Strange Bedfellows: How and When to Work with Your Enemy

Aaron D. Jaggard[1] and Rebecca N. Wright[2]

[1] U.S. Naval Research Laboratory aaron.jaggard@nrl.navy.mil*
[2] DIMACS and Department of Computer Science, Rutgers University.
rebecca.wright@rutgers.edu**

Abstract. There are many examples of parties that are seemingly in opposition working together. In this position paper, we explore this in the context of security protocols with an emphasis on how these examples might produce long-term benefits for the "good guys" and how a formal model might be used to help prescribe approaches to collaboration with the "bad guys."

Collaboration is usually thought of as a joint effort where the parties involved have the same or similar end goals. However, there are many examples of collaboration between parties that are actually or seemingly in opposition. We sketch some of these below. Parties may choose to participate in these types of collaborations even though they believe that their opponents are rational and thus must see some benefit from the collaboration. We suggest that a reason for this is that the parties may have different time horizons or discount rates. For example, law enforcement may value the capture of a major kingpin highly even though it requires years of work, while their informants may have a much shorter term focus. Beyond seeing this as a possible explanation, we propose that this should inform strategy. A party such as a police force can, and in some cases should, collaborate with parties that have short time horizons in an effort to attack common opponents with longer time horizons.

Collaborating parties have their own incentives and their own reasons for collaboration. They also know that the other parties are often behaving rationally (although the utilities that they are trying to maximize may be quite different and even opposed in some ways). In the language of game theory, the "good guys" collaborating with the "bad guys" can be viewed as the good guys trying to find bad guys who maximize the good guys' utility (perhaps in equilibrium) and then collaborating with those bad guys to defeat other bad guys whose presence hurts both the good guys and the bad guys with whom they collaborate.

There are a number of (not necessarily mutually exclusive) reasons that such collaborations might occur:

* Work supported by ONR and DARPA.
** Work partially carried out with support from the National Science Foundation under grants CNS-1018557 and CCF-1101690.

Appears in *Proceedings of the 22nd International Workshop on Security Protocols*, March 19–21, 2014.

1. Parties who normally have opposing goals find that they have common or similar goals in a particular context, even if for different reasons. For example, in a political race with three candidates, the two weakest candidates sometimes work together to attack the strongest candidates, because they both believe they can beat each other once the leading candidate is eliminated. Similarly, in criminal investigations, criminal suspects may be offered the opportunity to act as informants (or coerced into doing so, in some reported cases). The informants then help the police to investigate other suspects in the same case or other cases, in exchange for more lenient treatment in their own cases (such as avoiding arrest or being charged with less serious crimes than they otherwise could have been).
2. Due to incomplete or misleading information, a party may be able to take advantage of an opposing party who does not realize he is not acting in his own best interest. This, for example, is how con men operate.
3. If two parties have different risk tolerance or different time horizons (as in the purchase of insurance), then even if they have conflicting end goals, they might both be acting in accordance with their own preferences. For example, such a collaboration between two parties might allow party A to benefit in the short term and party B in the long term, with each choosing the arrangement because it fits their preferred time horizon. Or, in a situation where the most likely outcome for a given set of collaborative actions is moderately good for party A (but bad for party B) and a less likely outcome is very good for party B (but bad for party A), they may still be willing to work together to carry out the necessary collaborative actions.

As these and other examples illustrate, such collaborations are not without ethical perils, even outside the context of cybersecurity, and also have risks to their potential success. In a high-profile case related to cybersecurity, Albert Gonzalez reportedly became an informant for law enforcement in 2003 after being arrested for charges of ATM and debit card fraud. He is said to have provided information that led to the arrests of 28 people related to an identity-theft ring trafficking in 1.5 million stolen credit card and ATM card numbers. However, Gonzalez was later sentenced for 20 years in prison for continued work as a criminal hacker even while cooperating with law enforcement. He was charged with running an identity theft ring involving more 130 million card numbers and personal information stolen from five large companies via Internet attacks, much larger than the operation he had helped investigate [9].

Starting with technical issues, we consider how such methods might be effective in the setting of cybersecurity. As a oversimplification, consider a world in which there are two classes of parties: the good guys (defenders) and the bad guys (attackers). One strategy the good guys might employ would be to work with some of the bad guys to defeat the other bad guys. For example, the good guys could work with the kingpin bad guys to drive the smaller bad guys out of business, perhaps by driving up the market price for zero-day exploits to the point that only the kingpin bad guys can afford them. This could actually be good for both parties—the kingpin bad guys could increase their market share of

the attack market (thereby potentially increasing their profits) while the size of the overall market could be decreased (thereby potentially decreasing the overall impact of attacks). This seems more plausibly workable but less desirable than the related option of working with the smaller bad guys to drive the kingpin bad guys out of business. Even absent such collaboration, the good guys might still assess some adversaries as preferable to lose to over others. In such cases, doing things to foster their rise to the top (relative to adversaries who would be worse) might be worthwhile.

A promising approach to this seems to be modeling the utilities over time of the various participants, including some discounting of future utility over some time horizon. We suggest viewing the good guys as being able to take a longer view (i.e., caring about utility further into the future) than the bad guys. For example, good guys might care both about their utility now and when only one adversary is left or when all have been defeated, far into the future, while the bad guys care only about the present and near future.

