On the Application of Paired Comparison to Trust

Rob Brennan[1], Karl Quinn[2], Declan O' Sullivan[2], Dave Lewis[2] and Vincent P. Wade[2]

[1] Ericsson R&D Ireland,
Research Department,
Clonskeagh, Dublin 4,
Ireland.
rob.xa.brennan@ericsson.com
[2] Knowledge & Data Engineering Group,
Trinity College Dublin,
Ireland.
{Karl.Quinn, Declan.OSullivan, Dave.Lewis, Vincent.Wade}@cs.tcd.ie

Abstract. This paper describes a flexible approach to the internal representation, tracking and inference of trust in ubiquitous computing environments in the presence of heterogeneous trust models. Our method is based on knowledge engineering techniques and the application of modern paired comparison statistical methods from operations research.

1 Introduction

Mechanisms for the estimation and propagation of trust through the Semantic Web, and previously electronic commerce or other secure relationships, have been for many years a subject of intense debate. Despite this dialog there is still little agreement on fundamental issues such as the properties of trust as a concept, the types of operations that may be usefully applied to a trust value, or even what kind of domain trust values could or should be drawn from. Nonetheless trust has been adopted by many as a useful concept for a wide variety of communications network applications. For example, the Semantic Web research community has embraced it enthusiastically.

In this paper we present some initial ideas on how current developments in operations research and knowledge engineering may be applied to the problem of representing and propagating trust in ubiquitous computing environments where multiple, inconsistent trust models are likely to hold sway for the foreseeable future.

Our role here is not to define a full formal model for trust; it is to define a tractable solution for those parts of the concept of trust between software entities that may find immediate applicability in our problem domain; the dynamic selection and composition of (web) services from atomic service components in ubiquitous computing environments.

2 Trust Models

Trust has many definitions that are applicable to different areas of computer security and distributed systems. These definitions tend to use synonyms or trust inspiring terms in their explaination. “Belief” [McKnight96], “Credibility or Reliability” [Golbeck03], “Confidence or Faith” [Shadbolt02], “Reputation” [GH04], “Competence and Honesty” [Grandison00] have all been used in this way. Definitions generally try to convey that trust has some quantitative value associated with it such that A trusts B, but only by so much. Trust is multidirectional in that B may not trust A at all. Trust can be made specific even more by stating that A trusts B in relation to specific procedures. Trust and its synonyms can be applied to many facets of the A and B relationship. It can be interpreted that A trusts the information that B (BBC News service) provides if A finds that the BBC’s information is both credible and reliable. From this A can assign some level of confidence to the BBC’s information and act in good faith upon it. The BBC can build up its reputation (expectation of behaviour based on past observations/information) by using competent reporters so that A has a directed and weighted belief in the BBC and its statements. The BBC can at the same time hold (to some degree) confidence in its audience.

Trust is a hard issue to contemplate because it is such a human idea that has so many uses and general meanings. It is for this reason that the scoping of trust (much like security) can enable a more specific meaning that can be more efficiently translated into mission goals and statements. For the purposes of our work trust can be seen as the aggregation of many of the synonyms used above in conjunction with the ideals that they convey. Web services must be reliable and inspire confidence in their users. Their information must be credible (or from credible sources) and honest so that it can be believed. Together all these elements can create a reputation of trust that can help in such areas as the orchestration of composite WS. In the real world we tend to make choices about who we buy services like health insurance from based on trust because it gives us peace of mind, so it seems plausible that we should have similar notions of trust in the computing world in order to help make choices with online services.

Generally trust models can be split into socially based and observationally based models of trust. Our current model attempts to bridge this gap through the use of statistical methods for observation and knowledge engineering techniques for trust and policy modelling. Of particular influence on this work were the information-theoretical approach to trust proposed by Gerck [Gerck97], the more rudimentary statistical models of Shi *et al.* [SBA04] (although this work has interesting additional support for decision making through the use of utility functions) and the bold formal model of Carbone *et al.* [CNS03]. The latter model has interesting parallels to our own work through the concepts of trust ordering and information ordering which parallel our development of internal trust scales and reliability metrics.

