Collusion among Adversaries

Matt Malis

Texas Triangle IR Conference, UT Dallas
25 January 2025

Slides: `mattmalis.github.io/slides`

The Puzzle of Provocation

Intuitive logic of provocation:

- A wants conflict, but wants B to attack first
- A says or does something (essentially costless) to "provoke" B
- B attacks, fighting ensues

The Puzzle of Provocation

Intuitive logic of provocation:

- A wants conflict, but wants B to attack first
- A says or does something (essentially costless) to "provoke" B
- B attacks, fighting ensues

Historical examples:

- Franco-Prussian War, 1870
  - Ems Dispatch: "effect of a red rag upon the Gallic bull"

The Puzzle of Provocation

Intuitive logic of provocation:

- A wants conflict, but wants B to attack first
- A says or does something (essentially costless) to "provoke" B
- B attacks, fighting ensues

Historical examples:

- Franco-Prussian War, 1870
  - Ems Dispatch: "effect of a red rag upon the Gallic bull"
- South Korean martial law episode, 2024
  - "provoke the North's attack at the NLL [Northern Limit Line]"

The Puzzle of Provocation

Intuitive logic of provocation:

- A wants conflict, but wants B to attack first
- A says or does something (essentially costless) to "provoke" B
- B attacks, fighting ensues

Historical examples:

- Franco-Prussian War, 1870
  - Ems Dispatch: "effect of a red rag upon the Gallic bull"
- South Korean martial law episode, 2024
  - "provoke the North's attack at the NLL [Northern Limit Line]"
- US entry into WWII; Gulf of Tonkin; Iraq invasion

The Puzzle of Provocation

Puzzle:

- How can it be both in A's interest to *provoke* B...
  - ...and in B's interest to *be provoked*?

The Puzzle of Provocation

Puzzle:

- How can it be both in A's interest to *provoke* B...
  - ...and in B's interest to *be provoked*?
- How can costless communication between adversaries be informative and influential?

Overview

This paper:

- Formal model of diplomacy and domestic politics

Overview

This paper:

- Formal model of diplomacy and domestic politics
- Demonstrates how *provocation* can be:
  - in the interest of both leaders
  - achieved through cheap-talk communication (private or public)

Overview

This paper:

- Formal model of diplomacy and domestic politics
- Demonstrates how *provocation* can be:
  - in the interest of both leaders
  - achieved through cheap-talk communication (private or public)
- Intuition:
  - A and B collude against A's domestic opposition, to extract a divisible benefit

Overview

This paper:

- Formal model of diplomacy and domestic politics
- Demonstrates how *provocation* can be:
  - in the interest of both leaders
  - achieved through cheap-talk communication (private or public)
- Intuition:
  - A and B collude against A's domestic opposition, to extract a divisible benefit
  - B attacking first can be (under certain conditions):
    - advantageous to B, for security reasons
    - advantageous to A, for domestic political reasons

Overview

This paper:

- Formal model of diplomacy and domestic politics
- Demonstrates how *provocation* can be:
  - in the interest of both leaders
  - achieved through cheap-talk communication (private or public)
- Intuition:
  - A and B collude against A's domestic opposition, to extract a divisible benefit
  - B attacking first can be (under certain conditions):
    - advantageous to B, for security reasons
    - advantageous to A, for domestic political reasons
  - communication allows them to coordinate their actions to realize these benefits

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Sequence:

1. A and B types drawn by nature, observed privately

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Sequence:

1. A and B types drawn by nature, observed privately
2. B: mobilize for conflict $(z = 1)$ or not $(z = 0)$

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Sequence:

1. A and B types drawn by nature, observed privately
2. B: mobilize for conflict $(z = 1)$ or not $(z = 0)$
3. D observes: B's action z, and signal x of A's type

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Sequence:

1. A and B types drawn by nature, observed privately
2. B: mobilize for conflict $(z = 1)$ or not $(z = 0)$
3. D observes: B's action z, and signal x of A's type
   D: retain the incumbent leader $(r = 1)$ or replace her $(r = 0)$

Game setup, without communication

Three players:

- leader A
- domestic audience D (in A's country)
- leader/state B (unitary actor)

Sequence:

1. A and B types drawn by nature, observed privately
2. B: mobilize for conflict $(z = 1)$ or not $(z = 0)$
3. D observes: B's action z, and signal x of A's type
   D: retain the incumbent leader $(r = 1)$ or replace her $(r = 0)$
4. War occurs (or not), payoffs realized

