Able but not willing?

Information Technology & People

Able but not willing? Exploring divides in digital versus physical payment use in China

Ashley D. Lloyd Mario Antonioletti Terence M. Sloan

Received 24 October 2014
Revised 17 June 2015
Accepted 22 March 2016

Able but not willing? Exploring divides in digital versus physical payment use in China

Ashley D. Lloyd
Business School, The University of Edinburgh, Edinburgh, UK, and
Mario Antonioletti and Terence M. Sloan
EPCC, The University of Edinburgh, Edinburgh, UK

Abstract

Purpose – China is the world's largest user market for digital technologies and experiencing unprecedented rates of rural-urban migration set to create the world's first "urban billion". This is an important context for studying nuanced adoption behaviours that define a digital divide. Large-scale studies are required to determine what behaviours exist in such populations, but can offer limited ability to draw inferences about why. The purpose of this paper is to report a large-scale study inside China that probes a nuanced "digital divide" behaviour: consumer demographics indicating ability to pay by electronic means but behaviour suggesting lack of willingness to do so, and extends current demographics to help explain this.

Design/methodology/approach – The authors report trans-national access to commercial "Big Data" inside China capturing the demographics and consumption of millions of consumers across a wide range of physical and digital market channels. Focusing on one urban location we combine traditional demographics with a new measure that reflecting migration: "Distance from Home", and use data-mining techniques to develop a model that predicts use behaviour.

Findings – Use behaviour is predictable. Most use is explained by value of the transaction. "Distance from Home" is more predictive of technology use than traditional demographics.

Research limitations/implications – Results suggest traditional demographics are insufficient to explain "why" use/non-use occurs and hence an insufficient basis to formulate and target government policy.

Originality/value – The authors understand this to be the first large-scale trans-national study of use/non-use of digital channels within China, and the first study of the impact of distance on ICT adoption.

Keywords Digital divide, Technology adoption, E-inclusion/exclusion, E-science, Grid computing

Paper type Research paper

Introduction

In 2004 China commenced the "Village Access Project" to provide basic telephone services to rural areas and by 2008 had achieved connection of at least two workable telephone lines to 99.5 per cent of its total administrative villages (Xia and Lu, 2008). In the same year China became home to the world's largest national population of mobile phone users (Cellular News, 2008) and the world's largest national population of internet users (Macfie, 2008).

Information Technology & People
Vol. 29 No. 2, 2016
pp. 250-279
Emerald Group Publishing Limited
0959-3845
DOI 10.1108/ITP-10-2014-0243

The geographical contrasts implied by the above figures have made China the subject of numerous "digital divide" studies. Though the origins of this term are found in the design of digital communications (Kronmiller and Baghdady, 1966), its use in relation to societal access to digital communications and computers was first made popular in 1998 in the US National Telecommunications and Information Administration "Digital Divide" report (McConnaughey *et al.*, 1998). This was long after the importance of such access and use by all members of society had been recognised in the legislature, most notably in the "High Performance Computing Act of 1991" (US Congress, 1991). This Act included significant appropriations to support a National Research and Education Network and supported the work of the US National Centre for Supercomputing Applications from which emerged the Mosaic browser widely credited with accelerating diffusion of the web (Schatz and Hardin, 1994).

Academic studies of regional digital divides have emphasised inter-state and intra-state differences in access to infrastructure as an explanatory variable. Inter-state comparisons between countries at different levels of economic and technical development have tended to confirm a positive relationship between investment in infrastructure and social and economic benefits (Marton and Singh, 1991; Brown *et al.*, 1995; Castells, 1996; Chen and Wellman, 2004; Warschauer, 2010). Studies that probe intra-state divides have shown that these macro-level relationships break down as we start to resolve local contexts, where a positive investment in infrastructure may be either negatively or positively correlated with the scale of the divide. This divergence is ascribed to more finely grained socio-economic factors that affect both organisational (Forman *et al.*, 2002) and community uptake of digital technologies, ranging from barriers such as computer literacy (Light, 2001; Cushman and McLean, 2008; Zheng and Walsham, 2008; Bach *et al.*, 2013; Choi and DiNitto, 2013) to the perceived value of the information products and services (Hoffman *et al.* 2000).

Such "digital divides" are often unhelpfully treated as binary divides between "haves" and "have nots" (Norris, 2001) where empirical observations sit in categories defined by a matrix combining multiple definitions of "*who*, with *which* characteristics, connects *how* to *what*?" (Hilbert, 2011). As this lens becomes more discriminating, for example, by more detailed segmentation of the population into smaller, more numerous groups, the permutations necessarily increase exponentially. This allows very finely grained distinctions between what defines membership of one side of the binary divide vs the other, but potentially obscures the far more nuanced inequalities and associated behaviours that arise from membership of the "have less" category (Qiu, 2009).

Not only will a binary lens miss these more nuanced forms of digital exclusion, they cannot be countered with policy instruments derived from that initial assumption. Policy instruments based on a binary lens will focus on what is generally true for most of the population rather than exploration of barriers affecting those at the margins of society who may be found at the extremes of the population distribution. This should improve the situation for the majority but this may increase the relative divide for those who are missed.

For example, whilst computer access and computer literacy are important explanatory variables for modelling a binary digital divide, even if these factors are universally addressed a digital divide will not only remain, it may grow (Bach *et al.*, 2013). Indeed the increases in socio-economic inequality and social exclusion in the USA (Bach *et al.*, 2013) since the High Performance Computing Act of 1991 (US Congress, 1991) raises similar cautions about the evidence base from which policy assumptions about computer use are derived to those posed by Solow (1987) in relation

to industry's use of computers that "you can see the computer age everywhere but in the productivity statistics". Bach *et al.* (2013) highlight this disconnect in relation to broadband policy, citing Eubanks' (2011) conclusion that the validity of key assumptions were "dangerously unexamined".

Once the most obvious inequalities of access are addressed for the majority, there is the further problem that the binary lens may prove too crude to even resolve populations for whom a digital divide remains. There is some evidence of this in the UK where 90 per cent of the population are now classed as internet users (World Bank, 2015). Here detailed studies of media use and attitudes reported by the UK Communications Regulator, Ofcom, as part of their obligations under the UK 2003 Communications Act, note "one in eleven children do not use the internet at all, in any location" but that "no particular socio-economic group is more likely not to use the internet at all" (Ofcom, 2012). That such a large population of non-users cannot be discriminated on any socio-economic criteria suggests that the current demographic analysis needs to be refined.

In this paper we focus on the challenge of trying to "see" and understand these more nuanced forms of digital exclusion in very large populations. We approach this in two ways: through computational methods that scale efficiently as we study larger populations, and also by extending the socio-economic lens we use to try and improve the resolution with which we identify more nuanced behaviours.

To access the required quality and quantity of data we work with a very large commercial company in China. Though this introduction has framed the research in a policy domain, this choice recognises that exclusion is countered by a better understanding of user needs by both policymakers and by business. In particular the latter, as business should be able to respond very quickly to the opportunity to convert a "non-user" to a "new user" as they re-focus business processes from what one of the pioneers of internet search engines, Joe Kraus, described as "dozens of markets of millions of consumers" to "millions of markets of dozens of consumers" (Day, 2013).

Preserving the "visibility" of excluded user needs as markets globalise and suppliers grow in scale and efficiency

The fastest route to inclusion is identification of a new and unsatisfied combination of user requirements that characterise a group large enough to be commercially viable to design and produce for.

However, as domestic suppliers search for the more nuanced segments that characterise non-use, these groups tend to get progressively smaller and hence diminishing returns eventually limits the proportion of a national population who will have their needs addressed in this way (Verdegem and Verhoest, 2009). This proportion has some geometric dependency on the size of the domestic market as when the market segments of "dozens of consumers" anticipated by Kraus (Day, 2013) become commercially viable in a country of the size of the UK, the same characteristics may define a more profitable segment containing thousands of consumers within a country of the size of China (ONS, 2014; World Bank 2014).

Globalisation of markets therefore offers the promise of "scaling up" domestic demand to include similar user needs in much larger populations, giving suppliers potential economies of scale, reducing pressure on prices and hence accelerating inclusion. However, greater economies of scale in global markets that are open to competition also increase the minimum efficient scale required to be competitive and results in a market "shakeout" that reduces the spread and diversity of competitors (Teece, 1993).

Though the dot.com bubble provides an extreme example of such a "shakeout", with the NASDAQ composite historical peak still marking that event 14 years later (NASDAQ, 2014), the general point identified by Gort and Klepper's (1982) analysis of the diffusion of 46 new products introduced between 1887 and 1960 is that "the structure of markets (in terms of number and composition of producers) is shaped, to an important degree, by discrete events such as technical change and the flow of information among existing and potential producers". Such analysis today would draw on increasing recognition, and encouragement, of the active role of the consumer as a cooperative part of the production process, resonating with Parsons' (1951) "relationships of cooperation with others in the same 'productive' process", and reflected in the "professional-consumer dialectic" of Gartner and Reissman (1974) (Huber, 1976), Toffler's (1980) "prosumer" and Prahalad and Ramaswamy's (2000) value co-creation.

The impact of active consumers on market structure and diffusion of products and services is "most obvious in the digital domain" (Ritzer, 2014) as in a global digital economy, technical change and the flow of information are strongly coupled. Active consumers are a highly visible part of the business process and act to ensure that the output of the co-production process takes account of their needs and values. Non-users however are not a visible part of the process, and if Ritzer's (2014) view of the future impact of the Internet of Things is correct, with ad hoc networks of machines acting as negotiating proxies for digitally engaged human "prosumers", then non-users are at risk of increasing marginalisation across a widening digital divide.

This gives us two basic categories with which to explore the digital divide in large-scale consumer data: users who are "digital-visible" with varying degrees of activity, and non-users who are "digital-invisible". Note that, whilst all non-users are "invisible" to a digital economy in which consumption signals preferences, it does not follow that non-users are necessarily at risk of exclusion as they may be electing to use alternative technologies, and be able to afford to do so regardless of any increased cost. The digital-excluded are thus a subset of the "digital-invisible" and need to be distinguished from those who elect to be non-users if policy instruments are to be targeted effectively.

We must also take care with how we define use of a digital technology that leads to visible engagement with the digital economy. We define "digital visibility" as use that explicitly and synchronously establishes a personally identifiable digital trace in the consumption record. For example, a consumer may use mail or the internet to order the same product, from the same supplier, using the same payment method. In both cases a data trace is established in the invoice order processing system, but it is only by using the internet that the consumer synchronously establishes and validates that trace. With mail order the record is established and validated instead by the company.

As digital markets globalise, shakeouts increase the scale and efficiency of suppliers, and users become more actively engaged in co-production, a key policy concern is whether the needs of the potentially excluded become more or less "visible" in Gort and Klepper's (1982) "flows of information" and can therefore help shape the market?

