Work in progress – Not for wider dissemination

Managing diversity in privacy preferences: How to construct a privacy typology

Sören Preibusch
Microsoft Research
Cambridge, CB1 2FB
spr@microsoft.com

ABSTRACT

Privacy types carry the promise of breaking a population that is heterogeneous in their privacy preferences into a manageably small number of customer segments with similar privacy preferences. Westin's tripartite grouping is well-known, but suffers from methodological shortcomings. In this note, I first establish reliability and predictive power as the evaluation metrics for privacy typologies. Second, I outline a methodologically sound path towards identifying reliable groupings. The procedure is demonstrated on an existing dataset of N=1075 survey respondents. Whereas cluster techniques failed to yield groupings that pass statistical and practical quality measures, factor analysis uncovered the dimensionality of privacy concerns. Practitioners could use the insights to build more usable privacy-enhancing technologies.

1. UNDERSTANDING PRIVACY NEEDS

1.1 Balancing privacy and functionality

We live in a networked world where ubiquitous Web tracking and planet-scale government surveillance are not capabilities but realities. The resulting privacy challenges are calling for privacy-enhancing technologies, policies and practices. Delivering those can be difficult, especially when ignoring the true privacy needs of the intended audience. The failure, or at least lack of mainstream uptake of existing privacy-enhancing technologies (PETs) may not only be attributed to their low usability, but also to their low utility as they fail to address the privacy needs of the target user base.

Let's take the example of privacy in Web search. Search is the most ubiquitous Web activity; it permeates our personal and professional lives, at home, at work and on the move. A person's personality, lifestyle and aspirations unfold in the sequence of queries they issue. The consequential privacy concerns are even increasing as major search engines are now integrated with email, calendar, document storage or personal, location-aware digital assistants. Technical data protection approaches, such as query scrambling, come at the expense of search result quality. Similarly, suppressing cues such as location or search history have a detrimental effect on the search result quality.

A recent, 300+ participants' lab experiment into consumers' privacy choices revealed that search engine users do not want unconditionally high levels of protection. Although privacy-enhancing features were universally appreciated, users enabled them on a query-per-query basis, and very significantly more often if the search topics were sensitive. A search engine that disabled the search history altogether would thus not meet users' needs.

Behavioural economics allows studying the relative importance of different features or indeed between privacy and functionality. By introducing a nominal fee for the available search options, an experimenter can reproduce the real-world situation where one 'cannot have it all'. In the trade-off situation when privacy and functionality were priced equally, the demand for features that improved search result quality was more than three times higher than for features that enhanced privacy (e.g., no data sharing with third-party: 14% versus extra result quality: 52%). [1]

1.2 Managing heterogeneity

Service providers face a challenging threefold heterogeneity in privacy attitudes. First, not all users want the same privacy features. Second, not all of them balance privacy and functionality in the same way. Third, users' ultimate privacy choices are shaped by the context, such as the sensitivity of the search task at hand, so that the same user may exhibit variability in expressing her priorities. A first and influential approach to break down this heterogeneity into a small number of manageable buckets was proposed by Westin (Section 3.1). Interestingly, the resulting groupings were interpreted by locating them in the spectrum of privacy / convenience trade-offs. Westin wrote the so-called pragmatic majority would weigh the costs and benefits of disclosing personal details [2].

2. THE PROMISE OF PRIVACY TYPES

Privacy types are used to group together consumers with similar individual privacy preferences. A privacy typology is a categorisation that is often constructed from multiple attributes. As an aggregate, it is assumed to be a more reliable and more valid predictor than its underlying characteristics. Accordingly, the predictive power of a typology established on a sample should also generalise to the underlying population.

Privacy types promise to break down the diversity of privacy preferences into a manageable number of distinct groups. The reduction in complexity such a segmentation provides is appealing when building online services, marketing products, or conducting privacy surveys. For instance, privacy types enable companies to version their products by privacy in a way that maximises the total share of consumers satisfied with the respective privacy settings.

A privacy typology is typically built on stated privacy attitudes rather than observed behaviour. The type emerges from the attitudes as a latent construct, to explain observed behaviour. Established or novel survey instruments may be used for measuring privacy attitudes [3]. This reliance on self-reported attitudes may be surprising given the divergence between stated and actual

> Reliability in the grouping and a reduction in heterogeneity compared to the overall population.
>
> Predictive power of users' types for their privacy behaviours.