Game setup, without communication

Types:

- A types: $\theta \in \{L, H\}$, prior $Pr(\theta = H) = \frac{1}{2}$
- B types: $t \in \{\ell, m, h\} = \{\text{low, moderate, high}\}$ (any prior)

Game setup, without communication

Types:

- A types: $\theta \in \{L, H\}$, prior $Pr(\theta = H) = \frac{1}{2}$
- B types: $t \in \{\ell, m, h\} = \{\text{low, moderate, high}\}$ (any prior)

Interpretation:

- low type: dovish, low-resolve, risk-averse
- high type: hawkish, high-resolve, risk-accepting
  - or: low vs. high competence in managing conflict
- higher types make war more likely, but less costly (for themselves)

Game setup, without communication

Types:

- A types: $\theta \in \{L, H\}$, prior $Pr(\theta = H) = \frac{1}{2}$
- B types: $t \in \{\ell, m, h\} = \{\text{low, moderate, high}\}$ (any prior)

Interpretation:

- low type: dovish, low-resolve, risk-averse
- high type: hawkish, high-resolve, risk-accepting
  - or: low vs. high competence in managing conflict
- higher types make war more likely, but less costly (for themselves)

D's signal:

- $x \in \{\tilde{L}, \tilde{H}\}$, $Pr(x = \tilde{H}|\theta = H) = Pr(x = \tilde{L}|\theta = L) = \tau \in (\frac{1}{2}, 1)$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$
- $U_B(war; t, z = 0) = \beta^t, \quad \beta^\ell < \beta^m < \beta^h, \ \beta^m < 0$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$
- $U_B(war; t, z = 0) = \beta^t, \quad \beta^\ell < \beta^m < \beta^h, \ \beta^m < 0$

$$W_A(\theta, t, z = 0) = p_\theta^t \alpha_\theta \qquad W_B(\theta, t, z = 0) = p_\theta^t \beta^t$$

- higher types make war more likely, but less costly (for themselves)

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$
- $U_B(war; t, z = 0) = \beta^t, \quad \beta^\ell < \beta^m < \beta^h, \ \beta^m < 0$

$$W_A(\theta, t, z = 0) = p_\theta^t \alpha_\theta \qquad W_B(\theta, t, z = 0) = p_\theta^t \beta^t$$

- higher types make war more likely, but less costly (for themselves)

With mobilization:

- $Pr(war|\theta, t, z) = p_\theta^t + z\varepsilon, \quad \varepsilon \geqslant 0$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$
- $U_B(war; t, z = 0) = \beta^t, \quad \beta^\ell < \beta^m < \beta^h, \ \beta^m < 0$

$$W_A(\theta, t, z = 0) = p_\theta^t \alpha_\theta \qquad W_B(\theta, t, z = 0) = p_\theta^t \beta^t$$

- higher types make war more likely, but less costly (for themselves)

With mobilization:

- $Pr(war|\theta, t, z) = p_\theta^t + z\varepsilon, \quad \varepsilon \geqslant 0$
- $U_A(war; \theta, z) = \alpha_\theta - z\gamma_A, \quad \gamma_A \geqslant 0$

Game setup, without communication

Conflict payoffs, with no mobilization ($z = 0$):

- $Pr(war|\theta, t, z = 0) = p_\theta^t$, increasing in θ and t
- $U_i(peace) = 0, \quad i = A, D, B$
- $U_A(war; \theta, z = 0) = U_D(war) = \alpha_\theta, \quad \alpha_L < \alpha_H < 0$
- $U_B(war; t, z = 0) = \beta^t, \quad \beta^\ell < \beta^m < \beta^h, \ \beta^m < 0$

$$W_A(\theta, t, z = 0) = p_\theta^t \alpha_\theta \qquad W_B(\theta, t, z = 0) = p_\theta^t \beta^t$$

- higher types make war more likely, but less costly (for themselves)

With mobilization:

- $Pr(war|\theta, t, z) = p_\theta^t + z\varepsilon, \quad \varepsilon \geqslant 0$
- $U_A(war; \theta, z) = \alpha_\theta - z\gamma_A, \quad \gamma_A \geqslant 0$
- $U_B(war; t, z) = \beta^t + z\gamma_B, \quad \gamma_B > 0$

