a set W of archived websites we calculate yearly measures of the volume of LIC as follows:

$$\sum_{p \in W} \frac{1}{1 + r_g(p)}$$

[2]Northern Ireland was excluded as the available census geographies were not comparable in population.

2 Description of Data Files

We provide two data files that are based on different census geographies:

- *lic_2001.csv* is based on the boundaries of the 2001 census
- *lic_2011.csv* is based on the boundaries of the 2011 census

Each file has the following columns:

- *ID* is the unique MSOA/IZ identifier used by the census
- *20XX* are the values of LIC for the respective year ranging from 2001 to 2012

References

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns supplementary material. *Nature*, *453*. doi:10.1038/nature06958

Jackson, A. N. (2017). JISC UK web domain dataset (1996-2010) geoindex. *The British Library*. doi:10.5259/ukwa.ds.2/geo/1

Tranos, E. & Stich, C. (2019). Individual internet usage and the availability of online content of local interest: a multilevel approach. *CEUS*, in print.