Computing Trust Levels Based on User's Personality and Observed System Trustworthiness

Conference Paper · August 2015

DOI: 10.1007/978-3-319-22846-4_5

4 authors, including:

Michalis Kanakakis
Athens University of Economics and Business
12 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Shenja van der Graaf
imec
61 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Costas Kalogiros
Athens University of Economics and Business
19 PUBLICATIONS **56** CITATIONS
SEE PROFILE

Computing trust levels based on user's personality and observed system trustworthiness

Michalis Kanakakis, Shenja van der Graaf, Costas Kalogiros, Wim Vanobberghen

AUEB
GR1034, Athens, Greece
76, Patission Str.
T: 0030-2108203154
{kanakakis, ckalog}@aueb.gr

iMinds-SMIT, Brussels, Belgium
{a.van.der.graaf, wim.vanobberghen}@vub.ac.be

Abstract. In this article, we describe an approach for computing the current trust level of individual users towards an online system and present initial validation results from a small-scale experiment. This trust computational model relies upon survey research for identifying the set of key trust attributes and grouping users into four segments of expected behaviors. Each user's initial trust level is computed based on a set of assumptions tailored to the specific segment she belongs to, while the trust level evolution takes additionally into account the system outcomes she has experienced so far. More specifically, the trust update follows a machine learning approach, where during the training phase that consists of a small number of system outcomes, users are asked to report their actual trust levels. Finally, we demonstrate the trustors' segmentation validity and trust estimation accuracy by performing a small-scale experiment in the context of a fictitious online security service.

Keywords: trust computational model, trust, trustworthiness, trustor attributes, survey.

1 Introduction

The increasing complexity to attain trust in trustworthy Information and Communications Technology (ICT) systems and the conditions that affect it, has warranted continuous scrutiny from researchers in various domains. While trust is important in the real world too, it is said to be especially complex to achieve and sustain in Internet-based marketplaces due to the lack of the providers' physical presence and in certain settings the rare frequency of transactions between two entities [1], [2]. In this view, the need for models of trust and credibility in technology-mediated interactions can be detected, particularly, those that are not-domain specific and technology-independent [3]. These models can offer guidance for researchers across disciplines examining a

range of technologies and contexts, thereby highlighting multiple subcomponents, such as associated with antecedents (i.e. preconditions of trust), processes of trust building (e.g., interdependence), the context of shaping trust-building (e.g., social relations, regulation), decision-making processes in trust (e.g., rational choice, routine, habitual), implications and uses of trust (e.g., interpersonal entrepreneurial relations, moralistic trust), and lack of trust, distrust, mistrust and repair (e.g., risks, overtrust, trust violations) [4]. In addition, much of this research seems to mainly address how to optimize user trust.

In this study, we have taken the, arguably, complimentary perspective to examine how different trust-related user experiences are guided by different sets of trustor's attributes underpinned by aspects of well-placed trust and trustworthy behaviors. The reason for developing an approach conditioning the trust levels to individual entities, is that trust formation is a dynamic, or, contextual, yet subjective process, drawing attention to the presence of drivers, such as of a social, economic and legal nature [4], [5], [7]. More specifically, trust is approached as a property of an entity (known as the trustor) reflecting the strength of her belief that engaging in an online system (called the trustee) for some purpose will produce an acceptable outcome [8]. Here, a trustee's trustworthiness is defined as an objective measure (probability) of the provider's ability to produce an acceptable outcome, assuming consensus on the criterion for determining whether an outcome is acceptable or not. We argue that whenever such a criterion is not obvious it could be defined by a regulatory authority, or in the extreme case set by the dominant provider.

Estimating the current user's trust level can be useful for a provider of ICT systems/services both at design-time and run-time. In the former case, knowing the trustors' current trust level and the effects of both desirable and undesirable outcomes on them would allow her to predict the actual demand and set the optimal combination(s) of trustworthiness level and price. Obviously, failing to predict the true demand would result either in missed opportunities for higher revenues, or higher costs. In the latter case, a provider should meet users' expectations in order to avoid customer churn and do so in a cost-effective way at run-time. Thus, whenever the provider believes that a user's trust is lower than a certain threshold the former could make the necessary changes to system in order to regain its trust.

Our main contribution, therefore, is to propose a conceptual trust computational model that allows a provider to estimate the trust level of candidate trustors. Our approach differentiates among trustors based on their attributes and highlights their influence on trust. Our aim is to cover all the phases of the computation process: before engaging with the system and after observing evidences about its performance. For example, it is expected that a successful system outcome will not decrease the user's trust in the system, and similarly, an unseccessful outcome should not cause an increase. Thus, the trust computational model is based on system behavior instead of user behavior (such as eye gaze). Against the standard methodology of the well-known Bayesian models, which follows common initialization and evolution of trust among individuals, we introduce a modification to capture the attributes making trust subjective. The wide range of attributes affecting reactions of trust vis-à-vis ICT systems, motivated us to execute a user survey and identify the key drivers to be consid-

ered as trust indicators. Based on this analysis, the trustors were grouped into segments of expected behaviors and their properties are formulated via the variables of the modified Bayesian model. In order to demonstrate the validity of our approach we performed a small-scale experiment in the context of a fictitious online security service.

The remaining sections are structured as follows; Section 2 introduces the basic computational model that forms the basis for the proposed models. Section 3 presents the trustors' segments, using survey research, that were found to be statistically significant and Section 4 describes the initialisation and update process of the personalised trust computational model. Then, Section 5 presents the experiment setup and the validation results, while Section 6 motivates our work by providing an overview of trust computational models that explicitly consider trustor attributes. Finally, we conclude the paper and provide our future steps in Section 7.

