Why Do Users Trust The Wrong Messages? A Behavioural Model of Phishing

Paul A. Watters
Internet Commerce Security Laboratory (ICSL)
University of Ballarat VIC 3350, Australia
p.watters@ballarat.edu.au

Abstract – ***Given the rise of phishing over the past 5 years, a recurring question is why users continue to fall for these scams? Various technical countermeasures have been proposed to try and counter phishing, and none have yet comprehensively succeeded in preventing users from becoming victims. This paper argues that an explicit model of user psychology is required to understand user behaviour in (a) processing phishing e-mails, (b) clicking on links to phishing websites, and (c) interacting with these websites. Many users engage in e-mail and web activity with an inappropriately high level of trust: users are constantly rewarded by their online interactions, even where there is a low level of formalised trust between the sending and receiving parties, eg, if an e-mail claims to be sent from a bank, then it must be so, even if there has been no*** **a priori** ***exchange of credentials mediated by a trusted third party. Previously, mathematical models have been developed to predict trust established and maintenance based on reputation scores (e.g., Tran et al [1, 2]). This paper considers two inter-related questions: (a) can we model the behaviour of users learning to trust, based on non-associative models of learning (habituation and sensitisation), and (b) can we then locate this behavioural activity in a broader psychological model with a view to identifying potential countermeasures which might circumvent learned behaviour?***

I. INTRODUCTION

Historically, e-mail has been exchanged using open protocols with little regard for the establishment and/or verification of user credentials. This design was both intentional and well-intentioned, but the widespread abuse of e-mail now requires a more secure approach: the rise of phishing attacks, which are causing loss to consumers and banks, with the proceeds of crime being diverted to organised crime [15]. However, given that a wholesale replacement of e-mail clients and servers to make them more trustworthy appears unlikely in the short term, greater effort must be directed to (a) modelling the psychological bases of user interaction with e-mails and websites, specifically those used in phishing attacks, and (b) implement countermeasures which can be validated using such models.

One of they key characteristics of e-mail is that users trust that the stated sender is who they say they are. Two types of trust relationships are currently assumed during the sending and receiving of e-mail:

- Users trust the "system" that they are interacting with; and
- Sending and receiving entities trust each other

A phishing attack makes use of these relationships by betraying the trust that users place in their client software. Davis et al [3] have suggested that systems which operate in a manner which is consistent with their user's expectation of human behaviour will have the greatest user acceptance. Thus, extremely useful software (like e-mail and web clients) have become soft targets for organised crime.

This paper considers the broader case of how trust relationships can (and should) be developed between two or more entities that participate in a co-operative activity, such as exchanging e-mails, especially where the entities are not certified as trustworthy by an independent third party (the solution suggested by [4]). There can be many reasons why this certification is not available, including:

- Participation in an e-mail exchange may

978-1-4244-4626-1/09/$25.00 ©2009 Crown

need to be anonymous for fear of persecution or prosecution;

- The design of an e-mail system may enforce anonymity;
- The cost of purchasing certificates;
- Lack of agreed infrastructure (protocols, algorithms etc);
- The sheer scale and number of entities participating in the system makes it impractical to identify a single trust authority or chain of authorities who can vouch for the bona fides of all entities, especially where rekeying is required [16]

An understanding of how users develop and manage trust relationships in e-mail is critical if we are to defeat phishing – clearly, users who have a history of positive experiences in using e-mail (or web) have come to trust the technology "in the large", and also at the level of specific correspondents, such as banks, stockbrokers, auction houses etc – in other words, the cornerstone institutions of internet commerce.

One possible avenue in modelling trust would be to develop models of trust that entities can use to compute the trustworthiness of the parties that they interact with, and provide this information directly back to the user. Indeed, this is precisely what happens when one user identifies a phishing e-mail (including sender information) and submits a report to Phishtank – once verified, this information can potentially be used by other users to inform their decisions. In this example, Phishtank acts somewhat like a trusted third party - who stores trust data and forwards it on request.

In this paper, models of trust establishment and management are considered, that relate directly to phishing by providing a clear behavioural model for why users – especially those who have been using e-mail for a significant amount of time – become victims of phishing attacks. A broader model of user behaviour in phishing, including cognitive and perceptual inputs, is then proposed, and the process of using the model to propose and validate technical countermeasures is then described.

