14.129
Spring, 2025
Lecture 8

Mechanism Design & Incentives
vs.
Protocols & Notions of Trust

Private Information, Incentives to Report and Take Actions, Byzantine Generals Problem, Differences Between Economics and Computer Science Approaches to Algorithms and Trust

Robert M. Townsend

Elizabeth & James Killian Professor of Economics, MIT

Massachusetts Institute of Technology

Outline

- Information constrained allocation
  - An insurance example
  - Implementation without the planner, using the tools of computer science
- Algorithms and trust
  - Differences between economics and computer science
  - Illustrative, byzantine generals problem

Information Constrained Allocations: An Insurance Example

- Implementation without the planner
- Using the tools of computer science

Risk Sharing with Private Information on Crop Output

- Thus, imagine a pure exchange economy with one period; two agents, named 1 and 2; and a K-dimensional vector of goods as endowments.
- The endowment of agent 1, $e^1(\varepsilon)$, is seen by agent 1 alone. That is, shocks are private to agent 1
- Let agent 1's endowment be denoted parameter θ in some set Θ, realized with probability $p(\theta)$. Agent 1 is the stand-in for the villa
- Agent 2's endowment is presumed to be public, for simplicity, some constant K-dimensional vector W. Agent 2 is the stand-in for the central monastery.
- Agents 1 and 2 agree ex ante to some resource allocation rule $f(m)$ specifying a K-dimensional vector of commodity transfers from agent 1 to agent 2 given some message m sent by villa 1 to monastery 2, message m in a set of *a priori* feasible messages, M.
- Under this resource allocation scheme, villa 1 waits to see output vector θ before sending message m. Thus its decision problem is of the form, for every $\theta \in \Theta$, maximize $U^1[\theta \quad f(m)]$ by choice of $m \in M$

- Suppose there exists a unique maximizing solution to this problem, denoted $m^*(\theta)$. Then, given θ,

$$U^1\{\theta \quad f[m^*(\theta)]\} \quad U^1[\theta \quad f(m)] \tag{83}$$

for all possible messages $m \in M$.

- Evaluating the right hand side of (83) at $m^*(\widetilde{\theta})$, the maximizing message agent 1 would have sent if his endowment vector had been $\widetilde{\theta}$ even though it is θ,

$$U^1\{\theta \quad f[m^*(\theta)]\} \quad U^1\{\theta \quad f[m^*(\widetilde{\theta})]\} \tag{84}$$

- Now consider an *alternative* scheme in which villa 1 announces a value for its endowment vector directly, some arbitrary value of $\widetilde{\theta}$ in Θ, so that the message space is now the space of output possibilities Θ instead of arbitrary message space M (truth telling is not required).

- Suppose announced $\widetilde{\theta}$ effects transfers $g(\widetilde{\theta}) \equiv f[m^*(\widetilde{\theta})]$, so that the transfer function is the direct function $g(\bullet)$ rather than the composite function $f(\bullet)$.
- Transfer function $g(\bullet)$ define a new mechanism.
- By substitution of the notation of $g(\bullet)$ into (84), at any particular $\theta \in \Theta$, and for all alternative values $\widetilde{\theta} \in \Theta$,

$$U^1[\theta \quad g(\theta)] \quad U^1[\theta \quad g(\widetilde{\theta})] \tag{85}$$

- It is apparent from (85) that in the alternative mechanism, with message space Θ and transfer function $g(\bullet)$, villa 1 would "tell the truth," though, again, it is not required to do so.
- Announced values θ would coincide with actual values θ.
- Further, a parameter draw of θ thus would effect transfer $g(\theta) \equiv f[m^*(\theta)]$, so that the outcome under the original scheme would be sustained. Object $g(\theta)$ may be thought of as a parameter-contingent allocation, with θ now playing the role of the actual parameter value.

Summary

- Any particular resource allocation mechanism with abstract message space M and allocation rule $f(\bullet)$ is associated with another equivalent mechanism yielding parameter-contingent transfers $g(\theta), \theta \in \Theta$, satisfying *incentive compatibility constraints* (85)
- To find an optimal mechanism we can search over mechanisms M, $f(\bullet)$ directly,
- Or equivalently, we can search over these alternative parameter-contingent allocations $g(\theta)$ satisfying (85).