2 The Basic Trust Computational Model

We consider a system characterized by a wide range of trustworthiness factors, notated as J, e.g. reliability, availability, etc. In this work, we focus on factors resulting in outcomes of binary form, i.e. they may be characterized either as a success or a failure. The performance of the system for factor $j \epsilon J$, is determined by its actual trustworthiness notated as $w_{j,s}$; the probability of a successful transaction.

For each of them, any individual trustor estimates the trustworthiness in terms of a random variable θ which follows the Beta distribution and is determined by two parameters "α, β" for specifying the current beliefs. The choice of this particular distribution is inline with the related work (e.g., see [11] for more details). Given that these parameters can capture all factors that result in trust being subjective (for instance different trustor attributes) we use the notation $\alpha_{i,s}^{j}(t)$ and $\beta_{i,s}^{j}(t)$, where the indicators i, j, s, t stand for the trustor, the metric, the system and the time respectively. From now onwards, we simplify this notation by keeping only the indicators i, j and using s, t when necessary. Thus, for each trustworthiness factor there is an objective probability quantifying its trustworthiness level and a subjective trust level estimating the former. Such a fine-grained approach should give the provider the flexibility to identify, at run-time, the reason(s) for low trust and react accordingly.

Mathematically: $\theta | a_i^j, \beta_i^j \sim Beta\left(a_i^j, \beta_i^j\right)$, with probability density function (PDF):

$$f\left(\theta; a_i^j, \beta_i^j\right) = \frac{\theta^{a_i^j - 1}(1-\theta)^{\beta_i^j - 1}}{\int_0^1 y^{a_i^j - 1}(1-y)^{\beta_i^j - 1}\, dy}, a_i^j, \beta_i^j > 0 \quad (1)$$

Over this context, trust should be considered as the subjective probability that the system will provide a successful outcome in the next single transaction, and equals the expected value of the $Beta(\alpha_i^j, \beta_i^j)$:

$$\tau_i^j = E[\theta | \alpha_i^j,\ \beta_i^j] = \frac{\alpha_i^j}{\alpha_i^j + \beta_i^j} \quad (2)$$

Utilizing a PDF allows us to calculate not only trust, but also the confidence, i.e. the probability that the actual trustworthiness lies within an acceptable error range around trust. In general, higher value of α parameter indicates higher trust level (for equal values of β), while confidence depend on their respective sum (for equal trust values).

The values of these parameters can evolve over time, reflecting the trustor's ability to interact with the system further and use those outcomes for getting a more accurate idea of its trustworthiness. The Beta distribution is also appropriate for the update phase, mainly because the process results to the same prior-posterior distributions (before and after an outcome is observed). Indeed, if x stands for the binary outcome of a single transaction, then x follows the Bernoulli distribution with parameter θ, i.e.

$$x|\theta \sim Bern(\theta) \rightarrow f(x|\theta) = \theta^x(1-\theta)^{1-x}, x = 0,1 \quad (3)$$

Thus, for prior $Beta(a_i^j, \beta_i^j)$ the posterior distribution for parameter θ is as follows:

$$\theta|x, \alpha_i^j, \beta_i^j \sim Beta(x + \alpha_i^j,\ 1 - x + \beta_i^j) \quad (4)$$

Note that if the outcome is successful ($x = 1$), the α_i^j parameter will be increased by one, while β_i^j parameter will be increased by one in the opposite case, ($x = 0$).

3 Trustors Segmentation

For the examination of the conceptual dynamics underpinning trust-related user experiences and sets of trustor attributes, input from different stakeholders was sought. The focus was to yield insights into their trust perceptions and appetite towards digital technologies, in particularly the Internet. A two-step approach was followed. The first step consisted of survey and interview research where the stakeholders targeted were derived from members of the public (or, (end)users), the business community, and governmental institutions. Based on a thorough literature review focusing on designing ICTs supporting (mediated) transactions, the exploratory empirical investigation focused on drawing out several key aspects of trust, particularly, antecedents, processes of trust building, the context of shaping trust-building, decision-making processes in trust, implications and uses of trust, and lack of trust, distrust, mistrust and repair ([4]).

In doing so, we sought to draw out the combined underpinnings of relevant (socio-legal-economic) trust drivers, and which guided the main categories for which data were collected. Questions were asked about the disposition to and perceptions of trust, cost of trust, content and information quality, legal constraints, organisational trust, and demographics (user, organizations). These constructs were operationalized with using five-point rating scales open questions, checklist questions, and ranking questions.