3 Paired Comparison

In paired comparison experiments, the worth or merit of a unit is measured through comparisons against other units. Paired comparison statistical methods are the basis of lengthy development first by the psychology and then by the operations research communities.

3.1 Thurstone's Method of Comparative Judgement

Psychometric scaling is the science of determining measuring techniques for human judgement. Thus it deals with the means of capturing and reifying subjective experience into quantitative data values. Given the current confusion regarding trust definitions and the likely heterogeneous nature of any ubiquitous computing environment one approach to survival in such an environment is the application of these techniques by collaborating, communicating software entities to evaluate the responses of their peers.

In 1927 Thurstone [Thur27] presented a major advance in this field; a mathematical model for relating scale values of a set of stimuli to observable properties. Essentially the problem here is to determine the locations on a psychological continuum (e.g. of trust) of various stimuli (e.g. queries about other software entities) based on a series of responses to those stimuli. A stimulus (query) may occur more than once and the subject may generate a different response each time. This may be due to timing, context or issues that are entirely internal to the stimulated entity. Thus no assumptions are made about the internal makeup of our peer, except for the fact that responses may differ and that if we present the stimulus to the observer a large number of times these responses will have a normal distribution.

An entity cannot report directly in terms of the values on its psychological continuum however it can give relations between stimuli so we can deduce equations relating them, the underlying scale values and their dispersions. The mechanism used for this reporting is paired comparison of stimuli. This only depends on the subject being able to report on noticeable differences between pairs of stimuli. Hence we can say that for two "merits" (trust values in our case) V_i and V_j of objects, A_i and A_j, that we represent the observed merit of object A_i by a random variable X_i owing to observation-to-observation variation. The model then takes the form:

$$P\left(X_i > X_j\right) = \pi_{ij} = H\left(V_i - V_j\right) \quad \textbf{(1)}$$

where H is a monotonic, increasing function. In Thurstone's model H is given by a probit model i.e. the normal cumulative distribution function with zero mean. The method produces estimates v_i of V_i, $i = 1, ...,m$ from observations of comparisons. If they are present, other parameters of H must also be estimated. Estimation proceeds by tallying α_{ij}, the number of times object A_i is preferred to object A_j after n_{ij} comparisons. A sample estimate of π_{ij} is $\alpha_{ij} / n_{ij.}$ We define $H(d_{ij}) = \alpha_{ij} / n_{ij}$ and compute merit or scale estimates v_i by $d_{ij} = v_i - v_j$, $i, j = 1, ...,m$. It was shown by Mosteller that a least-squares estimate of V_i, is:

$$v_i = \frac{1}{m}\sum_{j=1}^{m} d_{ij} \tag{2}$$

When each pair is observed a fixed and finite (but not necessarily equal) number of times then the sums n_{ij} are fixed and the tallies, α_{ij} , are binomial random variables with a well-defined likelihood function of the form:

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \prod_{i<j} P(\alpha_{ij}) \\ &= \prod_{i<j} \binom{n_{ij}}{\alpha_{ij}} H\left(V_i - V_j\right)^{\alpha_{ij}} \left[1 - H\left(V_i - V_j\right)\right]^{n_{ij}-\alpha_{ij}} \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha} = [\alpha_{ij}]$, the matrix of preference counts.

Ideally each stimulus is compared with each other a large number of times. Alternatively the subject can report a rank order list of the stimuli (in order of "merit" or strength); such a list can then be converted into paired comparison data. Thus the raw data we collect specifies the number of times that a given stimulus was judged greater than another stimulus. The findings of the web-of-trust or friend of a friend researchers would suggest that most entities will interact most frequently with closely connected entities thus supporting a rapid accumulation of data.

Note that either paired comparison or rank order data can only give ordinal scale judgment. The method of Thurstone's comparative judgment provides a way to turn the ordinal scale data into interval scale data. The meaning of such interval data could be interpreted as the relative distance between all the stimuli on the attribute space.