- **Program 6**: Maximize by choice of the $g(\theta)$, $\theta \in \Theta$, the objective function

$$\lambda^1 \left(\Sigma_\theta p(\theta) U^1 [\theta \quad g(\theta)] \right) + \lambda^2 \left(\Sigma_\theta p(\theta) U^2 [W + g(\theta)] \right) \tag{86}$$

$$U^1 [\theta \quad g(\theta)] \quad U^1 [\theta \quad g(\widetilde{\theta})] \text{ for all } \theta, \widetilde{\theta} \tag{87}$$

- Note that to implement any particular solution $g(\theta)$ agent 1, the villa, need only choose one among the list of parameter-contingent transfers and offer it to agent 2, the monastery.
- If there is one good so that (87) seems to imply no trade, then what is necessary are lotteries
- Suppose that the risk aversion of agent 1 varies in a special way around realized endowment values θ. That is, agent 1 may be sufficiently more risk averse around high values of θ than around low values of θ that a lottery on (otherwise desirable) lower (or negative) transfers at low values of θ could prevent agent 1 at high values of θ from claiming to have a low value.
- Formally, to accommodate all possible beneficial exchange, let lottery $\pi(\tau \mid \theta)$ denote the probability of transfer τ as a function of announcement θ, and suppose, for simplicity, a finite number of values of τ. Then Program 6 becomes ...

- **Program 7**: Maximize by choice of the lotteries $\pi(\tau|\theta)$ the objective[9] function

$$\lambda^1 \left(\Sigma_\theta p(\theta) \Sigma_\tau U^1[\theta \quad \tau] \pi(\tau \mid \theta) \right) + \lambda^2 \left(\Sigma_\theta p(\theta) \Sigma_\tau U^2[W + \tau] \pi(\tau \mid \theta) \right) \tag{93}$$

subject to incentive constraints, for every actual value θ and counterfactual $\widetilde{\theta}$

$$\Sigma_\tau U^1[\theta \quad \tau] \pi(\tau \mid \theta) \quad \Sigma_\tau U^1[\theta \quad \tau] \pi(\tau \mid \widetilde{\theta}) \tag{94}$$

- Constraints (94) can again be derived as endogenous
- Program 7 is concave so its solution can be easily characterized. In fact, Program 7 is a linear program, so that solutions can be computed numerically.
- The only binding constraint would have high-θ agents claiming on the margin to be low-θ agents, so we would not need to worry about the feasibility of low-θ agents claiming to be high anyway. With multiple goods, however, there remains the possibility that something can be accomplished with pretransfer displays.

Optimal Multiperiod Tie-ins-1

- Fixed rentals in a given year need not imply fixed rentals over time. Rents can be changed as a function of past observables, such as past rents.
- The multiperiod analysis is straightforward for the obvious benchmark economy, a two-period, pure exchange environment.
- Let θ_t denote the endowment of agent 1, privately observed by agent 1 at the beginning of date t, $t = 1, 2$.
- Let W_t denote the known endowment of agent 2.
- Let $p(\theta_1)$ and $p(\theta_2 \mid \theta_1)$ denote known probabilities of these endowments
- The program for the determination of a two-period, private information Pareto optimal arrangement is

Optimal Multiperiod Tie-ins-2

- **Program 8**: Maximize by choice of lotteries over transfers τ at date 1, $\pi_1(\tau|\theta_1)$, and lotteries over transfers τ at date 2, $\pi_2(\tau|\theta_1,\theta_2)$, the objective function

$$\begin{aligned}
&\lambda^1(\Sigma_{\theta_1} p(\theta_1)\Sigma_\tau U^1[\theta_1 \quad \tau]\pi_1(\tau \mid \theta_1) \\
&+\beta\Sigma_{\theta_1} p(\theta_1)\Sigma_{\theta_2} p(\theta_2 \mid \theta_1)\Sigma_\tau U^1[\theta_2 \quad \tau]\pi_2(\tau \mid \theta_1,\theta_2)) \\
&+\lambda^2(\Sigma_{\theta_1} p(\theta_1)\Sigma_\tau U^2[W_1+\tau]\pi_1(\tau \mid \theta_1) \\
&+\beta\Sigma_{\theta_1} p(\theta_1)\Sigma_{\theta_2} p(\theta_2 \mid \theta_1)\Sigma_\tau U^2[W_2+\tau]\pi_2(\tau \mid \theta_1,\theta_2))
\end{aligned} \tag{95}$$

