Measuring the Digital Divide

Michael Paul Fourman*
School of Informatics
10 Crichton Street
Edinburgh, Scotland
michael.fourman@ed.ac.uk

Yao Chen
Richboy Echomgbe
Jing Yang

ABSTRACT

Digital exclusion compounds other forms of deprivation, by denying access to an increasing range of opportunities, in areas such as education, employment, and health.

Many nations now have funded programmes to reduce digital exclusion. The problem is how to set goals for such programmes. Digital exclusion is hard to define and identifying robust indicators of digital exclusion is challenging.

Generally, any increase in broadband uptake, or penetration, is taken as an indicator of success. The uptake statistic is also commonly used to make international comparisons because it is gathered in most regions. The Gini index has also been used as an indicator of the global digital divide.

However, increasing participation can mask increasing exclusion. Increased uptake among the most included sectors of the population increases participation, and can also reduce the Gini index, while still widening the gap to the excluded population.

Here we introduce a new index, a variation on the Gini index. We show that it has good theoretical properties, and argue that it provides greater insight into digital exclusion. By way of illustration, we report on two applications, one local, one global, in scale.

First, we use uptake data for Scotland to quantify digital deprivation, the link between digital exclusion and deprivation, both across Scotland, and locally within each of Scotland's 32 local authority areas. Digital deprivation across Scotland diminished from 2012 to 2013. However, in several areas where the exclusion–deprivation link was already strong, recent changes have served to reinforce this link.

We then quantify the global digital divide, represented by national variations in fixed broadband uptake, from 2000 to 2014. Our index highlights a sustained annual reduction in the global divide from 2000 to 2011, followed by a steady rise since 2011.

*Corresponding author

1. INTRODUCTION

The *Digital Divide* is the stratification of society resulting from inequality of opportunity to access and use digital technologies. Norris [13] introduces the term 'social divide' for inequalities of Internet access between groups within societies. We prefer to keep 'social divide' as a general term, covering a variety of aspects of deprivation and advantage, including digital exclusion and inclusion, and talk of local, national, and global digital divides.

Information is the life-blood of society. Digital technologies provide tools that transform our abilities to access, store, process, and communicate information. The increasing digitisation of society exacerbates other forms of deprivation. Those who cannot or do not use these tools are increasingly excluded from many economic, social, cultural and educational opportunities. They are digitally deprived.

Enlightened communities are concerned to identify, address and mitigate the various inequalities of opportunity that are commonly associated with socioeconomic factors such as income, employment, education, ethnicity, social class or residence in a deprived area. Digital inclusion is an unequally distributed opportunity of increasing social and economic significance. So it is important to be able to identify and monitor changes in the digital divide at every scale.

For example, Scotland's digital participation strategy will "encourage people and businesses to get online and enjoy all the opportunities of the digital age."[1] At a global scale, the ITU has identified an "important and urgent need to provide access to basic telecommunication/information and communication technology (ICT) services for everyone, and particularly for developing countries".[2]

If the increases are poorly targeted, increasing opportunities for inclusion can increase the digital divide, widening the gaps between different communities. How can we measure progress to ensure that we are narrowing the divide, both locally and globally?

Factors of deprivation.

Digital deprivation is linked to other factors of deprivation through intricate interactions between individuals, technology, economics and society. 'The Digital Divide' [14] provides a recent survey. It examines, primarily from a sociological perspective, "how various demographic and socio-

[1] `http://www.gov.scot/Topics/Economy/digital/Digital-Participation`

[2] `http://www.itu.int/en/ITU-D/Digital-Inclusion/`

economic factors including income, education, age and gender, as well as infrastructure, products and services affect how the internet is used and accessed,.' Norris [13] set out a model distinguishing different factors of digital engagement internationally. Recent policy discussions in Europe and the UK adopt a similar model and identify three such factors: access, motivation and skills [16].

Motivation and skills are hard to measure, so quantitative studies often focus on the access divide. Reducing the access divide should not be the only goal for any policy of digital inclusion. Motivation, skills, and a supportive environment are also required to reap the full benefits of digital inclusion. However, inequalities of access provide a lower bound for any comprehensive measure of the digital divide.

Our goal in this paper is modest. We do not address the skills, knowledge and motivation that are needed to make effective use of digital technologies. We do not assess technological differences in speed, symmetry, latency, etc. that may limit or enhance the value of a connection. Nor do we consider mobile connections. We focus on a fixed broadband connection, as a binary advantage that some households have, while others do not.

Our focus on access is analogous to using lack of access to domestic mains water as an indicator of health inequalities. Getting every home online is just a necessary first step towards effective digital inclusion.

The likelihood of a household having access to the internet can be affected by a complex mix of geographic, economic and social factors. The relative importance of these different geographic and economic factors varies from region to region and from country to country. However, many studies, from countries with diverse geographies and economies, have found that digital exclusion is strongly correlated with other aspects of deprivation.

Figures for access to the internet, at national and regional levels, are routinely reported by national governments, and tracked by international organisations such as the EU and the ITU. These, combined with population data, provide quantitative measures of regional and national differences, and reflect economically significant differences, but they do not measure the extent to which increasing digital participation may be subject to local variations that serve to reinforce or mitigate existing social divides.