As it was the aim to have a reliable question format and a good wording and order, the questions were pre-tested with a group of 142 respondents determining the effectiveness, the strengths and weaknesses of the questions. A principle factor analysis (PCA), therefore, was conducted to detect relationships within the data set generated

by the survey in order to yield insight into the underlying structure of trust elements. PCA works by revealing existing linear components in the data set and the way specific variables contribute to that component. First, 49 items were checked for their suitability by screening for high correlations (R<.9) and significance values over .05 (N = 142). This led to the removal of one item. The Kaiser-Meyer-Olkin value was .850 and Bartlett's Test of Sphericity was highly significant (p<.001), both indicating a good sampling adequacy. The PCA revealed 11 components with eigenvalues exceeding 1. The first component explained 14.7% of the total variance and all components combined, explained 61.1% of the total variance. A closer inspection of the scree plot and running the Monte Carlo parallel analysis indicated that the first few eigenvalues for the randomly generated data matrix scored below the observed eigenvalues from the reduced matrix of data. As a result, it was decided to retain five components based on their explained variance and the outcome of the reliability analysis (>.3). Together they accounted for 50.01% of the total variance. A Varimax rotation was used to help in interpreting the components: Disposition to trust (e.g. stance towards trusting another person or organization), trust management (e.g., tradeoff between personal information disclosure vs accessing an application), trust constraints (e.g., availability of legal guarantees, price), information and content quality (e.g., trust cues, transfer). The results from pre-testing were then used to adjust problematic questions in the questionnaire before releasing the questionnaires to the target groups. In February and March 2013, N= 203 responses served as input.

While the first step served mainly to learn about combined constructs in trust-related experiences and attributes [9], the second step was to conduct a 'segment-specific' analysis so as to learn about different types of subjective trust-related user experiences in this context. Examining the results of the (end user) survey (N = 90) linkages between different sets of trustor attributes could be associated with trust-related concepts of (1) Trust stance: the tendency of people to trust other people across a wide range of situations and persons; (2) Trust beliefs in general professionals; (3) Institution-based trust; (4) General trust sense levels in online applications and services; (5) ICT-domain specific sense of trust levels; (6) Trust-related seeking behavior; (7) Trust-related competences; and, (8) Perceived importance of trustworthiness design elements. And, which underpin the segmentation of trust-related user experiences on trustor attributes.

For the analytical exercise, a K-means clustering was performed for segmentation purposes and an Anova analysis was conducted to test for each item whether statistical significance differences could be retrieved between the uncovered trust-related user experience segments. Some iterative clustering and testing led us to a four segments solution to best explain differences in trust-related user experiences. These segments can be represented by the following terms, with the corresponding abbreviation to be used for the remaining of this article: "High trust" (HT), "Ambivalent (A) trust", "Highly active trust seeking" (HATS) and "Medium active trust seeking" (MATS). They differ on a number of aspects (see below), however, based on our analyses, three major concepts are sufficient to explain their core differences. The three underpinning concepts are 'trust stance' (e.g., 'I usually trust a person until there is a reason not to'), 'motivation to engage in trust-related seeking behavior' (e.g., 'I

look for guarantees regarding confidentiality of the information that I provide') and 'trust-related competences' (e.g., 'I'm able to understand my rights and duties as described by the terms of the application provider'). They could be measured on 3, 7 and 4 item-scale with a reliability coefficient of .69, .89 and .87 respectively. From this a few items could be further reduced to the summarized Table shown below:

Table 1. Segmentation results for the three underpinning concepts.

	Total (n=90)	HT (n=24)	HATS (n=28)	MATS (n=18)	A (n=20)	Anova	
	Mean	Mean	Mean	Mean	Mean	F	Sig.
Trust stance	3,22	3,85	3,15	2,86	3,50	7,260	**,000**
Trust related seeking behaviour	3,52	3,14	4,27	3,34	3,01	24,383	**,000**
Trust related competences	2,44	2,71	2,42	2,94	1,44	13,361	**,000**

The user experience for the "HT" segment can be characterized by a high level trust stance. This means an overall high trust level for the various online applications, such as social networks and online banking, accompanied by only few trust seeking behaviors, such as checking trust seals, even though the competences are present to cognitively assess the trustworthiness of online applications and services.

For the "HATS" segment, the user experience can be highlighted in terms of a high level of trust seeking behavior beyond the mere scanning of trustworthiness cues. It also suggests that individuals are informed about procedures in case of harms and misuse. It points to the capacity of certain competence level that facilitate the assessment of trustworthiness and to possess, at least, a minimal understanding of the rules and procedures to look for in case of complaints and misuse. Varied trust stance and trust levels could be observed including medium to low trust stance/trust levels.

For the "MATS" segment, the user experience is similar to the "Highly active" one, yet, here, trust seeking behavior is not so apparent. Thus, while drivers for trust seeking behavior, such as a low trust stance, are present as well as competences to assess trustworthiness, people's motivation may be absent to look for trustworthiness cues.

The "A" trust segment seems to highlight a clear perceived inability to assess the trustworthiness of online applications and services and which may be explained by the personal competence level. Hence, only few active trust seeking behaviors can be observed, yet do not equal low trust levels per se. Trust seems to be derived from either the general trust stance or basic heuristics, such as 'public organizations are more trustworthy than commercial companies'. It seems that the "Ambivalent" nature of this user experience can be explained by a failure to cognitively assess the trustworthiness and a certain need to trust in order to avoid, or to lower the omnipresence of cautious and other negative feelings, and which is a so-called 'forced trust' (that is,

trust without trustworthiness evidence and with a possible presence of cautious feelings). These findings point to understanding trustworthiness indicators based on the experience of others (referrals), as the main source of 'trustworthiness information' that is accessible for this cluster, and underlying the outcome of the trustworthiness assessment.

4 Model parameterization, based on segments' properties

In this section, we will present our methodology for transferring the fundamental properties of each segment into the Bayesian trust computational model, both in the initialization and evolution phases. Doing so will allow us to take into account user's personality when estimating it's trust level.