Hence, given incomplete relative information about stimuli and associated stochastic responses, we can map between a hidden relative continuum to discrete quantitative measures. When dealing with heterogeneous, internalised trust models and concepts this limits the type of information that needs to be transferred and minimises the assumptions required to leverage that information.

In practice Thurstone's formulation requires several simplifications to provide a tractable while still useful model. This is generally known as his Case V model, which assumes constant covariance between pairs of stimuli, constant correlation between response functions and response functions with a common standard deviation.

3.2 Further Development of the Method

In 1952 Bradley and Terry [Bra84] published an alternative version of the paired comparison model developed above. Instead of positing a normal distribution for responses (a probit model) they proposed that a logit function be used. Hence differences in merit are formulated as follows:

$$V_i - V_j = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) \tag{4}$$

This had the advantages of being much more easily manipulated, calculated and has a number of well-defined supporting test statistics. Given the likelihood of incomplete data for our software entities it is also significant that Bradley's analysis is explicitly based on calculations using only subsets of all the data. In fact a sufficient condition for a maximum likelihood is that each partition of the objects into two nonempty subsets such that some object in the second set has been preferred at least once to some object in the first set [For57]. Bradley also provides a method for calculating the confidence interval of estimated parameters.

The operations research community has developed this technique and both of these methods are now recognised as special cases of a general family of linear models of paired comparisons [CF91]. Recent developments in this area include the development of dynamic paired comparison models that explicitly allow for the merits of the objects to rapidly vary with time [Gli99], home-game advantage to be explicitly included [DB77], estimation methods for realtime updating and calculation of merits for very large systems [Gli99], allowances for scaled comparison and tied comparisons [Gli99]. The literature on paired comparison is now vast and spans fields such as statistics, marketing, psychology and decision sciences. The basic reference work in this field is now David [David88].

Of particular note to our work here is recent research on group decision support systems (GDSS) [LCWHT01] where comparative preference data from individuals in a group transformed into a single group composite interval scale. This can act as an aid to collaboration when direct physical scales (e.g. service performance data) are not available. It also helps to avoid the inconsistencies in group judgements generated by standard vote averaging approaches (i.e. Arrow's Paradox). Another major feature of this approach is that it supports dynamic voting whereby group members can change their responses and vote (respond) repeatedly until certain criteria are met.

The scales generated by such a method of course support inference on future likely outcomes of comparisons between objects for which we have data. For example, under the Bradley-Terry model for two objects with merit parameters V_i and V_j, object , A_i is preferred to A_j with probability $V_i / (V_i + V_j)$. Modern paired comparison models [Gli99] account for both the differences in estimated merit and the uncertainty in these estimations i.e. calculations are on the basis of both a scale and a reliability (standard deviation) and both of these values are tracked for all objects in the system. It is also possible to make a scaled comparison whereby the relative merit of compared objects is directly reported (when collecting data) or inferred (from previously collected data). These extensions will be of considerable utility when we try to apply this method to and environment of heterogeneous trust models and concepts.

4 Implications for Trust Propagation Protocols

Given the wide range of statistical methods for processing paired comparison data discussed in the previous section our goal is to outline how these techniques could impact on the implementation of entities in a ubiquitous computing environment that are collaborating at least partially on the basis of trust. Specifically it is interesting to

envision how these techniques can help to tame the inherent complexity and heterogeneity of trust mechanisms, concepts and metrics prevalent in the literature.

We see a number of important properties of entities that use paired comparison data to generate internal trust scales (metrics) and associated uncertainty (reliability) measures. These are discussed in the following sections.

4.1 Encapsulation of Heterogeneous Trust Models

Our first observation is that if an entity uses paired comparison data to generate internal trust scales (metrics) and associated reliability measures then, like an observer in a psychological experiment, it has the benefit of operating in a manner that is based on observation of and interaction with its environment rather than being dependant on prescient or pre-configured knowledge of the underlying models guiding other actors in that environment. This means that it can easily incorporate data from other entities with different internal representations of trust.