Figure 1: The two evaluation metrics of a privacy typology

privacy behaviour [4]. However, privacy types that are intended as a predictor for privacy behaviour should not be learned from it directly, to avoid a circular argument. It is noteworthy that affirmation of the privacy paradox does not necessarily preclude the usefulness of privacy types derived from stated behaviour: it is assumed that the grouping by privacy preferences augments and transcends the underlying self-reports.

3. WESTIN'S TYPOLOGY AND ITS SHORTCOMINGS

3.1 The mechanics of the grouping

A classic privacy typology is given by Westin, who posits that individuals can be classified as "fundamentalists", "pragmatists" and "unconcerned" in 1:2:1 proportions, in decreasing order of privacy concern [5]. Westin's classification is based on three statements with which respondents either agree or disagree on a four-point scale—reliability scores are not reported. Respondents who agree to "have lost all control over how personal information is collected and used by companies" and disagree that "businesses handle the personal information ...in a proper and confidential way" and that "[e]xisting laws and organizational practices provide a reasonable level of protection for consumer privacy today", are classified as fundamentalists [6]. Those with symmetrically opposite response behaviour are classified as unconcerned. Pragmatists are the remainders; this group is applied to six out of the eight possible answer combinations.

In retrospect, Westin notes that fundamentalists would be "high-privacy oriented proponents, [who] rejected consumer-benefit or societal-protection claims for data uses and sought legal-regulatory privacy measures." [7] At the other end of the spectrum, the unconcerned readily disclose their personal details to businesses and the government. The pragmatic majority, largely defined as those pertaining to neither camp, would decide upon disclosure on a case-per-case basis, considering safeguards, benefits from disclosure, as well as legal and organisational control [7].

3.2 Methodological critique of the tripartite privacy typology

In a 1991 "fresh analysis of the 1990 Harris-Equifax Privacy Survey data", Westin argues that the American public divides into three continuing groupings on issues of information privacy involving government, business, and personal morality. Coded as outlined above, these are the Fundamentalists (25%), the Unconcerned (15%) and the Pragmatists (57%) [5]. Together, they are supposed to give "the breakdown of the public" [8].

Whilst Westin acknowledged a "privacy dynamic" for this grouping, and conceded that individuals in the survey population may see their type change, only drifts of pragmatists towards the extremes were considered. At the same time, the original dominance of the pragmatics and the overall relative size of the segments can be attributed to the coding mechanism. Failure to take into consideration the dynamics of people's privacy concerns has attracted criticism [9].

Westin re-affirmed the tripartite privacy typology in 1999 [10], using data from a Privacy & American Business survey that year. Whilst the proportions of the groups changed slightly (11%, 13%, 75%), their number remained the same. Interestingly, the expanded pragmatists group is normatively renamed the "Golden Mean" and their characteristic is that "they want it all – notice, benefits and good privacy policies" [10] when asked whether they were happy providing personal information on the Internet in exchange for various benefits. However, only 457 of the 1014 adults taking part then were Internet users. Their responses are not scrutinised for an emerging alternative privacy typology; rather, they are sorted into the existing tripartite segmentation.

Both the 1991 and 1999 investigations use a survey methodology and results gathered from an American population, which may or may not be prototypical for a post-industrialised society. Moreover these survey results were gathered before the availability of social networks on desktop PCs or mobile phones.

Westin's interview waves have limited use as a longitudinal survey instrument: changes in sponsors for surveys result in changes to survey questions which in turn reduces the validity of inter-temporal comparisons [11, p. 25]. This variation in sponsors, each of which with their own interests, may also have biased the conclusions from the studies [6], as suggested by the Electronic Privacy Information Center (EPIC) as one of the "flaws in Westin's segmentation" [12]. Groups of individuals were repeatedly labelled as "privacy fundamentalists" despite major changes to the survey questions eliciting perceptions of personal privacy, to the answer options, and to the answer coding conventions. The same names of Westin's privacy groups thus refer to different concepts for each year [2]. This change in wording is further exacerbated by the over-sensitivity of the grouping mechanism to small changes in respondents' opinions [6].

