Small Fish in a Big Pond: An Architectural Response to Users Privacy, Rights and Security in the Age of Big Data

Journal:	*International Conference on Information Systems 2016*
Manuscript ID	ICIS-0556-2016.R1
Track:	10. IS Design and Business Process Management
Keywords:	Social Networking Sites, Identity, Architecture, Big Data, Small Data, Privacy, Rights, Security, Distributed Systems, 03. Data Science and Business Analytics
Abstract:	We focus on the challenges and issues associated with Big Data, and propose a novel architecture that uses the principles of Separation of Concerns and distributed computing to overcome many of the challenges associated with storage, analysis and integrity. We address the issue of asymmetrical distribution of power between the originators of data and the organizations and institutions that make use of that data by taking a systemic perspective to include both sides in our architectural design, shifting from a customer-provider relationship to a more symbiotic one in which control over access to customer data resides with the customer. We illustrate the affordances of the proposed architecture by describing its application in the domain of Social Networking Sites, where we furnish a mechanism to address problems of privacy and identity, and create the potential to open up online social networking to a richer set of possible applications.

Small Fish in a Big Pond: An Architectural Approach to Users Privacy, Rights and Security in the Age of Big Data

Completed Research Paper

IS Design and Business Process Management

Spyros Angelopoulos
Università della Svizzera italiana
Lugano, Switzerland
Spyros.Angelopoulos@usi.ch

Derek McAuley
University of Nottingham
Nottingham, UK
Derek.McAuley@nottingham.ac.uk

Yasmin Merali
University of Hull
Hull, UK
Y.Merali@hull.ac.uk

Richard Mortier
Cambridge University
Cambridge, UK
Richard.Mortier@cl.cam.ac.uk

Dominic Price
University of Nottingham
Nottingham, UK
Dominic.Price@nottingham.ac.uk

Abstract

We focus on the challenges and issues associated with Big Data, and propose a novel architecture that uses the principles of Separation of Concerns and distributed computing to overcome many of the challenges associated with storage, analysis and integrity. We address the issue of asymmetrical distribution of power between the originators of data and the organizations and institutions that make use of that data by taking a systemic perspective to include both sides in our architectural design, shifting from a customer-provider relationship to a more symbiotic one in which control over access to customer data resides with the customer. We illustrate the affordances of the proposed architecture by describing its application in the domain of Social Networking Sites, where we furnish a mechanism to address problems of privacy and identity, and create the potential to open up online social networking to a richer set of possible applications.

Keywords: Big Data, Small Data, Identity, Privacy, Rights, Security, Architecture, Distributed Systems, Data Science, Separation of Concern, Social Networking Sites

Introduction

Although the topic of Big Data has featured extensively in the discourse of IS academics and practitioners over the past decade (Chen et al. 2012; Davenport and Patil 2012; McAfee and Brynjolfsson 2012; Shmueli and Koppius 2011) there is no definitive characterization of the constitutive terms. In this paper, we unpack the distinction between issues associated respectively with "Big Data" and the more conventional challenges associated with handling large volumes of "Small Data" (e.g. the kind of data of well-defined – often pre-defined– granularity and dimensionality catered for by traditional data warehousing and data mining applications), and propose a novel architecture that bridges the two worlds. Our architecture uses the principles of Separation of Concerns (SoC) (Hürsch and Lopes 1995) and distributed computing to overcome many of the challenges associated with storage, analysis and integrity that are cited in the Big Data literature, whilst enabling individuals who provide personal data to exercise more control over access to their data. We illustrate the affordances of the proposed architecture by describing its application in the domain of Social Networking Sites (SNS).

Our first contribution is in elucidating the features that make the challenges and issues associated with Big Data and its use in organizations *distinctive* and *different in kind* from the technical issues that have always confronted the Computer Science community since the 1970s.

Our second contribution lies in the motivation and implementation of a novel architecture that can be deployed to address fundamental challenges of Big Data Analytics (BDA) with respect to handling volume, variety, currency, integrity and relevance without compromising users' security, rights and privacy. The architecture that we propose is based on the concept of SoC, and is informed by the developments in computer networking in the 1970s and 1980s, when the IT industry moved from many different, mutually incompatible, proprietary networking standards, towards a single common standard for inter-operation.

Our third contribution is to the growing body of the IS literature on BDA, and more specifically on the privacy issues and the systemic problems associated with the asymmetrical distribution of power between the originators of data and the organizations and institutions that make use of that data to generate information goods and services. The literature to date has largely focused on organizations' data management problems in relation to BDA, viewing one side as 'owning' the problem and the solution, and the other side as data sources, and beneficiaries of the outputs and outcomes of organizational analysis and leveraging of BDA. We take a more systemic perspective, and include both organizations and their clients in our architectural design, shifting from a customer-provider relationship to a more symbiotic one. The resulting architecture liberates organizations from the burden of maintaining costly infrastructures, and concurrently resolves some of the tensions associated with leveraging user-provided data whilst respecting and enhancing users' privacy, rights and security.

Our final contribution is in illustrating the utility of our architecture in the SNS domain, where not only do we furnish a mechanism to address problems of privacy and identity, we also open up online social networking to a much richer set of possible interactions and applications, with finer granularity of control over personal data. Ultimately, our work contributes to the broader IS literature, as a design study that leads to the development of a novel system architecture; a stream that has been at the core of IS Research, addressing how novel IS can be designed, and how IS can assist organizations in designing better organizational systems with improved business processes and workflows.

The rest of this paper is organized as follows: the next section delineates distinctive features of Big Data, the associated analytical techniques, and the issues in leveraging its potential. This treatment suggests a need to develop an alternative to centralized approaches to Big Data collection, storage and analysis. In the following section we propose a novel architecture to meet this need. We illustrate the merits of the proposed architecture by applying it in the case of SNS, and conclude the paper by discussing the implications of our approach for organizations, individuals and society, and reflecting on opportunities for further research.

Just how big is Big?

The term Big Data has been commonly used to refer to datasets large enough to require costly IT infrastructures. However, it has also been argued that in that parlance "Big" is a relational concept: what once required banks of mainframe machines can now be analyzed on desktop computers with standard

software (Manovich 2011). Similarly, it has been pointed out that the term "Big Data" would be superfluous if its sole defining characteristic was its immense size: "if 'Big Data' simply meant lots of data, we would call it 'Lots of Data'" (Williams 2012). Some characterizations link the measure of "Big" with the capacity required to process the data: the McKinsey Global Institute (2011) defines it as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze", while Gartner (2011) introduce the end-user perspective and the time dimension in their attempt to define Big Data as data that "exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population". Gartner are widely cited for defining the three dimensions of *volume*, *velocity*, and *variety* as defining Big Data, but as our next section shows, these are not in themselves sufficient for defining the distinctive nature of the challenges faced by Big Data analysts.

An oft-cited source of Big Data is 'Big Science', characterized by the production of large volumes of data, heavy on storage and computational resources. Venters et al. (2014) used an experiment in CERN to study digital coordination, reporting that such experiments produce data "equivalent to 15 million gigabytes of data per year or a DVD every 5 seconds" (Venters et al. 2014, p. 933). The resources that scientists in CERN use to analyze such data consist of 150,000 computers in 600 sites across 62 countries. Similarly, Chen et al. (2012) cite the Sloan Digital Sky Survey, which during the ten years of its operation has covered more than a quarter of the sky and collected images for the creation of 3D maps containing more than 930,000 galaxies and over 120,000 quasars, collecting data at the rate of 200GB per night. A "Big Science" project with more immediate societal applications is the Human Genome. Since the project's launch, the data set has grown rapidly to 200 terabytes, and the rate of growth is expected to accelerate with more widespread genetic profiling, and the widely held expectation that genomic studies will lead to advances in the diagnosis and treatment of diseases, and to new insights in many fields of biology, including human evolution.