Game setup, without communication

D preferences over types:

$$W_A(H, \ell, z) < W_A(L, \ell, z) \quad \text{and} \quad W_A(L, \hbar, z) < W_A(H, \hbar, z)$$

- given $t = \ell$, D prefers $\theta = L$
- given $t = \hbar$, D prefers $\theta = H$
- (preference not too strong either way given $t = m$)

Game setup, without communication

D preferences over types:

$$W_A(H, \ell, z) < W_A(L, \ell, z) \quad \text{and} \quad W_A(L, \hbar, z) < W_A(H, \hbar, z)$$

- given $t = \ell$, D prefers $\theta = L$
- given $t = \hbar$, D prefers $\theta = H$
- (preference not too strong either way given $t = m$)

Interpretation:

- not just crisis bargaining
- spiral model, with first-strike advantage
- long-term relationship trajectory, opportunities for cooperation

Game setup, without communication

Total payoffs:

Game setup, without communication

Total payoffs:

- $U_A = r(\psi + W_A(\theta, t, z))$
  - officeholding value ψ large

Game setup, without communication

Total payoffs:

- $U_A = r(\psi + W_A(\theta, t, z))$
  - officeholding value ψ large
- $U_D = W_A(\theta, t, z)$
  - θ is either original incumbent, or replacement

Game setup, without communication

Total payoffs:

- $U_A = r(\psi + W_A(\theta, t, z))$
  - officeholding value ψ large
- $U_D = W_A(\theta, t, z)$
  - θ is either original incumbent, or replacement
- $U_B = W_B(\theta, t, z) - z\kappa$
  - direct cost of mobilizing $\kappa \geqslant 0$

Equilibrium, without communication

Non-communication equilibrium

- B^h mobilizes ($z = 1$)
- B^ℓ and B^m do not mobilize ($z = 0$)
- D retains if signal of A's type matches B's action, i.e.:
  - $r = 1$ if $(x = \tilde{H}, z = 1)$ or $(x = \tilde{L}, z = 0)$
  - $r = 0$ otherwise

Equilibrium, without communication

Non-communication equilibrium

- B^h mobilizes ($z = 1$)
- B^ℓ and B^m do not mobilize ($z = 0$)
- D retains if signal of A's type matches B's action, i.e.:
  - $r = 1$ if $(x = \tilde{H}, z = 1)$ or $(x = \tilde{L}, z = 0)$
  - $r = 0$ otherwise

Intuition:

- D: if B mobilizes, must be B^h, so war is likely $\implies$ want A_H

Equilibrium, without communication

Non-communication equilibrium

- B^h mobilizes ($z = 1$)
- B^ℓ and B^m do not mobilize ($z = 0$)
- D retains if signal of A's type matches B's action, i.e.:
  - $r = 1$ if $(x = \tilde{H}, z = 1)$ or $(x = \tilde{L}, z = 0)$
  - $r = 0$ otherwise

Intuition:

- D: if B mobilizes, must be B^h, so war is likely $\implies$ want A_H
- B^h: war is likely, better to prepare

Equilibrium, without communication

Non-communication equilibrium

- B^h mobilizes ($z = 1$)
- B^ℓ and B^m do not mobilize ($z = 0$)
- D retains if signal of A's type matches B's action, i.e.:
  - $r = 1$ if $(x = \tilde{H}, z = 1)$ or $(x = \tilde{L}, z = 0)$
  - $r = 0$ otherwise

Intuition:

- D: if B mobilizes, must be B^h, so war is likely $\implies$ want A_H
- B^h: war is likely, better to prepare
- B^m and B^ℓ: mobilizing increases prob. of facing A_H

Equilibrium, without communication

"Inefficiency" of the non-communication equilibrium:

Equilibrium, without communication

"Inefficiency" of the non-communication equilibrium:

- B^m would want to mobilize if he knew A was A_H
  - doesn't know A's type, so defaults to $z = 0$

Equilibrium, without communication

"Inefficiency" of the non-communication equilibrium:

- B^m would want to mobilize if he knew A was A_H
  - doesn't know A's type, so defaults to $z = 0$
- A_H wants her audience to believe B is a threat
  - argument is undermined by B^m not mobilizing

Equilibrium, without communication

"Inefficiency" of the non-communication equilibrium:

