TRACE: A Dynamic Model of Trust for People-Driven Service Engagements

Combining Trust with Risk, Commitments, and Emotions

Anup K. Kalia, Pradeep K. Murukannaiah, and Munindar P. Singh

North Carolina State University, Raleigh, NC 27695-8206, USA
{akkalia,pmuruka,singh}@ncsu.edu

Abstract. Trust is an important element of achieving secure collaboration that deals with human judgment and decision making. We consider trust as it arises in and influences people-driven service engagements. Existing approaches for estimating trust between people suffer from two important limitations. One, they consider only *commitment* as the primary means of estimating trust and omit additional significant factors, especially *risk* and *emotions*. Two, they typically estimate trust based either on fixed parameter models that require manual setting of parameters or based on Hidden Markov Models (HMM), which assume conditional independence and are thus ill-suited to capturing complex relationships between trust, risk, commitments, and emotions.
We propose TRACE, a model based on Conditional Random Fields (CRF) that predicts trust from risk, commitments, and emotions. TRACE does not require manual parameter tuning and relaxes conditional independence assumptions among input variables. We evaluate TRACE on a dataset collected by the Intelligence Advanced Research Projects Activity (IARPA) in a human-subject study. We find that TRACE outperforms existing trust-estimation approaches and that incorporating risk, commitments, and emotions yields lower trust prediction error than incorporating commitments alone.

1 Introduction

People-driven service engagements involve how people interact to carry out collaborative business processes [6, 7]. Such business processes involve human judgment and decision-making [13]. Trust is a crucial element of achieving secure collaboration where people interact since it enhances the quality of a collaboration. We consider direct interaction between people, which can be used to inform the design of user agents to facilitate collaboration among people. As people interact, they estimate and continually revise trust in each other based on their mutual interactions. Trust is established as a crucial element of service selection, e.g., [11]. However, as service settings become more complex and intertwined with social interactions, we need to expand our understanding of trust in services to promote the human element.

Existing approaches to trust estimation consider commitments alone. Gambetta [4] interprets trust as a truster's assessment of a trustee for performing a specific task. Mayer et al. [12] define trust as the willingness of a truster to be vulnerable to a trustee for the completion of a task. Teacy et al. [18] consider trust as the truster's estimation

of probability that a trustee will fulfill its obligation toward the truster. Wang et al. [20] represent trust as the belief of a truster that the trustee will cooperate, and estimate trust by aggregating positive and negative experiences. Singh [16] provides a formal semantics for trust that supports various postulates on trust, including how trust relates to commitments. Kalia et al. [8] consider commitments to predict trust.

Considering trust as a dynamic variable, two major classes of trust models arise in the literature. First, fixed-parameter trust models, where the parameters of the model are manually fixed, typically, based on heuristics [18, 20]. Second, machine-learned trust models, typically Hidden Markov Models (HMM) [10, 19, 21], assume that input variables are conditionally independent of each other given the output variable.

Research Question. Our overarching question is: How can we improve trust prediction by incorporating (in addition to commitments) two attributes (1) *risk* taken by a truster toward a trustee, and (2) *emotions* displayed by a truster toward a trustee without presuming conditional independence? We consider risk because it depends on a truster's belief about the likelihood of gains or losses it might incur from its relationship with a trustee [12]. For example, a manager may trust a subordinate who performs a high-risk task more than another who performs a low-risk task, even if both subordinates succeed at the task. Conversely, the manager may assign high-risk tasks to a subordinate whom he or she trusts more than the other. We consider emotions because studies in psychology suggest that positive emotions (e.g., happiness, gratitude) increase trust, whereas negative emotions (e.g., anger) decrease trust [2]. Conditional independence may not hold in our setting. For example, consider the relationships between trust (output variable) and risk and commitments (two input variables). An HMM model would assume that risk and commitments are independent given the level of trust. However, the likelihood of gaining from risk is higher if commitments are satisfied.

Contributions. We propose TRACE, a model of trust based on Conditional Random Fields (CRF) [9]. TRACE avoids manual fixing of parameters and relaxes the conditional independence assumption. To create TRACE, first, we propose relationships between trust, risk, commitments, and emotions. Then, we train TRACE using the past observations between people. Once TRACE is trained, we use it to infer trust given new observations. Our claims are two fold: (1) by capturing complex relationships among output and input variables, TRACE estimates trust between people better than fixed–parameter and HMM-based trust models, and (2) by capturing risk, commitments, and emotions, TRACE performs better than models that capture only commitments. We evaluate our claims via data collected from a human-subject study conducted by the Intelligence Advanced Research Projects Activity (IARPA).