II. TRUST ESTABLISHMENT AND MAINTENANCE

Internet commerce depends on trust – users must trust the platforms that they use, and those of the merchants, auction houses and banks that they interact with online. As bandwidth availability and reliability becomes less of a concern, the core problem for users is the extent to which they can *trust* the content that is received. In many applications, users demand authentic, up-to-date data with guaranteed integrity, rather than sacrificing consistency to save a bit of bandwidth. This will increasingly be the case where web applications move from being simple catalogue lookup sites to more interactive and integrated applications, such as total wealth management and other sophisticated applications of internet commerce. Attacks like phishing threaten to erode public confidence in what are many new and exciting applications.

In this context, processes for modelling trust are presented in this paper, comprising two key biological learning processes:

(a) A mechanism for the <u>establishment</u> of trust, that is based on a generic model of non-associative learning (habituation), outlined in this paper; and

(b) A mechanism for the <u>maintenance</u> of trust relationships, that is based on sensitisation to malevolent events

The mathematical basis of the model is drawn from physiological models of non-associative learning in human and non-human species, like the giant slug (*aplysia*; [5]). The established benefits of using these two learning processes are:

(a) Many decades of experimental data have been gathered to support the processes of habituation and sensitisation; and

(b) Mathematical models have been fitted to this data with great accuracy; thus, future responses can be predicted on the basis of known data

In the sections that follow, habituation and sensitisation processes will be reviewed to show that they provide one model for understanding the processes required to establish trust relationships between e-mail receiving and sending entities, without reference to any external third parties.

In addition, the relationship between this purely behavioural level of explanation is contrasted with other elements, such as cognitive processing, which have an important role to play in preventing phishing attacks in the future. The goal is not to downplay the important role that entities like Phishtank play, but to better understand the basic psychological processes underlying trust establishment and management.

III. THE MODEL - HABITUATION

Two complementary processes are modelled in this paper: habituation and sensitisation. Before these concepts are reviewed in detail, the rationale for considering the development and maintenance of trust relationships between entities as a learning process will be presented.

Consider the situation where two e-mail sending and receiving entities on a network have no knowledge about each other (say, in the form of a reputation score supplied by a third party). How are these two parties to trust each other? During any initial exchange, the *maxima of distrust* occurs. However, if this first interaction is successful, then the level of distrust is likely to decrease at the second interaction, and so on. After some number of successful interactions, the two parties can be said to not distrust each other. This may imply trust, but most likely, an ongoing monitoring of the relationship integrity will continue.

For example, a user may become an Ebay user, and after having many successful and rewarding purchases, they are considered to have developed a trust relationship with Ebay. In contrast, they may be nervous about spending too much money on their first Ebay purchase because their relationship is at the maxima of distrust. If a user was not an Ebay user, but they receive an e-mail purporting from Ebay about a purchase that they have made, they are likely to be very suspicious about the content – in essence, the maxima of distrust has protected them from being phished. On the other hand, after being a happy Ebay user, their psychological guard is lowered, and they are more likely to click on any message which purports to come from Ebay.

This progressive reduction in distrust is a learning conceptualisation of habituation, i.e., where there is a decrease in a dependent variable (distrust) as a result of repeated, harmless (or beneficial) exposure to a stimulus (interactions with another party). Each party in an interaction is responsible for computing its own response – but one can imagine that, if two parties with zero knowledge about each other have equally satisfying interactions (in quantity and quality), then the distrust profiles for each other should be equivalent. For example, Paypal may place restrictions on user accounts until various verification checks have been successfully completed. This approach introduces more structure into the trust building process – more for the benefit of existing verified users, rather than new (and potentially malicious) users.

In biological terms, the initial approach of one entity to another that produces a distrust maxima, is known as the startle reaction [6]. In most biological systems, the startle response and habituation can be observed in so-called automatic (or autonomic) functions over which conscious control is not exercised.

Note that habituation processes are always stimulus-specific, meaning that a trust relationship developed between two entities may not necessarily be generalised to more than one service. For example, an entity may be trusted to download data, but not upload or an e-mail service may be trusted, but not a web service. Similarly, users may learn to trust messages from Ebay but not from another auction house, even if their businesses carry out a similar function.

Different thresholds on the distrust curve may be applied for different functions, and/or different rates of learning to trust may also apply. In short, one could imagine a two-dimensional matrix of trust association trust values for each user x service interaction. These may be related to the commercial value of the relationship involved – for example, for high value transactions like currency trading, users may be reluctant to commit too much funds to an untrusted service, while a micro-credit system with a limit of $10 per transaction would have a lower threshold to reach. In biological terms, this is known as salience – relative to other items of a similar type, how much does a particular item stand out?

