Assuring the Safety of Asymmetric Social Protocols

Virgil Gligor[1] and Frank Stajano[2]

[1] Carnegie Mellon University
[2] University of Cambridge

Abstract. Most studies of security protocols in the literature refer to interactions between computers. Nowadays, however, more and more fraud (such as phishing, Nigerian scams and the like) is carried out by abusing social protocols—that is to say, computer-mediated interactions between human subjects. We call a social protocol "asymmetric" when the initial sender benefits from execution of the protocol but the recipient is not guaranteed against dishonesty of the sender. Can a recipient ever safely engage in an asymmetric social protocol?

Over the past decade or two, computer-mediated communications and purchasing transactions have become pervasive among the general public. As a consequence, attacks on social protocols have grown in prominence and value. We need a principled and systemic response to this problem, rather than ad-hoc patches.

Our contribution is to introduce a framework, the "marketplace of social protocol insurers", in which specialised providers compete to offer safety guarantees, for a fee, to subjects who wish to engage in social protocols. Providers need to develop accurate classifiers for rating protocol inputs as safe or dangerous, and the providers with the most accurate classifiers can price their insurance premiums more competitively, thereby winning a greater share of the customers.

Our solution offers, through competition amongst providers, aligned incentives for the development and deployment of accurate classifiers to distinguish fraudulent and legitimate inputs and it offers a safe way for ordinary users to engage in asymmetric social protocols without having to become experts at detecting fraudulent proposals.

1 Introduction

People participate in a variety of socio-economic activities over computer networks such as social networking, commerce, crowdsourcing and so forth. We call "social protocols" these interactions of human subjects with networks that involve computers and other humans. People therefore engage in a variety of social protocols with subtle security and privacy consequences. Increasingly, these protocols lead to online manipulation, deception, and scams on an unprecedented scale. Particularly vulnerable to scams and deception is a growing aging population who engages in social protocols over the Internet daily for a variety of life-enriching activities. Protocols between computer-based principals have been

Authors' preprint. Appears in *Proc. Security Protocols Workshop 2017*, LNCS 10476, pp 38–48.

extensively studied. We focus instead on interactive social protocols in networks of computers and humans, and on the susceptibility of unsuspecting participants to psycho-social manipulation and deception by skilled adversaries.
The overarching questions that we ask are:

– How can computer systems and network services help people make critical decisions that affect the individuals' own security and privacy, that of organizations, and of entire social networks?
– How can network services protect themselves from manipulation, deception, and scams?

We offer the following contributions in this paper.

– We introduce the "marketplace of social protocol insurers" to assure the safety of asymmetric social protocols for end users.
– We show how competition in this marketplace provides a virtuous circle of incentives that results in safer online interactions for end users. (More specifically, competition among insurers rewards the development of more accurate classifiers to distinguish fraudulent inputs from innocuous ones—a task that is much better performed by dedicated expert operators than by naïve end users. Additionally, competition among insurers ultimately deters insurers from taking easy shortcuts to the detriment of the users.)
– We show how the classifiers employed by insurers don't need to be perfect in order to provide a tangible benefit.
– We show how non-technical end users can safely engage in asymmetric social protocols by paying an insurer to assess which interactions can be pursued and which ones should be avoided.
– We analyse ways of abusing the system and offer preventive countermeasures.

2 Social protocols

Past research identified the following three key characteristics of social protocols that lead to scams, deception, and manipulation [5,2,3].

Value perception: honest participants perceive that engaging in the social protocol will ultimately be beneficial to them, despite an implicit risk. For example, a person may willingly disclose private information to a service provider, accept a certificate of uncertain origin, or click on an unknown link based on the expectation that they will receive a better service. Or open an attachment to an email message from an otherwise trustworthy sender for timely response to a question of common interest.

Irreducible asymmetry: dishonest participants are better off after executing the protocol, while honest participants cannot a priori protect themselves from negative consequences of interacting with participants who may turn out to be dishonest. For example, an unscrupulous service provider may profitably misuse personal data disclosed by a user, who will have no recourse after the fact against the unauthorized disclosure of their private information. Or a fraudster may direct a participant's machine to connect to a server that offers malicious software for download, which surreptitiously accesses the unwitting participant's data.