- subject to incentive constraints at date 2, for every $\widetilde{\theta}_1$ announced at date 1, and for every actual θ_2 and announced $\widetilde{\theta}_2$ at date 2

Optimal Multiperiod Tie-ins-3

$$\Sigma_\tau U^1[\theta_2 \quad \tau]\pi_2(\tau \mid \widetilde{\theta}_1, \theta_2)$$
$$\Sigma_\tau U^1[\theta_2 \quad \tau]\pi_2(\tau \mid \widetilde{\theta}_1, \widetilde{\theta}_2) \tag{96}$$

and the incentive constraints at date 1, for every actual θ_1 and announced $\widetilde{\theta}_1$,

$$\Sigma_\tau U^1[\theta_1 \quad \tau]\pi_1(\tau \mid \theta_1) + \beta\Sigma_{\theta_2} p(\theta_2 \mid \theta_1)\Sigma_\tau U^1[\theta_2 \quad \tau]\pi_2(\tau \mid \theta_1, \theta_2)$$
$$\Sigma_\tau U^1[\theta_1 \quad \tau]\pi_1(\tau \mid \widetilde{\theta}_1) + \beta\Sigma_{\theta_2} p(\theta_2 \mid \theta_1)\Sigma_\tau U^1[\theta_2 \quad \tau]\pi_2(\tau \mid \widetilde{\theta}_1, \theta_2) \tag{97}$$

- Constraint (96) ensures that agent 1 will tell the truth at date $t = 2$ no matter what happened or what was announced at date 1.
- Working backward from this, constraint (97) ensures that agent 1 will tell the truth at date $t = 1$.
- Again, "Revelation Principle" arguments ensure that these constraints can be imposed without loss of generality in the search for private information efficient arrangements.

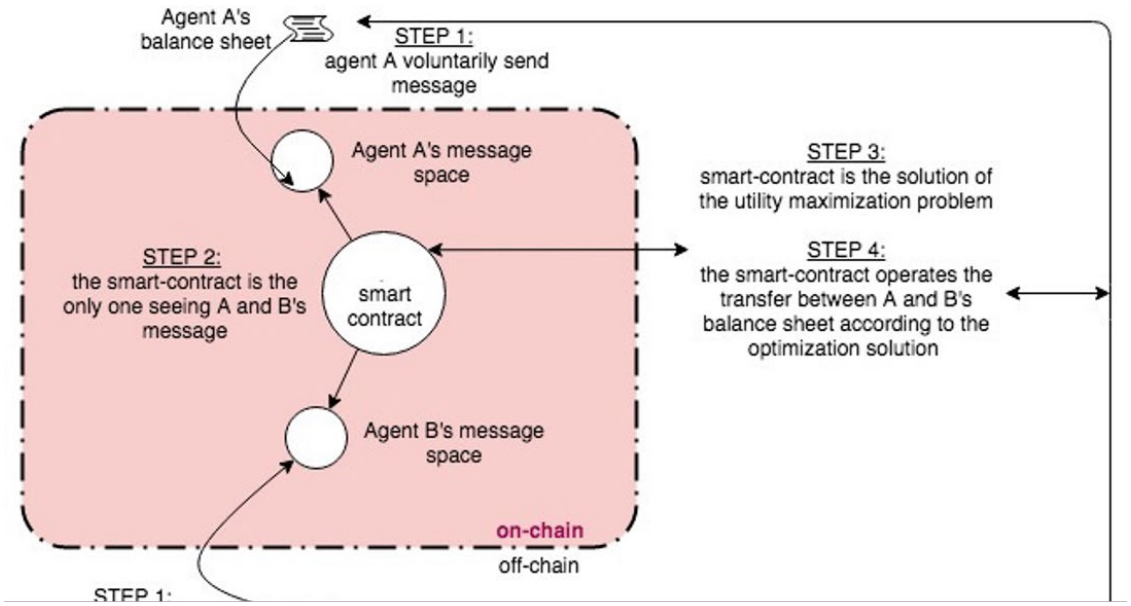
- What about solutions to Program 8? Without the incentive constraints, that is, in full-information efficient arrangements, current output $\theta_t + W_t$ is shared between agents 1 and 2 at date t, $t = 1, 2$, in such a way as to equate weighted marginal utilities. The contemporary endowment matters only, and there are no intertemporal tie-ins; the sharing rule at $t = 2$ is the same as $t = 1$.
- But in a private information efficient arrangement this may not be possible, especially if there is only one good, as the earlier analysis suggested
- Little can be done to improve risk sharing at the last date, $t = 2$
- A partial remedy at date $t = 1$, in the absence of multiple commodities, is to tie transfers at date 2 to current date $t = 1$ announcements.
- Hence, the notation $\pi_2(\tau \mid \theta_1, \theta_2)$ for transfers at date 2 as a function of current and past announcements θ_1, θ_2