Habituation is not just a qualitative phenomenon, but quantitative in the sense that models can be accurately fitted to real-world data. Some examples of habituation from the animal kingdom include:

- male sticklebacks, who mutually establish boundaries over a period of 30 minutes, with a 50% reduction in biting responses [7]
- *coelenterates*, like the *hydra*, whose probability of a response to an external stimulus drops from 1.0 to 0.2 after several hundred presentations [8]
- *protozoans*, such as *stentor coeruleus*, which display a similar response profile to the coelenterates [9]

Given its universality in the natural world, the response profile of habituation has been very broadly established. A number of general properties of habituation have been determined that may be relevant in the establishment of trust relationships between e-mail senders and receivers, such as a primacy effect, where initial interactions are treated with suspicion.

There are many implementations of habituation which differ in the number of natural phenomena that they attempt to reproduce. Figure 1 shows the distrust function, computed through a simple habituation model [10]. The decay of response to a stimulus presented through a number of sequential interactions shows how trust can be built up. In the absence of an aversive stimulus, the function is smooth – more sophisticated and non-linear behaviour can be introduced by integrating the outputs of cascading units, providing a quicker route to trust. As long a sensible sensitisation function can be formulated to reduce trust in the case of aversive event, the model replicates desired behaviour in sending and receiving agents.

What would make the model more directly useful for modelling the building of trust in e-mail interactions would be to correctly parameterise the rates at which distrust decreases, and the appropriate initial threshold at which the maxima of distrust occurs, either through the primacy effect, or the startle reflex.

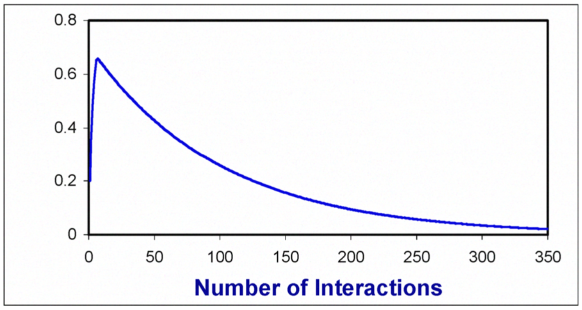

Figure 1. Distrust computed by habituation model

IV. PHISHING

How does the model of trust establishment relate to phishing? The answer is that it provides a clear behavioural model for why users – especially those who have been using e-mail for a significant amount of time – become victims of phishing attacks. Although it is possible that naïve users might be phishing victims, there is strong anecdotal evidence that even experienced users can become victims [17].

Users – through experience – clearly become habituated to the characteristics of a phishing e-mail that might potentially flag it as a phishing message. These characteristics include:

- The displayed URL being different to the URL embedded in the HTML code and usually visible when the user moves their cursor over the link
- Spelling mistakes – even in the subject line – of the bank's name from which the phishing message has been purportedly sent
- Being asked for additional information (such as license and passport numbers) in addition to normal banking login credentials

As long as the *shallow* features of the message appear to be genuine, the message tends to elicit a *behavioural* rather than a *cognitive* response. This means that most of the content in the message is not processed at anything other than a shallow level. Figure 4 shows a simplified version of a generic psychological model that shows how these different processing levels are related.

This outcome can be predicted from Craik and Lockhart's [14] classic work on *levels of processing.* In this approach, *depth* of processing is defined by the meanings extracted from the processing activity, rather than focusing on the number of times an item of information is processed. In this framework, shallow processing occurs when users focus on structural properties (such as how a word looks or sounds) versus deep processing, where the actual meanings (semantics) are extracted and understood in some way.

An example of the difference between shallow and deep processing in phishing would be users who take a cursory glance at the "Sender" or "Subject" field of an e-mail, and quickly action the item by (inappropriately) clicking on the link. Deep processing of the message would involve the user reading the contents carefully, cross-checking the claims made in the e-mail carefully, and then verifying whether the displayed link actually matched the known good link of the service in question.

If users really read every word of a phishing message, and checked the key structural elements such as the URL, then phishing would not occur at the same level that it currently does. So, while habituation is a positive and natural for e-mail users to trust each other, it may also lead to deeper processing at the cognitive level not being performed.