Whilst the Westin studies are unable to detect changes in privacy attitudes, it is not necessarily true these would have happened. For instance, data from the 2008 Eurobarometer into EU citizens' perceptions about data protection provides evidence that the level of concern has only changed slightly since the early 1990s [13]. However, it is important to re-evaluate the use of a particular typology at regular intervals. So far, the Eurobarometer series remains the only survey intended to support longitudinal comparisons over the four waves carried out in the expanding European Union between 1991 and 2008; the 2011 wave on "Attitudes on Data Protection and Electronic Identity in the European Union" no longer provides inter-temporal comparisons of privacy preferences.

The consistent choice of a pejorative label for consumers concerned about data protection has been criticised repeatedly [12] [6]. The term "privacy fundamentalist" has become decoupled from the original segmentation and proliferated in the academic literature. Other pejorative wording can also be found, for instance "too much privacy fuss" [7], which would be rejected by the unconcerned.

In summary, the tripartite privacy segmentation promoted by Westin appears to suffer from a number of *methodological shortcomings* that limit its scholarly applicability. These include: naïve partitioning with an unmotivated choice of three as the number of groups; conceptualisation of privacy types as a one-dimensional construct; undocumented or arbitrary procedures of how to construct the segmentation; and how group membership should be decided for a given subject in subsequent studies.

It is tempting to argue that the tripartite grouping has its merits as a 'practitioners' tool' despite its methodological shortcomings, if only it provided useful insight. This interpretation is questionable because empirical evidence suggests that the scale fails to provide reliable segmentation. Sometimes the supposedly homogeneous extreme groups, notably privacy fundamentalists, do not appear to

form a cohesive group for privacy-related decision making [14]; or the typology is just inapplicable [15] [16].

4. EMPIRICALLY MANIFESTED DIVERSITY

This section is devoted to a simple case-study into constructing a privacy typology from surveyed privacy preferences. Using an existing, recently collected dataset, I exemplarily investigate whether a privacy typology could be discovered in a data-driven manner. Continuing the earlier discussion of privacy economics in Web search, the typology shall be built on top of stated privacy/functionality trade-offs. I therefore use the consumers' willingness to prioritise functionality over privacy, which had been repeatedly emphasised in interpreting Westin's groupings.

4.1 Dataset

Microsoft recently sponsored an *N*=1075 online survey in the United States and four countries in Europe (Belgium, France, Germany and the UK) to measure people's perceptions regarding digital privacy [17]. Commissioned for Data Privacy Day 2014, with data collection in November 2013, the survey specifically recruited 'technology elites', characterised by owning a smartphone, tablet and/or computer, and self-identifying as an influencer on technology and as an early adopter of new technology. The geographically stratified respondents can thus be understood as a rather homogeneous population.

Considering device usage as an example, there are no significant differences across the five countries for PC / desktop (between 91% and 96% adoption). The differences for tablets or e-readers are also small (from 58% in Belgium to 75% in the UK, with France and Germany both at 61%; US: 65%). There are stronger difference in smartphone penetration rates, with 94% in France, but only 68% in Belgium, which is the country where a sizeable proportion still uses feature phones at most (22%) or no mobile phone at all (10%). There is no difference by gender for device use.

Respondents were balanced in terms of age and gender (Table 1).

Table 1: Respondent's demographics are not significantly different across the regions.

Age	**US**	**EU**
18 - 24	10%	12%
25 - 34	22%	28%
35 - 44	21%	23%
45 - 54	22%	17%
55 - 64	14%	13%
65 and older	11%	8%
Gender	**US**	**EU**
Male	49%	51%
Female	51%	49%

4.2 Key privacy attitudes

The survey delivered key findings into consumers' perceptions of data privacy issues [18]. The protection of consumers' online privacy is perceived as a shared responsibility. Tech elites expect companies to shoulder about 30% of this responsibility, and expect technology companies to deliver innovations that automatically protect individuals' privacy. Europeans see a greater role for government than Americans.

Consumers believe they have limited control over how personal data is used [18]. Tech elites estimate they control around half of the way their information is used online, with Europeans feeling less in control than their American counterparts. Furthermore, Americans are more willing to make privacy trade-offs for routine online activities (e.g., shopping and banking), while Europeans place lower value on ease-of-use (Figure 2). Let us use these reported priorities of function over privacy to investigate whether consumers can be segmented into a privacy typology.