At least in a static (but unrealistic) setting where no new adversaries arise, this could lead to exploring an approach to picking off the adversaries one-by-one in decreasing order of their time horizons. In such a setting, we suggest that the good guys should take a long view and then collaborate with adversaries who take a short view to defeat those with a long(er) view.

There are a variety of modeling issues involved with discounting future utility in general [2]. While we believe that capturing this will provide useful insight and even prescriptive guidance, there remain many questions to answer in constructing a useful yet workable formal model. Some natural assumptions and questions include:

- The modeling of time. Discrete time periods seem like a natural starting point.
- Aspects of time discounting that should be explicitly captured. We will want to consider at least preference for current consumption and uncertainty about the future. Are there others that should be considered in an initial model? What are reasonable effects of these?

A natural starting point is to assume the utilities of the good and bad guys are of the form

$$U^t(c_t, \ldots, c_T) = \sum_{k=0}^{T-t} D(k)u(c_{t+k}),$$

where $D(k)$ is an exponential discounting function and $u(c_i)$ describes the "instantaneous utility" derived from consumption c_i, as in Samuelson's discounted-utility model [6]. While this model plays a significant role in the economics literature, there are various issues with it, both theoretical and in comparison with experimental data; these have been surveyed by Frederick et al. [2]. Our perspective requires considering utilities over different time periods for different parties. We might also allow different discount rates for different parties, even if just one rate for the good guys and one rate (or a small number of rates) for the bad guys. Frederick et al. note that the use of multiple discount rates is a

natural extension to the basic discounted-utility model and that the correlation between greater discounting and greater risk or uncertainty (in the life of the discounting party) has been suggested throughout the study of intertemporal choice. While the discount rate may be difficult or impossible to prescribe, the potential difference in it between good guys and bad guys argues for enriching the basic model in this way.

As noted by others, there may also be settings in which the defenders and attackers have different but not necessarily strictly opposite security concerns. For example, after an attacker A steals defender D's data (e.g., customers' credit-card numbers), it is in A's interest that the data not be disseminated further, while it is in D's interest that the data not be disseminated except as D sees fit (for example, by purchase from D). Indeed, criminals are aware of this shared incentive, and can use it to offer D the chance to buy back D's data (but then requiring D to trust that A won't go ahead and sell the data elsewhere anyway).

A well-studied example of working with the bad guys (or trying to turn them into good guys by providing a desired pathway for their endeavors) is for the good guys to offer bounties for detected software vulnerabilities (such as put forth in [7] and later explored by others, e.g., [4, 5]). However, the cost to do so can be high, and just as in the previous case of purchasing data, there is no guarantee that the vulnerabilities won't still be sold to other bad guys in addition before they can be remediated. We note that the bad guys themselves suffer from lack of trust; researchers (e.g., [1,3]) have sought to better understand the underground markets used by cyberattackers and to use that understanding to suggest methods to disrupt those markets, including by introducing mistrust into them.

There are difficult questions of trust and incentives in all collaborations, but particularly so in collaborations between typically opposing parties. What is the role of trust? If the end goal of the good guys is to wipe out the bad guys, and the end goal of the bad guys is to disrupt the good guys, and the good and bad guys are all rational, why should either trust the other? What sort of partial trust might be reasonable? [8] Does "trust" imply trust to act irrationally?

The area of intertemporal choice has been of interest has been of interest to economists for well over a century. We have argued that a variety of time horizons should be assumed when studying collaboration between entities with opposing goals and that this might be of use, both descriptively and prescriptively, in studying security. The formal model can be enriched in a number of other ways, drawing on work in economics, to inform a richer analysis of this problem and identify beneficial approaches to collaboration that might be realistic to implement.

As noted above, there are non-trivial ethical issues involved in such collaborations that may be difficult or impossible to capture in a formal model. While models might prescribe approaches to collaboration—and we argue that such approaches should be investigated—careful consideration is needed before the adoption of any methods.

References

1. Franklin, J., Paxson, V., Perrig, A., Savage, S.: An inquiry into the nature and causes of the wealth of Internet miscreants. In: Proceedings of the 14th ACM Conference on Computer and Communications Security (2007)
2. Frederick, S., Loewenstein, G., O'Donoghue, T.: Time discounting and time preference: A critical review. Journal of Economic Literature 40(2), 351–401 (2002)
3. Holz, T., Engelberth, M., Freiling, F.: Learning more about the underground economy: A case-study of keyloggers and dropzones. In: Proceedings of the 14th European Conference on Research in Computer Security. Springer LNCS 5789 (2009)
4. Kannan, K., Telang, R.: An economic analysis of markets for software vulnerabilities. In: Proceedings of the Third Workshop of Economics and Information Security (2004)
5. Ozment, A.: Bug auctions: Vulnerability markets reconsidered. In: Proceedings of the Third Workshop of Economics and Information Security (2004)
6. Samuelson, P.A.: A note on measurement of utility. The Review of Economic Studies 4(2), 155–161 (1937)
7. Schecter, S.: Quantitatively differentiating system security. In: Proceedings of the First Workshop of Economics and Information Security (2002)
8. Syverson, P., Meadows, C., Cervesato, I.: Dolev–Yao is no better than Machiavelli. In: First Workshop on Issues in the Theory of Security WITS'00. pp. 87–92 (2000)
9. Zetter, K.: TJX hacker gets 20 years in prison. Wired Magazine (2010), www.wired.com/threatlevel/2010/03/tjx-sentencing