4.2 Minimised Information Requirements

The information to be collected from the other actors also makes very minimal assumptions about their capabilities for reasoning about trust. The only ability that they must possess is the ability to respond to simple queries regarding the relative trustworthiness of other actors, such as service providers, also present in the environment. Given advanced paired comparison models it can accept:

- Data that contains full trust-based rank orderings of sets of actors. This will be converted into a set of paired comparisons for all actors in the set.
- Paired comparison data about actor tuples. This may include tied value comparisons.
- Scaled comparison data about actor tuples.
- Trust metric values for sets of actors. This will be converted into a rank order data and hence into a set of paired comparisons for all actors in the set.

The above types of data give us great flexibility when incorporating other actors' responses.

It is also implicit in paired comparisons methods that incomplete pairings, within some limits, can be used to construct our set of trust scaling metrics. Even in the presence of such limited data we will have reliability estimates for the metrics available and so can use the data in combination with local policies or inference to aid decision-making.

4.3 Implicit Handling for Temporal Dependence of Trust

The most modern paired comparison models account for both slowly varying and rapidly varying merit (trust) values of the objects being monitored. For example Glickman [Gli99] adopts a state-space approach for modelling such data where the underlying merit parameters change as a Gaussian stochastic process and the variance of the

state process also changes stochastically. Thus there is handling for both scenarios where a popular (well-trusted) service suddenly becomes overloaded due to having a lack of resources to deal with the demand and where a new service gradually gains in popularity and dependability (and hence trust by its consumers). Both positive and negative changes in merit over time are of course possible

In addition when merit value reliability measures are constructed it is normal to include some sort of decay in the reliability measure over time. Hence if relatively current sample data is not available for a given object then a reduced reliability measure is calculated [Gli99].

4.4 Support for Long-lived Dialogs on Trust

Since each response to a given query represents the value of a random variable for a single trial within the stochastic model. The paired comparison technique can handle multiple (perhaps different) responses from the same actor. Each of these will be taken into account when estimating the final trust (merit) value and reliability value. Thus the entity implementing paired comparison methods may be considered to develop experience of its environment. This has led to the application of paired comparison in reinforcement learning research [SM01].

4.5 Isolation of Liars

An important problem in any unregulated ubiquitous computing environment that depends on collaboration is the potential presence of dishonest or faulty actors [GH04] and how to detect and deal with them. The stochastic nature of our model helps with this as any data received from such an actor is incorporated into our overall model rather than accepted as fact. Thus as long as there are a number of information sources for scaling a given trust metric the damage that such an actor can do is automatically limited.

In addition the Bradley-Terry family of models defines statistical tests to establish if there are differences between groups of responses [David88]. For large numbers of responses these techniques could allow us to isolate liars, especially if they are persistent ones.

4.6 Ability to Set Varying Levels of Initial Trust

In his treatise on models of trust [Gerck97], Gerck identifies as a requirement for all trust model that they should be able to set a level of basic or default trust for all interactions. He argues that at least sometimes in a trust-based system an entity is going to have to rely on such a default in order to operate. (Who do you trust first? How can you collaborate if you trust no-one? etc.). If we have an internal system of trust metrics based on paired comparison then a natural question is at what “level of trust do new entities enter the system (before they have any data collected on them). Since the outputs of scaling actions naturally fall on a distribution it is easy to define default values that correspond to a desired relative level of trust for the scaled population.

Many strategies could be adopted here, depending on local policy and perhaps the context of the interaction. For example a "trusting" entity might previously unknown actors as having an above median level of initial trust whereas a "distrusting" entity might default to giving all unknown actors a very low level of initial trust metric.

4.7 Allowance for "Semantic Distance" to Recommenders

When using recommendations from other actors as part of a trust decision it may be useful to consider that their internal trust priorities could be very different to local values. The most extreme case of this is of course a liar or dishonest source of recommendations. Abdul-Rahman [AH00] identifies this concept as the "semantic distance" between entities. If it had been established through experience or inference that another actor was giving recommendations that were inconsistent with an entity's world view then this could perhaps be incorporated into the paired comparison data generated from that source through the selective use of "home team advantage" model effects. This would take into account the source of the data as a bias towards one or other of the objects compared and hence allows better estimation of the underlying merits despite the bias (from this entity's point of view) in the reporting source.