- Generally, then, the past matters, an interpretation of the second conjecture of the historians that arrangements appeared to be "inefficient," bound by tradition. The past should matter. Indeed, one can find examples in which the *ex ante* optimal arrangement is not time consistent.
- Would want the parties to tear up the original date 1 agreement and start over.
- Another view of the same arrangement: the monastery could have been acting as a variable taxing center. That is, high yields and hence high payments at date $t = 1$ would have implied low payments at date $t = 2$ and vice versa.
- Points to consider:
  - Borrowing and lending is not optimal–why? And full insurance is not feasible–why?
  - Hence the information constrained optimum is a hybrid of credit and insurance.

Implementation Without the Planner, Using the Tools of Computer Science

- Implementation would seem to require the trusted planner, but it does not.
- Smart contract code can be validated and hashed. There is thus a commitment to carrying out the plan that is fully public and there can be no reneging through counter claims.
- Maximal payouts can be put in escrow as part of programmed dynamic contingent ledgers, as committed in advance, through tokenization and code.
- Messages about losses can be encrypted. The code can operate on ciphertexts to then implement the constrained optimal outcome, accesses escrow accounts.
- Past encrypted messages can be recorded as part of the data of the blockchain.

A Single Exchange and Contracting Platform: Layers 1 & 2



Protocols & Notions of Trust

- Alternative validation algorithms
- Validation algorithms for ledgers not necessarily compatible with incentives
  - The Byzantine Generals problem

Alternative Protocols for Transfers on Ledgers

Many are Faster and Less Costly than Bitcoin

Alternative Protocols can Replace a Bank's Telcom's, or Centralized Ledger of Account Balances

- Validators in a network must use a protocol that achieve consensus.
- Bitcoin use Proof of Work (PoW) the equipment and electricity used to solve the crypto puzzle. Anyone can do this mining; its open membership, and miners can join and/or leave.
- Byzantine Fault Toleration (BFT) is a property of an algorithm, guaranteed to converge or capable of reaching consensus, even if there are adversarial nodes or if nodes drop from the network.
  - The reason for such fails could be technological or nefarious, a node is honest or dishonest
- Practical Byzantine Fault Toleration (PBFT) chooses a leader in round-robin fashion from an agreed-upon "membership list" of nodes, originally picked by the company that designed the protocol
  - PBFT requires "3f+1" replicas to be able to tolerate "f" failing nodes.

Cont. Alternative Protocols

- Proof of Stake (PoS): the selected validators that suggest the next block for approval are chosen at random, followed by multi-round voting mechanisms based on the number of coins held.
- PoS BFT systems are significantly faster than PoW.
- Ripple and Stellar pioneered and further developed a decentralized alternative algorithm, Federated Byzantine Agreement, FBA.
  - Each entity decides on others it trusts, a so-called quorum slice. When these slides overlap sufficiently, it is a quorum, necessary to approve transactions.
  - Stellar's FBA is free entry or open membership into validation

Validation Algorithms for Ledgers are Not Necessarily Compatible with Incentives.

As in Morris and Shin (1997), we need to model the incentives of participants to follow the prescribed protocol in the Byzantine Generals problem

The Byzantine Generals Problem

- This is a classic and seemingly simple problem of coordinating a successful attack when the enemy may be prepared, in which case an attack even if coordinated will not be successful, and only one general is informed of the state of the enemy.
- Generals can send and messages back and forth but messages are noisy. That is, may not arrive albeit with a low probability.
- Morris and Shin (1997) show that strategic maximizing behavior in an explicit information game, without commitment, is inconsistent with the prescribed protocol.
  - An alternative protocol with commitment, specifically one that prevents the second general from replying with a validating message, achieves the optimum, ironically.
- The larger point: we cannot be sure validators will blindly follow the protocol
  - We must consider strategic behavior. The Morris and Shin (1997) example is particularly damning as it is in the public, group interest to coordinate.