4.1 Trust Initialization

The initial trust level of a user who has never interacted with the system in question before could be based on information present on the system's welcome screen, its past experiences in using other systems, the opinion of others users etc. Here we assume that the user has a glimpse of the actual system trustworthiness by looking at information present on the system's welcome screen (e.g., a page containing certifications, attractive layout, etc.). We call this information 'look and feel' elements. The users willing to invest sufficient amount of time in gaining information about system trustworthiness (or, equivalently those being extremely capable of finding evidences of trustworthiness) before using it would always have a good estimation of the actual trustworthiness. Furthermore, this will be the case regardless of how advanced strategies a provider had followed in order to deceive users (adopting for example techniques from social engineering).

Let d_l, m_l, c_l stand for mean values of trust stance, motivation and competence respectively, where $l = 1,2,3,4$ indicates the segment "HT", "HATS" "MATS" and "A" respectively. Additionally, let $e_l = \frac{m_l + c_l}{2}$ be the factor quantifying the aggregate impact of the two latter concepts.

In general, we consider that the closeness of initial trust to actual trustworthiness, depends on the combined impact of both the motivation to engage in seeking behavior and competences concepts, while trust stance determines whether it is under or overestimated. In order to compute the error magnitude and its sign we utilize the segmentation results (see Table 1). More specifically we follow a normalization approach using the second segment (HATS) as a benchmark, since it was found to achieve the highest "e" factor among all and thus users therein estimate trustworthiness accurately. Trustors in all other segments make an estimation error proportionally correlated to the normalized value of "e", i.e.: $\tilde{e}_l = \frac{e_2 - e_l}{e_2}$. Furthermore, under or overestimation is determined by the correlation of the trust stance values, e.g., if $(d_l - d_2) > 0$ the estimation error is added to the actual trustworthiness level.

In a mathematic formulation, the initial trust of user i in segment l is given by:

$$\tau^{j}_{i\epsilon l,s}(0) = \frac{\alpha^{j}_{i\epsilon l,s}(0)}{\alpha^{j}_{i\epsilon l,s}(0)+\beta^{j}_{i\epsilon l,s}(0)} = \begin{cases} \min\left(\max\left(0, w_{j,s} + \tilde{e}_l\right), 1\right), \text{if } d_l > d_2 \\ \min\left(\max\left(0, w_{j,s} - \tilde{e}_l\right), 1\right), \text{if } d_l \leq d_2 \end{cases}, l = 1,2,3,4 \quad (5)$$

, where we have restricted its value in the [0, 1] interval because it estimates the success probability.

Notice that 'trust level' alone, is not enough to calculate the exact values of a and β parameters, as an infinite number of their combinations may result to the same outcome. In Section 2, we mentioned that for equal trust values, their sum reflects the trustor's confidence. We reasonably assume that the level of confidence proportionally depends on the value of e_l coefficient and the number of look and feel elements with respect to factor j, notated as k^j_s. The equivalent mathematical expression is:

$$\alpha^{j}_{i\epsilon l,s}(0) + \beta^{j}_{i\epsilon l,s}(0) = e_l * k^j_s \quad (6)$$

Using (5) and (6) one can compute a pair of Beta parameters for each segment that depend on Table 1 and thus will reflect the personality of the users in that segment. Then, the initial trust level for each segment's users can be computed using eq (2).

4.2 Trust Evolution with Observations following a Machine Learning Approach

Contrary to the standard process where each outcome is equally weighted, here we consider that trustors apply greater importance to a success or failure: thus, biasing their trust to over or under estimate the corresponding trustworthiness respectively. The reason for doing so is that trust levels are subjective; two users having observed the exact same sequence of system outcomes can have significantly different estimation about the trustworthiness of the system in question. The subjectivity of trust will be demonstrated in the next section (see Figure 2 and Figure 3), where the averages of the trust levels being reported in a small-scale experiment varied significantly. Aligned with Equation (4), for each factor j a trustor in segment l updates her personal parameters as follows:

$$\alpha^j_l(t+1) = \alpha^j_l(t) + A^j_l \text{ and } \quad \beta^j_l(t+1) = \beta^j_l(t) + B^j_l \quad (7)$$

where A^j_l and B^j_l stand for the increment coefficients of segment l, after each success and failure observed with respect to trustworthiness factor j.

The parameters' values determining the trust evolution may be adjusted so that the theoretical model results to any given value "m", after a specific number of outcomes. This is easily feasible by setting:

$$\frac{\alpha^j_l(0) + s(t)A^j_l}{\alpha^j_l(0) + s(t)A^j_l + \beta^j_l(0) + f(t)B^j_l} = m(t) \quad (8)$$

, where $s(t)$ and $f(t)$ stand for the number of successes and failures observed until time t respectively. Note that if we apply this rule for the initial trust and two addi-

tional different time moments ($t_1 \neq t_2$), then we get a unique pair of increment coefficients, assuming that they remain constant for all observations.

The value of factor m, may be derived by any assumption concerning the impact of personal attributes on trust and trustworthiness correlation or may stand for actual measurements based on trustor's real responses. In this paper, we follow the latter approach: the trusts levels, as reported by participants, will be averaged per segment and fed into the theoretical model to reset the parameters so that they closely reflect the former. The initial trust may be either explicitly provided or may be derived by the relevant formula in the previous section. For completeness, we note that in this approach (three points equation), the estimated trust is unique and does not depend on the number of look and feel elements.