Here, geometric considerations offer a further insight. As the geographic boundaries of a market expand and scale returns reduce the number of competing producers, the result is a necessary increase in the average physical distance between suppliers to a market and between suppliers and consumers. Though consumers are presented with a range of direct and indirect market channels it is instructive to contrast Coughlan's

(1985) highly cited work on market channel structure that indicated profit maximisation by use of indirect "marketing middlemen" with the 2014 poll of 580 senior marketing executives by Accenture that found one third predicting digital market channels to account for 75 per cent of their media budgets by 2019 (Parsons, 2014). Coughlan's (1985) analysis had ignored the impact of economies of scope or scale whilst digital market channels offer both and are able to dis-intermediate other parties by using (Parsons, 2014): "data to target, data to personalise, data to analyse return on investment. In other words – direct marketing". Digital market channels are therefore evolving to become direct market channels.

There is no natural limit to the scale efficiencies offered by direct market channels in a global digital market. Ultimately this means that the physical distances between producer and consumer can be of the same scale as the market: global, with consequences for market structure that may be inferred from Chandler's observation, quoted in Teece (1993), that:

> The critical entrepreneurial act was not the invention-or even the commercialization-of a new or greatly improved product or process. Instead it was the construction of a plant of the optimal size required to exploit fully the economies of scale or those of scope, or both.

The same objectives are reflected in Oracle's consolidation of its 40 data centres in 2006 to two major facilities in the USA (Oracle, 2014), and the Range International Information Group's collaboration with IBM to build a data centre in China, planned for completion in 2016, to host China's global cloud computing services and set to be the size of the Pentagon (Range, 2013; IBM, 2011).

The potential significance of this increase in physical distance on the characteristics of the digital divide reflects its impact on Gort and Klepper's (1982) "flows of information" among existing and potential suppliers and customers. The introduction of such "factors preventing or disturbing the flow of information between potential and actual suppliers and customers" may be understood as increasing a "psychic distance" that impedes understanding of one party by the other (Evans *et al.*, 2000; Evans and Mavondo, 2002; Johanson and Vahlne, 1977; Johanson and Wiedersheim-Paul, 1975; Vahlne and Wiedersheim-Paul, 1973).

Though the value of psychic distance as a management tool has been contested (Stöttinger and Schlegelmilch, 2000), it is clear that it reflects impediments to understanding why certain behaviours are expressed. It is also generally acknowledged that geographical distance is one of the most significant antecedents of psychic distance: "Indeed, simple geographical distance turns out to be more than three times as important as cultural distance" (Håkanson and Ambos, 2010). Håkanson and Ambos (2010) also conclude that information flows, and hence competitiveness, can be highly sensitive to geographical proximity, citing the case of a German-domiciled pharmaceutical company that ascribed significant improvements to its competitiveness by locating its "Eastern Europe" hub a relatively short distance across a common border into Austria, rather than anywhere in the former East Germany.

If we accept the general points that increased geographical distance has a negative impact on information flow and that in a "direct digital economy" this is compensated for using "data to target, data to personalise, data to analyse return on investment" (Parsons, 2014), then it appears evident that as geographic distance increases, product development will be increasingly informed by "actual customers" on one side of the digital divide rather than the "potential customers" on the other.

A recent example of this is the development of Google Wave, a product designed to improve upon e-mail as a collaboration tool by combining features of shared document applications and social media that had already established significant user communities. "Wave" was released to developers and invited users by Google in 2009, opened to the general public in May 2010, with a decision to discontinue the product advertised just three months later (Wave, 2013). The relatively limited opportunity to bring in "potential users" who were new to these technologies reflected a strategy for market expansion that framed this product's development in terms of combining the expressed needs of existing users of competing and complementary products that might migrate to "Wave". The Google CEO Eric Schmidt articulated this strategy in the year "Wave" was first released (Arrington, 2009): "We've always taken the position of we want to do things that matter to a large number of people at scale. […] We don't want to work on problems that only affect a small number of people".

This strategic focus on large segments described by relatively homogeneous needs in a globalising market is logical for companies with the potential to be global-scale suppliers. It is also a strategy likely to receive positive feedback in a globalising market, as any sensitivity to physical distance that might negatively impact market segmentation and potential market share can be masked by sales growth due to that underlying market expansion. This has two potentially adverse consequences for the digital divide:

(1) A supplier will tend to promote growth through targeting more consumers that display the same demand behaviour, thus ignoring "non-users" who are excluded both at home and abroad.

(2) Since the supplier only samples one part of the distribution of demand characteristics, i.e. existing users of direct digital market channels, they are less able to understand "what" is distinctive about their customer group, "why" this consumption behaviour is being expressed and "how" it might evolve in future. This makes it harder for suppliers to identify new user requirements amongst existing users and unsatisfied requirements amongst non-users.

This latter point may appear counter-intuitive as companies such as Google and Amazon have demonstrated machine-learning approaches that deliver commercial value from establishing "inference free" associations between product purchases, and between product purchases and consumer demographics. However, despite claims that machine learning "provides the effective means for modeling and predicting a human subject's desires" (Chatzis and Demiris, 2012) such approaches focus on existing users in large groups for pragmatic and computational scalability reasons (Chatzis and Demiris, 2012; Zegarra and Efremenko, 2011). It is this focus on similarity and distinctiveness amongst existing users that limits any ability to infer that non-users have different demographics, nor that such differences that may exist necessarily explain, far less predict, why "a human subject" is a non-user of services like Google and Amazon. It follows that if non-users are not understood, then users are not fully understood and hence that product development is adversely impacted for both groups.

In summary, whilst globalising markets offer enhanced prospects for excluded segments to meet the test of commercial viability and hence for the digital divide to be reduced, we argue that this will not occur if scale efficiencies impede the flow of information required to identify non-user needs, or global-scale suppliers adopt

strategies that reduce relative investment in understanding more nuanced user requirements. Making the requirements of non-users more "visible" in a global digital economy is an important focus for policy interventions, with potential benefits for both user and non-user groups. We address this in the present study by focusing on digital payment technologies as a key enabler of access to, and visibility within, China's digital economy. In what follows we attempt to probe more nuanced user needs found in populations with demographics indicating an ability to pay by digital methods combined with behaviour that suggests a lack of willingness to do so.

Method and data

This paper sets out to explore nuanced consumer behaviour amongst non-users of digital payment technologies in the context of a globalising digital economy. Such behaviours may prove to be exhibited by a very small proportion of the overall population, described in statistical terms as "outliers". The difficulty of using small samples of an unknown distribution to identify outliers in the overall population has long been recognised (Hume, 1739) and hence the first requirement of the data for this study was that it should be large in scale relative to the population being studied.

The second requirement of the data is that it should capture the behaviour of users and non-users, contain demographic data that may allow these groups to be distinguished, and keep other factors as comparable as possible. In relation to use and non-use of digital payment technologies, this means that the data should arise from consumption of the same products and services from the same supplier, where both use and non-use of digital payment technologies are viable alternatives.

The location of the work in China addresses scale requirements in both market size and geographic extent, setting it in the context of the largest synchronous electronic market in the world: a single time-zone encompassing nearly one quarter of the world's population and the majority of current global economic growth (Figure 1).

The infrastructure required to combine expertise, confidential data and distributed computational resources from both Europe and China was established through complementary investments in grid computing facilities by the Chinese Academy of Sciences (CAS) and the UK Economic and Social Research Council (see Acknowledgements). This was enhanced by the development of the EU-funded Trans-Eurasia Information Network: the "first large-scale research and education network for the Asia-Pacific", giving more network bandwidth between the EU and China, and much shorter network latency (TEIN, 2014), both of which improve the efficiency of distributed computing interactions.

Building on an established collaboration with the Computer Network and Information Center of the CAS, access was brokered to a company within China that had sufficient scope and scale of operations to meet the above data requirements. The general characteristics of this company's operational data were:

(1) consumption and market channel data for millions of consumers across China over a period of years;

(2) a diverse range of alternative market channels, including "digital" options such as online sales and electronic payments, and "physical" options such as shops and cash, allowing consumption records of both users and non-users of digital channels to be captured; and

Note: Areas depicted in purple are outside the time zone but are included to aid country identification

Figure 1.
China and Western Australia define the extrema of the most populous time zone in the world (depicted in red) containing the majority of current global economic growth

(3) a set of products responding to diverse consumer preferences, including "premium" and "value" offerings across product lines that range from highly differentiated branded products to generic consumables, with promotional campaigns targeting a wide range of consumer demographics.

Though a similar scope and scale of data might be found in a number of countries, a particularly distinctive quality of this data arises from China's preference for "Face to Face" transactions (Haley, 2002), and in this case the requirement for all customers to have a validated account prior to purchase. This allowed all transactions to be unambiguously linked to a single consumer regardless of their timing, market channel or payment type, and hence for each individual consumption trace to be linked to a consumer's demographic data. These demographic data are listed in Table I, and include a new demographic: "Distance from home", that arises from the ability to associate each consumer with a government-registered address.

Though the collaboration offers access to very rich and large-scale data, it does not follow that a result from the study is generalisable. The data requirements articulated

above have the principal objective of improving the ability to resolve nuanced behaviours by setting them against a background where variance in other factors is deliberately minimised. It is for this reason when we introduce "Distance from home" as a new demographic that can range in any direction across the whole of China, that we deliberately focus on one urban location so that its role can be more clearly articulated. Commercial confidentiality prevents the precise location from being disclosed, however the city falls within the "Tier 2" category (Atsmon *et al.*, 2012) and is an example of a developed Chinese city with significant growth potential.

The following observations arise from studying all the active customers at this location, a population measured in tens of thousands, however, the preceding "ceteris paribus clauses" (Kincaid, 1996; Earman and Roberts, 1999) potentially limit the generalisability of resulting observations and hence we classify this as exploratory research.

Building a scalable, collaborative platform to access and analyse "Big Data"

The collaborating company's decision to give access to the above data was contingent on availability of a computing resource that was compatible with the scale of the data, and complied with corporate security requirements. Available computing models included highly scalable cloud solutions with third-party providers and an established grid facility in partnership with the CAS running across the TEIN infrastructure (TEIN, 2014). Though both models met the scale requirements, these distributed computing solutions move the data to the computation and hence challenge "data sovereignty": the requirement of the company to keep their data within set geographical boundaries and a single legal jurisdiction.

The solution was to move the computation to the data and build a collaborative platform that was embedded within the company's core data centre (Lloyd *et al.*, 2013). The platform was subjected to a process of testing and certification to ensure that it aligned with existing corporate information technology and communications policies, including International Organization for Standardization certification, and could be managed within existing business processes and reporting structures.