Beyond Technical Limitations

In comparison to Big Science, the demands imposed by the size of corporate data appear rather modest, but significant examples include the companies providing popular SNS. Whilst the size issue is often relegated to technological limitations, the more wicked problems associated with BDA in social systems are to do with:

a) The loss of original contextual information in combing data from diverse sources to construct large data sets, and

b) The complexity of socially situated systems-in-use, such as the context-dependent relevance and immediacy of information required by users, the behavioral inferences made on the basis of BDA, and the ethical and legal issues associated with the processing and use of personal data for purposes that its originators may not have envisaged, given permission for, or even been aware of.

Lazer et al. (2009, p. 722) highlight the multi-dimensional, multi-scale aspects when they argue that the term Big Data offers "the capacity to collect and analyze data with an unprecedented breadth and depth and scale". Following a more pragmatic approach, boyd and Crowford (2012) see Big Data as a sociocultural phenomenon on the interplay of technology, analysis and mythology that not only refers to large data sets and the tools and procedures used to manipulate and analyze them, but also to a paradigm shift in thought and research: they push for the maximization of algorithmic accuracy and computation power in order to achieve optimal collection, analysis, connection, and comparison of large data sets; draw on such data sets to extract patterns that will lead to economic, social, technical, and legal claims, but at the same time are based on the belief that such data sets generate a higher form of intelligence and knowledge that can generate previously-impossible insights, with the aura of truth, objectivity, and accuracy. We concur with boyd and Crowford's (2012) view that BDA is less about the size of data than it is about the capacity to search, aggregate, and cross-reference such datasets.

Whilst contemporary publications on Big Data often focus on management challenges associated with size, it is salutary to note that technical issues that surround the size and complexities of the databases have been in the spotlight for the Computer Science community for at least forty years. For example, one of the most important conferences in Computer Science, the VLDB (Very Large Databases) conference began as early as in 1975 in Framingham MA, and the proceedings of that first conference reveal that it

was focused on some of the very same problems that today surround BDA (Kerr 1975). While the concerns of BDA in academia and industry are currently articulated in debates about technologies and techniques, the researchers that attended the first VLDB conference in 1975 discussed the very same concerns under the thematic titles of: data description models, physical structures and implementation, database design tools, performance and restructuring, application and management issues, distributed databases, security and integrity as well as system and memory architecture, which encapsulate much of what appears in current debates on Big Data.

Before embarking to the opportunities and challenges of Big Data, however, it is worth stressing that dealing with the neighboring concept of Big Compute is beyond the scope of this paper. Big Compute is concerned with the issues emerging from the application of computational-resource-hungry approaches, such as genomic sequencing (Marx 2013), and weather modelling (Malakar et al. 2013). Big Data and BDA on the other hand are distinctive due to their scope and scale, as well as the complexity of the multilevel, multidimensional nature of the datasets. Whilst in some cases Big Data and Big Compute share approaches to deal with difficulties - such as High Performance Computing (HPC), the focus of Big Data is the rapid exploration of the complex, multi-dimensional, multi-level data, while the focus of Big Compute is on optimization of the software and hardware.

Opportunities and challenges of Big Data in the networked world

The networked context that we live in is one in which trillions of transactions and behaviors are being recorded daily by a large number of networked sensors and devices. Such data are related to ourselves, our friends, our preferences and our location, and can include intimate information about us, such as our sleeping or eating habits (World Economic Forum 2011). In addition to this, every day internet users from around the world collectively send more than 45 billion emails, and submit more than 95 million tweets (World Economic Forum 2013), and generate 2.5 quintillion bytes of data (IBM 2013), and this is only set to grow (World Economic Forum 2014). As the data accumulates, the management and strategic leverage of such information resources becomes a critical success factor in creating competitive advantage. This ability to harness vast collections of data is expected to give rise to new opportunities for economic and societal value creation (World Economic Forum 2011). It has already given rise to a total of 4.4 million IT-related jobs globally to support BDA (Gartner Group 2012), whilst the technology and services it spurs are expected to project growth rates at about seven times the value of the overall Information and Communication (ICT) market, reaching a market value of $16.9 billion by 2015 (European Commission 2014). Concurrently, it is expected that this will enhance the economy overall and enable the creation of novel business models that relate *inter alia* to the trading of such data. The literature, however, also highlights the need to develop requisite skills and capabilities for businesses to leverage the acquired data, and to evaluate the relevance of issues associated with contextual and social settings from which the data originate when analyzing and interpreting data derived from heterogeneous sources and timeframes.

Organizational Challenges

Big Data evangelists argue that escalation in *volume*, *velocity*, and *variety*, can be exploited to radically improve the overall organizational performance, and based on the idea that "you can't manage what you can't measure", organizations around the world have ultimately ended up collecting more data than they know what to do with (McAfee and Brynjolfsson 2012). Much of this data, however, comes in a messy and unstructured form, from diverse and incompatible sources (Davenport and Patil 2012), and the failure of investments in Big Data to deliver performance improvements is attributed to the lack of capabilities and skills required to leverage the data they acquire (Davenport and Patil 2012; McAfee and Brynjolfsson 2012; Ross et al. 2013). Kruschwitz and Shockley (2011) show that 30% of 4,000 executives said that their biggest analytics challenge lay in not knowing how to use the data they already have, while only a third of the participants said they had access to the information and analytics they needed to do their jobs successfully. McAfee and Brynjolfsson (2012) found that organizations that engaged with Big Data were confronted by challenges in five areas: leadership, talent management, decision-making, company culture, and technology. They assert that companies that succeed in the age of BDA are not the ones who have more or better data, but those that have better leadership, set clear goals, can define what success looks like, and have the ability to ask insightful questions: strategic decision-making will always be critical for organizations, and in the era of Big Data this is fortified with insightful analytics. They also suggest

that in an era when data are cheap and easy to acquire, and statistics is a skill that is easy to master, competitive advantage for organizations over the next decade will derive from recruitment and retention of data scientists who have the ability to work with large datasets, write code to run analytics on them, and are able to talk 'business' and support leaders to reformulate challenges in ways that Big Data can tackle.

Technical Challenges

Based on the extant literature in IS and Business Information Systems (BIS) on BDA, we identify three key technical challenges: i) data limitations, ii) hardware limitations, and iii) software limitations for Big Data and BDA in and around organizational settings. In this section we discuss these three critical limitations before proposing our architecture.

Data Limitations

The construction of large databases that combine data from a wide and diverse spectrum of sources can result in messy data with its attendant limitations. Data taken out of context, lose, to some degree both their meaning and value. Whilst there is value in data abstraction, it is critical to maintain knowledge of context in both data and analysis and their impact on the outcome of the inquiry (boyd and Crowford 2012). The scale of Big Data makes it harder to retain the context in the analysis and can result in misinterpreted outcomes. Moreover, the source of the data – apart from the context – needs to be taken into consideration when analyzing Big Data. Combining, for instance, several databases from a number of diverse data sources furnishing various daily measurements over a number of decades in order to create a longitudinal dataset, without taking into account where these datasets come from, what was measured in each dataset, and the conditions under which it was measured, can result in misleading information and possibly deleterious outcomes. Retaining the context and source of the data, thus, has acquired increasing importance in BDA discussions, and is an essential element of the novel architecture we propose.