Safety states: despite the asymmetry, honest participants must have ways of establishing well-founded beliefs in the honesty of other participants and the (positive or negative) value accruing from protocol interaction. For example, by relying on a network of trustworthy parties or commercially available services, a participant can gain visibility into the protocol originator's identity, network presence, reputation, and validity of their value offer.

Several open research questions arise about the safety of interactive social protocols. Under what conditions do safe protocols exist? And if they do, how can we identify them? If we can identify them, are they usable in practice? In other words, can they be understood and used by ordinary members of the public without undergoing extensive training? Can users engage in these protocols without relying on networks of trusted entities? For example, can users exploit "social collateral models" (e.g., all protocol participants putting some financial, reputational or other value at stake as a guarantee against their own misbehaviour) to derive safety conditions without relying on globally trusted authorities? Can we use decision science to help users behave securely in cyberspace? Can we improve interactive social protocols to guard not only against external adversaries but against misuse of legitimate privileges by insiders—an increasingly significant concern? Can we devise effective retribution mechanisms and formally prove that they would deter misuse by adversaries exhibiting rational behaviour[3]?

In this paper we address the subset of these questions that help frame (1) the dilemma faced by a receiver of an offer to engage in a social protocol from an unknown and possibly dishonest participant, and (2) the network services that assure a receiver that the protocol will reach a safe state.

3 Receiver's dilemma and past attempts to solve it

When engaging in an interactive social protocol, a receiver must first *assess the value proposition.* In particular, the receiver must determine whether the face value of an offer received represents a scam or an attempt to deceive/manipulate, or a mutually beneficial transaction.

Second, the receiver must *assess the trustworthiness* of the offer sender for the current protocol run. The assessment may be residual: it may account for the fact that an offer sender proved trustworthy in the past. In general, trustworthiness can be established only in the context of the receiver's beliefs and preferences. For example, the receiver may attempt to determine the offer sender's public reputation, or may rely on recommendations of others.

[3] Although we must accept the existence of irrational adversaries—the cyber equivalent of suicide bombers—whom such retribution mechanisms will not deter.

Third, the receiver may also want to lower his risk of using a social protocol by *finding ways to insure himself* in case of misbehaviour of the offer sender.

Fourth, the receiver may want to determine whether accountability mechanisms exist that can be used in conjunction with graduated punishment such that the offer maker[4] can be *deterred from perpetrating* a scam.

We identify three broad approaches to help with the assessments required to resolve the receiver's dilemma.

3.1 Asymmetric-protocol exclusion

One way to deal with the irreducible asymmetry is to detect it in a protocol and avoid using that protocol. This limits social interactions to symmetric protocols. For example, a third party that is trusted by both the offer maker and receiver may in fact act as an escrow agent and ensure that neither participant can cheat without losing their escrow deposit [3]. This approach is hard to adopt because trusted third parties are difficult to find in practice. Even when they can be found, agreement among participants on the appropriate escrow value may be difficult to reach.

3.2 Protocol compensation for asymmetry

Another potential solution to the receiver's dilemma is to add an explicit *asymmetry reduction step* to a protocol to compensate for the otherwise irreducible asymmetry. For example, protocols based on social collateral models [3] add a message exchange phase with an outside party who has collateral with the receiver and who is able to establish whether an offer sender is accountable. Although making a sender accountable does not remove protocol asymmetry, it can reduce the receiver's risk and placate his betrayal aversion. This approach has been used in practice in other domains, such as money lending in third-world countries. However, it is not always possible to find an outsider who can hold specific offer makers accountable and has sufficient receiver collateral. Even if such a party is found, whether the aversion reduction is sufficient to guarantee receiver engagement depends on whether a safe state can be established; i.e., whether accountability can lead to sufficient punishment to deter sender misbehavior.