Incentives to Follow Protocols, Morris and Shin (1997)

Halpern and Moses (1990)

Two divisions of an army, each commanded by a general, are camped on two hilltops overlooking a valley. In the valley awaits the enemy. The commanding general of the first division has received a highly accurate intelligence report informing him of the state of readiness of the enemy. It is clear that if the enemy is unprepared and both divisions attack the enemy simultaneously at dawn, they will win the battle, while if either the enemy is prepared or only one division attacks it will be defeated. If the first division general is informed that the enemy is unprepared, he will want to coordinate a simultaneous attack. But the generals can communicate only by means of messengers and, unfortunately, it is possible that a messenger will get lost or, worse yet, be captured by the enemy.

A probabilistic version of the coordinated attack problem.

Suppose now that with probability $\delta > 0$ the enemy is prepared, while with probability $1 - \delta$ the enemy is unprepared. Recall that the first division general knows which of these two contingencies is true, while the second division general does not. assume that each messenger gets lost with independent probability $\varepsilon > 0$.
We will be focusing on the case where ε is small and in particular is smaller than δ. We will make the unrealistic assumption that there is no upper bound on the number of messages that might be successfully sent (although with probability one a message will be lost eventually).

The naïve message protocol is general 1 receives signal and sends message if enemy is unprepared. General 2 may or may not get the message. If not the process stops. However, if received, general 2 replies back with confirmation message, and so on it goes.

The environment can be described by a state space as in Table I.

Thus state (n, m) refers to a situation where the first division general has sent n messages while the second division general has sent m messages

Table I. The naive communication protocal state space.

State	Enemy's preparedness	Probability
$(0, 0)$	Prepared	δ
$(1, 0)$	Unprepared	$(1-\delta)\varepsilon$
$(1, 1)$	Unprepared	$(1-\delta)(1-\varepsilon)\varepsilon$
$(2, 1)$	Unprepared	$(1-\delta)(1-\varepsilon)^2\varepsilon$
...	...	...

Suppose that a successful attack has a payoff of 1 for the generals, while an attack that is unsuccessful (either because the enemy is prepared or only one division attacks) has a payoff of $-M$, where M is a very large number. Both generals not attacking has a payoff of 0.

- **Theorem:**
  - If the communication system is sufficiently reliable, then the optimal action protocol (next) has coordinated attack almost always occurring when the enemy is unprepared.

Morris and Shin, 1997

The optimal action protocol, given the naive communication protocol and ε sufficiently small, has the first division general attacking whenever the enemy is unprepared (even if he has not received any message confirmation from the second division general); while the second division general attacks even if he has received only one message from the first division general. Thus coordinated attack occurs with probability $(1-\delta)(1-\varepsilon)$.

⋆ To illustrate this claim, consider three action protocols. (1) If the first division general attacks whenever the enemy is unprepared and the second division general always attacks, the expected payoff is $\delta(-M)+(1-\delta)(1) = 1-(M+1)\delta$. (2) If the first division general attacks whenever the enemy is unprepared, and the second division general attacks if he has received at least one message, the expected payoff is $(1-\delta)\varepsilon(-M)+(1-\delta)(1-\varepsilon)(1) = (1-\delta)(1-(M+1)\varepsilon)$. (3) If each general attacks if he has received at least one message, the expected payoff is $(1-\delta)(1-\varepsilon)(1-(M+1)\varepsilon)$. Protocol (2) is better than protocol (3) if $\varepsilon < (1/M+1)$. Protocol (2) is better than protocol (1) if $\varepsilon < (\delta/1-\delta)(M/M+1)$. Thus protocol (2) is better than protocols (1) and (3) for all ε sufficiently small. Similar arguments show that protocol (2) is better than *all* alternative protocols.

For claim 3

$$[(1-\delta)(1-\varepsilon)(1-\varepsilon)]1 + [(1-\delta)(1-\varepsilon)\varepsilon] \times (-M)$$

Morris and Shin, 1997

The analysis of this section showed that if the objective is only to achieve "coordination with high probability" and agents are not strategic, then it is enough that the communication system is usually accurate.