Hardware Limitations

Both the size of the data and the resource-hungry analytical approaches required for BDA give rise to a number of hardware limitations. One common issue is the need for storage, since real-life Big Data sets tend to not fit off-the-shelf storage solutions, and in some cases are difficult to migrate from one facility to another even over the internet. Whilst distributed computing is a one-way-street for BDA, the limits of CPU and RAM are pushed continuously with BDA, and storage increasingly becomes an issue with hardware bottlenecks limiting what data scientists can do with their coding abilities and wealth of data. For a more thorough and technical analysis of hardware limitations of BDA, which is beyond the scope of this paper, see Jacobs (2009).

Software Limitations

The needs of BDA go beyond the capabilities of off-the-shelf statistical software packages, and this constitutes another significant limitation. Data scientists tend to write code that pushes the hardware to its limits in programming languages such as Python and R. Python is a widely used programming language among data scientists and researchers dealing with BDA. It emphasizes code readability, and enables the expression of concepts in fewer code lines. Moreover, it provides constructs that enable clear programs on both a small and large scale. It features a dynamic type system and automatic memory management, has a large and comprehensive standard library, and supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. On the other hand, R is a statistical programming language, also widely used among data scientists for developing software, and for data analysis. Researchers and data scientists have used R to create packages that allow the creation of additional functions on top of existing ones, reducing in this way greatly the time spent on programming for BDA software. Both Python and R are currently widely used in both academia and industry amongst data scientists and researchers that deal with BDA. Whilst the rapid growth of BDA will result in libraries of code to be readily available for most routines in the years to come, at the moment, this still represents a significant limitation for BDA as for most routines new code need to be written by data scientists.

Ethical Challenges

Apart from the managerial and technical challenges highlighted above, there are significant ethical challenges associated with BDA. Whilst the value proposition of BDA can be attractive, the significant ethical questions it presents may fragment the public in troubling ways (boyd and Crowford 2012; Pariser 2011). Cohen (2014) notes that surveillance is becoming increasingly privatized, commercialized, and participatory. He suggests that it is no longer perceived as something to fear and regulate, and BDA turns it into a source of innovation. For example, in many cases gamification is deployed to induce users to hand over their data to companies. The notion, however, that public surveillance drives innovation in some way, does not make it acceptable in the face of the ethical questions that it gives rise to. Offering up part of one's private life to gain services such as personalization or even security is something that users need to be increasingly skeptical about. Whilst the collection of data and intimate information about peoples' sleeping or eating habits can give rise to new market models, it also represents a significant ethical challenge for BDA. As boyd and Crowford (2012) argue, the fact that we have access to such a vast amount of data does not mean it is ethical to use them. The dangers posed by BDA for the users is a critical issue, especially when the users do not understand what happens to their data after handing them to companies and governmental agencies. Acquisti and Gross (2009) for instance demonstrate how combining public databases can reveal an individual's Social Security Number, leading to serious privacy violations and generating critical ethical questions about the use of Big Data and BDA. Access to, and use of such data carry a number of implications for both academics and practitioners who increasingly have access to such data sets.

One of the first studies that helped in pinpointing this problem came from Lewis et al. (2008) who collected the Facebook profiles of 1,700 Harvard-based students in order to examine how their friendships and interests change over time. Whilst the dataset was initially anonymized when it was released openly to the public for other researchers to explore, it soon became apparent that it was possible for a significant part of it to be deanonymized, leading in this way to compromising privacy, rights and security of the students none of whom were aware that their data were being collected. The challenges of dataset deanonymization, however, have been troubling researchers for much longer: For example Latanya Sweeney (2000; 2002) demonstrated that very few characteristics are needed to uniquely identify a person: according to her findings on research using the 1990 U.S. Census summary data, 87% of the US population had reported characteristics that likely made them unique based only on their 5-digit ZIP code, gender, and date of birth, while about 53% were likely to be uniquely identified by only their place, gender, and date of birth. This poses number of critical ethical questions: should such data sets be used without explicit informed consent and without the users knowing exactly what the data will be used for? What exposure to danger exists or is permissible from the analysis of such a dataset and its release to the public sphere, and what are the implications for the individuals who are not able to withdraw details partly or in full from the dataset if they wish to not have them analyzed? Industry, appears to be more prone to misuse private data than academics, since the latter must meet strict rules through ethics committees on any research on human subjects, whilst in contrast, companies are largely free from such constraints, and already have wide latitude to snoop on, and mine, their user databases (Butler 2007).

The study by Kramer et al. (2014) reveals another aspect of the ethical challenges that surround BDA: Where the line should be drawn on what is ethical and what is not when it comes to research with BDA as, clearly, what is legal is not necessarily always ethical. Their study demonstrates that BDA can be easily used to manipulate SNS users' emotions and behaviors without them knowing that their data are being used for research purposes without their consent. Any data on human subjects inevitably raise privacy issues (Butler 2007), and the real risks of abuse of such data are difficult to quantify (Nature 2007).

As a response to the organizational, technical and ethical challenges we presented here, in the following section we present a novel architecture that ultimately addresses the fundamental challenges of BDA and preserves users' security, rights and privacy.

Divide and Conquer

As highlighted in the previous sections of the paper, there are two sets of problems associated with BDA: the first is concerned with storing and curating the data, and the second is concerned with processing the large amounts of diverse data whilst retaining contextual significance and preserving users' security,

rights and privacy. Both sets of problems result in increased costs of storage and computing resources. Big Data eventually come to data scientists in big volumes, but also in diverse formats from diverse sources. Most importantly, however, both problems are also implicated in compromising users' security, rights and privacy.

In order to deal with them, and taking into account the bottlenecks that computational resources can cause, data scientists need to employ analytical strategies and techniques that enable them to apply computational approaches to the data. Such strategies usually entail batching the data in smaller sets based on their nature (such as a data source) or the goal of the analysis (such as a period of time, or based on a location of interest) or by just using a clustering technique to minimize the data for the analysis. Another strategy is the simulation of the dataset using features of the Big Data set to create a smaller, more manageable one with similar properties, and run analytics on that, expecting results that would be close enough to the results one would have obtained by running the analytics on the Big Data set. Thus, the reality with Big Data analytics is that the wealth – in both size and variety – of the dataset is almost never actually used. But the goal always remains the same: turn data into information, and information into insight, whilst the ethical questions regarding users' security, rights and privacy remain critical. These observations provide the motivation for our proposed architecture which liberates organizations from the continual effort of maintaining costly infrastructures, and concurrently allows a higher degree of preservation of users' privacy, rights and security.

Public vs Private Data

Whilst the leveraging of Big Data and BDA has its benefits for businesses and governments, internet users are increasingly sensitized to the way in which companies use their data, and may feel like they are being carried along without fully understanding what is happening to their information. Companies can carry out detailed analyses of customer data to find new revenue streams, but the individuals who generated the data in the first place are excluded from the equation when their own data are used. Companies are able to obtain broad rights over the data they gather by using "take it or leave it" terms of service (ToS). If the users want the service an organization provides, they have to agree to their data being used in certain ways. Even when a company is socially responsible and offers reasonable ToS, there is always a possibility that it could merge or be bought by another company that is less scrupulous about what it does with the personal records.

In 2011, the US government established the Smart Disclosure programme, which is driving the release of both public and private sector data in a standardized, machine-readable format, aimed at helping the public make better choices about services such as healthcare, finance and energy. One of the projects underway as part of the programme is the *Green Button*. This is an initiative that allows customers to download their personal energy data in a standard, electronic and portable format. They can then use applications to manage their energy consumption, and, ultimately, lower their bills. Currently, 27 million households in seventeen US states have access to their own energy information. American consumers can also click on the *Blue Button*, a public-private initiative to give patients access to their health data electronically. Through the Blue Button, users can start compiling their personal medical history, check if the information is correct and have it ready in case of emergency, or when switching health insurance companies. They can also plug their own health information into applications and tools to help them set personalized health goals.