3.3 Conditional asymmetry acceptance

Perhaps the most common way to solve the receiver's dilemma is to enable the receiver to *assess the trustworthiness of an offer maker* when the value proposition is not in question. For example, in e-commerce services, both Ebay and Amazon (1) enable a receiver's belief formation by providing a reputation

[4] We maintain a subtle distinction between the *offer maker*, who creates the offer in the first place, and the *offer sender* who sends it to the recipient. These may be the same principal, as when a Nigerian operator sends a "419" to a victim, or not, as when a naïve user passes on the fake alert that the email password must be changed.

assessment system that is well-tuned to certain classes of services, (2) lower a receiver's risk aversion by providing reasonably-priced insurance against cheating by high-reputation offer senders, and (3) lower a receiver's betrayal aversion by banning known cheaters from using their services. However, the use of these protocols outside e-commerce settings may be limited by other factors such as service scalability, deployment cost and profitability.

Conversely, other protocols in this class attempt to help recipients *establish the face value of an offer* when the trustworthiness of the offer sender is not in question. These protocols are necessary because beliefs of offer sender's trustworthiness are intransitive, and also because honesty is not correlated with technical competence. For example, a recipient may trust an offer sender (e.g., a relative, a friend, a colleague, an employer) but not the offer sender's social contacts who might have scammed him into accepting deceptive email messages which he unwittingly forwards to the receiver—a common vector of malware propagation. Services offering trustworthiness assessments already appeared on the market and may become viable in the longer term as add-ons to popular services, such as email.

Finally, the last type of protocols in this class *balances cost of assessment of sender trustworthiness against value offered.* For example, if the cost of the trustworthiness assessment is impractically high and the value offered is comparatively low, the recipient's recommended response is to deny acceptance of the offer. This is an effective better-safe-than-sorry approach, which trades off value lost against safety.

4 Our solution: a marketplace of social protocol insurers

We argue, based on available evidence, that most people are demonstrably unable to perform trustworthiness assessments of a stranger, and equally unable to assess the face value of his offers accurately. In fact, most people seem to fall for the same type of scam repeatedly, even after appearing to understand the underlying scammers' protocols [4]. More alarmingly, even experts at detecting scams occasionally accept a too-good-to-be-true offer. The use of the Internet exacerbates this problem because possible telltale signs provided by a scammer's body language are unavailable; i.e., as Peter Steiner's 1993 cartoon in *The New Yorker* illustrates, "on the Internet, no one knows that you are a dog."

On that basis, we propose the creation of a framework in which Internet services can perform assessments of both sender trustworthiness and offer value. Our framework takes the shape of a *marketplace of social protocol insurers.*

One of the core ideas of our solution is for end users to *offload the burden of assessment to expert third parties*, in exchange for a modest fee. This allows specialisation, economies of scale and, from the third parties, investment in advanced techniques to improve such assessments. Another core idea is to *have several such third parties in competition with each other*, in order to keep fees low and to provide incentives for improving the accuracy of the assessments. The third core idea is for these third parties to *back their assessments by offering in-*

surance: if an assessor declares to a customer that a certain action is safe, the assessor promises to refund the customer if he incurs a loss as a consequence of taking that action.

Our framework offers the following advantages.

From the viewpoint of the user, social protocols are no longer dangerous. The user shows the available inputs to his preferred insurer, who issues a verdict about the risk of engaging in the protocol. If the verdict of the insurer says it's dangerous, the user does not proceed and stays safe. (In case of true positive[5], the user has avoided an attack. In case of false positive, the insurer has been overcautious and the user has needlessly given up on the value of that interaction. We'll revisit the case of false positives later, in section 5.2.) Otherwise, if the verdict said it was safe, the user proceeds and, assuming it was a true negative, enjoys the value deriving from the interaction. If it was a false negative, that is to say if the insurer said it was safe but it wasn't, and therefore the user incurs a loss, then the insurer steps in and offers a refund[6].

From the viewpoint of the insurer, payment is received for two conceptually different services: (1) offering an expert verdict on whether a potential interaction is dangerous; (2) offering insurance against false negatives for interactions that the insurer declared not to be dangerous. The insurer has an obvious incentive to keep a low rate of false negatives (because he must pay out for each of them). This in turn is an indirect incentive to make the classifier accurate (which reduces both false positives and false negatives). There is still the risk that a dishonest insurer would tackle the issue not by improving the classifier but by moving the threshold so as to have almost no false negatives even if this means many false positives (saying that everything is dangerous, even if it is not), but the fact that in our framework we competitively pit many insurers against each other acts as a deterrent for such dishonest behaviour, because the user may notice the too many rejections and may decide to switch to an insurer offering fewer false positives.