5 Validation Results

5.1 Experiment Setup

The experiment took place in October 2014 inviting participants to test and evaluate an online security service. A fictitious provider was offering a service, called Distributed Attack Detection and Visualization (DADV), for detecting virtual attacks on devices connected to the Internet, such as personal computers. The approach followed for attracting attackers was to deploy special decoy hosts in the subscribers' network that imitate vulnerable machines. All participants were assumed to be part of the same organization requesting protection and thus a single set of honeypots was deployed.

Real-time information about those incidents was sent to the provider for further processing so that the attack is prevented from expanding to other machines in the network. The experiment was performed for two versions of the online service; the Vanilla DADV where administrators are responsible for detecting and mitigating attacks and the Automated one where all tasks are performed by sophisticated tools [10].

The first step was for the participants to fill in the online segmentation-related (intake) questionnaire (See 5.2 below). In order to validate the trust initialization approach, participants were asked to report their initial trust towards the system before having any other evidence for its performance. To do so, each participant engaged with the DADV system, separately for each version during two different days, starting with the Vanilla DADV and then with the Automated one. After logging in to the online website (and before any attack was performed), they were given the opportunity to access the "about page" and familiarize themselves with the activated version. This webpage provided general information of the system functionality and a high-level description of its expected trustworthiness. Furthermore, users who had noticed and clicked on a distinguishable hyperlink were redirected to a more detailed webpage, which explicitly mentioned each system's actual trustworthiness in terms of the metric under interest. In this way we could validate the effects of "seeking motivation" on the initial trust level of each segment.

Afterwards, they observed the service performance for a sequence of 10 attacks that were identical for both DADV cases. During each attack, they could navigate to

the “health statistics page”, which was providing a holistic view of the system status. More specifically the subjects could judge whether an attack was taking place by observing the current CPU/memory/network load and observe the number of attempts initiated by a compromised sensor to the rest network hosts. At the end of each attack a message was appearing indicating whether the provider succeeded in preventing any network host from being attacked, or not. These pop-up messages also contained a link to a questionnaire where users were asked to indicate their current trust level that the provider would prevent future attacks from compromised honeypots to their computers. In other words, the metric of interest was the number of successfully mitigated attacks of each DADV system over the total number of attacks. This step provided the actual trust values, which after taking the average per segment, were utilized for training the trust computational model (see Section 5.3)..

The attacks resulted to the following sequences of outcomes, as depicted in Figure 1. The Automated DADV version outperfomed the Vanilla one in preventing a connection from being initialized since adminstrators had higher reaction times than their counterparts. Remember that all users observed exactly the same sequence of outcomes. This is essential to guarantee that the trust level was consistently monitored and, hence, any differentiations were guided by different sets of trustor’s attributes only.

Figure 1: The sequence of outcomes evidenced for each DADV version.

5.2 Validating Trustors’ Segmentation

In order to assess whether the four segmentation solution described in Section 3 could be deployed, additional empirical research was carried out. For this purpose the intake survey was dispersed using several Living Lab panels in September 2014. While 108 started the survey, 89 people from 11 European countries fully completed the survey and these were used for further analysis. Some 55% were aged between 25 and 34, followed by 32% that were aged between 35 and 44, and a few younger and older. Also, some 65% reported to have a university degree. The same steps were followed as in Section 3. Thus, a K-means clustering to segment different trust-related user experiences and an Anova analysis was performed to test the statistical significance for each item, thereby highlighting statistical differences between uncovered trust-related user experience segments. The results are shown in Table 2 below, where the absolute differences from Table 1 appear inside the parentheses.

Despite the minor variations between the two exploratory analyses presented below, the dominant drivers that seem to characterize users in each segment appear to be relatively constant. Thus, the findings seem to correspond to the previous ones indicating that the three underpinning users’ attributes appear as statistically significant difference. More specifically, we observe that the combined aggregate factor of “competences” and “seeking motivation” is again higher for the HATS segment. This

finding justifies our approach to correlate higher values of this factor with a more accurate estimation (equation 5). Furthermore, it is confirmed that a high level of "trust stance" results to trustworthiness overestimation (misplaced trust) and vice versa (presence of overcautious users).

Table 2. Intake survey segmentation results (n=89 participants)

	Total (n=89)	HT (n=25)	HATS (n=20)	MATS (n=32)	A (n=12)	Anova	
	Mean	Mean	Mean	Mean	Mean	F	Sig.
Trust stance	2,65 (-057)	3,42 (-0.43)	2,45 (-0.7)	2,33 (-0.53)	2,25 (-1.25)	27,053 (19.8)	**,000** (0)
Trust related seeking behaviour	2,38 (-1.14)	2,14 (-1)	3,02 (-1.25)	2,16 (-1.18)	2,44 (-0.57)	28,361 (3.98)	**,000** (0)
Trust related competences	3,65 (1.21)	3,88 (1.17)	4,29 (1.87)	3,63 (0.69)	2,17 (0.73)	53,592 (40.2)	**,000** (0)

5.3 Validating the Trust Computational Model

In order to validate the trust computational model described in Section 4 we employ two additional variations and compare the evolution of the computed trust levels with the actual ones, as reported by the participants. Before proceeding, we mention that while N = 89 were asked to fill in an online segmentation-related questionnaire, a subset N = 27 decided to also take part in the experiment. Table 3 below shows the output of the segmentation process and the mean values of the three trust-related concepts that were used for setting the initial values of the Beta parameters $\alpha_l^j(0), \beta_l^j(0)$ for each segment l, as described in Section 4.1.