Extending this capability across different aspects of people's daily lives, including their weekly shop and the number of phone calls they make, can enable parties holding such data to build a very accurate picture of who individuals really are (Angelopoulos et al. 2008). In reality, however, consumers are unlikely to want to handle large amounts of data, even if it came from them in the first place. The flip side of the consumers' concerns is the concern of the organizations that currently hold such data: examples include banks, credit card companies, retail chains, communication providers, and SNS, who may incur major reputational risks and liabilities in handing over data to third parties. Thus, privacy remains a serious concern, with many companies still struggling to display the needed transparency, thoughtfulness and rigor to win the trust of their users.

Sometimes smaller is better

Boyd and Crawford (2012, p. 668) ascribe as the core myth of BDA "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy", and assert that: "[t]oo often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions". Correlations that are intuitively appealing may be preferentially selected or over-interpreted to generate misleading causal explanations. In a notable example, Leinweber (2007) demonstrated that data mining techniques could show a strong but spurious correlation between the changes in the S&P 500 stock index and butter production in Bangladesh. It is thus possible for BDA to provide "destabilising amounts of knowledge and information that lack the regulating force of philosophy" (Berry 2011, p. 8), making in this way computational sociology an *ontotheology* that sparks "a new ontological 'epoch' as a new historical constellation of intelligibility" (Berry 2011, p. 12), representing a "profound change at the levels of epistemology and ethics" (boyd and Crawford 2012, p. 665).

Big Data is thus characterized not only by its immense size, but, more critically, by the complexity of its multi-level, multi-dimensional composition. The loss of context can lead to misleading results and the management of the large volumes of data collected under various conditions and using diverse parameters for defining the population from which it is collected is costly, and prone to moral hazard. Moreover, this gives rise to power imbalance between the originator of the data and the organizations that make use of it.

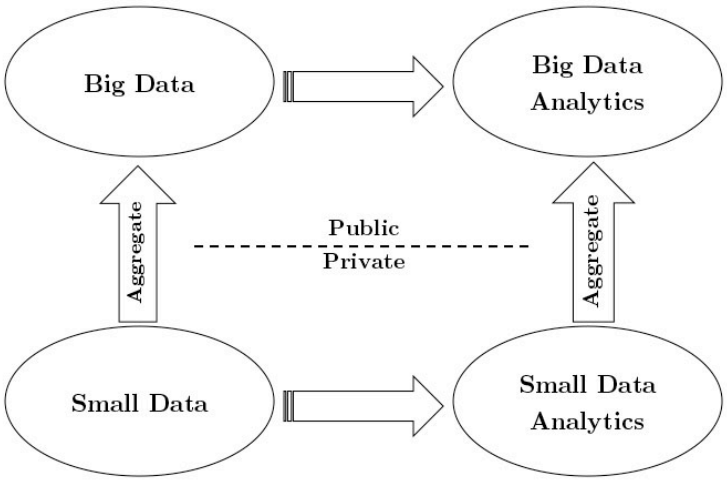

Figure 1. Big Data Analytics vs Small Data Analytics

BDA require costly resources and the creation of new analytical approaches in terms of both modeling and writing the code for them, as well as costly resources in terms of computational power and storage capacity. Analysts to date –be it in academia or in the industry– use Small Data and Analytics (SDA) to understand the world by collecting small samples of data, applying common analytical techniques while using off-the-shelf statistical packages, and aggregating the results of their analysis to the whole population from which the sample came from, using the notion of statistical significance of their results. Data Scientists use the dataset of the whole population to create smaller samples and understand parts of the population (such as topical analysis on twitter data for a certain event that took place at a certain place in a certain time). BDA, thus, address the whole population through the use of a sample from a dataset.

As we demonstrate in Figure 1, the results of BDA can be achieved with high enough accuracy and much lower cost by using SDA. In this way, data are kept private at a user level, and there is no need to store all the data for all users in costly infrastructures. If the users keep their data in a private level, an organization that wants to conduct analytics for certain attributes can simply locate individuals by using

the required set of attributes, and run analytics directly on their private dataset, which is stored on the user-side, without the need to pay for storing all these data on their own servers. Such an approach that bridges the merits of BDA and SDA is fundamentally different to random sampling from within a population, as it can target directly the users that are of interest, regardless the size of the set: can be applied for just one user, for personalization of a service based on his preferences, or to a whole population with a certain attribute for conducting analytics. Such an approach enables the ability to create multidimensional datasets of known provenance by retaining the originator information and context, and reduces the burden of storage and management by deploying distributed, high granularity storage which is maintained by the originator and thus mitigates against the existing power imbalance. A significant advantage of this approach is that the users opt at their discretion to provide or not their data to the service for this. In order to achieve this there is a need for a shift in power from the originators of the data to the organizations and institutions that make use of that data. A systemic perspective that includes both sides, shifting from a customer-provider relationship to a more symbiotic one, would address the issue of asymmetrical distribution of power. In the following section we present a novel architecture for BDA based on the concept we described here.

A Novel Architecture for Big Data Analytics

We propose a novel architecture for BDA that takes into account at its very core the fundamental challenges of the concept without compromising users' security, rights and privacy. Our architecture does not require modification of existing services; instead it is predicated on providing a connecting layer over them. It is informed by the development of the core internet networking protocols that provide a way for proprietary computer networks to interoperate. The core of the internet architecture as we know it today is founded on the separation between the Transmission Control Protocol (TCP) and the Internet Protocol (IP), with the former responsible for addressing and transferring data between hosts, implicitly identified with computer network interfaces, and the latter responsible for transferring data between processes within the computer network. This architecture had a profound commercial impact. The simplicity of the IP meant that it could easily be ported to run over almost any underlying technology (Waitzman 1990). Over the course of the 1970s, 80s and 90s this led to the diminishing in importance of proprietary Local Area Network (LAN) technology in favor of support for IP. On the flip side, software developers could cease caring what particular flavor of network they were operating over, and simply assume that IP, and their choice of transport protocol, were available. Referring to our outline of the internet's evolution and building on the idea of SoC, we view traditional commercial systems as analogous to proprietary computer networking solutions: whilst initially they may have been seen as successful and satisfactory for many of their users, they were too restrictive to enable the explosion of use that the internet subsequently saw. Only by interposing a simple interconnection could the complexity of development for these proprietary computer networks be contained, and eventually their utility be accessed.

Decentralized approaches address that weakness to some extent, but a subtler problem remains. By baking the data types handled by the system into the data exchange protocols, users must either cast the data they wish to exchange into the formats supported, or install expensive, brittle and bug-prone gateways to interconnect different networks. Again, referring to our outline of the internet's evolution, this is analogous to the early development of the internet protocols. Only by the SoC between IP and TCP could the combination of absolute flexibility and a simple and uniform protocol interface be provided to the users. Finally, a key requirement for providing the levels of access control, communication privacy and authenticated identity required by such disparate and personal interactions is the ability to securely and coherently generate, manage and distribute secrets. Only by using consistent encryption mechanisms to process information in a way that prevents it from being read by an interceptor (Goldreich 2009, p. 374) can we begin to meet the complex needs of users without compromising security, rights and privacy.