2 A Conceptual Model of Trust

TRACE enhances Mayer et al.'s [12] trust antecedent framework (TAF) as shown in Figure 1. The trust model contains four variables: trust (T), risk (R), commitments (C), and emotions (E). We describe each variable $\mathsf{V} = \langle \mathsf{T}, \mathsf{R}, \mathsf{C}, \mathsf{E} \rangle$ using Singh's [15, 16] formal notation $\mathsf{V}\langle \textit{debtor}, \textit{creditor}, \textit{antecedent}, \textit{consequent} \rangle$. In the notation, the

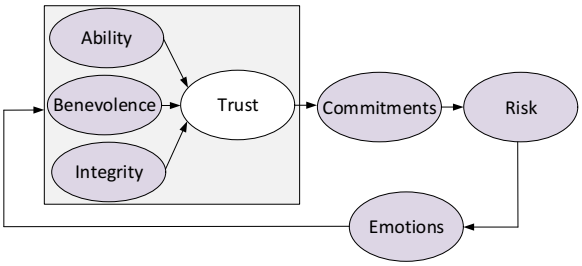


Fig. 1: The TRACE model enhances the trust antecedent framework [12] with emotion.

debtor and the creditor are the roles enacted by individuals. The antecedent represents *conditions* and the consequent represents *tasks*.

Commitments. We represent a commitment as $\mathsf{C}\langle$*trustee*, *truster*, *antecedent*, *consequent*$\rangle$. In a commitment, the trustee commits to the truster to perform the consequent. If the trustee performs the consequent, the commitment is *satisfied*. If the antecedent is true but the trustee does not perform the consequent, the commitment is *violated*.

Risk. We represent risk as $\mathsf{R}\langle$*truster*, *trustee*, *antecedent*, *consequent*$\rangle$, denoting that the truster takes a risk by accepting the trustee's offer to perform the consequent. If the trustee performs the consequent, the truster *gains*; else, the truster suffers a *loss*.

Trust. We represent trust as $\mathsf{T}\langle$*truster*, *trustee*, *antecedent*, *consequent*$\rangle$, denoting that the truster *believes* the trustee if the trustee performs the consequent. If the trustee does not perform the consequent, the truster begins to *doubt* the trustee. Based on TAF, trust has three dimensions: (1) *ability*, the trustee's competency to perform the consequent, (2) *benevolence*, the trustee's willingness to perform the consequent, and (3) *integrity*, the trustee's ethics and morality in performing the consequent.

Emotions. An emotion is a psychological response to an external or internal event [3, 17]. We introduce emotions as a response to commitment outcomes (satisfaction or violation) in the TAF (Figure 1). Similar to trust and risk, we denote emotions as $\mathsf{E}\langle$*truster*, *trustee*, *antecedent*, *consequent*$\rangle$. The truster displays a *positive emotion* if the trustee performs the consequent, else a *negative emotion*.

Postulates. Next, we propose postulates that capture relationships between the variables above. In these postulates, V_t represents the state of the variable V at time t.

$\mathbf{P}_1$: $\mathsf{T}_t \rightarrow \mathsf{T}_{t+1}$. The trust T_{t+1} is influenced by the past trust T_t. This postulate is consistent with the HMM trust models [10, 19, 21] since they assume that a truster computes its current trust T_{t+1} for a trustee based on its past trust T_t with the trustee.

$\mathbf{P}_2$: $\mathsf{C}_t \rightarrow \mathsf{T}_t$. The current commitment outcome C_t influences the current trust T_t. We consider this postulate since Kalia et al. [8] suggest that a truster trusts a trustee if the trustee satisfies the trustee's commitments toward the truster.

$\mathbf{P}_3$: $\mathsf{R}_t \rightarrow \mathsf{C}_t$. The risk taken influences the commitment outcome C_t or the gain or loss realized in the risk R_t . The postulate is supported by TAF [12].