Although "numerous studies have been conducted on the international and national level or country, government, policy levels — the macro perspective — few have attempted to evaluate these issues from a micro perspective" [7].

We introduce an index of inequality, with a simple probabilistic interpretation, as a new measure of the access divide. We propose that it should be monitored to ensure that increases in uptake do not widen the digital divide by excluding the more deprived sectors of society.

We use it first to assess how local access divides, within each of Scotland's 32 local authority areas, are associated with deprivation. We then apply it to ITU data to measure the global access divide. In each case, this allows us to quantify the effects of recent changes in broadband penetration on the digital divide.

Our index is closely related to the well-known Gini index for income inequality. It differs from previous applications of the Gini index to digital participation in that it focusses on inequality of opportunity amongst those who are still offline.

In § 2 we give a brief and self-contained introduction to our index and a summary of our results concerning the local evolution of the digital divide in Scotland. § 3 gives a more-detailed description of our index and its relation to the Gini index. In § 4 we apply our index, using ITU data, to assess recent changes in the global divide. In § 5 we discuss related work. In § 6 we describe the data we have used,[3] and some limitations on the interpretation of our results. Finally, in § 7, we conclude with a brief outline of some avenues for future work.

2. SCOTLAND'S DIGITAL DIVIDE

Ofcom provides postcode-level data on the uptake of fixed broadband across the UK. This shows that, even within regions with relatively high uptake figures, there are pockets of severe digital exclusion within almost all sections of society. These might result from a variety of social, geographical or technical factors.

The analysis carried out by a recent Royal Society of Edinburgh Inquiry [6] shows that these local divides make a significant contribution to digital inequality. That inquiry also found that there is a strong correlation between digital exclusion and the Scottish Index of Multiple Deprivation (IMD). Our index provides a non-parametric quantification of the link between these two forms of deprivation.

An index of inequality.

We consider the population of households in Scotland. The IMD provides a rank ordering of households, so we can ask whether one household, for example, the Browns, is more deprived than another, e.g. the Andersons. Suppose the Browns are offline and the Andersons are online, what are the odds that the Browns live in a more deprived neighbourhood than the Andersons?

If digital inclusion were independent of deprivation then, the odds would be even. Knowing that one household is offline and the other offline, would tell us nothing about their relative deprivation.

Definition 1. *Consider selecting at random a pair (b, a) of households such that b is offline, and a is online.*

We ask, for an offline-online pair (b, a),
What are the odds that the offline household, b,
is more deprived than the online household, a? (1)

Our inequality index *is based on the answer.*

If the odds are $B : A$,
we define our coefficient of inequality, φ,
to be the fraction $(B - A)/(A + B)$, (2)
and the corresponding inequality index
to be this value expressed as a percentage.

In Scotland, as in every country where this has been studied, if we randomly select two households, one online, the other offline, the odds are that the offline household, the Browns in our example, is more deprived than the online household, the Andersons. These odds give a measure of the association between digital exclusion and deprivation.

[3] This paper contains results derived from information licensed by the Office of Communications. The data supporting this research can be accessed via the following URL:
.

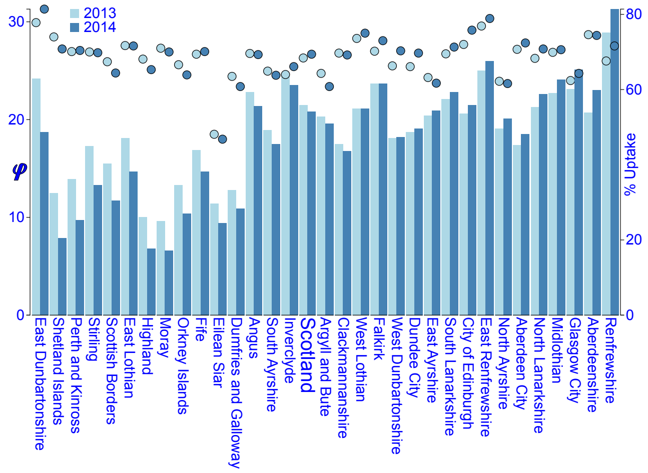


Figure 1: Scotland's Digital Divide 2013 − 14

Our index is defined in terms of the odds. We introduce the index in order to make a close connection, in § 3, to the well-known Gini index [2]. Working the other way, we can derive the odds from an index. An index of N means that the odds of the offline household being more deprived are $100 + N : 100 - N$.

For example, for Scotland, in 2013, the odds were roughly 3 : 2. These are the same odds as 120 : 80. This corresponds to a coefficient of 1/5, or an index of 20%.

Mathematically, our index could take any value, from −100% to +100%. However, in general the odds are on the offline household to be more deprived, so we normally see only positive values.

Results.

Figure 1 shows, graphically, both the broadband uptake statistic, and our inequality index, φ, computed from the Ofcom data for 2013 and 2014, for each Scottish local authority area, and for the whole of Scotland.

Colour codes the year: 2013 light; 2014 dark. The bars show index values, using the φ scale presented to the left of the chart. The small circles show the uptake figures, the percentage of households online, using the *% Uptake* scale presented to the right of the chart.