Our approach is based on the idea of SoC, which refers to the correlation and delineation of system elements to ultimately achieve order. Through this, the complexity of the system becomes manageable offering both robustness and resilience. The core idea behind SoC states that the elements of a system should have both singularity and exclusivity of purpose, which in practice means that within the system there should be no element that shares responsibilities of another element. Therefore, the establishment of logical or physical constraint delineating a given set of responsibilities achieves SoC in the system. SoC involves the division of a set of responsibilities to organize the system into elements of non-repeating sets

of cohesive responsibilities. The overall goal is to establish a well-organized system where each part fulfills an intuitive as well as meaningful role, maximizing in this way the ability of the system to adapt to change. This can result in a number of benefits for both organizations as well as the end users, since it can aid in the management of complexity of the system by eliminating duplication and provide responsibility allocation.

There are five main residual benefits of applying such an approach in the design of the system:

- Singularity of purpose of the components renders the system easier to maintain
- Stability of the system as a byproduct of the increased maintainability
- Extensibility of the system as a byproduct of element singularity
- Adoption of the system as a byproduct of maintainability and extensibility
- Decoupling of the system as a byproduct of element singularity

The architecture we propose inspires a paradigm shift on the way that user data are stored, accessed and analyzed by organizations: instead of storing personal data in costly-to-maintain corporate datacenters distributed around the world that are also subject to different legal systems, all user data are stored on the user side, providing full control to the user on who has access to the data and for what purposes. Any type of personal data can be stored on the user-side, such as credit card transactions, medical records, super market loyalty cards data, as well as data related to their SNS accounts and communication. This can give rise to novel business models in relation to the use of the data, their storage, as well as to novel business models associated with the user-based server.

In the following section we discuss the implications of our architecture through which we advocate such a move in the world of online social networking, in order to both, address problems of privacy and identity and to open up online social networking to a much richer set of possible applications.

An application on Networked Organizations

Online social networking is now in its second decade, and has become a central activity for a large proportion of the global population, as SNS are among the most used web sites today. This shift is not surprising, as humans are social animals with a need to connect and communicate with each other. SNS augments individuals' existing offline social networks, allowing them to keep in touch with people over great distances, share their experiences and associated content, organize their social lives and discover new contacts beyond physical reach. Many users maintain multiple online identities through which they actively manage their social interactions on the various SNS (Golbeck and Rothstein 2008). As social beings, people tend to participate in different, overlapping social groups, and adjust their identities to match the contexts, and adjust their use of SNS to match the constraints imposed by the various SNS. The reasons why people choose to explicitly manage the overlap among social networks, even keeping some networks completely distinct from others, are commonplace and usually not sinister. For example, teenagers wishing to discuss sensitive health matters in online forums (van der Velden and El Emam 2013), employees complaining about treatment at work (O'Brien 2014), or those engaged in political commentary in uncomfortable or dangerous situations (Attia et al. 2011). Similarly, one's identity, and the way one behaves and communicates, is likely to be different in a professional context from how it is in the context of a group of friends, though one may share individuals between professional and personal social networks.

As a result, users may choose to distribute, or not, certain content within certain SNS due to the technological affordances as well as the terms imposed by those SNS. The need of SNS users to incorporate multiple context-dependent identities emerges from the fact that they are inevitably actors in overlapping social networks that extend both online and offline (Angelopoulos and Merali 2015), but also expand horizontally to include their multiple roles such as their professional life, their personal life, their family life, or their hobbies and interests (Guha and Birnholtz 2013). The disembodiment imposed by the computer-mediated environment transforms online social interactivity amongst users on SNS into a constant negotiation of identities. Through their social interactivity, users on SNS project their public social identity in a way they would do in an offline setting (Angelopoulos and Merali 2015; boyd 2010; Lindqvist et al. 2011). Users, however, only present a limited number of their multiple identities at a time, based on the context in which they are situated (boyd, 2010). The self-management of users' multiple

context-dependent identities implies a process in which the users control how the other users with whom they are socially connected perceive them within a certain setting (Baumeister and Leary 1995; Leary et al. 1995).

The extant computer-mediated communication literature highlights the needs of a diverse range of groups to incorporate and maintain multiple identities on a plethora of media, such as hobbyists like cigar smokers (Angelopoulos and Merali 2015) and bodybuilders (Ploderer et al. 2010), as well as professionals who need to separate their professional from personal lives (Peluchette et al. 2013). The self-management of multiple context-dependent identities represents a topic that deserves greater attention within IS and needs to be further explored and elucidated, since it incorporates the entanglement of online and offline interactivity within and around computer-mediated environments, which can have significant implications for the overall sociability of SNS users (Angelopoulos and Merali 2015). The IS literature is increasingly interested in how users manage their multiple identities as actors in overlapping social networks that extend both online and offline. Although the need for users' multiple context-dependent identities to be further explored and elucidated is highlighted in the literature (Karl and Peluchette 2011; Peluchette et al. 2013; Talamo and Ligorio 2001) to date there are very few studies exploring the issue directly (Talamo and Ligorio 2001).

Proposed Architecture for SNS

Internet users commonly use several email addresses to distinguish personal and professional communication, and in some cases this is imposed on them. For instance, internet users might have their personal email address that provides ease of use and longevity, as well as having to use their corporate or university email for their professional / academic communication, and it is not uncommon for people to make use of separate personal email addresses to keep their personal interests separate, for instance maintaining their SNS and Online Communities accounts separate from their friends and family email account. Most importantly, the only point at which these identities must exist together is within the user-clients in the private context of our personal devices such as mobile phone, tablet and personal computer.

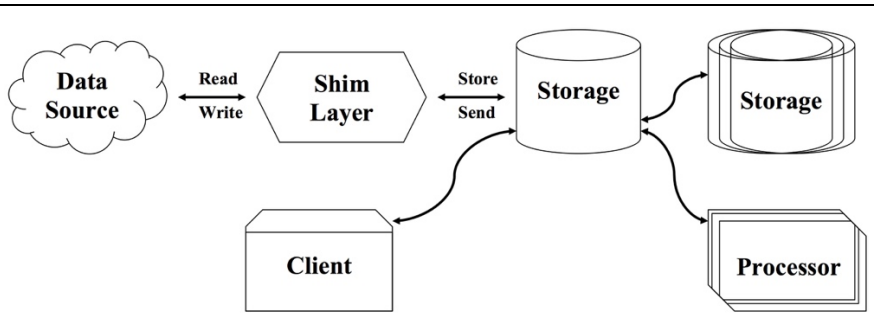

Figure 2. Information architecture for the prototype application

We propose implementing a simple structure for user-provided content that would enable contributors to continue discussions in a natural timeline style, while still providing sufficient cues for effective automated processing of the content. The novel approach we propose here to understand and manage this, as shown in Figure 2, is based on the SoC architecture, and is built upon three independent layers: a shim layer, a storage layer and a processing layer. With interoperability in mind, the architecture works in the same way as an email client would operate on any device and operation system, and aggregates the social networking interactivity of users on a single interface through the secure use of email protocols.

Mail servers make use of two protocols for sending and receiving email messages over the internet. For sending messages most email software is designed to use the Simple Mail Transfer Protocol (SMTP), which is a set of communication guidelines that allow applications to send emails. The Multi-Purpose Internet Mail Extensions (MIME) is an extension of the SMTP that lets users to exchange different kinds of data files on the internet such as audio, video, images, and software, as well as text. Through MIME an email can store almost any type of data along with metadata relating to addressing and provenance. At the

same time, MIME is understood by an enormous range of software clients allowing us to experiment with the storage of social network message types as MIME messages, while reusing existing software for message display.

As an open standard, using MIME as a storage mechanism allows the creation of new applications for the display of messages *without requiring changes to the underlying format*. This enables us to build, for example, clients supporting multi-party encryption of messages across a range of social network transports, and providing intelligent Carbon Copy (CC) lists. There are two other protocols that are used for retrieving and storing email: The Post Office Protocol 3 (POP3) and the Internet Message Access Protocol (IMAP), which was developed as an alternative to POP and enables users to synchronize their email across multiple devices, providing in this way a feature that is extremely important today, when most people use at least two devices to read and respond to their email communication with colleagues and / or friends.