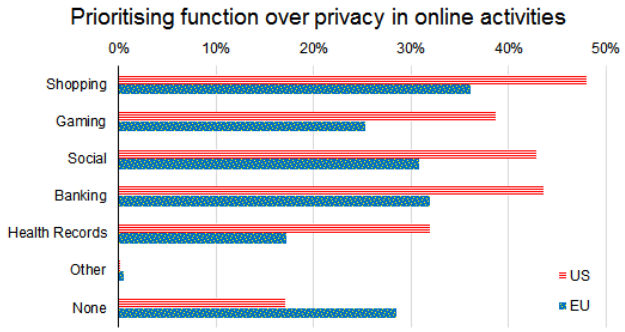

Figure 2: "Are there certain activities you engage in online for which you might prioritize function and ease-of-use over privacy?"—Differences between US and EU respondents in their willingness to trade privacy for functionality

4.3 Segmenting users by their privacy-functionality trade-offs

The ability to segment users by their privacy-functionality trade-offs would be very useful to service providers, as it would unlock the ability to tailor the user experience for these often conflicting goals.

The first question is: how many customer segments to use? This number should emerge naturally from the dataset—rather than being chosen a priori. Rules of thumbs exist to choose cluster counts. A more evidence-based criterion is the Calinski-Harabasz statistic that reaches a maximum for the suggested number of clusters. Here, k=2 is the first maximum of the variance ratio criterion (VRC, Figure 3). At the same time, it seems that two segments is not necessarily the wisest choice: there is an upward trend for the VRC as the number of clusters increases, suggesting that a larger number of segments would be better. The next best option would be k=11. Obviously, choosing higher numbers defeats the purpose of a typology to deliver a manageably small number of segments.

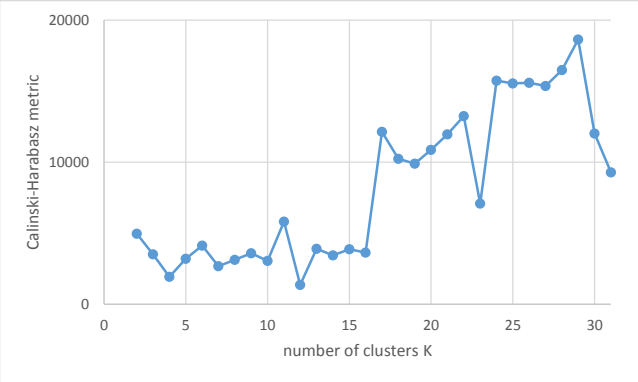

Figure 3: Using the Calinski-Harabasz variance ratio criterion (VRC) for choosing the appropriate number of clusters (plotted for *k* between 2 and 31).

Respondents are k-means clustered into *k*=2 clusters. Almost two thirds would rather not prioritise privacy over functionality for any online activity (Table 2: Cluster 2). The remaining users would be willing to renounce privacy for better functionality for almost every online activity (Table 2: Cluster 1).

Table 2: Cluster centres

Online Activity	*Cluster 1 privacy priority*	*Cluster 2 functionality priority*
Shopping	Yes	No
Gaming	No	No
Social	Yes	No
Banking	Yes	No
Health Records	Yes	No
Respondents	35%	65%

Service providers will typically want to guess users' cluster membership to be able to offer a personalised experience and to greet the customer with smart defaults. The users' location, which may be inferred from their IP address through geo-location, from sensor readings or from language settings, is a good indicator: respondents from the US are very significantly ($p < 0.0001$, Chi-squared test) more likely to be in Cluster 1 than EU respondents, meaning that they are more likely to prioritise functionality over privacy.

Table 3: Cluster size by country / region

Country / Region	*Cluster 1*	*Cluster 2*
US	42%	58%
Belgium	24%	76%
France	28%	72%
Germany	28%	72%
UK	32%	68%
EU	28%	72%

4.4 Reliability and coverage of the clustering

The usefulness of a privacy typology is first assessed by its ability to manage the heterogeneity of the population: along the principles of clustering, the heterogeneous user population should be broken down into few, homogeneous segments. Cronbach's alpha is a well-known measure for internal consistency (other statistical dispersion metrics for multivariate nominal data may also be used, such as the generalised variance analog).

However, the clustering failed to carve out users with similar privacy preferences. Starting with a low α=0.56 for the overall population, which reflects the inherent diversity in users privacy preferences, the resulting clusters have α=0.15 and α=0.05, respectively.