Once system testing and certification had been completed, data were extracted as a sequence of tables from the master record to construct a separate image of the

Table I. Data on demographics of sample population with the "Tier 2" city (Atsmon *et al.*, 2012) used in the present study

Attribute	Value range
Demographics	
Age	18-88
Gender	M(ale)/F(emale)
Spouse age	18-88 N/A recorded if no spouse listed
Spouse Gender	M(ale)/F(emale) N/A recorded if no spouse listed
Education	Middle school, high school, junior college and up N/A recorded if no education level disclosed
Occupation	Retired, unemployed, engineer/technician, sales personnel, freelance, self-employed, general staff, manager
Distance from home	0-3,300 kilometre (see Figure 6 for distribution)

corporate database. This facilitated data abstractions via SQL queries or more complex transformations via PERL scripts, and enabled all data relationships to be verified.

Data cleaning initially proceeded by searching for inconsistencies, such as calendar errors, or malformed/missing characters, such as postcodes that had inappropriate values or were too short. Whilst this is an important step in data cleaning, and can be largely automated, it is important to note that the analytical approach being used here (discussed in the following section) searches for meaningful patterns in the data. If the data is incorrect then the patterns will be meaningless, however if patterns are dominated by the way the data have been collected and collated, rather than the underlying behaviour, then their meaning is obscured and no inferences about why behaviour occurs should be drawn. Such patterns can be very strong, for example, when a company grows through acquisition, the customers it inherits in new geographical locations may exhibit very different behaviours to those it has attracted to its own brand. After systems integration this data provenance can be lost, yet is key to understanding why geographical location explains behavioural differences statistically, but offers no meaningful insight into why such behaviour exists.

To help identify such artefacts required a review of the business processes within the company, from sales to order fulfilment. This involved interviews with analysts and senior management, including the Chief Information Officer, to get a top-level view of business processes and, in particular, any historical changes that may affect the current data record. These views were then resolved against "bottom up" perspectives on the same business processes through process walk-throughs and interviews with staff in production, dispatch and in a shop setting where the customer experience of sales promotion and the purchasing process could be understood.

This exploration of the link between the data and the business processes was highly iterative, because anomalies could arise at any stage of the data cleaning and analysis, and necessarily collaborative, because such anomalies needed to be resolved quickly. A 7-8 hour time difference between collaborating sites gave an opportunity to reduce cycle times with overlapping working day slots used for conferring and to pass queries over for resolution by the next working day. This was supported by on-site face-to-face training in systems management and use of the data analysis tools, with every step in the analysis logged for audit or replay, allowing the identification of any anomalies to be replicated and analysed independently.

Once this data cleaning was complete we were able to test whether the data set would allow the modelling and prediction of the nuanced forms of behaviour that were the focus of the project, using analytical methods that that scaled up to "Big Data".

Mining "Big Data" and modelling behaviour

"Big data" and "data mining" have been presented as components of a new, "fourth paradigm" of science (Hey *et al.*, 2009) that elevates the importance of the empiricist epistemology. Unfortunately this elevation has been used by some to frame a polarised debate in which inductive and deductive research methods are criticised as if they were mutually exclusive substitutes rather than complements, leading to some extreme conclusions about data mining (Anderson, 2008): "There is now a better way. Petabytes allow us to say: 'Correlation is enough'. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot".

Correlation can be "enough" to deliver business value, such as the 700 per cent increase in sales of both toaster pastries and beer prior to hurricane activity observed by Wal-Mart (Hays, 2004) that allows Wal-Mart to use weather forecasts to reduce stock-outs and improve total sales performance. However it is the ability to infer something from such correlations that engages a wider range of corporate expertise in formulating a potentially more effective organisational response.

This is exemplified by the often-quoted, story of Wal-Mart discovering a correlated increase in sales of diapers and beer on Friday night. This was explained by Hayes (1996) as "men in their 20s who purchase beer on Fridays after work are also likely to buy a pack of diapers" and reportedly led to a decision to promote both products in the same aisle location that, in turn, resulted in improved sales. Though entirely apocryphal (Power, 2002), the story illustrates how a correlated "what" and "when" becomes an opportunity to promote sales only after the socio-demographic contexts of age, gender and employment are added so that management can draw inferences about "why" the behaviour occurred.

In the present study we take a complementary approach, using inductive methods but employing deduction to steer them. For example, deduction is evident in the way the sample has been constructed to help focus the study, reflected in the "ceteris paribus clauses" discussed earlier. Deduction is also used in the final stages of data cleaning to test hypotheses about behaviours derived from interviews with experienced senior management. This tests whether recognised behaviours can be modelled and predicted using a "data-mining" approach, and whether such models can be contextualised and interpreted by management to support decision making. Examples from earlier large-scale studies in the financial services and telecommunications sectors range from simple calculation of the average length of time a mortgage is held, to modelling the demographics of "high-churn" customer groups. In all cases the analyses are performed without access to the earlier studies that they are expected to confirm to allow a collaborator to independently validate the results.

Following this validation process we take a more inductive research approach to explore behaviour in the sample, employing a highly scalable (multi-threaded) data-mining tool, C5.0 (Rulequest Research, 2012), to discover whether a range of use/non-use behaviour can be classified in terms of distinctive socio-demographics.

C5.0 is, in computing terms, a highly scalable development of C4.5, ranked as the most influential data-mining algorithm at the IEEE International Conference on Data Mining in 2006 (Wu *et al.*, 2008). C5.0 has been validated in a very wide range of application domains at large scale, including remote weather assessment by the US Naval Research Laboratory (Bankert and Hadjimichael, 2007), satellite image processing by the US Geological Survey (Fry *et al.*, 2009), supply chain management (Emerson *et al.*, 2009), the "value" of air passengers (Chiang, 2012), economic modelling (Chen *et al.*, 2015), fraud detection in banking (Song *et al.*, 2014), pricing in financial markets (Basti *et al.*, 2015), bioinformatics (Chu *et al.*, 2014; Shao *et al.*, 2015), geoscience (Akkaş *et al.*, 2015) and numerous diagnosis and patient management applications in medicine (Sanga *et al.*, 2009; Gupta *et al.*, 2010; Gilchrist *et al.*, 2011; Son *et al.*, 2012).

The use of this de facto standard is a deliberate methodological choice given the commercial confidentiality associated with the data source. Since confidentiality prevents direct replication of the results a novel algorithm would tend to increase the number of hidden degrees of freedom and hence reduce the validity of any conceptual replication (Pashler and Harris, 2012).

A key concept in the construction and interpretation of decision trees using C5.0 is "information gain". If, for example, we are looking at the predictors of the uptake of a digital service like eHealth and find (Kontos *et al.*, 2012): "the most significant differences in eHealth use were across SES [socioeconomic status] (either by education, income, or both) and by age and sex". Then it is clear that these attributes must contain information that helps explain behaviour across the sample.

Though Kontos *et al.* (2012) used logistic regression to model the observed behaviour, C5.0 also has to evaluate how much of the observed behaviour is explained by each attribute. C5.0 then splits the sample according to the attribute that provides the most information gain, then splits the subsample by the attribute that provides the most information gain and repeats until the sample cannot be split any further.

The decision tree that results may be very large as the only measure used to steer the process is how well the tree fits the data. A tree that is allowed to grow as large as the data set may classify known objects perfectly but be of little predictive accuracy for new cases as it necessarily reflects any statistical irregularities or idiosyncrasies of the original data (Quinlan and Rivest, 1989). This outcome is described as "over-fitting". For example, if the sample used to build the decision tree contains 100 patients and a tree with 100 "leaf nodes" results, then the ability to predict an outcome for each known patient is perfect, but the model may not be able to predict a likely outcome for a previously unseen patient and hence has little practical value.

With C5.0 therefore, the objective is to minimise the error rate when previously unseen cases are presented and so candidate trees are pruned using heuristics derived from the Minimum Description Length Principle (Rissanen, 1978). This penalises over-parameterisation, and may be thought of as a cautious application of Occam's Razor (Kohavi and Quinlan, 2002) to "shave away" parts of the model that add complexity but have little information value. Returning to the study by Kontos *et al.* (2012), this means that a C5.0 decision tree is likely to include socio-economic status, age and gender, but "shave away" splits based on race/ethnicity given their observation that (Kontos *et al.*, 2012): "Among online adults, there was little evidence of a digital use divide by race/ethnicity"[1].

Pruned, smaller trees are not only expected to have greater predictive accuracy, the tree structure is also easier to follow and the segments (the population represented by each "leaf") showing similar behaviour are necessarily larger. From an economic perspective such segments represent distinctive market segments for which the combination of product/service design criteria is unique, and the more each segment can be grown then the more likely it is to become commercially viable to pursue.

The fully pruned tree not only has the attributes with most information gain at the top, by the same virtue it also tends to have the largest groups defined by higher leaf nodes, leaving the lower leaf nodes to describe smaller groups with more nuanced behaviours. The weight given to each attribute and the resulting order of most significance represented by the tree may change as more data and/or data over longer periods are included, either through improvements in the discriminatory power of the model or gradual changes in market structure in which new forms of behaviour become dominant and old ones die away.

An example of the emergence of new behaviours is the "EU Effect" studied by Bekaert *et al.* (2012) where one impact of market integration and convergence is that industry sector rather than geography has become the most significant determinant of valuation differentials. This reverses a traditional view of market segmentation in Europe and if this causes changes in actual investment behaviour, then applying C5.0

to sequential samples of investment behaviour would yield a series of models in which splits by geography were initially in the position of most significance at the top of the tree, but gradually move down the tree as industry sector gradually moves up.

The implication of this for interpreting tree structures in a single time frame is that the lower branches of a model capture significant behaviours today that may become dominant or die away. Distinguishing between the two is critical for a supplier's product development and strategic positioning, and illustrates why Anderson's (2008) argument that "Correlation is enough" falls short. Emergent and obsolete behaviours may be correlated with equal strength and hence cannot be discriminated in the short term without an attempt to understand "why" these behaviours exist.

Exploring willingness and ability: physical vs digital payment methods

From our sample we selected consumers who had established a pattern of repeat purchases, allowing us to separate out consumers that may have rejected the products being offered rather than the channels.

The company uses cash and cheques as "physical" payments methods and a wider variety of "digital" payment methods that include (Figure 2): bank transfer, credit cards (cleared predominantly through China's Central Bank regulated UnionPay), Electronic Point of Sale and Alipay. Alipay is a privately owned third party online payment service providing services analogous to PayPal in other parts of the world, but with over 700 million user accounts reported at the end of 2012 (Wee, 2012) and claiming more than 8.5 million transactions daily (Alipay, 2013)[2].

This variety of payment methods allowed us to split the population into two sub-sets: "Digital" and "Physical" as proxies for the "Digital Visible" – those who necessarily leave an authenticated digital data trace, and "Digital Invisible" consumers, those whose interactions require physical exchanges and for whom any digital record is created by the company. The behavioural traces over time are illustrated in Figure 2 where the relationship between product choice and payment method is shown.