- B^m would want to mobilize if he knew A was A_H
  - doesn't know A's type, so defaults to $z = 0$
- A_H wants her audience to believe B is a threat
  - argument is undermined by B^m not mobilizing

$\rightarrow$ mitigated by communication between A and B

Game setup, with communication

Sequence:

1. A and B types drawn by nature, observed privately
2. A: send private, costless message to B
   - conciliatory ($s = 0$) or hostile ($s = 1$)
3. B: mobilize for conflict ($z = 1$) or not ($z = 0$)
4. D observes: B's action z, and signal x of A's type
   D: retain the incumbent leader ($r = 1$) or replace her ($r = 0$)
5. War occurs (or not), payoffs realized

All other game features same as before

Game setup, with communication

Sequence:

1. A and B types drawn by nature, observed privately
2. A: send private, costless message to B
   - conciliatory ($s = 0$) or hostile ($s = 1$)
3. B: mobilize for conflict ($z = 1$) or not ($z = 0$)
4. D observes: B's action z, and signal x of A's type
   D: retain the incumbent leader ($r = 1$) or replace her ($r = 0$)
5. War occurs (or not), payoffs realized

All other game features same as before

Influential communication

Communication equilibrium

- A strategy:
  - A_H sends hostile message ($s = 1$)
  - A_L sends conciliatory message ($s = 0$)
- B strategy:
  - B^h always mobilizes ($z = 1$)
  - B^ℓ never mobilizes ($z = 0$)
  - B^m mobilizes iff receives hostile message ($z = s$)
- D strategy (same as before):
  - retain if signal of A's type matches B's action, i.e.:
    - $r = 1$ if $(x = \tilde{H}, z = 1)$ or $(x = \tilde{L}, z = 0)$
    - $r = 0$ otherwise

Influential communication

A_H incentive-compatibility:

Influential communication

A_H incentive-compatibility:

- provoking B into mobilizing:
  - strictly harmful, for conflict payoffs
  - but demonstrates to audience that B is a threat, so they need a leader like A
- beneficial for A_H if office-holding value ψ large

Influential communication

A_H incentive-compatibility:

- provoking B into mobilizing:
  - strictly harmful, for conflict payoffs
  - but demonstrates to audience that B is a threat, so they need a leader like A
- beneficial for A_H if office-holding value ψ large

B^m incentive-compatibility:

Influential communication

A_H incentive-compatibility:

- provoking B into mobilizing:
  - strictly harmful, for conflict payoffs
  - but demonstrates to audience that B is a threat, so they need a leader like A
- beneficial for A_H if office-holding value ψ large

B^m incentive-compatibility:

- if he learns that A is type A_L (given message $s = 0$):
  - war is unlikely, preparation is unnecessary
  - by not mobilizing ($z = 0$), also helps A_L survive politically

Influential communication

A_H incentive-compatibility:

- provoking B into mobilizing:
  - strictly harmful, for conflict payoffs
  - but demonstrates to audience that B is a threat, so they need a leader like A
- beneficial for A_H if office-holding value ψ large

B^m incentive-compatibility:

- if he learns that A is type A_L (given message $s = 0$):
  - war is unlikely, preparation is unnecessary
  - by not mobilizing ($z = 0$), also helps A_L survive politically
- if he learns that A is type A_H (given message $s = 1$):
  - war is likely, so better to be prepared ($z = 1$)
  - weighed against cost of helping A_H stay in office

Influential communication

A_H incentive-compatibility:

- provoking B into mobilizing:
  - strictly harmful, for conflict payoffs
  - but demonstrates to audience that B is a threat, so they need a leader like A
- beneficial for A_H if office-holding value ψ large

B^m incentive-compatibility:

- if he learns that A is type A_L (given message $s = 0$):
  - war is unlikely, preparation is unnecessary
  - by not mobilizing ($z = 0$), also helps A_L survive politically
- if he learns that A is type A_H (given message $s = 1$):
  - war is likely, so better to be prepared ($z = 1$)
  - weighed against cost of helping A_H stay in office
- if both conditions satisfied: message is influential, $z = s$

Influential communication

To summarize:

- A and B collude against A's domestic opposition, to extract a divisible benefit

Influential communication

To summarize:

- A and B collude against A's domestic opposition, to extract a divisible benefit
  - A gives B a security benefit:
    - lets B know whether he should or shouldn't mobilize