A more practical approach to assess the usefulness of the resulting clustering would be to examine how many users would be satisfied if offered the choice between the two settings represented by the cluster centres. A user is satisfied if the service balances functionality and privacy in a way that is compatible with her own preferences. However, the first cluster only satisfies 5% of users, the second another 23%. If given the choice between the two settings represented by the cluster centres, 73% of the population would not find their preferences matched.

4.5 Factor analysis: determining the dimensionality

It is plausible that respondents' tendency to prioritise functionality over privacy will be similar across the various online activities. In this case, the dimensionality of the privacy typology would be lower than the number of underlying variables. Examining the co-occurrence matrix between the different online activities for which respondents prioritised functionality indicates that their responses are indeed systematically associated.

A factor analysis can be applied here to extract the principal components (Kaiser-Meyer-Olkin criterion: 0.66). The scree plot of Eigenvalues suggests extracting two factors (Figure 4), which is also the number of factors with Eigenvalues above one.

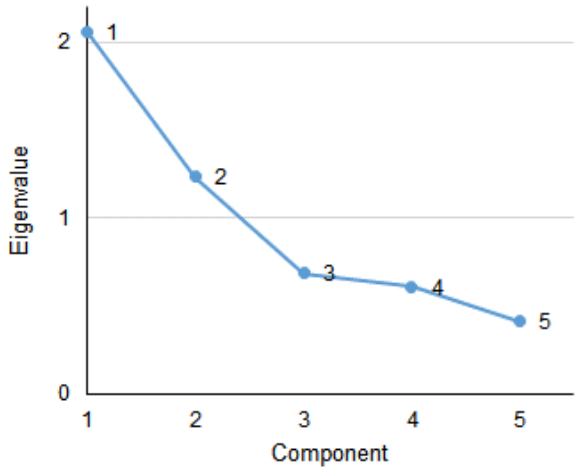

Figure 4: Scree plot; the knee is at 3, suggesting the extraction of two factors. Also, only two components have Eigenvalues above 1.

The factor loadings of the Varimax-rotated components are given in Table 4. The different activities load unequivocally on two factors: Factor 1 stands for 'serious stuff', where one's money and health records are at stake; Factor 2 comprises pastime activities.

Table 4: Factor loadings

Online Activity	*Factor 1: 'serious stuff': money and health*	*Factor 2: gaming and socialising*
Shopping	**0.753**	0.132
Gaming	-0.200	**0.816**
Social	0.285	**0.751**
Banking	**0.856**	0.003
Health Records	**0.787-**	0.053

4.6 Practical implications of dimensionality

From a practical point of view, knowing the dimensionality is important for at least two engineering and design challenges. Online activities which load on the same factor will solicit similar behaviour from consumers.

Firstly, user interfaces that allow consumers to choose appropriate levels of privacy versus functionality can alleviate the configuration burden by grouping activities that load on the same factor. An example is third-party content that is loaded by Websites. These resources (e.g., images, scripts) are often used for tracking users, but sometimes disabling those breaks the browsing experience. Browsers could be configured to be more lenient on gaming or social networking sites, whilst blocking more aggressively on banking sites. The browser options only need to surface two configuration settings instead of five, while still offering enough granularity for users to make meaningful choices.

Secondly, the factor loadings indicate that users who care about privacy in banking will probably also care for privacy in shopping. If the user has made several privacy-enhancing choices while banking online (e.g., secure networks and connections, not remembering passwords for these sites), then the browser could offer a smart default to make similar choices for online retailing sites. The browser could also alert the user when shopping over an unsecure connection, whilst avoiding such warnings for gaming Websites—and thereby reducing the warning fatigue.

5. CONCLUSION

Privacy types are used to group together consumers with similar individual privacy preferences. They carry the promise to make the heterogeneity in consumers' privacy preferences manageable. From this follows the desideratum that the resulting segmentation should yield groups of users with more homogeneous privacy preferences.

In this note, I set out to demonstrate how to establish a possible privacy typology using cluster techniques and factor analysis. Users are segmented by their willingness to prioritise privacy over functionality—a trade-off that has been at the heart of interpreting Westin's tripartite group, which however suffers from numerous methodological shortcomings.

Even for the simple scenario examined here, the segmentation reached through clustering is not convincing: it fails to reduce the homogeneity amongst users, and service providers would be better off by simply tailoring their user experience by country or region (e.g., US versus EU).