[5] To avoid misunderstanding, the obvious definitions are as follows. *True positive:* the input was dangerous and was flagged as such. *True negative*: the input was not dangerous and was flagged as such. *False positive:* the input was flagged as dangerous despite not being dangerous. *False negative:* the input was flagged as not dangerous despite being dangerous.

[6] Note that insurance and refund are only able to "undo the evil deed" for certain types of threats, such as those that result in financial loss, where the victim may be fully refunded. For others, such as those that result in confidentiality loss, compensation may still be offered but it is impossible to undo the disclosure and restore the state of the world to that before the occurrence of the attack. This is an inherent limit of any approach involving remedial compensation and is not specific to our framework. It should also be noted that our approach only resorts to compensation in the case of false negatives, but that in the case of true positives it employs prevention (not engaging in the protocol that would result in, say, confidentiality loss), which is much better. Note also that, as detailed in the following paragraph, the financial incentives for insurers in our framework are aligned to favour true positives against false negatives, which is precisely the intended outcome.

5 The details: how the marketplace protects end users

5.1 An example: email protection

Since so many computer-mediated frauds and scams arrive through email, let's consider, by way of example, how our marketplace framework would protect end users from fraudulent emails.

A basic service could offer receivers a number of security features based on measurement of the receiver's untrusted inputs (email messages) according to proprietary metrics. The general idea here is that the end user (customer) buys a service from one of several competing insurers (provider). The customer forwards all his emails to the provider[7] and receives a risk rating for each message. The risk rating indicates the provider's opinion about whether it is safe for the customer to engage in the interaction offered by the sender (opening an attachment, sending money to a stranger, purchasing a blue pill and so forth). The risk rating is backed by an insurance, in that the provider will compensate the customer if engaging in an interaction that had been deemed safe results instead in a loss[8].

The insurance needs to be paid for with a premium. One might at first imagine that the cost itself of the premium could be used as the rating (if a message is safe then it's cheap to insure and vice versa) but this simple-minded strategy backfires and opens the system to abuse. The user would normally be expected to pay the premium for each message in order to feel safe in opening the message. However the user might simply consider that a high premium means "risky" and therefore choose not to open the message (without paying), and similarly consider a low premium as a signal that the message is "safe" and therefore open it without buying the insurance. The user would not pay in either case, and would free-ride on the signalling provided by the price set by the insurer.

It is difficult for the insurer to extract payment of the premium without revealing its amount: why would the user want to commit to paying an unknown

[7] We are glossing over the obvious confidentiality problems, which would have to be addressed by an appropriate service level agreement. On the other hand, we observe that a not insignificant fraction of battle-hardened security researchers nonchalantly forward all their emails to Google or Yahoo without batting an eyelid.

[8] The trust* protocol by Clarke et al [1] also attempts to protect the recipient against spam by paying an insurance broker, but there it's the sender who pays, rather than the recipient: the sender, who wishes her own mail to get through, pays the broker to offer a guarantee to the recipient that the mail is not spam. It is assumed that a recipient may choose not to open any emails that arrive without such guarantees. Until such a system becomes widespread, however, most emails will arrive without guarantees anyway (because senders won't even know about the existence of the trust* scheme), so the recipient will have to decide for himself whether to open them without protection from the broker. In our system, by contrast, once the recipient establishes an insurance contract with a broker, the recipient is protected against all incoming emails, whether the senders play the game or not. This means the scheme in this paper offers its benefits to its early adopters even before it becomes mainstream.

price? If the insurer attempts to aggregate messages and price them as a bundle (such as a batch of 100 messages, or a batch of one day's worth of messages), the user experiences unacceptable delays: who would want to get all their emails one day late? Moreover, if the insurer offered a premium for the bundle, with the semantics that he will compensate the user if the user is damaged by opening one of the messages in the bundle, then the user would be entitled to be careless and open any messages without precaution (moral hazard). Attempting to counter this situation by charging a premium equal to the maximum possible loss would make the insurance unattractive and pointless.