$\mathbf{P}_4$: $\mathsf{R}_t \rightarrow \mathsf{T}_t$. The current risk taken R_t influences the current trust T_t . The postulate is supported by TAF [12].

$\mathbf{P}_5$: $\mathsf{C}_t \rightarrow \mathsf{E}_t$. The commitment outcome C_t influences the current emotion E_t. Smith and Ellsworth [17] suggest that a truster's emotions depend on the truster's appraisal of a trustee's commitments toward the truster.

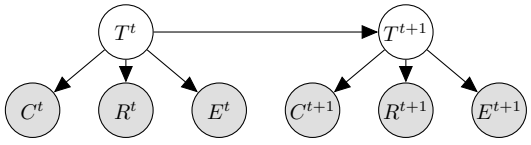
$\mathbf{P}_6$: $\mathsf{R}_t \rightarrow \mathsf{E}_t$. The risk taken R_t influences the truster's emotion E_t. We consider this postulate to capture the indirect effect that the risk taken influences the commitment outcomes (P_3), which influence emotions (P_5).

$\mathbf{P}_7$: $\mathsf{E}_t \rightarrow \mathsf{T}_t$. The current emotion E_t influences the current trust T_t. Psychological studies suggest that a truster makes trust-based judgments toward a trustee based on his or her emotional relationships with the trustee [2].

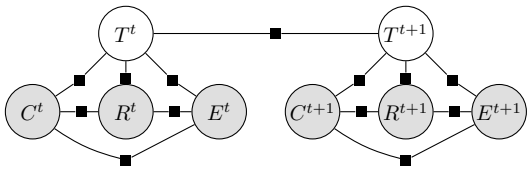
3 The TRACE Model

To compute trust, we propose the TRACE model using dynamic Bayesian models. In these models, we consider T, R, C, and E as random variables. Using the variables and the relationships proposed above, we construct two dynamic Bayesian models as show in Figure 2a and Figure 2b, respectively. Figure 2a represents the HMM model (the state-of-the-art-model) whereas Figure 2b represents the TRACE model.

HMM-based solutions and their limitations. Dynamic Bayesian models such as HMMs can be adapted to compute trust as shown in Figure 2a. Here, input variables are considered as a sequence of observations $\boldsymbol{x}$=$\{\mathsf{C}, \mathsf{R}, \mathsf{E}\}_{t=1}^{T}$, and output variables are considered as a sequence of states $\boldsymbol{y}$=$\{\mathsf{T}\}_{t=1}^{T}$ where T is the length of a specific sequence. Then, a HMM represents the joint distribution $p(\boldsymbol{y},\boldsymbol{x})$, making two independence assumptions: (1) the current state y_t is independent of y_1, y_2, ..., y_{t-2}, given y_{t-1}; (2) observations x_t are independent of each other, given y_t. Given these independence assumptions, the joint distribution can be computed as $p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1}) \times p(x_t|y_t)$. However, a downside of making these assumptions is that the corresponding models ignore some of the trust dependencies postulated in Section 2. For example, the HMM shown in Figure 2a assumes C^t to be independent of E^t given trust T^t, which may not be true according postulate P_5.



(a) HMM



(b) CRF

Fig. 2: Graphical representation of HMM and TRACE trust models (two time slices).

TRACE. TRACE employs CRFs to overcome the limitations of HMM-based trust models. As shown in Figure 2b, our CRF-based model considers all the dependencies postulated in Section 2. As Lafferty et al. [9] describe, unlike HMMs, CRFs are agnostic to dependencies between the observations. Further, the conditional probability of the label sequence can depend on arbitrary, nonindependent features of the observation sequence without forcing the model to account for the distribution of those dependencies. CRFs capture relationships between input and output variables ($\boldsymbol{x}$, $\boldsymbol{y}$) as *feature functions* (undirected edges in the graphical model shown in Figure 2b). A feature function can be computed by considering the entire input sequence.