In 2014, uptake varies from 81%, in East Dunbartonshire, to 47%, in Eilean Siar (the Outer Hebrides). However, in East Dunbartonshire our index is 19%, while in Eilean Siar it is only 9%. We interpret this as indicating that in East Dunbartonshire digital exclusion is more strongly associated with other forms of deprivation, and that social issues may not be the primary cause of low uptake in Eilean Siar — which will not surprise those who know the local geography.

The index values for 2014 range from 6.6% in Moray, to 31.3% in Renfrewshire. This means that in Moray the odds that an offline household was more deprived than an online household were only slightly above evens, roughly 107 : 93. While, in Renfrewshire the odds that the offline household is the more deprived were roughly 131 : 69 — just under 2 : 1.

In our chart, East Dunbartonshire, on the left, shows the largest decrease in inequality over this period, while Renfrewshire, on the right has the largest increase. Uptake improved in both East Dunbartonshire, from 77.8% to 81.4%, and Renfrewshire, from 67.6% to 71.6%. However, in Renfrewshire the odds of the offline household being more deprived increased, inequality grew from 28.9% to 31.3%, while in East Dunbartonshire it fell, from 24.2% to 18.7%.

The remaining authorities are ordered according to the degree to which the divide has changed from 2013–2014. In West Lothian, the index did not change (although uptake increased from 73.6% to 75.0%). In those authorities to its left, the divide has reduced. To its right we have those authorities, where the divide has increased. The entry for Scotland as a whole is found in the middle of the chart, since it shows the average change, a (slight) decrease.

We see that in general the areas with low inequality indices (say, indices less than 15%) have improved. However, in many of the areas that started with higher indices (say, indices greater than 20%), with East Dunbartonshire, Dumfries and Galloway, Inverclyde, and Argyll and Bute as notable exceptions, increases in broadband uptake have served to increase the digital divide between the more and less deprived sectors of society.

These results have clear policy implications, and should serve to direct the focus of future interventions.

3. QUANTIFYING INEQUALITY

The Gini coefficient is a well-established measure of inequality of distribution, introduced, and best-known, for its application to inequalities in the distributions of income and wealth. However, increasing uptake amongst those already well-served may lower the Gini index, so it is not an appropriate measure.

The Gini index, as used in [17], can never take a value greater than q when applied to a binary advantage such as a domestic broadband connection. The extreme case, in which *one takes all*, that is conceivable for a distributed quantity such as income or wealth, cannot occur for a binary benefit, such a broadband connection. It would make no sense to concentrate all the connections in one household. In the extreme case, the connections are divided among some privileged class, who *all take one*, while the rest have none. The Gini coefficient for this extreme case is q, the proportion of households offline.

In this section, we show that the φ-index introduced in § 2 is equivalent to a Gini index, scaled so that the extreme case corresponds to 100% inequality. We also show that this adaptation corresponds to a Gini index for the inequalities of opportunity facing those who are still offline.

We use a fixed broadband connection as our running example of a binary advantage. The *haves* are online, the *have-nots* are offline. An index of deprivation is an ordering $\prec$ of the population where $a \prec b$ means that a is more deprived than b. The index of an individual, b is the proportion of the population represented by $\{a \mid a \prec b\}$.[4] Typically, the more deprived members of the population have a lower chance of benefitting from an advantage. In our example, they have a lower chance of being online.

Figure 2 shows four graphs, all referring to the same model population.[5] These all plot probabilities against index of de-

[4] For a theoretical treatment it is simplest to assume that $\prec$ is a total ordering. For practical applications, where ties occur, we assume that ties are resolved 50 : 50.

[5] For this model population, uptake is given by $p(x) = 0.45 + 0.5x^2$, where x is the coefficient of deprivation. It rises from

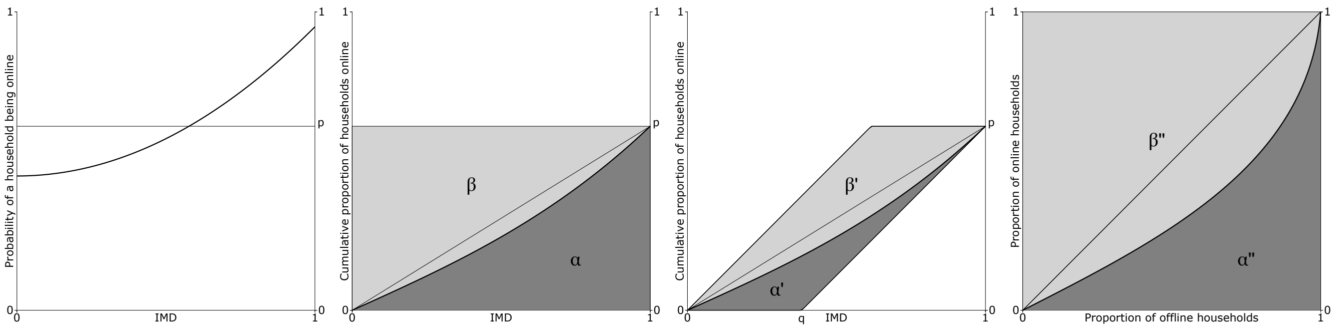


Figure 2: Quantifying Inequality: pdf and Lorenz curves for a model distribution.

privation. In the first, left-most graph, the heavy, rising line plots a probability density function (pdf) $p(x)$, giving the probability of being online for various levels, x, of the Index of Multiple Deprivation (IMD). The area under this curve represents p, the mean probability of being online, the usual broadband penetration statistic. For this model population, $p = 0.6$. The light line represents the equal distribution that would result if the opportunities for broadband connection were shared equally across society, so all households have the same probability, p of being online..