Figure 2 shows a simple psychological (structural) model that explains how phishing results from behaviour over-riding cognition, as a result of information which is visually acquired during perceptual processing of e-mail messages. If a user is habituated, they are more likely to be phished, than if they cognitively process the phishing message *at sufficient depth.*

Starting from the top of the diagram, when a user receives a phishing message, the first level at which they process the message's information is perceptual, i.e., the characters of a message are extracted. If the user has a very high trust level (or expectancy) of a certain type of e-mail – such a message from a taxation department – then they might simply click on the link and be phished.

On the other hand, if they do not trust the message, they are more likely to deeply process the contents of the message and further investigate its provenance and authenticity.

While this model is very general, it does open up the possibility of using the model to suggest new countermeasures or understand why some existing countermeasures actually work. For example, newer browsers which bold the fully-qualified domain name of a webserver host are designed to operate at the perceptual level, *before* behavioural responses are initiated.

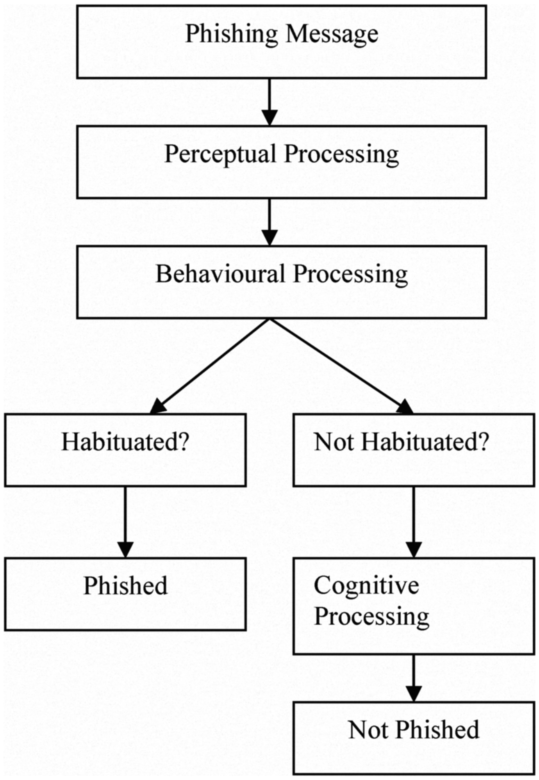

Figure 2. Psychological model of phishing

V. THE MODEL – SENSITIZATION

From Figure 1, you can see that habituation converges on a *distrust minima* after a number of stimulus presentations. This represents a mature trust relationship – or one where phishing is more likely. However, are there any biological correlates of situations where this trust relationship is broken? One can imagine many situations where this might occur during e-mail exchange, including phishing attempts that seek to take advantage of previously established trust relationships[2].

Sensitization provides one mechanism for rapidly eliminating habituation in non-associative learning. Sensitization arises when an aversive stimulus is presented in place of the stimulus which is anticipated. In the case of *aplysia*, this may mean that habituation has been achieved by a gentle linear stroking, leading to a distrust minima, and the aversive stimulus is delivered in the form of a sharp tap. The immediate reaction of *aplysia* is that habituation is minimized, i.e., the habituation process needs to be initialized once again before the response is minimized with respect to the non-aversive stimulus [5]. Figure 3 shows a sensitisation event following a previous habituation.

In terms of trust, *aplysia*-style sensitization provides an "all-or-nothing" route to maximize distrust after previously minimizing trust. Clearly, this does not provide a generalizable model for all scenarios involving a transition from a totally trusted to a less trusted relationship. Penalties may be applied between parties according to quite different policies, e.g., "three strikes" policies against spammers, or immediate termination of relationship if child pornography was being distributed (and police notified). So, sensitization matches some use cases but not all.

However, it should be noted that sensitization should not necessarily be viewed as the opposite of habituation. Indeed, for some species, sensitization is a progressive amplification of a response which builds up gradually – especially for more complex species like human beings. An example is the amplification of neural response to stimulation of some peripheral nerves in response to repeated rubbing – after some time, the response will be amplified to the degree that pain is experienced [11].

Most importantly for internet commerce security, sensitization predicts that – for users who have previously been habituated and who are sensitized – they will become habituated once again, given sufficient time, and further positive interactions. This is important, since a single poor experience with phishing should not deter users from engaging in e-commerce as a client.

However, the risk is that – if continued press coverage about phishing begins to sensitize a sufficiently large number of users, they may begin to avoid internet commerce security applications, as their cognitive processes take over their behavioural responses. While cognitively processing messages is encouraged, developing a generalized aversion to using the web for banking would be a very poor outcome.