While proprietary corporate email systems and cloud-based webmail providers use their own non-standard protocols to access email accounts on their own servers, all use SMTP and IMAP when sending or receiving email from outside their own systems. The standard email distribution and access protocols of SMTP and IMAP allow us to use a range of existing software services for the transmission and storage of information.

The approach is presented schematically in Figure 2, where 'Data Source' represents the various SNS from which the application reads/writes updates to the 'Shim Layer' that ensures compatibility of the ongoing updating of the various SNS Application Programming Interfaces (API). Then the data are stored to the 'Storage Layer' and to the 'Replication Repository', which is responsible for the redundancy of the system and both reside on the users' side. The 'Processors' are responsible for the intelligence behind the application, and connect the 'Shim Layer' with the 'Storage Layer' before the information arrives to the user through the client. In the following section we discuss the three layers of the approach in detail.

Shim Layer

In mechanical engineering, a shim is a very thin piece of material used to fill gaps, make something level, or adjust something to fit properly, and that is what pretty much a shim layer does in a software application. Shims are compatibility layers that help with the overall fit when there are compatibility issues in an application: the shim acts as a compatibility layer that resolves issues that may arise during the development of newer versions of applications, or to execute applications on software platforms for which they were not originally designed. It is mainly used to provide code extensibility, and prevent code duplication, which can result to lower deployment costs and less maintenance. The shim layer in our application is responsible for connecting with existing SNS in order to read and write data. Principally this means aggregating social networking data (messages, contacts, media, etc.) from multiple SNS into a common storage format as well as acting as a distribution channel for the transmission of data. Moreover, the shim layer is also responsible for the collection of the special data, such as the ToS of each SNS that the user has an account on, that are used later by the processing layers of the applications. The Shim layer, thus, is an important element of our application as it is responsible for the communication with the SNS, and the translation of all incoming data in a unified way to be used by the other two layers of our application.

Storage and replication layer

Storage of the data is a significant element of the system, and its redundancy through replication is as important. Data replication involves the sharing of information to ensure consistency between redundant resources, to improve the reliability and fault-tolerance of the system, as well as data accessibility. The storage and replication layer in our application is a distributed repository for data collected by the shim layer and acts as an intermediary for client applications to communicate with, obviating the need for clients to communicate directly with SNS. Such a combination enhances the overall resilience of the system by providing a layer that transparently ensures the redundancy and efficacy of the overall system use. The storage and replication layer of our application enables it to operate both in real-time, getting data from the Shim layer and feeding them to the processing layer, as well as by storing data for future use, based on the user preferences. In this way, special data such as ToS can be saved once and be updated

when needed instead of being updated on every communication. Moreover, data regarding past communication can be stored for use by the processing layers to optimize communication in the future connections.

Processing layer

The processing layer is conceptually an input and output pipe, which connects the shim layer to the storage layer: data flowing through this layer, is scanned and optionally transformed by *processors*, which are essentially small, agent-like applications. The processing layer is intended to be pluggable with published API so enabling third-party developers to extend and experiment with the platform. The proposed processors represent the element of *intelligence* in our architecture, enabling greater control on the appearance and presentation of published content to specific recipients and on specific SNS.

Our architecture has two processors. The first processor is an adaptive filter for message reception. It is common for people to have accounts on different SNS, and to have overlapping groups of contacts on them. It is also common to send the same message through different SNS. Typically, this means that a user can receive the same message multiple times. An adaptive filter detects duplicate messages from different sources and filters them out so that the receiver gets the message only once. In a similar way, if a user is sending a message targeted at a specific user or group of users, an adaptive filter could potentially send the message through the SNS in which the recipient user is most likely to respond on. By using social network analytics at the user-level, the processor can conduct real time analysis of past communication between users, and based on the results identify the optimal channel that a message should be sent through to the recipient in order to be seen and responded to faster.

The second processor helps users understand better the consequences of posting copyrightable content onto different SNS that allow them to post content like photos and videos where the media itself is hosted by the SNS. In order to do this, the ToS of the SNS specify that the user must transfer some rights of the media file use to the service. Whilst this may be acceptable for a lot of users, there are also a lot of situations under which transferring some rights to the service may not be acceptable. To address this, we propose an intermediate processor in our architecture through which a user can specify which rights they are happy to share, and which they do not wish share with the SNS; the intermediate processor could then identify instances in which license 'collisions' may occur and warn the user before posting the content to the SNS. The application of our proposed architecture in the domain of online social networking can be enhanced by the implementation of more pluggable processors in the processing layer based on the specific user needs, providing in this way the element of intelligence in our architecture and concurrently enabling third-party developers to extend and experiment with the platform.

Key Features

The proposed architecture enables the user to determine the rights, privacy levels, as well as the use of their data, and to redress some of the extant asymmetries in the balance of power between the user and the service provider as discussed in the earlier sections. The key features are outlined below.

Identity, Integrity and Privacy

The key features for an inter-SNS layer equivalent to the IP network layer are transport-independent, addressing format standardization for referring to data distributed through a particular SNS, and flexible –but standardized– support for use of asymmetric encryption for per-service, per-recipient authentication and privacy. The linking of users' multiple identities is only performed in the client side, and relies on information received out-of-band to indicate that the different identities belong to the same individual. In cases where identities refer to multiple distinct pods, they could be completely opaque, having semantic meaning only for those who already know how to resolve them. The service provider is at liberty to structure their messages as much as they see fit. In some cases, a message might be nothing more than text, or an image; in others, a message might have very rich structure, with extensive metadata in addition to the raw content. This naturally enables asymmetric communication through simple key management APIs, and services might also support authentication of message exchange, with only trusted clients able to interpret incoming messages, becoming trusted through out-of-band mechanisms or face-to-face interaction.

User Selective Filtering and Streamlining Receipt and Delivery of Communications

The proposed architecture can provide integrated filtering of communication through the various SNS in such a way that each message that is distributed through the various SNS is received only once from an end user. At present it is common to receive the same status update from the same person through multiple channels as many people choose to link their various SNS accounts for convenience, and post the same content on all of them. The standardization across alternate transport-layer names and content metadata that a social inter-network layer provides, makes it far easier to create clients that can intelligently manage content presented across multiple transports, and to filter based on content or on transport, or on a combination of these two. Moreover, the proposed architecture can provide intelligent CC lists. It is common for a user to wish to distribute content to multiple recipients, and different transports provide different mechanisms for doing so. Given the frailty of human memory, it can be difficult to recall for a given piece of content to which exactly it would be relevant. There are many cases, however, where exercising control in this regard is extremely valuable: family members may not be interested in seeing updates concerning hobbies; it can become embarrassing to share family-specific photos with work colleagues; and it may be career limiting to allow your manager to see what you did last weekend (Watson et al. 2012). Our approach enables an intelligent contacts application that remains aware of the distribution properties of the various SNS, while coalescing the multiple identities that the users' contacts may have. This allows the application to make suggestions concerning who should receive the message, based on its content and the inferred interests of contacts, as well as the properties of the transports available to reach the recipient and inferred knowledge about which contacts should not receive the message.