In the second graph, the heavy line is the *Lorenz curve* given by taking the population in order, from most deprived to least deprived, and plotting the cumulative sum, $P(x)$, of households connected against the cumulative sum of households — both normalised wrt the total number of households. These proportions range from 0 at the origin to p, the population level of broadband penetration. The light diagonal line shows the Lorenz curve we would see if opportunities for broadband connection were shared equally across society.

In Gini's construction the Lorenz curve plots the cumulative sum of incomes against population, ordered by increasing income. Gini uses the the area λ enclosed by the Lorenz curve and the diagonal — which is half of the difference between α and β — scaled so that the maximum possible inequality is 1, as his measure of inequality. His coefficient is given by,

$$\gamma = \frac{\beta - \alpha}{\beta + \alpha} = \frac{\lambda}{\lambda + \alpha}\ , \quad \text{where } \lambda = \frac{\beta - \alpha}{2}\ . \tag{3}$$

The Gini index is simply this coefficient represented as a percentage.

The classical treatment scales this diagram vertically, by a factor of $1/p$, so that the shaded area fills the unit square, the Lorenz curve is a path from $(0, 0)$ to $(1, 1)$, and the Gini coefficient is the difference between the areas above and below the Lorenz curve. Our construction is similar, but not the same.

For Gini, taking the population in order of increasing income, the Lorenz curve is convex (its slope never decreases). In the extreme case, where all of the income is concentrated in one individual, $\alpha = 0$.

For our construction, the Lorenz curve need not be convex, since we can choose any order as an index of deprivation. Furthermore, the area λ may include both positive and negative contributions — our index can be negative or positive; so we take $\lambda = (\beta - \alpha)/2$ as a definition. In both constructions, the Lorenz curve, since it is given by a cumulative sum, is monotone increasing (in whatever order we take the population).

Most importantly, in our present context it makes no sense to consider concentrating all the broadband connections within one household. The slope of our Lorenz curve represents the rate of broadband uptake, at each level of deprivation. Since the rate of uptake can never be more then 100%, this slope can never be greater than 1.

So, the Gini index can never be greater than $q = 1 - p$. As penetration, p, increases, the Gini index will inevitably go down, even if we increase inequality by serving those at the top of the pile first. The actual extreme would produce a Gini coefficient of q; we scale this to give an extreme value of 1.

Our third graph illustrates this. For a given population level of broadband uptake, p, our Lorenz curve will be a monotone, non-decreasing function, whose slope is ≥ 0 and ≤ 1, tracing a path from the origin to the point $(1, p)$. We show four such paths. To the two paths from the previous graph — the Lorenz curve, and the fine line of perfect equality — we add paths showing two possible extremes of inequality.

One extreme, where only those most deprived are online, traces a path along the top of the parallelogram. The real world is not like this: we do not see it in practice.

The other extreme, closer to reality, traces the bottom edge of the parallelogram. In this extremal situation the population is partitioned so that all households less deprived than some critical unfortunate household are connected, while the remainder are not. The Lorenz curve follows the axis, up to and including this critical household, and then rises, with slope 1, from $(q, 0)$ to $(1, p)$. The value $\alpha = 0$ is not achievable, the minimum value of α, corresponding to our line of maximum inequality, is $p/2$. We must use a different scaling.

The Lorenz curve for *any* distribution of connections with an overall uptake of p will lie within the parallelogram. All values of $\alpha', \beta' \geq 0$ are possible, subject only to the constraint that $\alpha' + \beta' = pq$, the area of the parallelogram.

Where Gini scales the shaded area in our second diagram to fill the unit square, we scale the parallelogram in our third diagram, to fill the unit square. The fourth, rightmost graph shows the result of this scaling.

45% for the most deprived ($x = 0$), to 95% for the least deprived ($x = 1$). It has a mean population uptake $p = 60\%$ and a φ-index of roughly 35%.

Theorem 1. *Our index satisfies, and could alternatively be defined by, the equation, $\varphi = \beta'' - \alpha''$.*

Proof. By construction, we have the equations,

$$\frac{\beta' - \alpha'}{\beta' + \alpha'} = \frac{\lambda}{\lambda + \alpha'} = \frac{\beta - \alpha}{pq} = \beta'' - \alpha'' \ . \tag{4}$$

By definition (Equation 2), it suffices to show that $\beta' : \alpha'$ is the odds that, given an (offline, online) pair, the offline household is more deprived.

