

# Linguistic Harbingers of Betrayal



*me, Vlad Niculae, Cornell University  
with Srijan Kumar, University of Maryland College Park  
Jordan Boyd-Graber, University of Colorado Boulder  
and Cristian Danescu-Niculescu-Mizil Cornell University*

Linguistic  
Harbingers  
of Betrayal  
is everywhere.

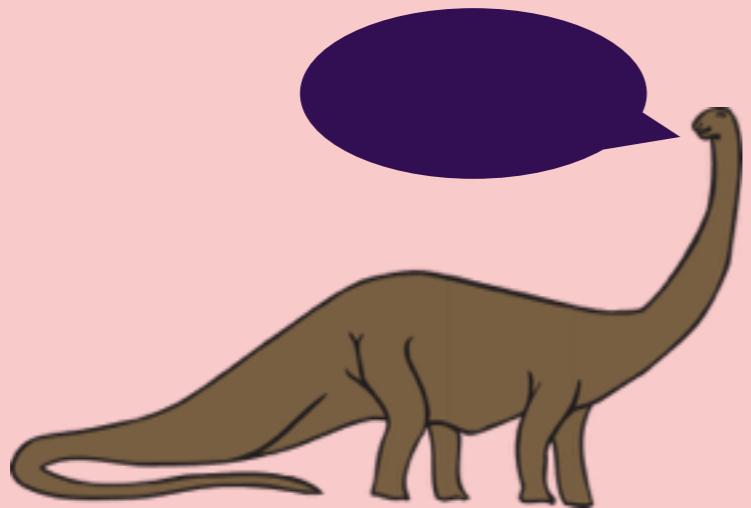




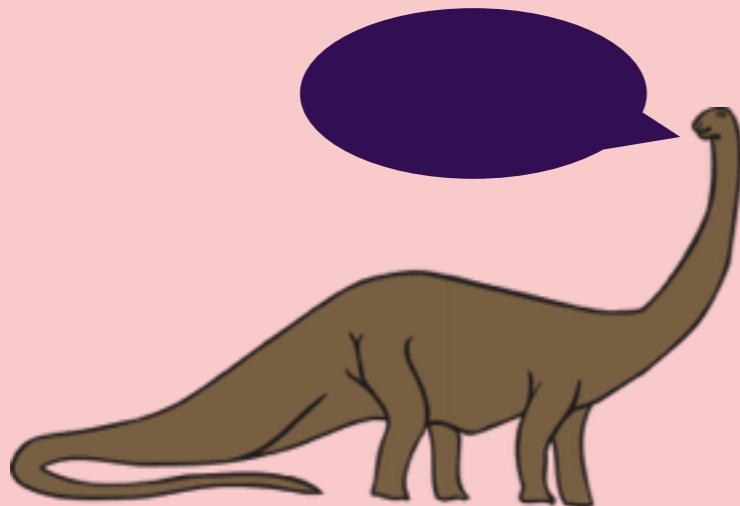
Are there any  
*linguistic* cues that  
foretell betrayal?

# **What is betrayal?**

(And how does it  
differ from just lying?)



**Can this be  
a betrayal?**



# Can this be a betrayal?



Thomas\_0608  
Hong Kong, China  
Level 6 Contributor  
153 reviews  
57 hotel reviews  
81 helpful votes

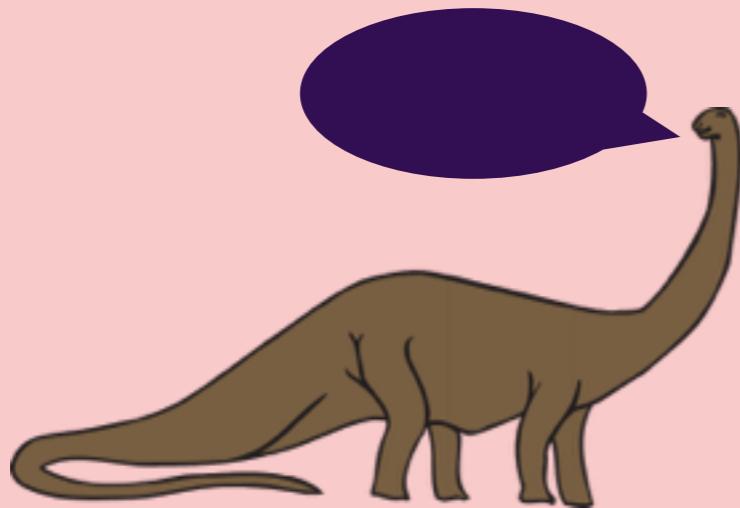
***A good congress hotel***  
 Reviewed May 4, 2015

If you're target is the convention center or the bird nest, the hotel is walking distance away from both. Besides this, there is not much to say other then: the rooms are nice & clean, everything works, FOC Wifi within its premises. Nice bar area, friendly staff. In summary, an ideal congress/convention hotel

Was this review helpful?