An HMM model simplifies the computation of the joint probability by assuming conditional independence. In contrast, a CRF model employs discriminative modeling, where the distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is learned directly from the data (not requiring to learn the parameters of the entire joint distribution). The most important aspect of CRFs is to relate $p(\boldsymbol{y}|\boldsymbol{x})$ and feature functions $f_k(y_t, y_{t-1}, x_t)$. Each feature function covers either a state-state pair (y_t, y_{t-1}), e.g., (T_{t+1}, T_t) or a state-observation pair (x_t, y_t), e.g., (C_t, T_t), (E_t, T_t), and (R_t, T_t). Suppose we have K feature functions that represent state-state and state-observation pairs from $\boldsymbol{x}$ and $\boldsymbol{y}$. Then, $p(\boldsymbol{y}|\boldsymbol{x})$ can be computed starting from the joint distribution $p(\boldsymbol{y},\boldsymbol{x})$ as follows.

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{y},\boldsymbol{x})}{\sum\limits_{\boldsymbol{y}} p(\boldsymbol{y},\boldsymbol{x})} = \frac{\exp\left\{\sum\limits_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right\}}{\sum\limits_{\boldsymbol{y}} \exp\left\{\sum\limits_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right\}} \tag{1}$$

Training. To estimate the parameters λ_k in Equation 1, we consider the training data $\mathcal{D}$=$\{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=1}^N$. The parameters can be estimated by maximizing the log-likelihood $\mathcal{L}$ on the training data $\mathcal{D}$, i.e., $\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{N} \log p(\boldsymbol{y}|\boldsymbol{x})$.

Inference. To find the best possible state sequence $\boldsymbol{y}$ for observations $\boldsymbol{x}$, we use the Viterbi algorithm [14]. According to the algorithm, we define a quantity $\delta_t(i)$ that indicates the highest score (highest probability) of a path at time t as $\delta_t(i) = \max_{y_1,y_2,\ldots,y_{t-1}} p(y_1,y_2,\ldots,y_t{=}i, x_1, x_2, \ldots, x_t|\lambda)$ where i represents the state at time t.

4 Evaluation

We evaluate TRACE on data collected from subjects executing the Checkmate protocol [5] adapted from the iterated investment or dictator economic decision-making game [1]. The subjects assessed each other's trustworthiness as they played the game.

The checkmate protocol. The protocol involves two roles: *banker* and *game player*. The banker's task is to loan money to a game player from an initial endowment of 50 USD. The game player's task, in a single round of the protocol, is to complete a virtual maze of desired difficulty and collect as many cash boxes hidden in the maze as possible within the allotted time. The game player requests a loan from the banker to play a maze, promising to play a maze of certain difficulty and return (1) the loan with all gains, (2) the loan with 50% of all gains, (3) 50% of the available money, or (4) a fixed amount. After the game player's request, the banker chooses a loan category: small (1–7 USD), medium (4–10 USD), or big (7–13 USD). Then, a dollar amount, randomly

generated within the banker's chosen category, is loaned to the game player. The game player does not know the category chosen by the banker. Next, the game player plays a maze of a certain difficulty (not necessarily what he or she had promised). The banker will not know the actual maze played. The difficulty of the maze determines the risk involved: low risk (75–150%; i.e., the player could lose up to 25% or gain up to 150% of the loan amount in this maze), moderate risk (50–200%), or high risk (0–300%). Finally, the game player returns some money to the banker (not necessarily what he or she had promised).

A pair of subjects (one banker; one player) executed the protocol for up to five rounds. After each round, the subjects answered questions about their (individual) emotions and perceptions of the opponent's trustworthiness. All the money involved was real—that is, subjects kept the money they were left with at the end of all rounds.

Data. The data consists of 431 rows collected from 63 subjects, where each row corresponds to the sequence of rounds played between two subjects. The data we obtained reflects only the banker's perspective. Thus, our observations and predictions are from the banker's perspective. We compute the variables of our interest for each round in a sequence as follows. (1) We treat the commitment from the player to the banker, $\mathsf{C}\langle \mathit{player}, \mathit{banker}, \mathit{loan}, \mathit{return}\rangle$, as satisfied if the player returned at least the amount he or she had loaned, and as violated, otherwise. (2) We compute the gain or loss in the risk, $\mathsf{R}\langle \mathit{banker}, \mathit{player}, \mathit{loan}, \mathit{return}\rangle$, based on the difference between the loaned and returned amounts. (3) The dataset represents the banker's trust for the player after the round, $\mathsf{T}\langle \mathit{banker}, \mathit{player}, \mathit{loan}, \mathit{return}\rangle$, as a three-tuple $\langle \mathsf{A}, \mathsf{B}, \mathsf{I}\rangle$, indicating the banker's perception of player's ability, benevolence, and integrity, respectively, each a real value (0–1) derived from the post-round questionnaire. (4) The dataset represents the banker's emotion after he or she receives a return from the player, $\mathsf{E}\langle \mathit{banker}, \mathit{player}, \mathit{loan}, \mathit{return}\rangle$, as real-valued (1–10) state anxiety scores derived from the post-round questionnaire.