The areas $\alpha, \beta, \alpha', \beta'$ in our diagrams correspond to sets of pairs of individuals, in the sense that each area represents the probability that a randomly chosen pair will belong to the corresponding set. In particular,

$$\alpha \sim \big\{(b,a) \mid b \prec a \ \&\ b \text{ is online}\big\} \quad \text{and,} \tag{5}$$

$$\alpha' \sim \big\{(b,a) \mid b \prec a \ \&\ a \text{ is offline } \&\ b \text{ is online}\big\} \tag{6}$$

The first of these (5) results from the construction of the Lorenz curve as a cumulative sum; the second follows, as the difference between these two areas is a triangle with area $p^2/2$ that represents the difference between the two sets,

$$\alpha - \alpha' \sim \big\{(b,a) \mid b \prec a \ \&\ a \text{ is offline } \&\ b \text{ is online}\big\} \tag{7}$$

The parallelogram itself, has area pq:

$$pq \sim \big\{(b,a) \mid a \text{ is offline } \&\ b \text{ is online}\big\} \,. \tag{8}$$

$$\text{So,} \quad \beta' \sim \big\{(b,a) \mid a \prec b \ \&\ a \text{ is offline } \&\ b \text{ is online}\big\} \,. \tag{9}$$

Thus $\beta' : \alpha'$ is the odds that $a \prec b$, given that a is offline and b is online. □

Our index is given by the difference between the areas above and below the transformed Lorenz curve. This allows us to plot, and compare, on the same diagram, the curves for populations with different levels of overall uptake.

Interpretation.

The classical Lorenz curve plots points $(x, P(x))$, where $P(x)$ is the cumulative proportion of the entire population online, and x is the cumulative proportion of the entire population.

The distance of such a point from the left-hand edge of the parallelogram is $x - P(x)$. But $x - P(x) = Q(x)$, the cumulative portion of the population offline. So, if we consider coordinates given by position within the parallelogram, then the Lorenz curve is given parametrically by points $(Q(x), P(x))$, where x is an index of deprivation.

In fact, the parallelogram can be viewed as the set of (offline, online) pairs, (b, a), arranged so that the pairs are placed in increasing $\prec$ order in each component: the offline component, b left-to-right, and the online component a, bottom-to-top. Below the Lorenz curve, $b \prec a$, while above it $a \prec b$.

We redraw the parallelogram as (i.e. transform the parallelogram linearly to) a 1×1 square, to give a picture that looks just like the classical Gini diagram, but that represents our index. The abscissa now represents the total offline population, while the ordinate represents the total online population.

If the population is totally ordered by IMD, so that we consider the households one-by-one, then our transformed curve makes no large jumps. It traces a 'Manhattan' path, stepping right or up, through the offline and online populations, from $(0, 0)$ to $(1, 1)$. In the case of extreme inequality, it first traces right through the offline population, to the point $(1, 0)$, and then continues up through the online population, rising finally to the point $(1, 1)$.

The transformed Lorenz curve is given parametrically, in terms of a deprivation index, x, by the points $(\mathbb{Q}(x), \mathbb{P}(x))$, where,

$$\mathbb{P}(x) = \frac{P(x)}{p} \qquad\qquad \mathbb{Q}(x) = \frac{Q(x)}{q} \tag{10}$$

$\mathbb{Q}(x)$ and $\mathbb{P}(x)$ are the proportions, of the offline and online populations (respectively), that have index $\preceq x$.

In this presentation, our index can be viewed as a classical Gini index for *the inequality in the opportunities available to those who are still offline*, where $p(x)$ represents the opportunities available to an individual with IMD x. In effect, we are using the proportion of online households no less deprived than a given threshold, x, as an estimator for the cumulative sum of the opportunities available to an offline household with index x.

We can now compute the population index from population data, just as for the Gini, using Brown's formula [1]. For a population stratified by ranks $i \in [1, N]$ we can compute,

$$\varphi = 1 - \sum_{i=1}^{N} (\mathbb{Q}(i) - \mathbb{Q}(i-1)).(\mathbb{P}(i) + \mathbb{P}(i-1))$$

The slope of the curve represents the odds ratio relating the odds of being online for a household at some particular level of deprivation to the odds of being online for the population as a whole. Our index thus depends on the way this ratio varies over the offline population. Two populations with different levels of uptake will have the same curve, and therefor the same index, if the relative advantage or disadvantage represented by this ratio varies in the same way.

This observation is the analogue for our index of the scale-independence of the Gini index.

If the index is positive then converting an offline-online pair (b, a) with $b \prec a$ into one with $a \prec b$, by transferring the connection from one household to the other, will reduce the index. This is our analogue of the Pigou-Dalton transfer principle. We also have a property of population independence: our index is determined by the probability distribution, and is thus independent of the size of the population.

In summary, our index has a simple statistical interpretation, which we used as our definition. Like the Gini, it has good theoretical properties, can be used at varying scales, and is simple to calculate. Finally, as we have just shown, it is the Gini index for the inequality of distribution of opportunity over those still offline, the have-nots.

Binary advantage.

Applying the Gini index directly to a binary advantage such as broadband access computes the same inequality as does ours, but it effectively assumes that the haves, those who are already online, share the pain of this inequality; which they do not. By contrast, our index is a measure of the weight of the inequality of opportunity borne by the have-nots.

We have presented our model in terms of the concrete example of access to a fixed broadband connection. It can clearly be used to quantify inequalities in the distribution of any binary advantage, such as a vote, or a university education.

4. THE GLOBAL DIVIDE

To compute our index we require data on numbers of connections and numbers of households. Figure 3 shows φ, our index of inequality, and uptake, as percentage of households connected, across the 68 countries for which we have been able to determine these figures for the years 2000 –2014. We see that our index of inequality reduced annually in the period 2000 –2011, but has since increased in the period 2011–2014.