Table 3. Intake survey segmentation results for experiment participants (n=27)

	Total (n=27)	HT (n=5)	HATS (n=4)	MATS (n=10)	A (n=8)	Anova	
	Mean	Mean	Mean	Mean	Mean	F	Sig.
Trust stance	2,65	3,40	2,63	2,30	2,63	4,519	**,012**
Trust related seeking behaviour	2,16	2,06	2,82	1,89	2,25	6,879	**,002**
Trust related competences	3,53	3,80	4,13	3.58	3.58	3,067	**,048**

In Figure 2 and Figure 3 we juxtapose the actual trust values with those derived by the three variations of the trust computational model (T1, T2 and T3), for the Vanilla

DADV experiment. Similar results are obtained for the Automated DADV, but omitted for brevity).

The approaches used for the initialization and update phase for each of the three variations T_o (where $o = 1,2,3$ denotes the number of actual trust values used as input to the model) are described below:

the T1 model computes the initialization parameters $\alpha_l^j(0), \beta_l^j(0)$ for each segment using the average of the actual trust values, as reported by their members before using the system. Note that in this case, the number of "look and feel" elements affects the initial values of the Beta parameters and consequently the graph oscillations. After observing the actual trust values and especially the significant trust degradation following each negative outcome we have set their number to one ($k = 1$). Furthermore, T1 relies on the standard unitary update coefficient for all segments ($a = 1, \beta = 1$) and thus follows the basic Bayesian model for the update (see Section 2).

the T2 model uses equations (5) and (6) for deriving the initial trust value and thus follows the approach described in Section 4.1. For the update process, the respective coefficients A_l^j and B_l^j are computed based on two measurements only using equation (8). More specifically, we used the actual trust values after the 2nd and 8th outcome. These four pairs of values, one for each segment, are denoted as (2, 8).

the T3 model requires three input values from the actual responses and can be seen as a hybrid of T1 and T2. More precisely, T3 follows the same initialization process as with T1, while the update process is similar to T2.

We observe that the models are aligned with the expected user reactions for most segments; namely trust should not decrease after a success and should not increase after a failure. The only exception is T2 for the "MATS" segment, which appears to constantly increase with the number of trials. This can be attributed to the error in estimating that particular initial value; in such cases the system of equations (7)-(8) may result in negative values for one or both update coefficients. Notice that T2 succeeds in computing a very accurate initial trust value for the High Trust and Highly Active Trust Seeking segments, while the relative error for the Ambivalent and Medium Active Trust Seeking segments is 10% and 20%, respectively. Before proceeding further, recall that T1 and T3 are initialized explicitly from the initial values, thus the effect described above is avoided over these two methods.

Additionally, observe that T2 and T3 manage to closely estimate the average trust of the HATS and Ambivalent segments, while for the rest segments the deviations tend to vanish as the number of observed system outcomes increases. Concerning the T1 graph, it is easy to see that this naïve approach fails to capture the segment differentiations in the trust evolution and thus its estimation is outperformed by both T2 and T3. Intuitively the common update coefficients of T1, result in all segments converging to the same value (which equals the actual trustworthiness) despite the personalized initial trust values. Thus, any potential different reactions among the segments are not captured on the trust evolution computation and the impact of the different initial values fades out as the number of observations increases

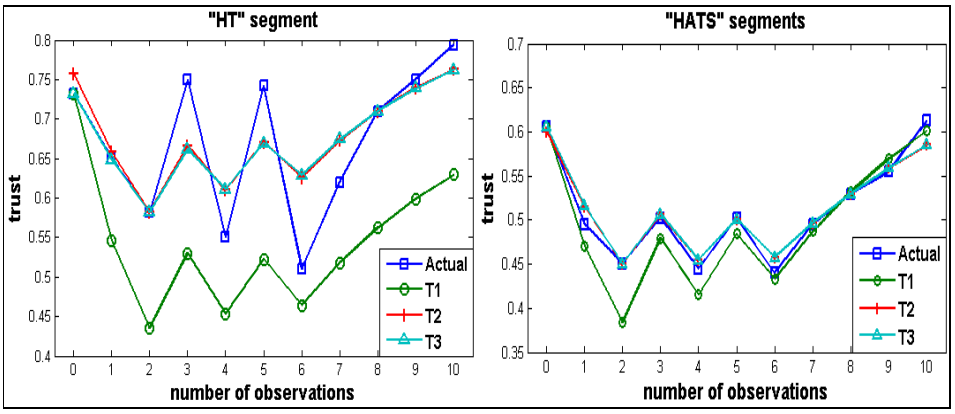

Figure 2: Actual and estimated trust values for the "HT" (left) and "HATS" (right) segments.

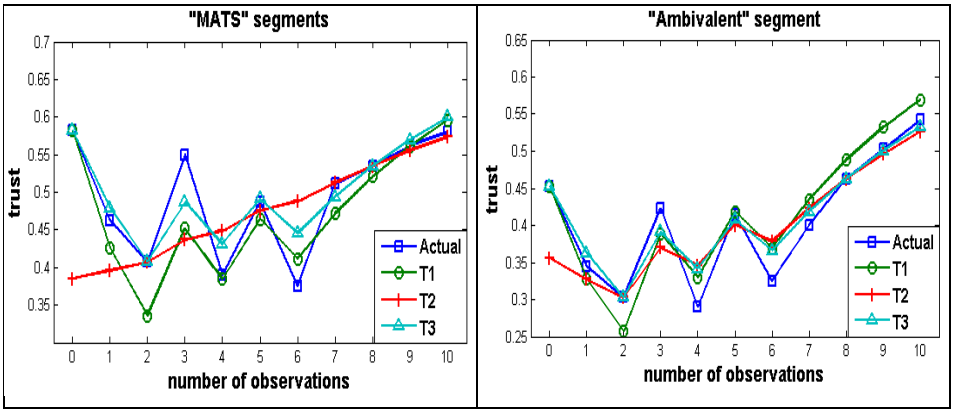

Figure 3: Actual and estimated trust values for the "MATS" (left) and "A" (right) segments.