The confidentiality agreement required for this collaborative analysis of a large commercial database prohibits identification of the precise numbers in each category, however we encountered a common problem in the diffusion of technology: that the populations are heavily skewed towards specific, usually long-established, channels. Accordingly we took care in the model building process to ensure that the samples of both behaviours were balanced.

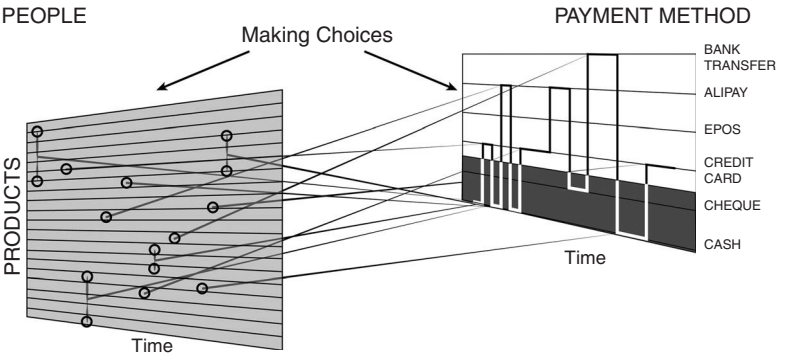


Figure 2.
Relating consumption and payment method over time: establishing "pathways" through sales channel technologies that determine whether a digital trace is established

Other skews were also evident in the sample, for example age and gender, which is plotted in Figure 3, against the demographics for China overall. Here we see a sample in which women between 20 and 50 years of age are significantly over-represented, with men over 40 generally under-represented. From an inductive data mining perspective this type of skew in the sample makes age and gender strong sample descriptors and hence attributes that the C5.0 data-mining algorithm might be expected to pick up as having significant information gain and place high up in the tree resulting from the modelling process.

In fact, C5.0 analysis showed neither age nor gender to be highly significant in terms of explaining observed behaviour, and part of the reason for that is evident when these demographics are viewed in terms of the target behaviour: use of digital vs physical payment methods. This abstraction is shown in Figure 4, where we see that age and gender have much less information value in describing population splits than those shown in Figure 3 and hence are candidates for pruning away from the resulting decision tree.

It is important to note however that age describes more variation in the sample than gender, and interesting to observe that younger men appear to prefer digital payment methods in comparison to women of the same age, a preference that reverses at age 50.

In terms of the remaining traditional demographics of education and occupation, the variation was of a similar order, however, it is not the relatively low value of traditional demographics in explaining behaviour that requires explanation here, rather it is the relatively high explanatory value of a new demographic: "Distance from Home".

Distance from home: hukou

China's hukou system was introduced in the 1950s to regulate population mobility, and is still used to determine access to social benefits, including healthcare and children's

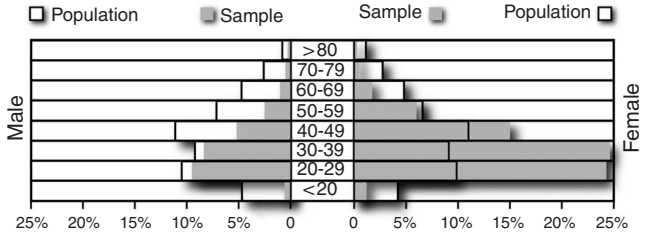
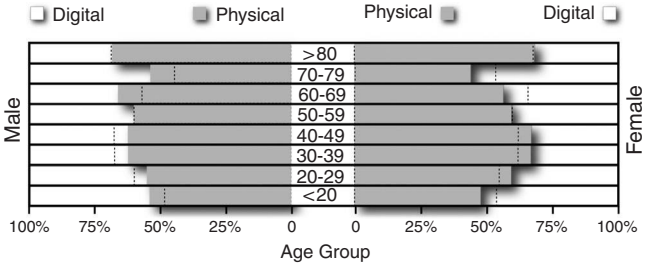


Figure 3.
Comparison of city sample with overall population



Note: Each gender is overlaid with an imprint of the opposite split (dotted line) to show differences by gender in each age bracket

Figure 4.
Behavioural groups: those that prefer to use physical payment methods (cheque, cash) vs those that prefer to use digital payment methods (bank Transfer, AliPay, EPoS, credit card) by gender

education. As China's economy has grown, increasing economic incentives for workers to migrate has led to China's urban population exceeding its rural population for the first time in 2011, with recent estimates of the population separated from their hukou households now standing at 271 million (NBSC, 2012), a figure equivalent to one third of China's economically active population (ILO, 2012).

It should be noted that the hukou system does not prohibit movement and hence the economic attraction of rural migration is countered by the social benefits of maintaining a "hukou" home, reflected in large numbers of "Circular Migrants" who move back and forth between their hometown and employment. In Hu *et al.*'s (2011) study of migrant workers in China the proportion of migrant workers classified as "circular migrants" in their sample was 92 per cent. As circular migrants are, by definition, economically active, this would equate to some 30 per cent of the economically active population overall.

With such a high proportion of any population sampled through consumption being likely to have split geographical ties and high mobility, the hukou address becomes a potentially important demographic. We plot the hukou addresses for our sample in Figure 5 and this clearly shows a pattern of migration that extends across the whole of China. To account for this in our model we define a "Distance from Home" for each customer calculated using the Haversine formula from the longitude and latitude of the postcodes associated with the shop and the customer's registered hukou address. There are two important points to note here:

(1) There is only one shop in the city and the shop is the administrative unit for all sales across all market channels in that city, regardless of whether they arise from physical presence of a customer within the shop or from a direct sale

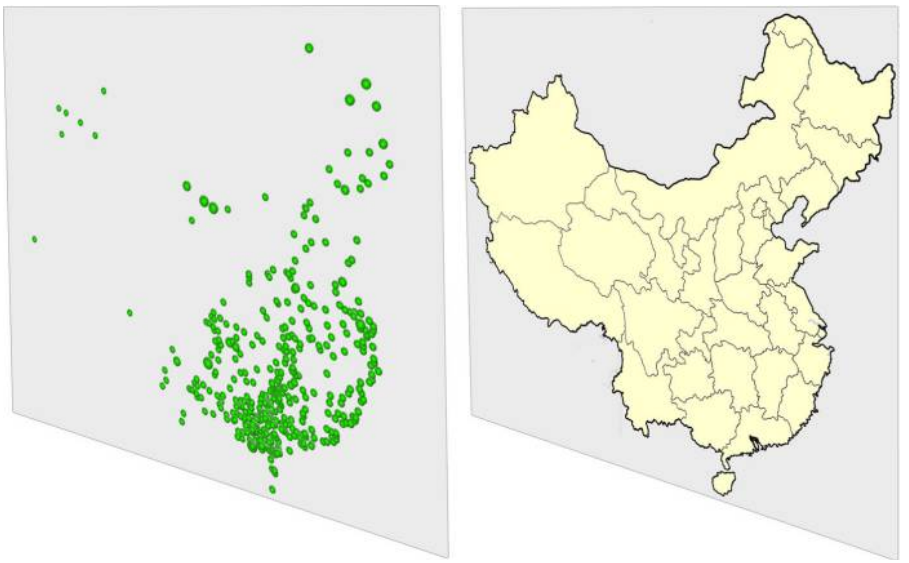


Figure 5.
Sampling a customer population in one city in China

Notes: Plotting the registered home addresses of each customer shows a very wide distribution extending over 3,000 kilometres. The precise distribution is shown in Figure 6 and relates to customer accounts measured in tens of thousands

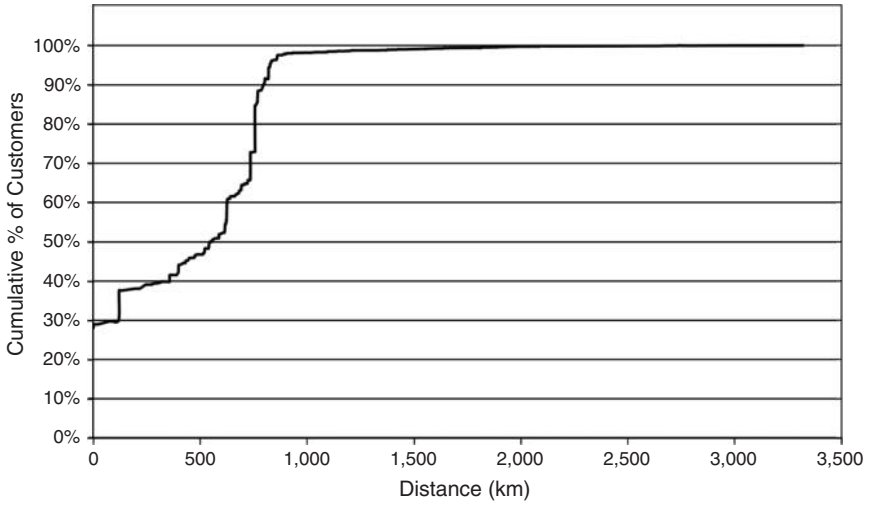
through the internet, fax, telephone or mail order. The shop data therefore includes all customers in the city.

(2) The postcodes associated with the shop and the customer's registered hukou address are large in terms of population cover, and preserve anonymity of the individual consumer. However, they are relatively small in comparison to the scale of the distances being plotted, allowing a "Distance from Home" to be meaningfully associated with each customer.

This "Distance from Home" is plotted for the city sample in Figure 6 where we see that this demographic distributes the population over the state space very well, with over 70 per cent of the sample having a registered hukou address that is greater than 100 kilometre from the shop. We should also note at this stage that while we are sampling consumption through the shop, the shop is not a reason for migration and nor are the consumers expected to reside very far from the shop they have chosen as there are equivalent shops in cities across China operated by the same company and multiple shops in the largest (typically "Tier 1") cities. We therefore assume that the shop at which each customer is registered is proximate to where the consumers spend most of their time, i.e. where they are ordinarily resident.

A C5.0 model of physical vs digital payment method use

In this study of payment method use in one city in China, the decision tree resulting from the modelling was found to correctly classify the behaviour of 73 per cent of the whole sample on the basis of customer demographics and order characteristics. This model is shown in Figure 7 where the attribute that explains splits in the behaviour of the population better than any other, the attribute with the highest information gain, is shown furthest to the left. This is described as the "root" node, though generally depicted as the top of the tree.