In terms of applying the phishing model to sensitization, anti-phishing plug-ins appear to provide

[2] Or perhaps repeated media reports of how easily you can be phished would also have this effect!

the necessary sensitization event, where they flag a message as potentially being a phishing message. Reviewing the model, the plug-ins may be more effective if multiple sensory modalities are utilized, eg,, flashing a warning visually on the screen while playing an alert sound.

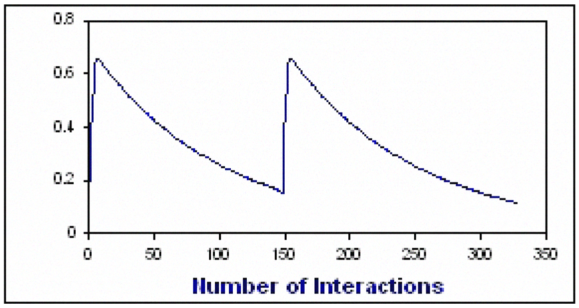

Figure 3. Sensitization occurring at interaction 150

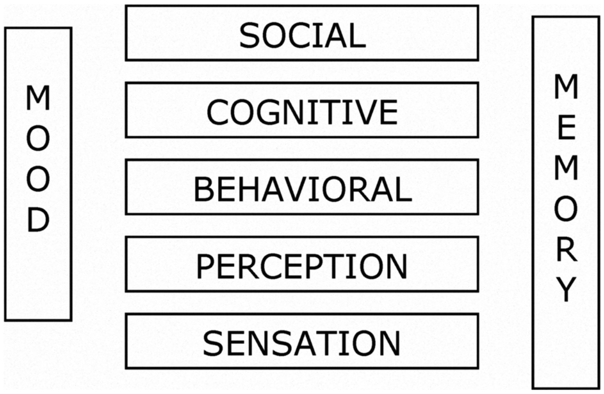

Figure 4. Generic psychological model

VI. DISCUSSION

In this paper, an attempt has been made to illustrate how some simple ideas from non-associative learning theory can be used to provide a model for behavioural responses in phishing attacks, based on the notion that phishing succeeds when users trust a purported e-mail sender too much. This trust starts at zero, and is built up over successful, repeated interactions.

Models for habituation and sensitization have been applied to understand current issues in distrust of e-mail messages, such as phishing, and a first sketch of a cognitive model to explain how phishing success can result from habituation has been demonstrated. In summary, users relying on habituation to process cues about phishing in e-mail messages are not processing messages cognitively at sufficient depth to detect some fairly obvious clues.

Further work is required to implement cascaded versions of the model, and a real-world implementation with accompanying experiments will clarify whether principles from the biological world are relevant to e-mail technology. What is required to parameterise the model is further data on which types of phishing e-mail are more successful at luring users who have different types of experience. This data could be used to test model predictions based on various demographics. Also, there seem to be some fairly obvious characteristics of phishing e-mails that any deep analysis by a user should identify, such as mis-spellings.

Vendors are already using an implicit model of user psychology to defeat phishing – both Microsoft Internet Explorer and Mozilla Firefox have now implemented anti-phishing measures at the *perceptual* level, by bolding the fully-qualified domain names of hosts when a users connects to a website using their browser. By attempting to intervene between user's learned responses, and the presentation of the message, the vendors hope to subvert the semi-automated responses of users who have learned to trust e-mail.

Moving beyond e-mail exchanges at the user level, what role do intermediaries – such as distributed reputation managers, certification agencies etc - have to play in trust development and maintenance? Is there a role for these agents in preventing phishing? Can they adequately deal with zero-day exploits? What type and scope of verification is required for phishing reports?

The use cases considered in this paper are based around the pairwise computation of trust (or distrust) between two parties, but the mechanisms (habituation and sensitization) do not immediately scale to cater for more complex and/or distributed cases, that might include a trusted third party. Saliency – mentioned earlier – may be one factor, although repetition (as seen in the habituation model) is also a factor.

Where reputation scores are computed in distributed trust environments, it is certainly possible that distrust scores computed from habituation and sensitization processes could be used as one element of scoring. I.e., where an email recipient scores a recipient at maximal distrust (e.g., as a phishing message), other e-mail clients could consult other sources who have previously dealt with the sender, or users could combine their individual scores centrally. This type of reputation model has worked quite successfully for Ebay, for example, and it's interesting to note the restrictions that were necessary to put in place in recent years, including the prevention of sellers adding negative ratings to buyers.

In the case of email, senders who have high distrust scores could also be prevented from scoring down those recipients who fail to trust them [12]. The specific dynamics of these relationships – especially where they involve assymetries and nonlinearities – could be modelled using dynamical systems models [13].