These figures quantify inequality between the countries for which the necessary data is available, and ignore the remaining countries — many of which have poor or no broadband, as well as the inequalities that will exist within each country. Thus, they provide a lower bound for the global digital divide.

Our Lorenz curve for each year is also drawn, with increasing opacity for successive years. Five years, 2000, 2004, 2007, 2011, and 2014, are highlighted, in bold in the table, and with thicker strokes for their curves. The curve for 2000 shows the greatest inequality; the curve for 2011 shows the least.

Each successive curve for 2001–2004 dominates the 2000 curve and its predecessors.

The curves for 2005–2007 show a decrease in φ, and a changing pattern of inequality, with no Lorenz domination. From 2007– 2011 we again see successive curves that dominate their predecessors.

Finally, from 2011 to 2014 inequality increases, as more people go online in the more connected countries.

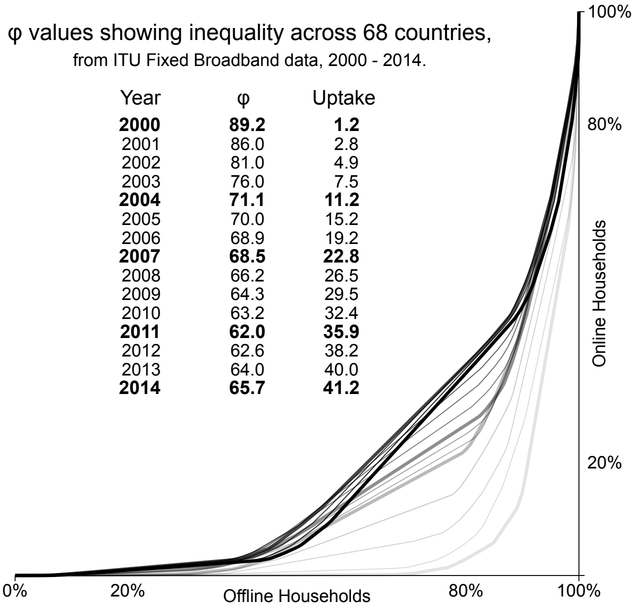
In the 2014 curve we see three clear groups of countries. Our Lorenz curve is approximated by three straight line segments. Their different slopes show different levels of opportunity.

Roughly 50% of the online households are in a group of well-connected countries that accounts for only 10% of the offline households. In these countries the odds of being online are over 3 : 1. The next 45% of the online households are found in a group of moderately-connected countries that accounts for around 40% of those offline. Their odds of being online are roughly 3 : 4. The final, poorly connected group includes around 5% of the online households, and 45% of those offline. In one of these countries, your odds of being online are roughly 3 : 40.

5. RELATED WORK

In 2002, a report from the National Telecommunications and Information Administration (NTIA), *A Nation Online: How Americans are expanding their use of the Internet* [17], introduced a version of the Gini coefficient derived by plotting, for example, a "Lorenz Curve for Households with Computers vs. Income," and comparing the area between that curve and the diagonal with the area of the triangle below the diagonal.

The NTIA report uses this Gini coefficient to justify the assertion that, "while inequality remains . . . inequality has been declining, by the standard measure of inequality used by economists." For example, it tells us that, between 2000 and 2001, the percentage of US households with Internet rose from 41.5% to 50.5% (*op cit.* Table 1.1), while the Gini index for Internet Connection of any type by Family Income fell from 0.309 to 0.270 (*op cit.* Table 9.1). Chakraborty and Bosman [3] use the same procedure to compute a coefficient of 19.7% for home PC ownership (in September 2001 in the US) as a function of household income. They also show that this coefficient varies from state to state, and between different racial groups. More recently, Howard et al. [8] have used a similar procedure.



Year	φ	Uptake
2000	**89.2**	**1.2**
2001	86.0	2.8
2002	81.0	4.9
2003	76.0	7.5
2004	**71.1**	**11.2**
2005	70.0	15.2
2006	68.9	19.2
2007	**68.5**	**22.8**
2008	66.2	26.5
2009	64.3	29.5
2010	63.2	32.4
2011	**62.0**	**35.9**
2012	62.6	38.2
2013	64.0	40.0
2014	**65.7**	**41.2**

Figure 3: The Global Digital Divide 2000-2014

Cho [4] has also suggested that Gini coefficients should be used to measure the various aspects of the global digital divide.

Kelly, in [10], reports Cho's finding, that the global divide for Internet users, as measured by Gini index, reduced from 0.728 in 1999 to 0.618 in 2003. "However," he says, "other evidence [15] suggests that the progress in reducing the digital divide has occurred mainly as a result of middle-income countries catching up, whereas some of the least developed countries have actually been falling behind." Kelly advocates the use of other measures.

Nevertheless, the Information Economy Report 2009, from the United Nations Conference on Trade and Development uses the Gini Index to justify a claim that, "Inequality is shrinking" ([9] p. 16).

Reprise.

As the NTIA report points out (p. 88), the justification of the Gini index rests on the fact that,

> In a situation of absolute inequality (in which only one person had all income), the Lorenz curve would run from (0,0) along the X axis until virtually (1,0) and then abruptly rise to (1,1).