Note: Distance is calculated using the Haversine formula from the longitude and latitude of the postcodes associated with the shop and the customer's registered hukou address

Figure 6. Cumulative number of customers plotted as a function of distance from registered home address to the shop

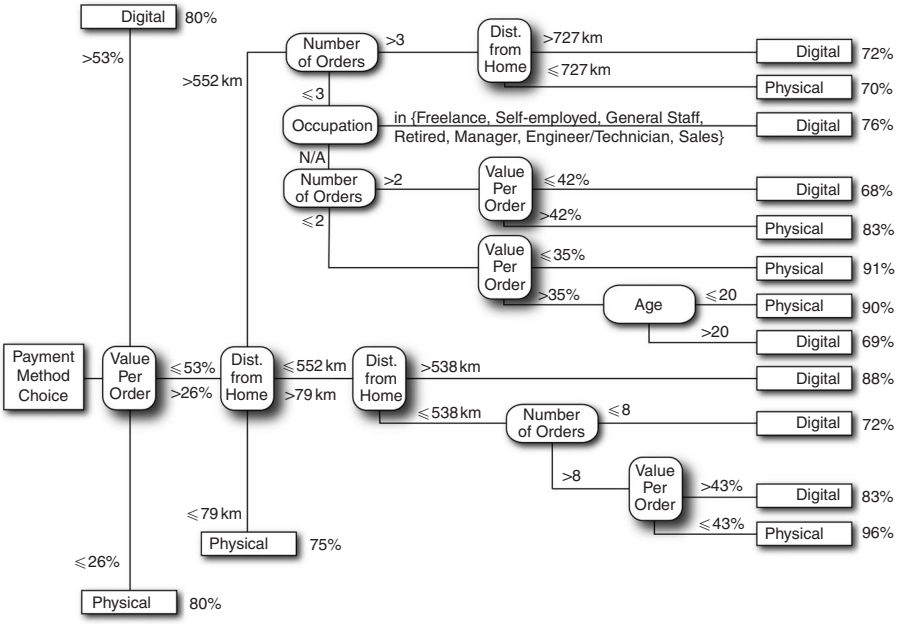


Figure 7.
C5.0 Decision tree model of consumer choice of payment method

Notes: Note that all references to value per order are given in percentage terms of the average Value Per Order for this city. See text for details

As we proceed across the tree from the left, each node contains an attribute that explains most of the behaviour in the subpopulation to its right. After each split this is recalculated and means that attributes at the "top" of the tree can reappear at multiple levels.

When further population splits add no significant value to classification accuracy we reach the leaf node. The leaf node contains the predicted behaviour for this group, and is shown with the associated classification accuracy next to it, expressed as a percentage.

As Figure 7 shows, value per order is most significant split in the population, where high order values (greater than 53 per cent of the average order value) explain digital payment use with a classification accuracy of 80 per cent, whilst low order values (less than or equal to 26 per cent of the average order value) explain use of physical payment methods with an equally high accuracy.

Between these two extremes of value per order, the behaviour is more nuanced, and demographics emerge to explain behaviour. "Distance from Home" is the first demographic to appear, indicating that it is not only the most significant explanatory demographic for the subpopulation to its right, but that it is also ranked as the most significant demographic for the population as a whole. As we read the tree from this point we see the following dependencies on this demographic articulated:

(1) Group I: where the home address is within 79 kilometres of the shop, consumers tend to use physical payment methods (classification accuracy 75 per cent).

(2) Group II: where the home address is greater than 79 kilometres but less than or equal to 552 kilometres from the shop, we see a general preference for digital

payment methods (Group IIa), except where the number of orders is relatively high for this subgroup but the value per order is relatively low (Group IIb).

(3) Group III: where the home address is greater than 552 kilometres and the number of orders is greater than three, we see a general preference for those who live further away to use digital payment methods.

(4) Group IV: where the home address is greater than 552 kilometres and the number of orders is less than or equal to three, we see those who declare an occupation preferring to use digital payment methods (Group IVa).

Note that at this position in the model we encounter a group of people whose home address is very distant from the city, whose numbers of orders are very low and who do not declare an occupation. Up to this point in the model all layers have been generally consistent, with higher order values and longer distances within each sub-group correlating with use of digital payment methods. Within this group however, we start to see some highly nuanced and counter intuitive behaviour, such as younger customers preferring to use physical payment methods for relatively high value orders (Group IVb).

Discussion

The model presented in Figure 7 both confirms expectations and offers new insights. At the top of the tree we see highly significant splits in the population that can be thought of as "rules" describing behaviour. For example, one "rule" that applies to the whole population is:

> If "Value Per Order" is greater than 53% of the average order value then consumer will use a digital payment method with a probability of 80% (Rule 1).

This rule is explicitly associated with a probability, and hence is not absolute. In the context of an exploratory study it may be helpful to think of this as a "general rule" for the overall population that describes common behaviour and provides a background against which more nuanced forms of behaviour lower down in the model may be contrasted. In this case Rule 1 might be re-expressed as:

> With high value orders, digital payment is preferred (General Rule 1).

Two other rules applying to the whole population may also be extracted from Figure 7:

> If "Value Per Order" is less than or equal to 26 per cent of the average order value then consumer will use a physical payment method with a probability of 80 per cent (Rule 2).

> If "Value Per Order" is between 27 and 53 per cent of the average order value and "Distance from Home" is less than or equal to 79 kilometre then consumer will use a digital payment method with a probability of 75 per cent (Rule 3).

That may also be re-expressed as general rules:

> With low value orders, physical payment is preferred (General Rule 2).

> When customers live close to the shop, physical payment is preferred (General Rule 3).

Rule 1 confirms earlier work in China by Worthington *et al.* (2007) that identified "purchase trigger points" above which credit card use is preferred. Rule 2 shows that converse is also true for this population: that physical payment methods are preferred

when order values are low. These behaviours do not appear exceptional and might be understood as simply reflecting practical considerations that are generally true of the entire population, all of whom may be able to elect to use one or the other.

Alternatively, given the generally strong association reported between wealth and credit card use in Asian countries (Khare *et al.*, 2012) and notably the strong association reported for an early study in Hong Kong which, with the exception of Australia, is the longest established market for credit cards in the Asia Pacific region (Chan, 1997), these behaviours may reflect economic barriers that constrain parts of society who cannot access credit to operate within the cash economy.

Discriminating between the two is difficult given the typically qualitative nature and small samples of published studies that explore the "why" of individual credit card use rather than characterising the "what" of actual use across the whole population. In contrast, the model in Figure 7 expresses an empirically grounded general rule about the "what" of digital payment use (including credit cards) but at a level of aggregation that makes it impossible to discern whether either explanation of "why" explains behaviour in this case.

At the next level in Figure 7, the principal splits in behaviour are not by traditional demographics nor by value of order, but by a new demographic: "Distance from Home". This may be interpreted as a proxy for a person's distance from the centre of their social network and, given the link of social and economic interests through hukou registration, the centre of their financial interests. We are not aware of any previous study in which the impact of this distance on behaviour has been explicitly tested, and in Jeyaraj *et al.*'s (2006) extensive review of predictors, linkages and biases in information technology innovation and adoption research no distance of any type appears in the 135 independent variables listed as previously examined.

Distance does appear as a consideration in more recent work on social networks, with Hipp and Perrin (2009) seeking an empirical base for equivalence between physical distance and social distance, and Mok *et al.* (2010) exploring the impact of the internet on interpersonal contact. Both studies acquired data on telephone and e-mail use, and though both recognised the importance of physical distance to the strength of social interactions, Hipp and Perrin (2009) did not report any relationship between technology use and distance, and Mok *et al.* (2010) found technology use to be insensitive to distance. This prior work stands in contrast to the distance-dependent behaviours identified in the second level of Figure 7 that defined Groups I-IV discussed in the previous section.

The behaviours displayed at this level of the model are more nuanced and raise an open set of questions for further study. For example, it is interesting to note that physical payment methods such as cash are preferred by people in Group I whose hukou residency, social benefits and presumably social network, are relatively close to the shop's location (Rule 3). In the context of the preceding Rule 2 that indicates a preference for physical payment methods for all orders below 26 per cent of the average order value for the whole population, this means the threshold at which the benefits of using digital payment methods exceeds that of physical methods is very much higher for this group.

Remember that the relative position of Rule 3 in this tree is determined by its information value not simply its accuracy, which means that it relates to a significant proportion of the population and yet the modelling process does not detect any significant socio-demographic distinctiveness that explains this behaviour.

Whilst an analysis of the cash economy is beyond the scope of the present paper, a detailed study of 19 rural migrants in Shenzhen by Wang and Tian (2014) showed that there were clear economic benefits for people who have easy access to their social network to use physical payment methods and operate principally within a cash economy. Keeping money away from the banks was seen as more convenient, safer, and allowed easier access to borrowing and saving, with loans attracting more interest than the banks paid on savings and less interest than the banks charged on borrowings: "We manage our own money in our way and we have reasonable interest. It's much better than banks". Such consumers may therefore elect to be "Digital Invisible" even when there may be no traditional demographic to suggest that they are at risk of being "Digitally Excluded" and hence Group I may include a significant proportion of non-users who are "able but not willing".

We might expect those who are at risk of Digital Exclusion: smaller groups of non-users who are "willing but not able", to be more clearly discernable lower down in the model. Here the high level "general rules" helps throw anomalous group behaviour into relief. For example, Group IIb exhibits a preference for physical payment methods even though it combines a value per order that may be up to 65 per cent higher than the threshold for Rule 2, a number of orders that is 170 per cent higher than the threshold value for this attribute anywhere else in the model, and a distance from home that may be nearly 600 per cent higher than the threshold for Rule 3.

In Group IVb we see an age group typically associated with digital engagement, with a home address more than 552 kilometres from the city, some 600 per cent greater than the threshold for Rule 3, preferring to use physical payment methods with order values that are 35 per cent higher than the threshold for Rule 2.

Such behaviours contrast with the common behaviours expressed in General Rules 1-3. It is this contrast that suggests that the "choice" made to use physical payment methods is not elective, but reflective of opportunity. Customers in this category are necessarily "Digital Invisible" as they do not use digital payment methods, but this contrast suggests that they are at risk of being "Digital Excluded".

It is important to recall, as exemplified by the apocryphal tale of beer and diapers discussed earlier, that the ability to draw such inferences is constrained by the presence, or not, of demographic descriptors in the model. Whilst it is interesting that traditional demographics start to appear at lower levels of the model to help describe more nuanced behaviours, a more important point is to note the absence of traditional demographics from the higher levels. This absence suggests that traditional demographics do not describe the most significant antecedents of behaviour within the sample population as a whole.

This observation clearly separates the inductive data-mining approach, where the strongest correlations emerge from the analysis, from deductive approaches that hypothesise and seek to find correlations between reported behaviour and traditional demographics. For example Zhu and Chen (2013) highlight a lack of empirical studies in China: "Those researching the use of ICTs in China, however, have conducted few empirical studies, relying instead on descriptive research and theoretical discussions", but inherit assumptions from earlier studies outside China that: "Demographic characteristics such as age, gender, marital status, and race/ethnicity are significant indicators of e-commerce activities in many countries". These assumptions necessarily become embedded in the objective of Zhu and Chen's (2013) study: "To what extent do demographic characteristics, socio-economic attributes, and migration and residency status affect the use of the Internet and e-commerce in China?" from which they

hypothesise relationships between traditional demographics and access to the internet and participation in ecommerce, and then test for them.

