Shared Secrets Come Cheap

Robin Hanson

November 17, 1995

The Basic Idea

Imagine two people share a secret which would hurt them each $1000 worth if it got out. You offer to pay them each $1 to (verifiably) tell you their secret. If this is a one-shot simultaneous game, there are two pure-strategy equilibria: one where they both tell and another where neither of them tell. But since the no-tell equilibria makes them both better off, your chances aren't good.

What if, instead, you commit to paying only the first person who tells you, and to paying more the longer you have to wait? You'll pay $2000 a year from now, but you won't pay anything after that. Now they both know they will each tell a year from now, if no one has told before then. But then they each would tell you the day before, to be first, for only $1990. Following this reasoning back day by day, they would each tell you the very first day, even if you only offered them $1! Shared secrets can be bought very cheaply.

A Formal Model

There is a set I of agents, a set Ω of states, and a state-dependent subset $K_\omega \subset I$, for $\omega \in \Omega$, of agents who know (and can verifiably reveal) a certain secret. Assume agent i's utility is of the form $u_{i\omega}(r, s, t)$, where $r \in [0, \infty)$ is a reward given privately for revealing the secret, $s \in [0, \infty)$ is the date of reward payment, and $t \in [0, \infty]$ is the date on which the secret is revealed, with $t = \infty$ indicating the secret is never revealed.

Assume that the $u_{i\omega}$ are continuous in all arguments, are strictly increasing in r, have limits $u_{i\omega}(r, s, \infty) \equiv \lim_{t\to\infty} u_{i\omega}(r, s, t)$, and satisfy $u_{i\omega}(0, s, t) = u_{i\omega}(0, 0, t)$ for all s, t. Let us say that agent i's harm from the secret getting out is *bounded* by $(\hat{r}, \hat{t})$ at ω if $u_{i\omega}(\hat{r}, \hat{t}, \hat{t}) > u_{i\omega}(0, 0, t)$ for all $t \geq \hat{t}$.

Consider the game where any agent can tell at any time $t = n\delta$, for step δ and integer n. A reward $r(t)$ is paid promptly ($s = t$) to the first person who tells. Ties are broken randomly, by independently generating some random ranking over I, not encoded in ω, paying the highest ranked agent to tell.

Theorem 1 *If it is common knowledge that there exists at least two agents who know, with harm bounded by $(\hat{r}, \hat{t})$, then for any $\epsilon > 0$ there is a positive step δ and a reward schedule $r(t)$ such that in all sequential equilibria all who know tell at $t = 0$ for reward $r = \epsilon$.*

PROOF: Call J the assumed set of agents. Let $r(t)$ be increasing and continuous on $[0, \hat{t}]$, with $r(0) = \epsilon$, $r(\hat{t}) = \hat{r}$, and $r(t) = 0$ for all $t > \hat{t}$. Require $\hat{t}/\delta$ to be an integer.

If someone else has already told, then telling has no effect on one's utility. So the strategic choices are whether to tell at each valid time t, assuming no one else has yet told, and given some information set $\pi_i(\omega) \subset \Omega$. Consider the choice, by some $j \in J$, of whether or not to tell at $\hat{t}$. For each $\omega \in \pi_j$, not telling gives a mixture of $u_{j\omega}(0, 0, t)$ for various $t \geq \hat{t}$, and telling adds some chance of $u_{j\omega}(\hat{r}, \hat{t}, \hat{t})$, which we've assumed is strictly greater. Thus all $j \in J$ will tell at time $\hat{t}$ at all π_j, and so all who know, K_ω, would also tell for a chance at $\hat{r}$.

Now consider the choice of any agent $k \in K_\omega$ of whether to tell at some $t = n\delta \leq \hat{t} - \delta$, given that all who know would tell at $t + \delta$, and given a subjective probability $\lambda_{k\pi_k(\omega)} = \lambda_{i\omega}$ that someone else will tell at t. If a higher ranked agent would tell at t, then telling makes no difference. So assume this isn't the case. If agent k is the highest ranked agent who knows, the value of telling is

$$\Delta U^1_{k\omega} = u_{k\omega}(r(t), t, t) - \lambda_{k\omega} u_{k\omega}(0, 0, t) - (1 - \lambda_{k\omega}) u_{k\omega}(r(t + \delta), t + \delta, t + \delta)$$

Given $\lambda_{k\omega} > 0$, $r(t) \geq \epsilon > 0$ and all $u_{i\omega}$ continuous and increasing in r, then for small enough positive δ, we must get $\Delta U^1_{k\omega} > 0$. If $\lambda_{k\omega} = 0$, then $\Delta U^1_{k\omega} \to 0$ as $\delta \to 0$.

If agent k is not highest rank, and no higher ranking agent tells, then the value of telling is

$$\Delta U^2_{k\omega} = u_{k\omega}(r(t), t, t) - \lambda_{k\omega} u_{k\omega}(0, 0, t) - (1 - \lambda_{k\omega}) u_{k\omega}(0, 0, t + \delta)$$

For any $\lambda_{k\omega}$ and small enough positive δ, we must get $\Delta u^2_{k\omega} > 0$. If $\lambda_{k\omega} = 0$, so no one else tells, this scenario must have positive probability, giving $\Delta U^2_{k\omega}$ a positive weight in the total value of telling $\Delta U_{k\omega}$.

Thus for any $\lambda_{k\omega}$ and small enough positive δ, we have $\Delta U_{k\omega} > 0$ for all k, ω, and so all k who know tell at t, given that they all tell at $t + \delta$. Recursing, all K_ω will tell at $t = 0$ for $r(0) = \epsilon$. QED.

Discussion

This analysis depends sensitively on assuming sequential equilibrium. And experiments on centipede-like games cast doubts on subgame perfection for games like this. Is there a way to structure $r(t)$ so that similar behavior results even with a noisy game theory?

Notice we have assumed that agents can not punish each other for telling. While this might be justified by imagining that the agents are never told which other agent told, this lack of information does not imply such a lack of punishment. Players could, for example, play a game where all who ask get some reward if all but one person asks, but all get punished if they all ask. Then there could be a "guilty" equilibria where the person who told knows not to ask, and is thereby punished.