Where Gini, as applied by the NTIA report, would divide λ by $p/2$, we divide by $pq/2$. The additional factor of $1/q$, relating our φ to the Gini coefficient, is vital when we compare inequalities between populations with different levels

Year	1985	1990	1994	1998	2002
Uptake	8.2%	15.9%	22.6%	36.6%	56.5%
Gini index	0.44	0.40	0.39	0.31	0.23
φ inequality	47.9%	47.6%	50.4%	48.9%%	53.1%

Figure 4: Inequalities in home computer uptake

of broadband penetration, or inequalities between different years for the same population as penetration increases.

For example, the NTIA report tells us that the Gini index for Households with a Computer, plotted against Income, fell over the period 1985–2002. However, combining the NTIA uptake data for Households with a Computer (*op. cit.* Fig. 1.1) with their figures for the corresponding Gini index (*op. cit.* Fig. 9.3), we see in Figure 4, that from 1985 to 2002, our inequality index, φ, rose and fell, with an overall *increase* from around 48% in 1985 to almost 53% in 2002.

From a mathematical perspective, our index is closely related to the Gini coefficient, as discussed. It is also related to the Mann-Whitney-Wilcoxon rank sum test [12] applied to test whether the SIMD rank of an offline household is stochastically larger from that that of an online household, and the Lorenz curve is the ROC curve [5] for the use of SIMD rank to classify online/offline status.

We have already criticised earlier uses of the Gini coefficient to quantify the digital divide. Now we consider another measure, which, it turns out, suffers from the same defect.

The digital divide is not the only inequality of interest to policy makers. Health, for example, normally takes priority over digital inclusion. So others have considered how to measure the level of health inequality locally. Low & Low [11] recommend a relative form of the slope index of equality (SII).

This is computed from a regression line for the dependence of the probability of inclusion on deprivation. The regression gives a line characterised by two extreme values for the probability of inclusion: a for the most deprived sections of society, with rank 0, and b for the least deprived, with rank 1. The relative SII index is given by $\text{SII} = (b - a)/p$, where $p = (a + b)/2$ is the population probability of inclusion.

For this model distribution, the Lorenz area can be calculated to be $\lambda = (b - a)/12$. The Gini index would be $\gamma = (b-a)/6p$, whereas our index is given by $\varphi = (b-a)/6pq$, where $q = 1 - p$. Apart from a constant factor, Gini and SII are equivalent for the regression model. Our index introduces a further factor of $1/q$. This means that our index recognises the strength of any dependence on deprivation even when the overall volume of exclusion is small.

A regression model sometimes fits reasonably well for the distribution of broadband uptake. For example, the Royal Society of Edinburgh's inquiry on Digital Inclusion [6] used Ofcom's 2013 postcode-level data on numbers of connections, combined with census and other demographic data,[6] to show that there was a strong correlation between digital exclusion and various other factors of deprivation. Their regression line for broadband uptake against SIMD decile, using the Ofcom data for 2013, with $R^2 = 0.9945$, corresponds to a model with $a = 53\%$ and $b = 83\%$. This regression line has a φ-index of 23%, which approximates our population value, computed from the same data, of 21.5%.

In such cases, we might well compute an approximate value for φ directly from the regression parameters, using the simple equation above (although a direct computation of φ from the underlying data is simple, and preferable). However, a regression line cannot capture the difference between a sigmoid step and a linear ramp, two quite different models for the dependency of uptake on deprivation.

6. DATA

Detailed data on broadband connections is recorded by service providers, for their own business purposes. In the UK some of this data is published by Ofcom, and available for analysis.[7]

The Scottish Index of Multiple Deprivation identifies small area concentrations of multiple deprivation across all of Scotland in a consistent way. It is intended to allow effective targeting of policies and funding where the aim is to wholly or partly tackle or take account of area concentrations of multiple deprivation. The SIMD ranks small areas (called datazones) from most deprived (ranked 1) to least deprived (ranked 6,505).

The Information Services Division (ISD) of the Scottish Government produces periodically a postcode reference file, which combines SIMD data with census data, in particular household counts, down to postcode level. Ofcom postcode-level data for 2013 and 2014 provides counts of households online for most postcodes. We join these two datasets to derive proportions of households online.

For any region, for example a particular local authority, we can then group the postcodes within that region by SIMD rank and for each grouping compute the counts of households on- and off-line. This allows us to compute our index directly from the data.

The data for Figure 3 is taken from published sources. Numbers of connections are taken from ITU Fixed Broadband Subscriptions data,[8] with linear interpolation of missing data. Numbers of households are computed, by division, from World Bank Total Population data[9], combined with household size data for 68 countries from 2000–2012 assembled and interpolated by TekCarta[10], which we have extrapolated to 2013-14.

Caveats.

The data available for this study has various shortcomings, the most obvious of which we describe briefly. First, although the Ofcom data refers to specific months in 2013 and 2014; while both the census data giving household counts, and the SIMD data, date from 2012. Furthermore, Ofcom data is not given for postcodes with very few households.

We have restricted our analysis of the digital divide in Scotland to postcodes for which Ofcom gives data in both years, and have used the 2012 data for SIMD and household counts to compute penetration and deprivation for both years.