We now compare the accuracy of the three versions of the computational model for different input pairs. More specifically, we fix the first part of the input data (always after the second trial) and vary the second one. We consider the evaluation metric "AAD" standing for the average absolute difference of estimated and actual values. In order for the comparison to be fair the "AAD" is computed over the non-provided points in each case, meaning that it is the average of 10, 9 and 8 points for T1, T2 and T3 respectively. In **Error! Reference source not found.**, we report the measurements for the Vanilla version only, for both T3 and T2 (when meaningful). For T1, this metric has a single value, as the update coefficients are static and thus the input pair is not considered.

First note that the average absolute difference decreases, as we delay the second input value for all segments over both T2 and T3. This is because the second trust value provides collective knowledge about the user's reaction at the intermediate trials, even though the actual trust at these moments is not explicitly given in the model. Although "AAD" is not always decreasing (meaning that we don't always achieve a more accurate trust estimation with more experimental trials), it seems to converge at acceptable levels for input pairs where users have observed adequate

evidence from the system performance and consequently their trust appears with small variations (last four trials). We expect that in a larger-scale experiment with increased number of trials, "AAD" will reach even lower values, as more trust measurements will be available.

Table 4: Comparing the accuracy of the 3 versions of the Trust Computational Model for the Vanilla DADV version using the Average Absolute Difference (AAD) of estimated and actual values.

SEGMENT & MODEL	*INPUT PAIR*						Average
	(2,5)	(2,6)	(2,7)	(2,8)	(2,9)	(2,10)	
"HT"- T3	.0866	.1334	.0706	.0549	.0549	.0604	.0768
"HT"- T2	.0788	.1168	.0643	.0509	.0510	.0559	.0696
"HT"- T1	.1399						
"HATS"–T3	.0118	.0224	.0106	.0107	.011	.0221	.0148
"HATS"-T2	.0109	.0210	.0099	.0101	.0102	.0202	.0137
"HATS"-T1	.0206						
"MATS"–T3	.0286	.0790	.0386	.0299	.0287	.0311	.0393
"MATS"–T1	.0340						
"A" –T3	.0246	.0500	.0285	.0222	.0222	.0236	.0285
"A"-T2	.0419	.0721	.0448	.0372	.0370	.0387	.0453
"A"-T1	.0302						

Concerning the comparison between T2 and T3, we can observe that the former is more accurate for "HT" and "HATS", despite the fact that it requires fewer user responses. This seems to be due to the close estimation of initial trust in these two segments, and which is also justified by the "A" segment where this property does not hold and consequently T3 outperforms T2. Thus, T2 and T3 have similar performance, meaning that our methodology for the trust initialization not only achieves to capture the segment properties but may also be utilized to estimate the actual trust values, when limited input is available, or more desirable.

When looking at T1, on average, it outperforms T3 for the "MATS" segment. Despite this fact, for all segments there is at least one input pair for which T3 provides better estimations, with this observation being particularly apparent during the latest pairs. Similarly, comparing T1 with T2 we observe that the latter outperforms the former for the "HT" and "HATS" segments. From the average values of T2 and T3 over all input pairs, we notice an improvement reaching up to 50% for "HT" when using T2 compared to T1. The reason that the highest improvement appears for this segment, is that its trust is clearly higher than the actual trustworthiness and our models capture this deviation. This fact is less intense for the other segments, thus the improvement is less impressive, but still remarkable: Notice that even though trust of "HATS", is the most accurate estimation of trustworthiness among all segments, our approaches provide interesting results in this case also. This is because the evolution

of estimated trust levels closely matches the actual ones, another important property apart from the accuracy in the long-run.

Thus, we may conclude that our approach to cluster users into segments and update their trust level according to the segment they belong to seem to provide valuable results towards a more accurate trust estimation.

6 Related Work

Significant research effort can be evidenced to understand the factors that affect a trustor's trust and build trust computational models that can be configured to make autonomous decisions that mimic a personalized mental process. The rationale behind this is that trust formation has been found to be a rather subjective and dynamic process. Such computational models are usually initialized using reputation systems that aggregate experiences of other trustors. Later, as users interact with the system/service and get direct observations, their trust levels are updated. Below, we provide an overview of trust computational models that explicitly consider trustor attributes and how these differ from our model. For a comprehensive overview of such models the interested reader is redirected to, for example, [11].

In [12] the following personal trust factors are considered when initializing a trust level: a) the effects of stereotypes such as appearance, the context and existence of certificates proving expertise b) trustor characteristics like general propensity to trust, user expertise and user need, as well as, c), similarity between persons (and empathy when the trustee is a system). Even though the authors did not quantify the effect of these personal factors, their importance has been validated via experiments. Furthermore, the resulting trust level is a single value (not a probability density function, or PDF) and thus the confidence cannot be determined.

In [13] a computational model is provided that allows to reason about the produced trust level by analyzing and formalizing the dynamics of trust in the light of experiences. Furthermore, they hypothesize that trustors can be grouped into sets based on their attributes, which, however, were not produced following a statistical approach nor were associated with trustor attributes (such as trustor expertise).