5 Application Scenario

Our goal in developing useful trust scales is to facilitate the dynamic composition of services from atomic services in a ubiquitous computing environment. Atomic service selection in this scenario is based on trust, which is in turn based on: trust scales and reliabilities provided by other actors (as above), experience, knowledge of service properties and internal policies.

5.1 Ontology-based Internal Representation of Trust

So far we have only considered trust scales based on external recommendations of other actors, however our own experience and local policies are also important factors in making decisions based on trust. To this end we propose that entities will need to use some internal model of trust and how it applies to a given decision. We have already proposed a detailed ontology of the trust domain, which is described elsewhere [IM05]. This OWL based trust model is a rich semantic representation of trust concepts, relationships, and constituent service elements.

This trust structure enables developers to semantically annotate their web services with an enumeration of trust model classes (or properties) that are pertinent to a particular service. This helps an entity to quantify its experience of a given service or to infer a likely suitability of such a service of which it has no prior knowledge. It is also hoped that this model will be a useful aid in determining the semantic distance between an entity and other actors.

5.2 Dynamic Service Composition

The application use case scenario can be described as follows. Upon entering a network domain, an entity 'Y' discovers service 'X' and decides that it is of use to it. However, 'Y' doesn't have any information about 'X' on which it can make a choice on whether to trust it or not. It is assumed that other actors already present do have knowledge or experience about 'X'. Therefore, 'Y' can query them in relation to service 'X'. Once queried, the other actors can choose to share their data with 'Y'. This data may be service profile of 'X' that has semantically annotated with trust data or paired comparison data. Such data captures the experience a node had with service 'X'. From these recommendations, internal processes and policies a composition decision can be made.

5.3 Calculating Trust Scale for Composite Services

If we wish to make a decision about which composite service to use, then it may be useful to infer a trust scale and reliability metric for it. We will take the simplest composite service as an example i.e. one that is composed of two atomic services, S_1 and S_2. For each elemental service we have a merit estimate v_i and a reliability metric σ_i^2 where i = 1, 2. Now calculate the probability of a likely preference for S_1 as the outcome of a comparison between the two services, p_{12}. Since the weakest link will tend to dominate the combined service we use the inverse of this value as a weight for the contribution of S_1 and (1-1/ p_{12}) as the contribution of S_2. Hence the combined scale metric is given by:

$$v_{12} = \frac{v_1}{p_{12}} + \left(1 - \frac{1}{p_{12}}\right) v_2 \tag{5}$$

Similarly we can calculate the composite reliability metric, however in this case we must make allowance for the fact that these reliability metrics are in fact standard deviations and that in this case the most unreliable atomic service will dominate so we no longer need to invert p_{12}.. This is calculated as follows:

$$\sigma_{12}^2 = \sqrt{\left((\sigma_1^2)^2 \; p_{12}\right) + \left((\sigma_2^2)^2 \; (1 - p_{12})\right)} \tag{6}$$

Thus we have a composite trust scale and reliability metric for our composite service. Larger numbers of atomic services can be accommodated by recursive application of this method. Similar inferences can be made to decompose the results of paired comparisons of composite services into their constituent atomic services.

6 Conclusions and Further Work

A survey of the sophisticated methodology available for paired comparison in the operations research literature has been provided. The applicability of these methods to

the evaluation and propagation of trust in a ubiquitous computing environment has been illustrated. Further investigation of these techniques in terms of simulation and actual deployments needs to be performed. However it seems that some adaptation of these techniques to this problem space could produce significant benefits and insights for the collaborative computing community. Open issues include investigating the computational costs of maintaining trust scaling data, the volume of data transferred between entities to support these calculations and the necessary conditions for reliable scale generation.