Mean absolute error (MAE). We treat trust estimation as a classification problem. Thus, we discretized each trust dimension (A, B, and I) into three categories (low, medium, and high) of almost equal frequency, making sure that no trust value is repeated across categories. The sizes of the resulting categories were A: [114, 155, 162], B: [109, 163, 159], and I: [118, 136, 177]. We measure the performance of a trust model via $\mathrm{MAE} = \frac{\sum_i^N |actual^i - predicted^i|}{N}$.

Comparison. For comparing HMM and TRACE, we perform a three-fold cross validation and compare the average MAE of the three folds. For each model, we considered the following feature combinations to predict trust: (1) C: only commitments, (2) C+R: commitments and risk, (3) C+E: commitments and emotions, (4) R+E: risk and emotions, and (5) C+R+E: commitments, risk and emotions. For each of these settings, we hypothesize that TRACE yields a lower MAE than HMM.

5 Results and Discussion

Table 1 compares HMM and TRACE, considering different feature combinations for predicting trust. When we consider only C, TRACE yields lower MAEs than HMM for each trust attribute. The primary reason for the result might be that CRF employs discriminative modeling whereas HMM employs generative modeling. Considering all

Table 1: MAEs of HMM and TRACE considering different feature combinations.

Input Variables	HMM			TRACE		
	A	B	I	A	B	I
C	1.1220	0.8564	1.0917	**0.8744**	**0.7576**	**0.7988**
C + R	0.8974	**0.7655**	0.8484	**0.7463**	0.7685	**0.7876**
C + E	0.8619	**0.7433**	**0.7184**	**0.8617**	0.7656	0.7580
R + E	**0.8468**	0.8376	0.7992	0.8949	**0.6815**	**0.6568**
C + R + E	0.8870	0.7977	0.7714	**0.7878**	**0.7427**	**0.7141**

features (C + R + E), TRACE again yields lower MAEs than HMM for each trust attribute (A, B, and I). We attribute this result to dependencies between C, R, and E, given T, which HMM ignores but TRACE incorporates.

Next, for C + R, TRACE performs better than HMM in predicting A and I (MAEs for B are quite similar). Thus, not assuming C and R as conditionally independent given T (as TRACE does) is beneficial to trust prediction than assuming so (as HMM does). However, for C + E, HMM performs better than TRACE for B and I (MAEs for A are quite similar). Thus, treating C and E as conditionally independent is beneficial to trust prediction than not treating so. This result suggests that changes in a truster's emotions are not limited to the appraisal of commitments, but can depend on other factors present in the truster's environment [17]. The result for R + E is mixed: TRACE performs better than HMM for B and I, whereas HMM performs better than TRACE for A.

In summary, these observations suggest that dependency relationships between commitments, risk, and emotions vary depending on whether the observed trust attribute is ability, benevolence, or integrity. Thus, our finding can be valuable in choosing the right set of dependencies given input and output variables of interest.

Threats to Validity. We identify three caveats about our evaluation. First, our dataset, although real, consists of short sequences. We expect both HMM and TRACE to perform better given longer sequences. Second, the dataset is skewed toward positive trust values and our conclusions may not hold since the trust values have a different distribution. Third, the dataset represents emotions using anxiety scores only, thereby lacking realistic emotion responses along multiple dimensions such as anger and joy.

Discussion. Despite these limitations, TRACE illustrates that a probabilistic model of trust that incorporates commitments, risk, and emotions can produce trust estimates with fairly good accuracy. Collaboration inevitably involves one party making itself vulnerable to another and inherently involves negotiation. The negotiation may be explicit, as in the dataset we studied, or implicit, such as when one party decides whether to take up any offer from another, including commonplace situations such as accessing a weblink or an email attachment. Our findings therefore open up the possibility of developing user agents that promote secure collaboration by helping a user calibrate the perceived trust with the risk undertaken in light of available measures of risk and gain from commitments. We defer investigating such agents to future research.