The core mistake of this naïve approach is the attempt to "price" and therefore insure *all* messages—even the ones that are blatantly fraudulent. Both provider and subscriber are better off if the provider refuses to insure the messages it deems too risky. At a conceptual level it is advantageous to separate the service of rating the messages from that of insuring the customer's actions. The provider can then safely sell a rating service[9] whereby each email for the subscriber is given a risk rating. Then, as a complementary service that would normally go together, the provider can also sell insurance, possibly at several levels with different prices. For a given premium[10], the provider would guarantee to compensate the subscriber for engaging with any message that had been rated as not exceeding a defined risk threshold—the threshold being a function of the premium.

From a marketing viewpoint, this conceptual separation could be totally hidden: the subscriber simply pays to have each message rated and guaranteed. The guarantee is that, if he acts upon the rating as recommended (i.e., if he only opens messages that have been deemed to be below the agreed risk threshold for the paid premium), then he will be safe, either because nothing bad will happen or because he will be made whole by the insurer.

5.2 False positives

One remaining problem for the subscriber is that an unscrupulous insurer could honour the above agreement simply by being overcautious and labelling almost all messages as too risky. This would indeed provide a safe experience for the subscriber (and a lucrative one for the provider, who would pocket the premiums without hardly ever having to pay on insurance claims) but the subscriber would lose out on the value proposition of many false-positive innocuous messages that had been unnecessarily marked as dangerous.

In our framework, this is couterbalanced by the competition between the providers. Given a common fee, from the user's viewpoint the best provider is the one that labels more messages as safe. A user, or more likely an association of consumers, could subscribe to several providers simultaneously and rate them on the proportion of messages that they label as safe. Providers cannot simply mark

[9] Conceptually with a per-message charge, though commercially this might be more easily sold as flat-rate subscription with reasonable-use quotas.

[10] Again, ideally per-message but potentially as a flat-rate subscription with quotas.

all messages as safe because they would be liable for enormous payouts when users open malware; therefore market forces would oblige insurers to compete on classification accuracy, with the most accurate classifier getting the most business and making the greatest profits.

5.3 Service offered

In summary, the service offered by the provider gives a risk rating to each message and defines a threshold level above which a user's actions on rated input messages are insured by the service; i.e., a level of compensation a service subscriber receives for his losses incurred when accepting a malformed or deceptive input from a sender incorrectly assessed to be trustworthy by the service. The level of protection is defined according to state-of-the-art detection of malformed and deceptive input and known-to-be-malicious sender domains; e.g., spear phishing, encoded malware in email attachments and possibly text, or links to malicious servers. Zero-day attacks may be ruled out by the service, which limits the scope of the assessment to known attacks and due state-of-the-art diligence. Just as with anti-virus tools, the level of protection offered increases over time, as the service's database of known attacks increases. The security features offered are tailored to a subscriber's expectation of trustworthiness, his risk tolerance, and betrayal aversion. The assessment service learns these levels as it analyses subscriber-selected security features and subscription levels, which may change in time.

The assessment service needs to deliver a state-of-the-art level of due diligence to the point that it becomes a useful, profitable business. However, it need not be perfect. It would only need to ensure that its subscribers are better off than those who do not use the service. The existence of multiple competing services in the Internet assures that a service provider offers good value to its subscribers. In the absence of ground truth for complete solutions to the receiver's dilemma, competition in the marketplace will undoubtedly end up establishing which services are the de facto holders of ground truth.

The availability of multiple assessment services also assures subscriber choice in the marketplace. Independent *Consumer Reports*-like agencies could offer ratings of different services and provide users the opportunity to make informed choices in selecting a service. In addition, for subscribers who use multiple services that offer different assessments for the same input messages, these ratings may provide help in selecting the best assessment, since services' ratings reflect different levels of accuracy for common security metrics.

Services can also offer additional benefits, such as deterrence mechanisms against scam perpetrators, which include social protocol auditing and naming and shaming miscreants. They could also create honeypots for scammers, which would reveal the scammers' strategies, and publish those, along with naming and shaming them.

5.4 Subscriber Protocols

Subscribers pick the level of protection they desire and issue a subscription request to the service. This provides an initial indication of the subscriber's need for accuracy in trustworthiness and value proposition assessment to the service provider. Before acting on an input message, the subscriber sends it to the service, which performs the assessment. The assessment is returned to the subscriber, who then decides how to act on it; e.g., accept the message, read its contents, open an attachment, follow a link, accept a certificate, or discard the message.