Partitioning of Social Identities

The proposed architecture can be used to separate data that users do not wish to be combined for the construction of their identity within a certain SNS. The reasons why users choose to explicitly manage the overlap among their social networks, even keeping some networks completely distinct from others, are usually completely normal and not in the least clandestine. Attempts to support this richness in the social networks of users via access control mechanisms (e.g., Facebook lists, Google+ Circles), have proved largely inadequate, since whilst users understand how these mechanisms work, the cognitive effort required for creation and maintenance results in either their mis- or non-use. Moreover, such access control mechanisms do not allow the user to have power over their data. With SNS being increasingly eager to combine diverse data sources in order to construct the identity of users, collating all of one's social interactions and data into a single service gives rise to serious security risks such as identity theft.

Implications

The implications that emerge are trifold: for the users, for organizations, and for the society as a whole. The proposal that all user data are stored on the user-side enables the user to control who has access to the data and for what purposes. Users thus regain their privacy, rights and security over both their data and their private lives. Furthermore, in delivering services, the proposed architecture enables providers to have lightweight, agile applications tailored to be relevant to real-time individual contexts. Thus, it can give rise to novel business practices in and around networked organizational settings, with regard to the strategies for storage and use of the data, and in spawning business models related to the affordances of the user-based server. Moreover, changing the paradigm of the way data are stored and accessed may act as a catalyst to precipitate a "change the entire social theory that goes with them" (Latour 2009, p. 9).

Limitations

The application of our architecture in the SNS domain admittedly is faced with a number of limitations arising from real-world difficulties, concerning the users and the providers of such services. First of all, the main limitation of the application of our architecture in the SNS domain is that it requires the users to take responsibility of their own data for their storage and redundancy. We acknowledge that this is an optimistic endeavor, since users tend to resist changes in the ways they do things (Lapointe and Rivard 2005; Markus 1983; Sykes et al. 2014). We envision, however, that this limitation of the architecture will give rise to novel business opportunities that will provide solutions for the user-side data storage and

redundancy. Moreover, such a need will inspire the open source community to come up with innovative, secure, reliable as well as cost-effective applications for the for the user-side storage and redundancy of data. As an example, we can envision user-side Raspberry Pi based Network Attached Storage (NAS) servers running on Linux.

A second limitation arises from within the application of our architecture on the SNS domain. The central focus of social networking is the message (e.g. text, picture, video), that is, some data that is conveyed by the network from the originating user to a defined list of other users. The various existing SNS have specific message structures that broadly fall into the same generic structure of metadata, text and attachments, a generic structure also shared by email. Each one of the various existing SNS, however, have their own message structures, and therefore a critical component in developing the application of our architecture in the SNS domain is to implement the mapping of these message structures to email. Whilst this is part of the design of our architectural approach in order to ensure interoperability of the service, it does represent a limitation, and we need to look further into mapping the various SNS message structures to email.

Last but not least, a third limitation for the application of our architecture in the SNS domain arises from the ease of adoption from the existing SNS. The existing SNS generate their revenue through targeted advertising based on their users' preferences as extracted from content creation. The application of our architecture in the SNS domain will most probably be seen as a barrier, as it could negate the provision of user data to the SNS, and thus, the generation of revenue through targeted advertising. However, there is growing advocacy and public pressure for users' privacy, rights and security to be protected, and our approach addresses the asymmetrical distribution of power between the originators of data and the organizations and institutions that make use of that data by taking a systemic perspective to include both sides in our architectural design, shifting from a customer-provider relationship to a more symbiotic one in which control over access to customer data resides with the customer. Whilst the architecture we propose might not be directly implemented by the organizations behind the existing SNS themselves, for the reasons outlined above, it can inspire novel business endeavors through the use of the various SNS APIs for implementation from third-parties.

Conclusions

In this paper we focused on the challenges and issues associated with BDA, and proposed a novel architecture that uses the principles of SoC and distributed computing to overcome many of the challenges associated with storage, analysis and integrity, that are directly linked to users' privacy, rights and security.

We addressed the issue of asymmetrical distribution of power between the originators of data and the organizations and institutions that make use of that data by taking a systemic perspective to include both sides in our architectural design, shifting from a customer-provider relationship to a more symbiotic one in which control over access to customer data resides with the customer. Last but not least, we illustrated the affordances of the proposed architecture by describing its application in the domain of SNS, where we furnish a mechanism to address problems of privacy and identity, and create the potential to open up online social networking to a richer set of possible applications, referring also to the limitations that this presents.

We believe that our study is timely and important for both IS research and practice, and contributes to the broader IS literature as a design study that leads to the development of a novel system architecture; a stream that has been at the core of IS Research, addressing how novel IS can be designed, and how IS can assist organizations in designing better organizational systems with improved business processes and workflows.

In this paper we claim three contributions. First, we provide a clear perspective on the distinctive features of Big Data and BDA, and their use in organizations. Second, we motivate and implement a novel architecture based on the idea of SoC and informed by the development of computer networking, able to address the fundamental challenges of BDA without compromising users' security, rights and privacy. Third, by demonstrating the utility of our architecture in SNS we advocate a similar move towards symbiotic models for the preservation of user rights on SNS, both to address problems of privacy and

identity and to open up online social networking to a much richer set of possible interactions and applications.

Going forward, there are two significant directions for future research emerging from our study. The first is to evaluate the architecture we propose and compare it to the existing model. The second is to apply our architecture to domains other than SNS, and to see applications in more conventional domains such as healthcare, banking, commerce, energy use, and governance. When it comes to the application on SNS, it is necessary to revisit the ways that we architect and build online social networking platforms. The approach we propose in this paper will enable greater flexibility, creativity and utility in the exploitation of our social networks, while also providing SNS users with greater control over that exploitation. We believe that doing so will open online social networking up to richer application development, and thus enable the same kind of explosion in the use of SNS that the internet caused with computer networking.

Much is made of the dangers associated with Big Data and BDA, and this is often confounded with issues of personal data, but in fact many of the benefits accrue through Big Data having to any personal identification information. The future of data analytics is admittedly big, and the ability of analytics to support decisions and improve performance is of great importance. In this light, we should keep our focus where we believe our efforts can make the most meaningful impact: on using intelligent data analytics to make smaller bodies of data powerful and actionable. With the proposed architecture we are on track to make a success story of personalized services, as long as they are engineered with "privacy-by-design".

Acknowledgements

We would like to thank the two anonymous reviewers, as well as the organizers of the Track "IS Design and Business Process Management" for their remarks, suggestions, and constructive comments. This work was supported by the Horizon Digital Economy Research, RCUK grants EP/G065802/1 and EP/M02315X/1; and by CREATe, the Centre for Copyright and New Business Models, RCUK grant AH/K000179/1.

References

Acquisti, A., and Gross, R. 2009. "Predicting social security numbers from public data," *Proceedings of the National Academy of Sciences*, (106:27), pp. 10975–10980.

Angelopoulos, S., and Merali, Y. 2015. "Bridging the Divide Between Virtual and Embodied Spaces: Exploring the Effect of Offline Interactions on the Sociability of Participants of Topic-Specific Online Communities," *48th HICSS Conference*, Kauai, Hawaii.

Angelopoulos, S., Kitsios, F. and Babulac, E. 2008. "From e to u: Towards an innovative digital era". In Kotsopoulos, S. and Ioannou, K. (Eds) *Heterogeneous Next Generation Networking: Innovations and Platform*, pp. 427-444, Idea Group Publishing.

Attia, A. M., Aziz, N., Friedman, B., and Elhusseiny, M. F. 2011. "Commentary: The impact of social networking tools on political change in Egypt's "Revolution 2.0"," *Electronic Commerce Research and Applications*, (10:4), pp. 369-374.

Baumeister, R. F., and Leary, M. R. 1995. "The need to belong: desire for interpersonal attachments as a fundamental human motivation," *Psychological bulletin*, (117:3), pp. 497.