Acknowledgements
We would like to thank Sandra Collins for helpful discussions on an earlier draft of this paper. This work was carried out in the Ericsson Ireland Research Department as part of an Internship by the principle authors, Rob Brennan and Karl Quinn, during 2004.

References

[CNS03] Carbone, M., Nielsen, M., Sassone, V.: A Formal Model for Trust in Dynamic Networks. First International Conference on Software Engineering and Formal Methods (SEFM'03), Brisbane, Australia. September 22 – 27 (2003)

[Thur27] Thurstone, L. L.: A law of comparative judgement. Psychological Review, 34, pp.273-286 (1927)

[Brad84] Bradley, R. A.: Paired comparisons: Some basic procedures and examples, Handbook of Statistics, Vol. 4, Krishnaiah, P. R. and Sen, P. K., eds., Elsevier Science Publishers, pp. 299-326 (1984)

[CF91] Critchlow, D. E. and Fligner, M. A.: Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. Psychometrika, 56, 517-533 (1991)

[Gli99] Glickman, M. E.: Parameter estimation in large dynamic paired comparison experiments. Applied Statistics, 48, 377-394 (1999)

[For57] Ford, Jr., L. R.: Solution of a ranking problem from binary comparisons, American Mathematical Monthly, 64, pp. 28-33, (1957).

[LCWHT01] Li, Z., Cheng, K., Wang, Y., Hiltz, S. R., and Turoff, M.: Thurstone's Law of Comparative Judgment for Group Support, in Proceedings of the 2001 Americas Conference on Information Systems, Boston, MA (2001)

[DB77] Davidson, R. R. and Beaver, R. J.: On Extending the Bradley-Terry Model to Incorporate Within-Pair Order Effects. Biometrics, 33, 693-702 (1977)

[David88] David, H. A.: The method of paired comparisons (2nd ed.), London: Chapman and Hall (1988)

[SM01] Strens, M. and Moore, A.: Direct Policy Search using Paired Statistical Tests. In Proceedings of the 18th International Conference on Machine Learning (2001)

[GH04] Golbeck, J. and Hendler, J.: Accuracy of metrics for inferring trust and reputation in semantic web-based social networks. In Proceedings of 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW'04), 5-8th October, Northamptonshire, UK (2004)

[Gerck97] Gerck, E.: Toward Real-World Models of Trust: Reliance on Received Information. Available online (2004) at http://www.safevote.com/papers/trustdef.htm (1997)

[AH00] Abdul-Rahman , A. and Hailes , S.: Supporting Trust in Virtual Communities. In Proceedings Hawaii International Conference on System Sciences 33, Maui, Hawaii, 4-7 January (2000)

[McKnight96] McKnight, D. H. and Chervany, N. L.: The Meanings of Trust. Technical Report 94-04, Carlson School of Management, University of Minnesota, 1996.

[Golbeck 03] Golbeck, J., Parsia, B. and Hendler, J.: Trust networks on the semantic web. In Proceedings of Cooperative Intelligent Agents 2003, Helsinki, Finland, August (2003).

[Shadbolt02] Shadbolt, N. R.: A Matter of Trust. IEEE Intelligent Systems -(-):pp. 2-3. (2002)

[Grandison00] Grandison, T. and Sloman, M.: A survey of Trust in Internet Application , IEEE Communications Surveys Tutorials (Fourth Quarter), vol. 3 (4) (2000)

[SBA04] Shi, J., Bochmann, G. and Adams, C.: A Trust Model with Statistical Foundation, Workshop on Formal Aspects in Security and Trust (FAST '04), Toulouse, France, Kluwer Academic Press, August 26-27, 2004 (to appear).

[IM05] Quinn, K., O' Sullivan, D., Lewis, D., Brennan, R., Wade, V. P.: deepTrust Management Application for Discovery, Selection, and Composition of Trustworthy Services. In Proceedings of Ninth IFIP/IEEE International Symposium on Integrated Network Management (IM 2005) 16-19 May (2005) To be published.