Zhu and Chen's (2013) hypotheses were tested in "national study" of 1,288 respondents drawn from "20 provinces, 28 cities, and 33 urban districts or counties". Though their observations were based a sample 1-2 orders of magnitude less than the sample used here for a single location in China, the authors were still able to conclude that: "Age, gender, education, and residency were identified as significant predictors for individual e-commerce use", and that: "The research can also be used in designing effective policies to reduce China's digital inequality". Though Zhu and Chen's (2013) results may be viewed as simply confirming earlier work from which they drew their hypotheses, it is this extrapolation from small samples to policy interventions formulated using the lens of traditional demographics, and then targeted using demographic-based media segmentation, that is challenged by the low significance of traditional demographics in Figure 7. If traditional demographics do not describe most use within a digital economy, then forcing a sample to fit within this lens is unlikely to improve a study's ability to resolve digital inequalities for the relatively small groups for which digital divides persist and grow. Policies informed by such studies are at risk of missing the most important factors and targeting the wrong people.

Conclusions

The move from local, indirect and physical market channels to global direct digital market channels is predicted to accelerate. In 2014, McKinsey and Company reported that 90 per cent of the US population lived within ten minutes of three or more different banks, but by 2020 they predict that more than 95 per cent of banking transactions will take place though direct or digital channels (Bollard *et al.*, 2014).

In China this timescale coincides with predictions of 900 million digital banking customers (Chen *et al.*, 2014), some 79 per cent of the predicted adult Chinese population (United Nations, 2012).

Given evidence for the persistence of a digital divide in the USA (Bach *et al.*, 2013) it is clear that ensuring society is as inclusively served by digital channels in 2020 as the US consumer was served by physical channels in 2014, presents a significant policy challenge.

We have argued that policy effectiveness is constrained by methodological issues that obscure more nuanced behaviours. These include treating digital divides as if they were binary splits in the population, missing parts of society that "have less"; deductive approaches to research that assume behaviour is predominantly described by traditional demographics and then reads any significant correlation in the results as confirming this; and studies that are either too small in scale to characterise "outlier" behaviours, or attempt to draw inferences about non-users by studying only users.

In the exploratory work reported here we have concentrated on the largest user market for digital technologies: China, and built on work in the high-performance distributed computing field to enable what we understand to be the first trans-national collaborative access to, and analysis of, commercial "Big Data" inside China capturing behaviour across a wide range of physical and digital channels.

Sampling the behaviour of tens of thousands of users and non-users of digital payment methods in one urban location we have applied a widely tested and well understood, industry-standard data-mining algorithm to demographic and use data to develop a predictive model of behaviour that characterises digital vs physical payment method use.

The analysis is novel in its formulation of "Distance from Home" as a new demographic that can be attached to each consumer, and the results indicate that

"Distance from Home" explains more use behaviour than the traditional demographics included in the study. We understand that this paper reports the first use of such a "distance" demographic in any large study of the adoption and use of technology.

The prevalence of circular migrants amongst China's economically active and an unprecedented rate of net migration that may see the world's first urban billion in China by 2026 (Miller, 2013), suggest that "Distance from Home" is an important demographic to consider in any study of China's socio-economic development. Principally it gives a way of distinguishing migrant from non-migrant workers in large-scale studies. Additionally, since "Home" may represent the centre of an individual's social network and financial interests, this distance may prove a useful proxy for aspects of migrant workers' social and economic imperatives that helps explain distinctive behaviours within this group. Accordingly, such a demographic may prove important in understanding the impact of hukou reform inside China and current policies to increase urbanisation (Chan and Buckingham, 2008; Chen, 2011; Wang and Tian, 2014).

At lower levels of the model shown in Figure 7, the more nuanced behaviours that are a specific focus of this work emerge. We suggest that contrasting these with general rules about behaviour that apply to the majority of the sample may help distinguish non-use of digital technologies that is elective from that reflecting exclusion. If such distinctions can be generalised they may help to discriminate between the "digital invisible" who are "able but not willing" to use such technologies from those "digital invisible" who are "willing but not able" and hence at risk of "digital exclusion". Such a distinction is important if governments are to develop effective policy instruments to reduce the digital divide, and industry is to accelerate this process by designing the products and services that are valued by the former and needed by the latter.

Returning to methodological considerations in the study of behaviour in a global digital society. The empirical distribution of distance in Figure 6 highlights a tension in the application of ethnographic methods arising from sociology's inheritance of holism from anthropology, where "we cannot understand what goes on within particular situations unless we can locate these within a larger picture" (Hammersley, 2006). Hammersley draws attention to a divergence between research funding trends and research challenges in a global context. Constraints imposed by the former have tended to reduce the physical distance over which the field research is conducted whilst the latter has increased the distance over which a "holistic" context needs to be characterised. Together this leaves a "physical" gap between local, holistic studies and attempts at "virtual" ethnographies where distance is much harder to interrogate.

This diffusion of the subject matter of anthropology from the local to the global was related by Comaroff (2010) to a deliberately contentious statement made by the eminent anthropologist Marshall Sahlins that "anthropology appears to have become little more than the production of 'thin' ethnographic accounts of the myriad, dispersed effects of global capitalism" (Comaroff, 2010). In the case of Figure 6, the long tail suggests that, particularly within China, care needs to be taken with assumptions about homophily and propinquity when attempting to justify the extrapolation of observations from small samples of cultural values, beliefs and norms to the rest of the local population. Distance as a demographic should thus be explored as an extension of traditional demographics that may allow better control for sample dispersion, as well as providing a richer and more nuanced characterisation of behaviour.

Finally, we note that the absence of distance as a factor related to individual behaviour in the information technology innovation and adoption research literature

(Jeyaraj *et al.*, 2006) may help explain Burns' (2007) observation that "Research on innovative behavior has persistently focused on determining the correlations between numerous demographic and psychographic variables and specific external actions. Unfortunately, such research has consistently produced unreliable results". More research in this area is clearly needed, and especially in China, where small-scale studies need to be set in the context of large-scale studies if their detailed insights are to inform effective policies as China approaches the world's first urban billion.

Acknowledgements

The authors gratefully acknowledge the support of the UK Economic and Social Research Council (award RES-149-25-0005) for the initial phase of the INWA Grid and its "Follow-On Funding" (award RES-189-25-0039); the UK EPSRC (award EP/H006753/1 on "Building Relationships with the 'Invisible' in the Digital (Global) Economy") for supporting the reported work with collaborators in China and Thailand; the Scottish Funding Council (edikt2 grant HR04019); the Australian Research Council in partnership with Singapore Telecom for its support (awards LP0454322 and SR0567388) and the endowed SingTel Optus Chair of eBusiness held by Lloyd at Curtin University; Sun Microsystems, Oracle and the Australian Academic and Research Network (AARNet) for continued support since the INWA Grid became operational in 2003, qand colleagues at the Computer Network and Information Center of the Chinese Academy of Sciences who have enabled and hosted the connections within China since 2004.

Notes

1. It is worth observing at this point, in relation to earlier arguments, that the Kontos *et al.* (2012) sample was restricted to users of digital technologies, limiting the value of this insight into the Digital Divide.
2. Note that following the initial public offering of Alipay's parent, the AliBaba Group, in 2014, Alipay was reporting more than 80 million transactions daily and was rebranded as Ant Financial Services Group (Shih, 2014).

References

Akkaş, E., Akina, L., Çubukçua, H.E. and Artuner, H. (2015), "Application of decision tree algorithm for classification and identification of natural minerals using SEM-EDS", *Computers & Geosciences*, Vol. 80, pp. 38-48.

Alipay (2013), "Alipay business, Alipay, Hangzhou, China", available at: http://global.alipay.com/ospay/home.htm (accessed 20 January 2013).

Anderson, C. (2008), "The end of theory: the data deluge makes the scientific method obsolete", *Wired Magazine*, available: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 9 July 2014).

Arrington, M. (2009), "Google CEO Eric Schmidt interview: his thoughts on search, books, news, mobile, competition and more", TechCrunch, 31 August, available at: http://techcrunch.com/2009/08/31/google-ceo-eric-schmidt-interview-his-thoughts-on-search-books-news-mobile-competition-and-more/ (accessed 26 August 2012).

Atsmon, Y., Magni, M. and Lihua, L. (2012), "From mass to mainstream: keeping pace with China's rapidly changing consumers", Annual Chinese Consumer Report, McKinsey Consumer & Shopper Insights, McKinsey & Company, Boston, MA, pp. 1-36.

Bach, A., Shaffer, G. and Wolfson, T. (2013), "Digital human capital: developing a framework for understanding the economic impact of digital exclusion in low-income communities", *Journal of Information Policy*, Vol. 3, pp. 247-266, available at: http://doi.org/10.5325/jinfopoli.3.2013.0247

Bankert, R.L. and Hadjimichael, M. (2007), "Data mining numerical model output for single-station cloud-ceiling forecast algorithms", *Weather and Forecasting*, Vol. 22 No. 5, pp. 1123-1131.

Bastı, E., Kuzey, C. and Delen, D. (2015), "Analyzing initial public offerings' short-term performance using decision trees and SVMs", *Decision Support Systems*, Vol. 73, May, pp. 15-27.

Bekaert, G., Harvey, C.R., Lundblad, C.T. and Siegel, S. (2012), "The European Union, the Euro, and equity market integration", AFA 2012 Chicago Meetings Paper, available at: http://ssrn.com/abstract=1573308; http://dx.doi.org/10.2139/ssrn.1573308 (accessed 26 August 2012).

Bollard, A., Doshi, N. and Maxwell, M.N. (2014), "The future of US retail-banking distribution", McKinsey and Company, available at: www.mckinsey.com/~/media/McKinsey/dotcom/client_service/Financial%20Services/Latest%20thinking/Consumer%20and%20small%20business%20banking/Future_of_US_retail_banking_distribution.ashx (accessed 21 August 2014).

Brown, R.H., Barram, D.J. and Irving, L.J. (1995), "Falling through the net: a survey of the 'Have Nots' in rural and urban America", National Telecommunications & Information Administration, United State Department of Commerce, available at: www.ntia.doc.gov/ntiahome/fallingthru.html (accessed 26 August 2013).

Burns, D.J. (2007), "Toward an explanatory model of innovative behaviour", *Journal of Business and Psychology*, Vol. 21 No. 4, pp. 461-488.

Castells, M. (1996), *The Information Age: Economy, Society and Culture Vol. I: The Rise of the Network Society*, ISBN: 1-55786-616-3/1-55786-617-1, Blackwell Publishers, Cambridge, MA and Oxford.