Berry, D. 2011. "The computational turn: thinking about the digital humanities," *Culture Machine*, (12:0), pp. 1-22.

boyd, d. 2010. "Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications," in Z. Papacharissi (Ed), *Networked Self: Identity, Community, and Culture on Social Network Sites*, pp. 39-58.

boyd, d., and Crawford, K. 2012. "Critical Questions for Big Data," *Information, Communication & Society*, (15:5), pp. 662-679.

Butler, D. 2007. "Data sharing threatens privacy," *Nature*, (449), pp. 644-645.

Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, *36*(4), pp. 1165-1188.

Cohen, J. E. 2014. *The Surveillance-Innovation Complex: The Irony of the Participatory Turn*. The Participatory Condition (University of Minnesota Press, 2015, Forthcoming).

Davenport, T. H., and Patil, D. J. 2012. "Data Scientist: The Sexiest Job of 21st Century," *Harvard Business Review*, (90:10), pp. 70-76.

European Commission, 2014. "Communication from the Commission to the European Parliament," *The Council, The European Economic and Social Committee and the Committee of the Regions: Towards a thriving data-driven economy*. Available at: goo.gl/1fbwm4

Gartner. 2011. *CEO Advisory: "Big Data" Equals Big Opportunity*. Available at: http://www.gartner.com/id=1614215

Golbeck, J., and Rothstein, M. 2008. "Linking Social Networks on the Web with FOAF: A Semantic Web Case Study," *AAAI*, (8), pp. 1138-1143.

Goldreich, O. 2009. *Foundations of Cryptography: Volume 2, Basic Applications (Vol. 2)*. Cambridge university press.

Guha, S., and Birnholtz, J. 2013. "Can you see me now?: location, visibility and the management of impressions on foursquare." In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pp. 183-192. ACM, 2013.

Hürsch, W. L., and Lopes, C. V. 1995. *Separation of concerns*. Technical report, North Eastern University.

Jacobs, A. 2009. "The pathologies of big data," *Communications of the ACM*, (52:8), pp. 36-44.

Karl, K. A., and Peluchette, J. V. E. 2011. "Friending professors, parents and bosses: a Facebook connection conundrum," *Journal of Education for Business*, (86:4), pp. 214-222.

Kerr, D. S. 1975. *Proceedings of the International Conference on Very Large Data Bases*, September 22-24, Framingham, Massachusetts, USA. ACM.

Kramer, A. D., Guillory, J. E., and Hancock, J. T. 2014. "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, (111:24), pp. 8788–8790.

Kruschwitz, N., and Shockley, R. 2011. "First Look: The Second Annual New Intelligent Enterprise Survey," *MIT Sloan Management Review*, (52:4), pp. 87-89.

Lapointe, L., and Rivard, S. 2005. "A Multilevel Model of Resistance to Information Technology Implementation," *MIS Quarterly*, (29:3), pp. 461-91.

Latour, B. 2009. "Tarde's idea of quantification," in M. Candea, (Ed.), *The Social after Gabriel Tarde: Debates and Assessments*, Routledge, London, pp. 145–162.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. 2009, "Computational social science," *Science*, (323:5915), pp. 721-723.

Leary, M. R., Tambor, E.S., Terdal, S. K., and Downs, D. L. 1995. "Self-esteem as an interpersonal monitor: The sociometer hypothesis," *Journal of personality and social psychology*, (68:3), pp. 518.

Leinweber, D. 2007. "Stupid data miner tricks: overfitting the S&P 500," *The Journal of Investing*, (16:1), pp. 15–22.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. 2008. "Tastes, ties, and time: a new social network dataset using Facebook.com," *Social Networks*, (30:4), pp. 330-342.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. 2011. "I'm the Mayor of My House: Examining Why People Use Foursquare - a Socialdriven Location Sharing Application." In *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 2409-2418.

Malakar, P., George, T., Kumar, S., Mittal, R., Natarajan, V., Sabharwal, Y., Saxena, V., and Vadhiyar, S. S. 2013. "A divide and conquer strategy for scaling weather simulations with multiple regions of interest," *Scientific Programming*, (21:3-4), pp. 93-107.

Manovich, L. 2011. "Trending: the promises and the challenges of big social data," in M. K. Gold (Ed.), *Debates in the Digital Humanities*, The University of Minnesota Press, Minneapolis, MN.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. *Big data: The next frontier for innovation, competition and productivity*, Report, McKinsey Global Institute.

Markus, M. L. 1983. "Power, politics, and MIS implementation," *Communications of the ACM*, (26:6), pp. 430-444.

Marx, V., 2013. "Biology: The big challenges of big data," *Nature*, (*498*:7453), pp. 255-260.

McAfee, A., Brynjolfsson, E. 2012. "Big Data: The management revolution," *Harvard Business Review*, (90:10), pp. 61-67.

Nature Editorial. 2007. "A matter of trust," *Nature*, (449), pp. 637-638, doi:10.1038/449637b

O'Brien, C. N. 2014. "The top ten NLRB cases on Facebook firings and employer social media policies," *Oregon Law Review*, (92:2), pp. 338-379.

Peluchette, J. V. E., Karl, K., and Fertig, J. A. 2013. "Facebook 'friend' request from the boss: Too close for comfort?," *Business Horizons*, (56), pp. 291-300.

Ploderer, B., Howard, S., Thomas, P., and Reitberger, W. 2008. "Hey World, Take a Look at Me!": Appreciating the Human Body on Social Network Sites. *Proceedings of Persuasive Technology*, (5033), pp. 245-248.

Ross, J. W., Beath, C. M., and Quaadgras, A. 2013. "You May Not Need Big Data After All," *Harvard Business Review*, (91:12), pp. 90.

Shmueli, G., and Koppius O. R. 2011, "Predictive Analytics in Information Systems Research," *MIS Quarterly*, (35:3), pp. 553-572

Sweeney, L. 2000. *Uniqueness of Simple Demographics in the U.S. Population*, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000

Sweeney, L. 2002. "K-anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, (10:5), pp. 557-570.

Sykes, T. A., Venkatesh, V., and Johnson, J. L. 2014. "Enterprise system implementation and employee job performance: Understanding the role of advice networks," *MIS Quarterly*, (38:1), pp. 51-72.

Talamo, A., and Ligorio, B. 2001. "Strategic identities in cyberspace," *CyberPsychology & Behavior*, (4:1), pp. 109-122.

van der Velden, M., and El Emam, K. 2013. ""Not all my friends need to know": A qualitative study of teenage patients, privacy, and social media," *Journal of the American Medical Informatics Association*, (20:1), pp. 16-24.

Venters, W., Oborn, E., and Barrett, M. 2014. "A Trichordal Temporal Approach to Digital Coordination: The Sociomaterial Mangling of the CERN Grid," *MIS Quarterly*, (38:3), pp. 927-949.

Waitzman, D. 1990. "Standard for the transmission of IP datagrams on avian carriers," *RFC 1149, IETF.*

Watson, J., Besmer, A., and Lipford, H. R. 2012. "+your circles: sharing behavior on google+". *Proceedings of ACM Symposium on Usable Privacy and Security*, pp. 12:1–12:9.

Williams, D. 2012. "If 'Big Data' Simply Meant Lots of Data, We Would Call It 'Lots of Data'," *Forbes.com*, Available at: http://goo.gl/VHIonn

World Economic Forum. 2011. "Personal Data: The Emergence of a New Asset Class," available online at: http://goo.gl/OWdFPK

World Economic Forum. 2013. "Unlocking the Value of Personal Data: From Collection to Usage," available online at: http://goo.gl/qGD1A3