A remaining difficulty is that the optimal action protocol turns out to be sensitive to strategic concerns. The optimal action protocol described is optimal as long as the generals can be relied on to choose to follow it. Perhaps their battle orders instruct them to follow that protocol and generals always follow their battle orders. But suppose that the first division general knows that the enemy is unprepared, sends a message to the second division general, but receives no confirmation. He then believes with probability $1/(2-\varepsilon)$ that his own message never arrived, and thus that the second division general will not attack.See equation below. For all ε, this probability is more than $1/2$. Perhaps he would be tempted not to commit his division to the battle in these circumstances. Anticipating this possibility, the second division general may hesitate to attack if he has not received a re-confirmation from the first division general. The unraveling argument may start all over again.

$$\frac{\varepsilon}{\varepsilon+(1-\varepsilon)\varepsilon}=\frac{\varepsilon}{2\varepsilon-\varepsilon^2}=\frac{1}{2-\varepsilon}>\frac{1}{2}$$

Morris and Shin, 1997

To understand this argument formally, we must treat the situation as an "incomplete information game" played between the two generals. It is a game because each general, in seeking to maximize his expected payoff, must take into account the action of the other general.

We must specify the generals' payoffs. Suppose that each general gets a payoff of 0 if his division does not attack. If his division participates in a successful attack, he gets a payoff of 1; if his division participates in an unsuccessful attack (either because the enemy is prepared or the other division does not attack), he gets a payoff of $-M$. Thus if the enemy is in fact prepared (i.e., the state is $(0,0)$), the payoffs can be represented by the matrix as in Table II.

Table II. Payoffs if the enemy is prepared.

	Attack	Don't attack
Attack	$-M, -M$	$-M, 0$
Don't attack	$0, -M$	$0, 0$

Table III. Payoffs if the enemy is unprepared.

	Attack	Don't attack
Attack	$1, 1$	$-M, 0$
Don't attack	$0, -M$	$0, 0$

If the enemy is unprepared (i.e., the state is anything other than (0,0)), the payoffs are as given in Table III.

- In the strategic coordinated attack problem with the naive communication protocol, both generals never attack if the communication system is sufficiently reliable.

The argument is as follows. Clearly the first division general will never attack if he knows the enemy is prepared. Now suppose $\varepsilon < \delta$ and the second division general never receives a message. Ther seond believes that with probability $\frac{\delta}{\delta+(1-\delta)\varepsilon} > 1/2$, the enemy is prepared. Whatever he believes the first division general will do if the enemy were unprepared, his optimal action must be not to attack: not attacking gives a payoff of 0, while attacking gives an expected payoff of at most $\left(\frac{1}{2}\right)(-M) + \left(\frac{1}{2}\right)(1) = -(M-1)/2$ (recall that M is very large, and in particular greater than 1).

Now the first division general knows that the second division general will never attack if he does not receive any messages (i.e., in states (0,0) and (1, 0)). Suppose that the first division general knows that the enemy is unprepared (and so sends a message) but never receives a confirmation from the second division general. Thus the first division general believes the true state is either (1,0) or (1,1). He believes that with probability

$$\frac{(1-\delta)\varepsilon}{(1-\delta)\varepsilon+(1-\delta)\varepsilon(1-\varepsilon)}=\frac{1}{2-\varepsilon}>\frac{1}{2},$$

the second division general did not receive any message (i.e., the true state is (1,0)) and so will not attack. By the same argument as before, this ensures that the first division general will not attack even if he knows the enemy is unprepared, but has received any confirmation.

Morris and Shin, 1997

Table V. An equilibrium action protocol under the simple communication protocol.

State	Enemy's preparedness	Probability	First division general's action	Second division general's action
No message	Prepared	δ	Don't attack	Don't attack
Message sent but not received	Unprepared	$(1-\delta)\varepsilon$	Attack	Don't attack
Message sent and received	Unprepared	$(1-\delta)(1-\varepsilon)$	Attack	Attack

The strong conclusion of the previous section, that coordinated attack never occurs, is not robust to the communication protocol. The generals would indeed be exceptionally foolish to attempt to coordinate their attack using the naive communication protocol. Consider the following "simple communication protocol". Suppose that if the enemy is unprepared, the first division general sends one message to the second division general informing him of this state of affairs.

- For sufficiently small ε, there exists an equilibrium of the strategic coordinated attack problem with the simple communication protocol where coordinated attack almost always occurs whenever the enemy is unprepared.

Morris and Shin, 1997