Cellular News (2008), "Asian markets April 2008 update", available at: www.cellular-news.com/story/31265.php?source=newsletter (accessed 26 August 2013).

Chan, K.W. and Buckingham, W. (2008), "Is China abolishing the hukou system?", *The China Quarterly*, Vol. 195, September, pp. 582-606.

Chan, R.Y. (1997), "Demographic and attitudinal differences between active and inactive credit cardholders – the case of Hong Kong", *International Journal of Bank Marketing*, Vol. 15 No. 4, pp. 117-125.

Chatzis, S.P. and Demiris, Y. (2012), "A sparse nonparametric hierarchical Bayesian approach towards inductive transfer for preference modeling", *Expert Systems with Applications*, Vol. 39 No. 8, pp. 7235-7246.

Chen, F.-H., Chi, D.-J. and Wang, Y.-C. (2015), "Detecting biotechnology industry's earnings management using Bayesian network, principal component analysis, back propagation neural network, and decision tree", *Economic Modelling*, Vol. 46, April, pp. 1-10.

Chen, J., HV, V. and Lam, K. (2014), "How to prepare for Asia's digital-banking boom", McKinsey and Company, available at: www.mckinsey.com/insights/financial_services/how_to_prepare_for_asias_digital_banking_boom (accessed 21 August 2014).

Chen, W. and Wellman, B. (2004), "The global digital divide – within and between countries", *IT & Society*, Vol. 1 No. 7, pp. 18-25.

Chen, Y. (2011), "Rural migrants in urban China: characteristics and challenges to public policy", *Local Economy*, Vol. 26 No. 5, pp. 325-336.

Chiang, W.-Y. (2012), "Applying a new model of customer value on international air passengers' market in Taiwan", *International Journal of Tourism Research*, Vol. 14 No. 2, pp. 116-123.

Choi, N.G. and DiNitto, D.M. (2013), "The digital divide among low-income homebound older adults: internet use patterns, ehealth literacy, and attitudes toward computer/internet use", *Journal of Medical Internet Research*, Vol. 15 No. 5. p. e93. doi: 10.2196/jmir.2645. available at: www.jmir.org/2013/5/e93

Chu, C.-M., Yao, C.-T., Chang, Y.-T., Chou, H.-L., Chou, Y.-C., Chen, K.-H., Terng, H.-J., Huang, C.-S., Lee, C.-C., Su, S.-L., Liu, Y.-C., Lin, F.-G., Wetter, T. and Chang, C.-W. (2014), "Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees", *Disease Markers*, Vol. 2014, p. 634123. doi: 10.1155/2014/634123.

Comaroff, J. (2010), "The end of anthropology, again: on the future of an in/discipline", *American Anthropologist*, Vol. 112 No. 4, pp. 524-538.

Coughlan, A.T. (1985), "Competition and cooperation in marketing channel choice: theory and application", *Marketing Science*, Vol. 4 No. 2, pp. 110-129.

Cushman, M. and McLean, R. (2008), "Exclusion, inclusion and changing the face of information systems research", *Information Technology & People*, Vol. 21 No. 3, pp. 213-221.

Day, P. (2013), "Imagine a world without shops or factories", *BBC News Magazine*, 11 October, available at: www.bbc.co.uk/news/magazine-23990211 (accessed 11 October 2013).

Earman, J. and Roberts, J. (1999), "Ceteris paribus, there is no problem of provisos", *Synthese*, Vol. 118 No. 3, pp. 439-478.

Emerson, D., Zhou, W. and Piramuthu, S. (2009), "Goodwill, inventory penalty, and adaptive supply chain management", *European Journal of Operational Research*, Vol. 199 No. 1, pp. 130-138.

Eubanks, V. (2011), *Digital Dead End: Fighting for Justice in the Information Age*, MIT Press, Cambridge, MA.

Evans, J. and Mavondo, F.T. (2002), "Psychic distance and organizational performance: an empirical examination of international retailing operations", *Journal of International Business Studies*, Vol. 3 No. 33, pp. 515-532.

Evans, J., Treadgold, A. and Mavondo, F.T. (2000), "Psychic distance and the performance of international retailers: a suggested theoretical framework", *International Marketing Review*, Vol. 17 Nos 4/5, pp. 373-391.

Forman, C., Goldfarb, A. and Greenstein, S. (2002), "Digital dispersion: an industrial and geographic census of commercial internet use", in Wildman, S. and Cranor, L. (Eds), *Rethinking Rights and Regulations: Institutional Responses to New Communication Technologies*, NBER Working Paper No. 9287, MIT Press, Cambridge, available at: www.nber.org/papers/w9287 (accessed 26 August 2013).

Fry, J.A., Coan, M.J., Homer, C.G., Meyer, D.K. and Wickham, J.D. (2009), "Completion of the National Land Cover Database (NLCD) 1992-2001 land cover change retrofit product", US Geological Survey Open-File Report No. 2008-1379, p. 18.

Gartner, A. and Reissman, F. (1974), *The Service Society and the Consumer Vanguard*, Harper and Row Publishers, New York, NY.

Gilchrist, J., Frize, M., Ennett, C.M. and Bariciak, E. (2011), "Neonatal mortality prediction using real-time medical measurements", *Medical Measurements and Applications Proceedings (MeMeA), 2011 IEEE International Workshop*, pp. 65-70.

Gort, M. and Klepper, S. (1982), "Time paths in the diffusion of product innovations", *The Economic Journal*, Vol. 92 No. 367, pp. 630-653.

Gupta, R.R., Gifford, E.M., Liston, T., Waller, C.L., Hohman, M., Bunin, B.A. and Ekins, S. (2010), "Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties", *Drug Metabolism Disposition*, Vol. 38 No. 11, pp. 2083-2090.

Håkanson, L. and Ambos, B. (2010), "The antecedents of psychic distance", *Journal of International Management*, Vol. 16 No. 3, pp. 195 -210.

Haley, G.T. (2002), "E-commerce in China: changing business as we know it", *Industrial Marketing Management*, Vol. 31 No. 2, pp. 119-124.

Hammersley, M. (2006), "Ethnography: problems and prospects", *Ethnography and Education*, Vol. 1 No. 1, pp. 3-14.

Hayes, H.B. (1996), "The knowledge banks", Government Executive, 1 March, available at: www.govexec.com/technology/1996/03/the-knowledge-banks/206/ (accessed 9 July 2014).

Hays, C.L. (2004), "What Wal-Mart knows about customers' habits", *The New York Times*, 14 November, available at: www.nytimes.com/2004/11/14/business/yourmoney/14wal.html (accessed 26 August 2013).

Hey, T., Tansley, S. and Tolle, K. (2009), "Jim Grey on eScience: a transformed scientific method", in Hey, T., Tansley, S. and Tolle, K. (Eds), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA, pp. xvii-xxxi, available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf (accessed 9 July 2014).

Hilbert, M. (2011), "The end justifies the definition: the manifold outlooks on the digital divide and their practical usefulness for policy-making", *Telecommunications Policy*, Vol. 35 No. 8, pp. 715-736, available at: http://dx.doi.org/10.1016/j.telpol.2011.06.012

Hipp, J.R. and Perrin, A.J. (2009), "The simultaneous effect of social distance and physical distance on the formation of neighborhood ties", *City & Community*, Vol. 8 No. 1, pp. 5-25.

Hoffman, D.L., Novak, T.P. and Schlosser, A. (2000), "The evolution of the digital divide: how gaps in internet access may impact electronic commerce", *Journal of Computer-Mediated Communication*, Vol. 5 No. 3, Blackwell Publishing Ltd. doi: 10.1111/j.1083-6101.2000.tb00341.x.

Hu, F., Xu, Z. and Chen, Y. (2011), "Circular migration, or permanent stay? Evidence from China's rural-urban migration", *China Economic Review*, Vol. 22 No. 1, pp. 64-74.

Huber, M.J. (1976), "The service society and the consumer vanguard (book)", *Journal of Consumer Affairs*, Vol. 10 No. 1, pp. 103-105.

Hume, D. (1739), *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, *Volume I, of the Understanding*, *Part III of Knowledge and Probability*, John Noon, London.

IBM (2011), "IBM and range technology collaborate to build Asia's largest cloud computing data center in Hebei Province", available at: www-03.ibm.com/press/us/en/pressrelease/33447.wss (accessed 18 February 2013).

ILO (2012), "Total and economically active population, by age group: China 1999-2008", available at: http://laborsta.ilo.org/ (accessed 14 May 2013).

Jeyaraj, A., Rottman, J.W. and Lacity, M.C. (2006), "Research review: a review of the predictors, linkages, and biases in IT innovation adoption research", *Journal of Information Technology*, Vol. 21, pp. 1-23. doi: 10.1057/palgrave.jit.2000056 (accessed 10 January 2006).

Johanson, J. and Vahlne, J.E. (1977), "The internationalisation process of the firm – a model of knowledge development and increasing foreign market commitment", *Journal of International Business Studies*, Vol. 8 No. 1, pp. 23-32.

Johanson, J. and Wiedersheim-Paul, F. (1975), "The internationalization of the firm: four Swedish cases", *Journal of Management Studies*, Vol. 8 No. 1, pp. 23-32.

Khare, A., Khare, A. and Singh, S. (2012), "Factors affecting credit card use in India", *Asia Pacific Journal of Marketing and Logistics*, Vol. 24 No. 2, pp. 236-256.

Kincaid, H. (1996), *Philosophical Foundations of the Social Sciences*, Cambridge University Press, Cambridge.

Kohavi, R. and Quinlan, J.R. (2002), "Data mining tasks and methods: classification: decision-tree discovery", in Klösgen, W. and Zytkow, J.M. (Eds), *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, New York, NY, pp. 267-276.

Kontos, E., Blake, K.D., Chou, W.Y.S. and Prestin, A. (2012), "Predictors of eHealth usage: insights on the digital divide from the health information national trends survey 2012", *Journal of Medical Internet Research*, Vol. 16 No. 7, p. e172, available at: www.jmir.org/2014/7/e172/ (accessed 26 August 2014).

Kronmiller, G.C. and Baghdady, E.J. (1966), "The Goddard range and range rate tracking system: concept, design and performance", *Space Science Reviews*, Vol. 5 No. 2, pp. 265-307.

Light, J.S. (2001), "Rethinking the digital divide", *Harvard Educational Review*, Vol. 71 No. 4, pp. 709-733.

Lloyd, A.D., Sloan, T., Antonioletti, M. and McGilvary, G. (2013), "Embedded systems for global e-social science: moving computation rather than data", *Future Generation Computer Systems*, Vol. 29 No. 5, pp. 1120-1129.