The Ofcom data is derived from reports by the largest UK

[6] `http://www.isdscotland.org/Products-and-Services/GPD-Support/Deprivation/SIMD/`

[7] There are roughly 1.5 million UK postcodes, about 220,000 of which lie in Scotland.

[8] `http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2015/Fixed_broadband_2000-2014.xls`

[9] `http://data.worldbank.org/indicator/SP.POP.TOTL`

[10] `http://www.generatorresearch.com/tekcarta/databank/households-average-household-size/`

Internet Service Providers, who accounted for 90% of connections in 2013 and 89% in 2014, across the whole UK.[11] We have not made any adjustments for this, so our results refer to inequalities of domestic access provided by the largest providers. Other providers may have disproportionate significance in some localities, particularly in rural settings, and this data cannot capture such effects. A small number of postcodes are split across datazones. The ISD data includes only the largest component of any split postcode. We ignore the remaining fragments.

The alert reader will have observed that the Ofcom statistics show *reductions* in broadband uptake from 2013–2014 for some local authorities. We surmise that these may represent defections from the larger providers to smaller ISPs that do not report data to Ofcom, or result from population changes that are not reflected in the data available. To the extent that any such changes are strongly correlated with deprivation, they may also have an effect on φ.

The data we have used to assess the global divide come from various sources, which will inevitably give rise to differences and errors in the estimation of the numbers of both households and fixed broadband connections. These will result in systematic errors in our estimation of φ. In four countries, Taiwan, Philippines, Bahrain, and Lebanon, the data for 2014 shows more connections than households, so we show no households offline. To the extent that such errors are systematic, our assessment of the changes in the digital divide will be more robust than the annual estimates of uptake.

7. FURTHER WORK

In further work we plan to extend this analysis in space, time and detail — to other countries, to the data from future years, and to examine the dependence of the digital divide on different factors of deprivation.

8. REFERENCES

[1] Brown, M. Using Gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data. *Soc Sci Med. 38*, 9 (1994), 1234–56.

[2] Ceriani, L., and Verme, P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J.Econ.Inequal* (2012).

[3] Chakraborty, J., and Bosman, M. M. Measuring the digital divide in the United States: Race, income, and personal computer ownership. *The Professional Geographer* (2005). doi: 10.1111/j.0033-0124.2005.00486.x.

[4] Cho, C. How to measure the digital divide? ITU/KADO Symposium on Building Digital Bridges, September 2004. `http://www.itu.int/osg/spu/ni/digitalbridges/presentations/02-Cho-Background.pdf`.

[5] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters 27* (2006), 861–874.

[6] Fourman, M., Alexander, A., Bechhofer, F., Brown, J., Macaskill, N., Moore, J., Osborne, N., Ritchie, I., Skerrit, S., Wade, M., and Yiu, C. Digital Scotland: spreading the benefits of digital participation. Royal Society of Edinburgh, April 2014.

[7] Hanafizadeh, M. R., Hanafizadeh, P., and Bohlin, E. Digital divide and e-readiness: Trends and gaps. *International Journal of E-Adoption 5*, 3 (2013), 30–75. doi:10.4018/ijea.2013070103.

[8] Howard, P. N., Busch, L., and Sheets, P. Comparing digital divides: Internet access and social inequality in Canada and the United States. *Canadian Journal of Communication 35* (2010), 109–128.

[9] ICT Analysis Section. Information economy report 2009: Trends and outlook in turbulent times. United Nations Conference on Trade and Development, 2009. `http://unctad.org/en/docs/ier2009_en.pdf`.

[10] Kelly, T. Twenty years of measuring the missing link. In *Maitland+20: Fixing the missing link*, G. Milward-Oliver, Ed. Anima Centre Limited, Bradford on Avon, 2005, pp. 23–33. `http://www.itu.int/dms_pub/itu-s/oth/02/0B/S020B0000014E09PDFE.PDF`.

[11] Low, A., and Low, A. Measuring the gap: quantifying and comparing local health inequalities. *Journal of Public Health 26*, 4 (2004), 388–395. doi:10.1093/pubmed/fdh175.

[12] Mann, H. B., and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics 18*, 1 (1947), 50–60.

[13] Norris, P. *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide.* CUP, September 2001. ISBN: 0521002230 `http://www.utwente.nl/gw/mco/bestanden/digitaldivide.pdf`.

[14] Ragnedda, M., and Muschert, G. W., Eds. *The Digital Divide: The Internet and Social Inequality in International Perspective.* Routledge Advances in Sociology. Routledge, June 2013. ISBN: 978-0-415-52544-2.

[15] Sciadas, G., Ed. *Monitoring the Digital Divide ... and beyond.* Claude-Yves Charron, in association with NRC Press, Canada, 2003. `http://orbicom.ca/upload/files/research_projects/2003_dd_pdf_en.pdf`.

[16] van Dijk, J. One Europe, digitally divided. In *The Handbook of Internet Politics*, A. Chadwick and P. N. Howard, Eds. Routledge, 2008.

[17] Victory, N. J., and Cooper, K. B. A nation online: How Americans are expanding their use of the Internet, February 2002. `http://www.ntia.doc.gov/legacy/ntiahome/dn/Nation_Online.pdf`.

[11] `http://media.ofcom.org.uk/facts/`