One may ask why the clustering approach failed here, although it seems to have worked so splendidly in earlier investigations. First, there are still only a few published works that truly try to uncover a natural segmentation in the population rather than binning respondents into three pre-conceived buckets. Second, statistical dispersion metrics have so far not been in the forefront when discussing privacy typologies.

The analysis into the dimensions of privacy attitudes, through factor analysis, delivers statistically more convincing and practically helpful results. Users were found to prioritise privacy for certain groups of online activities, which can be neatly dissected (in this case gaming/socialising versus money/health related activities). Such bundling of activities provides helpful support in engineering more usable privacy-enhancing technologies.

While this first analysis focused on reliability issues of a privacy typology, future work should also examine the predictive power of an established grouping for privacy behaviours. Given the divergence of privacy attitudes and behaviours, this will be a challenging task.

6. REFERENCES

[1] S. Preibusch, "The value of privacy in Web search," in *Twelfth Workshop on the Economics of Information Security (WEIS 2013)*, 2013.

[2] P. Kumaraguru and L. F. Cranor, "Privacy indexes: A survey of Westin's studies," Institute for Software Research International, School of Computer Science, Carnegie Mellon University, 2005.

[3] S. Preibusch, "Guide to measuring privacy concern: Review of survey and observational instruments," *ernational Journal of Human-Computer Studies*, vol. 71, no. 12, pp. 1133--1143, 2013.

[4] S. Preibusch, D. Kübler and A. R. Beresford, "Price versus privacy: an experiment into the competitive advantage of collecting less personal information," *Electronic Commerce Research*, vol. 13, no. 4, pp. 423--455 , November 2013.

[5] A. F. Westin, "Harris-Equifax consumer privacy survey 1991," Harris, Louis and Associates / Equifax Inc., 1991.

[6] R. Leenes and I. Oomen, "The role of citizens: What can Dutch, Flemish and English students teach us about privacy?," in *Reinventing Data Protection?*, S. Gutwirth, Y. Poullet, P. Hert, C. Terwangne and S. Nouwt, Eds., Springer Netherlands, 2009, pp. 139--153.

[7] A. F. Westin, "Social and political dimensions of privacy," *Journal of Social Issues*, vol. 59, no. 2, pp. 431--453, 2003.

[8] J. Baker, J. Goldman, M. Rotenberg, A. F. Westin and L. Hoffman, "Personal information privacy-i," in *Computers, Freedom & Privacy*, 1991.

[9] A. Morton, “Move Over, Westin,” in *Dagstuhl Seminar 13312: "'My Life, Shared' – Trust and Privacy in the Age of Ubiquitous Experience Sharing"*, 2013.

[10] O. R. Corporation and A. F. Westin, ““freebies” and privacy: What net users think,” Privacy & American Business, 1999.

[11] G. Iachello and J. Hong, “End-user privacy in human-computer interaction,” *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 1, pp. 1--137, 2007.

[12] Electronic Privacy Information Center (EPIC), “Public Opinion on Privacy (section “Alan Westin and Privacy ‘Fundamentalists’ ”),” 2009. [Online]. Available: http://epic.org/privacy/survey/.

[13] The Gallup Organization, “Data Protection in the European Union – Citizens' perceptions (Summary),” 2008.

[14] C. Jensen, C. Potts and a. C. Jensen, “Privacy practices of internet users: Self-reports versus observed behavior,” *International Journal of Human-Computer Studies*, vol. 63, no. 1-2, pp. 203--227, 2005.

[15] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert and P. Powledge, “Location disclosure to social relations: why, when, & what people want to share,” in *SIGCHI Conference on Human Factors in Computing Systems (CHI 2005)*, 2005.

[16] A. Khalil and K. Connelly, “Context-aware telephony: privacy preferences and sharing patterns,” in *20th Anniversary Conference on Computer Supported Cooperative Work (CSCW 2006)*, 2006.

[17] Microsoft, “Trustworthy Computing - Data Privacy Day,” January 2014. [Online]. Available: http://www.microsoft.com/en-us/twc/privacy/data-privacy-day.aspx.

[18] Edelman Berland; Microsoft, “Microsoft Trustworth Computing: 2013 Privacy Survey Results,” January 2014. [Online]. Available: http://download.microsoft.com/download/A/A/9/AA96E580-E0F6-4015-B5BB-ECF9A85368A3/Microsoft-Trustworthy-Computing-2013-Privacy-Survey-Executive-Summary.pdf.