McConnaughey, J.W., Lader, W., Chin, R. and Everette, D. (1998), "Falling through the net II: new data on the digital divide", National Telecommunications & Information Administration, United State Department of Commerce, available at: www.ntia.doc.gov/report/1998/falling-through-net-ii-new-data-digital-divide (accessed 26 August 2013).

Macfie, N. (2008), "China becomes world's largest internet population", Reuters News, available at: www.reuters.com/article/2008/04/24/us-china-internet-idUSPEK34240620080424 (accessed 26 August 2013).

Marton, K. and Singh, R.K. (1991), "Technology crisis for third world countries", *World Economy*, Vol. 14 No. 2, pp. 199-213, (0378-5920).

Miller, T. (2013), "AIC 2013 interview: Tom Miller – China's new urbanization initiative", *Asian Investment Conference 2013, Credit Suisse, Hong Kong, 18-22 March*, available at: www.youtube.com/watch?v=zfq_E5wpnHg (accessed 21 August 2014).

Mok, D.K., Wellman, B. and Carrasco, J. (2010), "Does distance matter in the age of the internet?", *Urban Studies*, Vol. 47 No. 13, pp. 2747-2783.

NASDAQ (2014), "NASDAQ index highs: NASDAQ composite", available at: www.nasdaq.com/markets/index-highs (accessed 9 July 2014).

NBSC (2012), "China's total population and structural changes in 2011", National Bureau of Statistics of China, available at: www.stats.gov.cn/english/newsandcomingevents/t20120120_402780233.htm (accessed 14 May 2013).

Norris, P. (2001), *Digital Divide – Civic Engagement, Information Poverty, and the Internet Worldwide*, ISBN: 0521002230, Cambridge University Press, Cambridge.

Ofcom (2012), "Children and parents: media use and attitudes report", Office of Communications (Ofcom), available at: http://stakeholders.ofcom.org.uk/binaries/research/research-publications/childrens/oct2012/main.pdf (accessed 11 October 2013).

ONS (2014), "Total population (UK) mid-2013 estimate", Office for National Statistics, available at: www.ons.gov.uk/ons/taxonomy/index.html?nscl=Population (accessed 9 July 2014).

Oracle (2014), "Corporate citizenship report: sustainability", Oracle, available at: www.oracle.com/us/corporate/citizenship/sustainability/datacenters/index.html (accessed 9 July 2014).

Parsons, R. (2014), "The rise of digital is the rise of direct", *Marketing Week* (online edition), Business Source Complete, Ipswich, MA, July, p. 3, available at: www.marketingweek.com/2014/07/07/the-rise-of-digital-is-the-rise-of-direct/ (accessed 10 October 2014).

Parsons, T. (1951), *The Social System*, Routledge & Kegan Paul Ltd, London.

Pashler, H. and Harris, C.R. (2012), "Is the replicability crisis overblown? Three arguments examined", *Perspectives on Psychological Science*, Vol. 7, pp. 531-536.

Power, D.J. (2002), "What is the 'true story' about using data mining to identify a relation between sales of beer and diapers?", *DSS News*, Vol. 3, No. 23, available at: www.dssresources.com/newsletters/66.php (accessed 6 July 2014).

Prahalad, C.K. and Ramaswamy, V. (2000), "Co-opting customer competence", *Harvard Business Review*, Vol. 78 No. 1, pp. 79-87.

Qiu, J.L. (2009), *Working-Class Network Society*, ISBN: 9780262170062, MIT Press, Cambridge, MA.

Quinlan, J.R. and Rivest, R.L. (1989), "Inferring decision trees using the minimum description length principle", *Information and Computation*, Vol. 80 No. 3, pp. 227-248.

Range (2013), "Project summary, Range International Information Group", available at: www.rangeidc.com/en/project.asp (accessed 20 February 2013).

Rissanen, J. (1978), "Modeling by shortest data description", *Automatica*, Vol. 14 No. 5, pp. 465-471.

Ritzer, G. (2014), "Automating prosumption: the decline of the prosumer and the rise of the prosuming machines", *Journal of Consumer Culture*, pp. 1-18. doi: 10.1177/1469540514553717, available at: http://joc.sagepub.com/content/early/2014/10/15/1469540514553717 (accessed 20 January 2015).

Rulequest Research (2012), "C5.0", available at: www.rulequest.com (accessed 9 July 2012).

Sanga, S., Broom, B.M., Cristini, V. and Edgerton, M.E. (2009), "Gene expression meta-analysis supports existence of molecular apocrine breast cancer with a role for androgen receptor and implies interactions with ErbB family", *BMC Medical Genomics*, Vol. 2 No. 59, available at: www.biomedcentral.com/content/pdf/1755-8794-2-59.pdf (accessed 20 January 2014).

Schatz, B.R. and Hardin, J.B. (1994), "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet", *Science*, Vol. 265 No. 5174, pp. 895-901, available at: www.jstor.org/stable/2884507 (accessed 20 April 2016).

Shao, C.-Y., Su., B.-H., Tu, Y.-S., Lin, C., Lin, O.A. and Tseng, Y.J. (2015), "CypRules: a rule-based P450 inhibition prediction server", *Bioinformatics*, Vol. 31 No. 11, pp. 1869-1871.

Shih, G. (2014), "UPDATE 1-Alibaba affiliate Alipay rebranded Ant in new financial services push", 16 October, available at: www.reuters.com/article/2014/10/16/china-alibaba-idUSL3N0SB3MX20141016 (accessed 16 October 2014).

Solow, R.M. (1987), "We'd better watch out", *New York Times Book Review*, 12 July, p. 36.

Son, C.S., Jang, B.K., Seo, S.T., Kim, M.S. and Kim, Y.N. (2012), "A hybrid decision support model to discover informative knowledge in diagnosing acute appendicitis", *BMC Medical Informatics and Decision Making*, Vol. 12 No. 17, pp. 1-14.

Song, X.-P., Hu, Z.-H., Du, J.-G. and Sheng, Z.-H. (2014), "Application of machine learning methods to risk assessment of financial statement fraud: evidence from China", *Journal of Forecasting*, Vol. 33 No. 8, pp. 611-626.

Stöttinger, B. and Schlegelmilch, B.B. (2000), "Psychic distance: a concept past its due date?", *International Marketing Review*, Vol. 17 No. 2, pp. 169-173.

Teece, D.J. (1993), "The dynamics of industrial capitalism: perspectives on Alfred Chandler's scale and scope", *Journal of Economic Literature*, Vol. 31 No. 1, pp. 199-225.

TEIN (2014), "TEIN3 media centre", available at: www.tein3.net/Media_Centre/Pages/Brochures_and_Maps.aspx (accessed 9 July 2014).

Toffler, A. (1980), *The Third Wave*, ISBN: 0688035973, William Morrow, New York, NY.

United Nations (2012), "World Population Prospects: The 2012 Revision", Department of Economic and Social Affairs, Population Division, available at: http://populationpyramid.net/china/2020/ (accessed 21 August 2014).

US Congress (1991), "High Performance Computing and Communications Act of 1991 (HPCA, Pub.L. 102-194, enacted 1991-12-09)", US Government Printing Office, Washington, DC.

Vahlne, J.-E. and Wiedersheim-Paul, F. (1973), "Economic distance: model and empirical investigation", in Hornell, E., Vahlne, J.-E. and Wiedersheim-Paul, F. (Eds), *Export and Foreign Establishments*, University of Uppsala, Uppsala, pp. 81-159.

Verdegem, P. and Verhoest, P. (2009), "Profiling the non user: rethinking policy initiatives stimulating ICT acceptance", *Telecommunications Policy*, Vol. 33 Nos 10-11, pp. 642-652.

Wang, J.J. and Tian, Q. (2014), "Consumer vulnerability and marketplace exclusion: a case of rural migrants and financial services in China", *Journal of Macromarketing*, Vol. 34 No. 1, pp. 45-56.

Warschauer, M. (2010), "Digital divide", in Bates, M.J. and Maack, M.N. (Eds), *Encyclopedia of Library and Information Sciences*, 3rd ed., Vol. 1, Taylor & Francis, London, pp. 1551-1556.

Wave (2013), "Final steps for Google Wave", available at: http://googlewave.blogspot.co.uk (accessed 14 February 2013).

Wee, W. (2012), "China's Alipay has 700 million registered accounts, beats PayPal?", TechinAsia, available at: www.techinasia.com/chinas-alipay-700-million-registered-accounts-beats-paypal/ (accessed 20 January 2013).

World Bank (2014), "China", available at: www.worldbank.org/en/country/china (accessed 9 July 2014).

World Bank (2015), "Internet users (per 100 people)", available at: http://data.worldbank.org/indicator/IT.NET.USER.P2?order=wbapi_data_value_2013+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed 20 March 2015).

Worthington, S., Stewart, D. and Lu, X. (2007), "The adoption and usage of credit cards by urban-affluent consumers in China", *International Journal of Bank Marketing*, Vol. 25 No. 4, pp. 238-252.

Wu, X., Vipin, K., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D. and Steinberg, D. (2008), "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14 No. 1, pp. 1-37.

Xia, J. and Lu, T.-J. (2008), "Bridging the digital divide for rural communities: the case of China", *Telecommunications Policy*, Vol. 32 Nos 9-10, pp. 686-696.

Zegarra, E.F. and Efremenko, T. (2011), "Using a recommendation engine to personalize your web application", *IBM WebSphere Developer Technical Journal*, Vol. 14 No. 6, pp. 1-20, available at: www.ibm.com/developerworks/websphere/techjournal/1109_zegarra/1109_zegarra-pdf.pdf (accessed 15 April 2014).

Zheng, Y. and Walsham, G. (2008), "Inequality of what? Social exclusion in the e-society as capability deprivation", *Information Technology and People*, Vol. 21 No. 3, pp. 222-243.

Zhu, S. and Chen, J. (2013), "The digital divide in individual e-commerce utilization in China: results from a national survey", *Information Development*, Vol. 29 No. 1, pp. 69-80.

About the authors

Ashley D. Lloyd has researched widely in the emerging technologies arena, from fundamental materials research to innovative ICT applications. Publications range from the *IEEE Journal of Quantum Electronics* to the *International Journal of Innovation Management* (top three articles cited >300 times). Research support received from industry and research councils in four continents. Ashley D. Lloyd is the corresponding author and can be contacted at: ashley@edinburgh.ac.uk

Mario Antonioletti works as a Software Architect at EPCC. Mario has a Mathematics and Physics background and a PhD in astrophysics. He has been involved in various HPC projects as well as the OGSA-DAI project (accessing and integrating databases using web services) and standards-based work at the Open Grid Forum (OGF).

Terence M. Sloan is a Software Development Group Manager at EPCC. He has extensive experience of managing projects in high-performance computing and distributed computing with both academia and business in the UK, Europe, Asia and Australia. He has published more than 40 articles in journals and conference proceedings.