6 Acknowledgments

Thanks to the NCSU Laboratory of Analytic Sciences and to the US Department of Defense for support through the Science of Security Lablet. Thanks to Zhe Zhang, Chung-Wei Hang, and the anonymous reviewers for useful comments.

References

1. Berg, J., Dickhaut, J., McCabe, K.: Trust, reciprocity, and social history. Games and Economic Behavior 10(1), 122–142 (1995)
2. Dunn, J.R., Schweitzer, M.E.: Feeling and believing: The influence of emotion on trust. J. Personality and Social Psychology 88(5), 736–748 (May 2005)
3. Friedenberg, J., Silverman, G.: Cognitive Science. SAGE Publications, 2nd ed. (2012)
4. Gambetta, D.: Can we trust trust? In: Trust: Making and Breaking Cooperative Relations, pp. 213–237. Blackwell (1988)
5. IARPA: The checkmate protocol (2014), https://www.innocentive.com/ar/challenge/9933465
6. Kalia, A.K., Motahari-Nezhad, H.R., Bartolini, C., Singh, M.P.: Monitoring commitments in people-driven service engagements. In: Proc. 10th IEEE Intl. Conf. Serv. Comput. pp. 160–167. (2013)
7. Kalia, A.K., Singh, M.P.: Muon: Designing multiagent communication protocols from interaction scenarios. J. Auton. Agents and Multi-Agent Sys. 29(4), 621–657 (2015)
8. Kalia, A.K., Zhang, Z., Singh, M.P.: Estimating trust from agents' interactions via commitments. In: Proc. 21st European Conf. Artificial Intelligence. pp. 1043–1044. (2014)
9. Lafferty, J.D., McCallum, A.K., Pereira, F.C.N.: Conditional random fields. In: Proc. 18th Intl. Conf. Machine Learning. pp. 282–289. (2001)
10. Liu, X., Datta, A.: Modeling context aware dynamic trust using Hidden Markov Model. In: Proc. 26th National Conf. Artificial Intelligence. pp. 1938–1944. (2012)
11. Maximilien, E.M., Singh, M.P.: Toward autonomic web services trust and selection. In: Proc. 2nd Intl. Conf. Service-Oriented Computing. pp. 212–221. (2004)
12. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. The Academy of Management Review 20(3), 709–734 (1995)
13. Motahari Nezhad, H.R., Spence, S., Bartolini, C., Graupner, S., Bess, C., Hickey, M., Joshi, P., Mirizzi, R., Ozonat, K., Rahmouni, M.: Casebook: A cloud-based system of engagement for case management. IEEE Internet Computing 17(5), 30–38 (2013)
14. Rabiner, L.R.: Readings in speech recognition. chap. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296. Morgan Kaufmann (1990)
15. Singh, M.P.: An ontology for commitments in multiagent systems: Toward a unification of normative concepts. Artificial Intelligence and Law 7(1), 97–113 (1999)
16. Singh, M.P.: Trust as dependence: A logical approach. In: Proc. 10th Intl. Conf. Auton. Agents and MultiAgent Sys. pp. 863–870. Taipei (2011)
17. Smith, C.A., Ellsworth, P.C.: Patterns of cognitive appraisal in emotion. J. Personality and Social Psychology 48(4), 813–838 (1985)
18. Teacy, W.L., Patel, J., Jennings, N.R., Luck, M.: Travos: Trust and reputation in the context of inaccurate information sources. J. Auton. Agents and Multi-Agent Sys. 12(2), 183–198 (2006)
19. Vogiatzis, G., MacGillivray, I., Chli, M.: A probabilistic model for trust and reputation. In: Proc. 9th Intl. Conf. Auton. Agents and Multiagent Sys. pp. 225–232. (2010)
20. Wang, Y., Hang, C.W., Singh, M.P.: A probabilistic approach for maintaining trust based on evidence. J. Artificial Intelligence Research 40, 221–267 (2011)
21. Zheng, X., Wang, Y., Orgun, M.A.: Modeling the dynamic trust of online service providers using HMM. In: Proc. 20th IEEE Intl. Conf. Web Services. pp. 459–466. (2013)