In [14] a trust computational model is proposed that takes into account the following personal attributes for trust update only: a) trust flexibility that expresses how much each system outcome counts, b) trust decay that defines how fast the trust level goes back to a neutral state in absence of new experiences, and c) autonomy that indicates whether the trust level to one trustee is affected by the trust level to other trustees. Even though the importance of these attributes has not been validated (using surveys etc.) in the sequel paper [15] the authors suggested and compared four techniques that could be used for estimating the values of these parameters from subjects' responses.

A different approach for estimating a user's trust level is based on user (as opposed to system) behavior. In [16] they performed an experiment to identify that different eye gaze and heart rate patterns could indicate different trust levels.

Our trust computational model builds upon a set of trust concepts that were found to be statistically significant; a) general propensity to trust, b) user expertise, and c) motivation to search for stereotypes that prove provider trustworthiness. Thus, although these concepts focus on trustor characteristics only, there is significant overlap with the findings in [12] and [13]. In addition, we utilize those trustor attributes to suggest how the trust level of each segment should be initialized, as well as, updated after successful or unsuccessful system outcomes. Thus, we argue that we follow a more holistic approach compared to papers [12], [13], [14] and [15].

7 Conclusions and Future Work

In this paper, we have drawn out the conceptual background for our proposed a trust computational model that allows a provider to estimate the trust level of candidate trustors, using a holistic approach. We also demonstrated the validity of our results via a small-scale experiment in the online security service context. More specifically, we have identified four segments with statistically significant differences which affect both the initial level but also the evolution of trust towards a system. These differences are captured by means of a modified Bayesian inference model, where the system outcomes have a weighted impact on the trust of each segment. We observed that our approach, i.e., to feed in the model with actual data so as to identify the individual weights, results to remarkably improved trust estimation compared to the standard process where the personal attributes are not considered in the trust update.

In the future we plan to revisit the initialization steps for the Medium Active Trust Seeking and Ambivalent segments and perform another experiment, possibly in another domain, where participants would engage with the system for more transactions. In this way, it allows to accurately estimate all users' trust level with a small subset of actual trust values provided by the trustors themselves. Furthermore, we will validate that the number of transactions necessary for the trust level to converge is limited (~10-15) and, thus, the trust computational model can afterwards be used for helping the provider to meet customers' expectations at run-time.

8 References

1. Habib, Sheikh Mahbub, et al. (2012). Trust as a facilitator in cloud computing: a survey. *Journal of Cloud Computing.* Vol. 1 (1): 1-18.
2. Riegelsberger, J., Sasse, M. A. and McCarthy, J. D. The Mechanics of Trust: A Framework for Research and Design. *International Journal of Human-Computer Studies*. Vol. 62(3), 2005, pp. 381-422.
3. McKnight, Harrison & Chervany, Norman (2001). "Trust and distrust definitions: one bit at a time" in: Falcone, Rino, Singh, Munindar & Tan, Yao-Hua (eds.). *Trust in cyber-societies.* Springer, Berlin-Heidelberg, pp. 27-54.
4. Lyon, F., Möllering, G., Saunders, M.N.K. (Eds.) 2012. *Handbook of Research Methods on Trust*. Cheltenham: Edward Elgar.

5. Li, F., Kowski, D.P, van Moorsel, A. and Smith, C. (2012). Holistic Framework for Trust in Online Transactions. *International Journal of Management Reviews*, Vol. 14, 85–103.
6. G. Möllering (2006). Trust: Reason, Routine, Reflexivity, Oxford, UK: Elsevier Ltd.
7. P. Sztompka, Trust: A Sociological Theory, Cambridge University Press, UK, 1999.
8. Gambetta, D. (1990). Can We Trust Trust? In Trust: Making and Breaking Cooperative Relations, pages 213-238. Basil Blackwell. Oxford.
9. M. Surridge, et. al. "OPTET D2.1 – Socio-economic requirements for trust and trustworthiness.," technical report, OPTET consortium, 2013.
10. N. G. Mohammadi et. al. "Maintaining Trustworthiness of Socio-Technical Systems at Run-Time" 11th International Conference on Trust, Privacy & Security in Digital Business. TrustBus 2014
11. Pinyol, Isaac, and Jordi Sabater-Mir. "Computational trust and reputation models for open multi-agent systems: a review." *Artificial Intelligence Review* 40.1 (2013): 1-25.
12. J. Masthoff, 'Computationally Modelling Trust: An Exploration', in Proceedings of the SociUM workshop associated with the User Modeling conference, Corfu, Greece, (2007).
13. Jonker, Catholijn M., and Jan Treur. "Formal analysis of models for the dynamics of trust based on experiences." In *Multi-Agent System Engineering*, pp. 221-231,1999.
14. Hoogendoorn, Mark, S. Waqar Jaffry, and Jan Treur. "Modeling dynamics of relative trust of competitive information agents." In Cooperative Information Agents XII, 2008.
15. Hoogendoorn, Mark, S. Waqar Jaffry, and Jan Treur. "An adaptive agent model estimating human trust in information sources." In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009.
16. Leichtenstern, K., Bee, N., Andre, E., Berkmuller, U., and Wagner, J. (2011). Physiological Measurement of Trust-Related Behavior in Trust-Neutral and Trust-Critical Situations. In: I. Wakeman et al. (Eds.). IFIPTM 2011, IFIP AICT 358, pp. 165–172.