Further empirical work is required to determine which type of psychological model is most appropriate for understanding phishing in different contexts. Habituation, for example, is not an appropriate model where there is an element of reward in the activity; thus, it works well for understanding the time-pressured reading of e-mails, but not so well for understanding highly-rewarding activities like social networking.

In addition, the behavioural level is just one level in a broader psychological understanding of how time-pressured decision making about phishing e-mails results in poor choices. Future work will examine how decision models based on the accumulation of information – such as a random walk model for two-choice reaction time [18] - can be used to understand the interplay between cognitive and behavioural processing in phishing. Indeed, there has already been some excellent progress in this area vis-à-vis human factors in security [19, 20].

V. REFERENCES

[1] H. Tran, P. Watters, M. Hitchens, and V. Varadharajan, "Trust and authorization in the grid: a recommendation model", Proceedings of the International Conference on Pervasive Services, pp. 433 – 436, 2005.

[2] H. Tran, M. Hitchens, V. Varadharajan, and P. Watters, "Trust based Access Control Framework for P2P File-Sharing Systems", Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.

[3] F. Davis, R. Bagozzi, and P. Warshaw, P., "User acceptance of computer technology: a comparison of two theoretical models", Management Science, 35, pp. 982-1003, 1989.

[4] S. Taylor, and P. Watters, "Trustworthy e-mail using secure XML Web services", Proceedings of Seventh IEEE International Conference on E-Commerce Technology, pp. 307- 312, 2005.

[5] E. Kandel, Behavioral Biology of *Aplysia*, San Francisco, CA: Norton, 1989.

[6] M. Davis, D. Gendelman, M. Tischler, and P. Gendelman, "A primary acoustic startle circuit: lesion and stimulation studies", The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 2(6), pp. 791–805, 1982.

[7] H. Peeke, and A. Veno, "Stimulus specificity of habituated aggression in the three-spined stickleback (Gasterosteus aculeatus)". Behav. Biol. 8, pp. 427–431, 1973.

[8] N. Rushforth, "Inhibition of contraction responses of Hydra", Integrative and Comparative Zoology, 5, pp. 503-513, 1965.

[9] D. Wood, "Habituation in Stentor produced by mechanoreceptor channel modification", Journal of Neuroscience, 2254, 1988.

[10] J. Staddon, and J. Higa, "Multiple time scales in simple habituation", Psychological Review, 103, pp. 720-733, 1996.

[11] I. Bell, E. Hardin, C. Baldwin, and G. Schwartz, "Increased limbic system symptomatology and sensitizability of young adults with chemical and noise sensitivities", Environ Res 70(2), pp. 84–97, 1995.

[12] D. Houser, and J. Wooders, "Reputation in Auctions: Theory, and Evidence from eBay", Journal of Economics & Management Strategy, Vol. 15, pp. 353-369, 2006.

[13] P. Watters, P. Ball, and S. Carr, "Social processes as dynamical processes: Qualitative dynamical systems theory in social psychology", Current Research in Social Psychology, 7(1), 1996.

[14] F. Craik, R. Lockhart, "Levels of processing: A framework for memory research", Journal of Verbal Learning and Verbal Behavior, 11, pp. 671-684, 1972.

[15] S. McCombie, P. Watters, A. Ng, A., and B. Watson, "Forensic characteristics of phishing - Petty theft or organized crime?", Proceedings of the Fourth International Conference on Web Information Systems and Technologies (WEBIST 2008), pp. 149-157, 2008.

[16] T. Pham, T., and P.A. Watters, "The efficiency of periodic rekeying in dynamic group key management", Proceedings of the 4th European Conference on Universal Multiservice Networks, Toulouse, France, 2007.

[17] A. Fowler, "Fear in the fast lane", Transcript of Four Corners, ABC TV, http://www.abc.net.au/4corners/content/2009/s2655088.htm, Broadcast 17/8/2009.

[18] S. Link, and R. A. Heath, "A sequential theory of psychological discrimination", Psychometrika, 40, 77-pp. 105, 1975.

[19] J. Grossklags, N. Christin, and J. Chuang, "Predicted and Observed User Behavior in the Weakest-link Security Game", Proceedings of UPSEC, 2008.

[20] J. Grossklags, N. Christin, and J. Chuang, "Secure or insure?: a game-theoretic analysis of information security games", Proceedings of WWW 2008, pp. 209-218, 2008.