Influential communication

To summarize:

- A and B collude against A's domestic opposition, to extract a divisible benefit
  - A gives B a security benefit:
    - lets B know whether he should or shouldn't mobilize
  - B gives A a political benefit:
    - helps convince A's audience that A is the right type of leader for the moment

Influential communication

To summarize:

- A and B collude against A's domestic opposition, to extract a divisible benefit
  - A gives B a security benefit:
    - lets B know whether he should or shouldn't mobilize
  - B gives A a political benefit:
    - helps convince A's audience that A is the right type of leader for the moment
- communication allows them to coordinate their actions to realize these benefits

Extension: Public communication

What if A's message to D is public (observable by D)?

Extension: Public communication

What if A's message to D is public (observable by D)?

- Qualitatively similar equilibrium supported
  - (semi-separating messages, rather rather than fully separating)

Extension: Public communication

What if A's message to D is public (observable by D)?

- Qualitatively similar equilibrium supported
  - (semi-separating messages, rather rather than fully separating)
- A_H does not have to "deceive" her audience
  - can provoke openly and still be politically rewarded

Extension: Public communication

What if A's message to D is public (observable by D)?

- Qualitatively similar equilibrium supported
  - (semi-separating messages, rather rather than fully separating)
- A_H does not have to "deceive" her audience
  - can provoke openly and still be politically rewarded
- Why? Because audience is *prospective*
  - doesn't matter who "started" the conflict
  - B's action reveals (probable) hostile intent, even if provoked

Cases

Franco-Prussian War, 1870:

- Bismarck (A) wanted smaller German states (D) to support unification under Prussia
  - needed to demonstrate that Napoleon (B) had hostile intent

Cases

Franco-Prussian War, 1870:

- Bismarck (A) wanted smaller German states (D) to support unification under Prussia
  - needed to demonstrate that Napoleon (B) had hostile intent
- France initiated the war ($z = 1$) on its own timeline
  - advantageous for France—power shifting towards Prussia

Cases

Franco-Prussian War, 1870:

- Bismarck (A) wanted smaller German states (D) to support unification under Prussia
  - needed to demonstrate that Napoleon (B) had hostile intent
- France initiated the war ($z = 1$) on its own timeline
  - advantageous for France—power shifting towards Prussia
- (Complication: domestic politics on both sides—Napoleon also wanted to provoke Bismarck into initiating)

Cases

South Korean martial law episode, 2024:

- Yoon government (A) wanted to overcome domestic gridlock
  - tried to demonstrate to public that NK (B) was a threat

Cases

South Korean martial law episode, 2024:

- Yoon government (A) wanted to overcome domestic gridlock
  - tried to demonstrate to public that NK (B) was a threat
- "provoke the North's attack at the NLL"
  - drone flights dropping propaganda leaflets
  - shooting down trash balloons

Cases

South Korean martial law episode, 2024:

- Yoon government (A) wanted to overcome domestic gridlock
  - tried to demonstrate to public that NK (B) was a threat
- "provoke the North's attack at the NLL"
  - drone flights dropping propaganda leaflets
  - shooting down trash balloons
- unsuccessful; Kim Jong Un not actually a high type?
- (details still coming to light)

Contribution

Distinction from **audience costs**:

- message is intended to *provoke*, rather than *deter* as in AC
- message can be private or public (but must be public for AC)

Contribution

Distinction from **audience costs**:

- message is intended to *provoke*, rather than *deter* as in AC
- message can be private or public (but must be public for AC)

Distinction from **diversionary war**:

- insufficient for A to show she is aggressive, hawkish, competent at managing conflict, etc.
- must also show that the international environment (i.e. B's type) makes these attributes valuable

Contribution

Distinction from **audience costs**:

- message is intended to *provoke*, rather than *deter* as in AC
- message can be private or public (but must be public for AC)

Distinction from **diversionary war**:

- insufficient for A to show she is aggressive, hawkish, competent at managing conflict, etc.
- must also show that the international environment (i.e. B's type) makes these attributes valuable

Novel mechanism of cheap-talk diplomacy between adversaries:

- coordinating action to collude against a third party

Thank you!

- Matt Malis, Texas A&M University
- Slides: `mattmalis.github.io/slides`
- Comments welcome and appreciated: `malis@tamu.edu`