The subscriber-service protocol needs to assure that a subscriber gets an assessment only if s/he has a required type of subscription and the subscriber's actions on the security-measured input are in accordance with the level of insurance provided. In effect, the service returns a sealed object to the subscriber with a certain set of actions that can be performed on the input message and logged by the service.

5.5 Thwarting further attacks on providers

The framework must ensure that neither the subscriber nor the service can deny the assessment request and response. Neither can cheat the other directly.

The subscriber (who may be the service's competitor) could send different versions of malformed inputs to a service to discover the service's "secret assessment sauce" and level of protection. Similarly, a subscriber's adversary may try to learn the assessment metric, adapt/bypass it, and attack the service's subscribers. For example, a subscriber may operate botnets that could issue multiple requests at a very high rate. The quota mechanisms and the cost of a subscription should be set such that an adversary's learning of the "secret sauce" by issuing multiple requests at a high rate becomes too high to be practical. The desired effect is that the adversary would not adapt and would instead target "non-insured" users, if any.

6 Conclusion

Past analyses of scam, deception or manipulation attacks via interactive social protocols have common characteristics; i.e., value perception by receivers of various types of offers, irreducible asymmetry in terms of the gains achieved by honest and dishonest participants, and the existence of safe states despite asymmetry. Past research in human psychology indicates that most people are generally unable to perceive the value of offers made via social protocols nor assess the trustworthiness of the offer senders correctly. Motivated by these observations, we propose a framework, the marketplace of social protocol insurers, in which users are able to delegate these assessments to specialists. This framework fosters the development of accurate services that measure a receiver's untrusted input (e.g., email message), and offer assessments of sender's trustworthiness and value offered. The service insures user actions on rated inputs according to a subscription scheme that compensates the user for losses incurred by acting on incorrectly rated inputs. Preliminary evidence of service viability appeared recently: new companies that discover the identity of email senders already exist[11]. We anticipate that rapid development of machine learning techniques will soon provide the ability to discern and understand the value proposition offered both in email text and encoded (e.g., *pdf*) attachments.

Subscriber-service protocols can be designed to protect services from miscreant subscribers and a service provider's adversaries. Market competition assures that security measurements of receivers' untrusted inputs by services can increase subscriber security at reasonable cost despite the absence of established ground truth.

7 Acknowledgements

We are grateful to Trinity College and the Computer Laboratory for hosting Virgil Gligor for part of his sabbatical in summer 2016 while much of this research was carried out.

References

1. Clarke, S.W., Christianson, B., Xiao, H.: Trust*: Using local guarantees to extend the reach of trust. In: Christianson, B., Malcolm, J.A., Matyas, V., Roe, M. (eds.) Security Protocols XVII, 17th International Workshop, Cambridge, UK, April 1-3, 2009. Revised Selected Papers. Lecture Notes in Computer Science, vol. 7028, pp. 171–178. Springer (2009), https://doi.org/10.1007/978-3-642-36213-2_21
2. Gligor, V.D., Wing, J.M.: Towards a theory of trust in networks of humans and computers. In: Christianson, B., Crispo, B., Malcolm, J.A., Stajano, F. (eds.) Security Protocols XIX - 19th International Workshop, Cambridge, UK, March 28-30, 2011, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7114, pp. 223–242. Springer (2011), https://doi.org/10.1007/978-3-642-25867-1_22
3. Kim, T.H., Gligor, V.D., Perrig, A.: Street-level trust semantics for attribute authentication. In: Christianson, B., Malcolm, J.A., Stajano, F., Anderson, J. (eds.) Security Protocols XX - 20th International Workshop, Cambridge, UK, April 12-13, 2012, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7622, pp. 96–115. Springer (2012), https://doi.org/10.1007/978-3-642-35694-0_12
4. Konnikova, M.: The Confidence Game. Viking (2016)
5. Stajano, F., Wilson, P.: Understanding scam victims: Seven principles for systems security. Commun. ACM 54(3), 70–75 (Mar 2011), http://doi.acm.org/10.1145/1897852.1897872

[11] www.agari.com