*Deceptive review spam*  
(Li, Huang, Yang & Zhu, 2011)  
(Ott, Choi, Cardie & Hancock, 2011)  
(Feng, Banerjee & Choi, 2012)

...



# Can this be a betrayal?

nas\_0608  
Hong Kong, China  
6 Contributor  
153 reviews  
57 hotel reviews  
81 helpful votes

**"A good congress hotel"**  
Reviewed May 4, 2015

If you're target is the convention center walking distance away from both. Beside other then: the rooms are nice & clean, within its premises. Nice bar area, friend congress/convention hotel

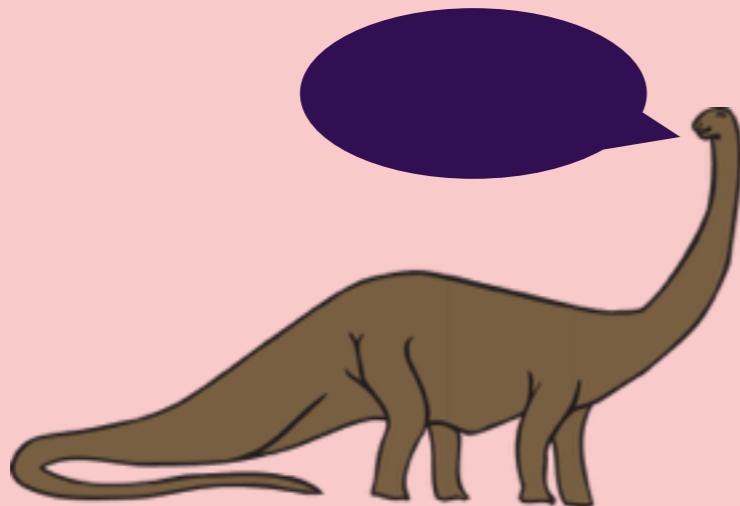
Was this review helpful? Yes



*Deceptive review spam*  
(Li, Huang, Yang & Zhu, 2011)  
(Ott, Choi, Cardie & Hancock, 2011)  
(Feng, Banerjee & Choi, 2012)

*Deception in court cases*  
(Bachenko, Fitzpatrick  
& Schonwetter, 2008)  
(Fornaciari & Poesio, 2013)

...



# Can this be a betrayal?

**"A good congress hotel"**  
5 stars Reviewed May 4, 2015  
nas\_0608 Kong, China  
6 Contributor  
153 reviews  
57 hotel reviews  
81 helpful votes

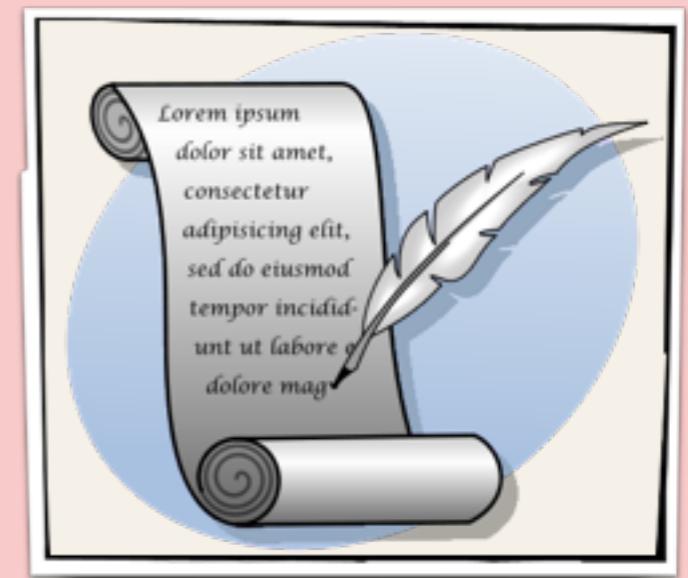
If you're target is the convention center walking distance away from both. Beside other then: the rooms are nice & clean, within its premises. Nice bar area, friend congress/convention hotel

Was this review helpful? Yes

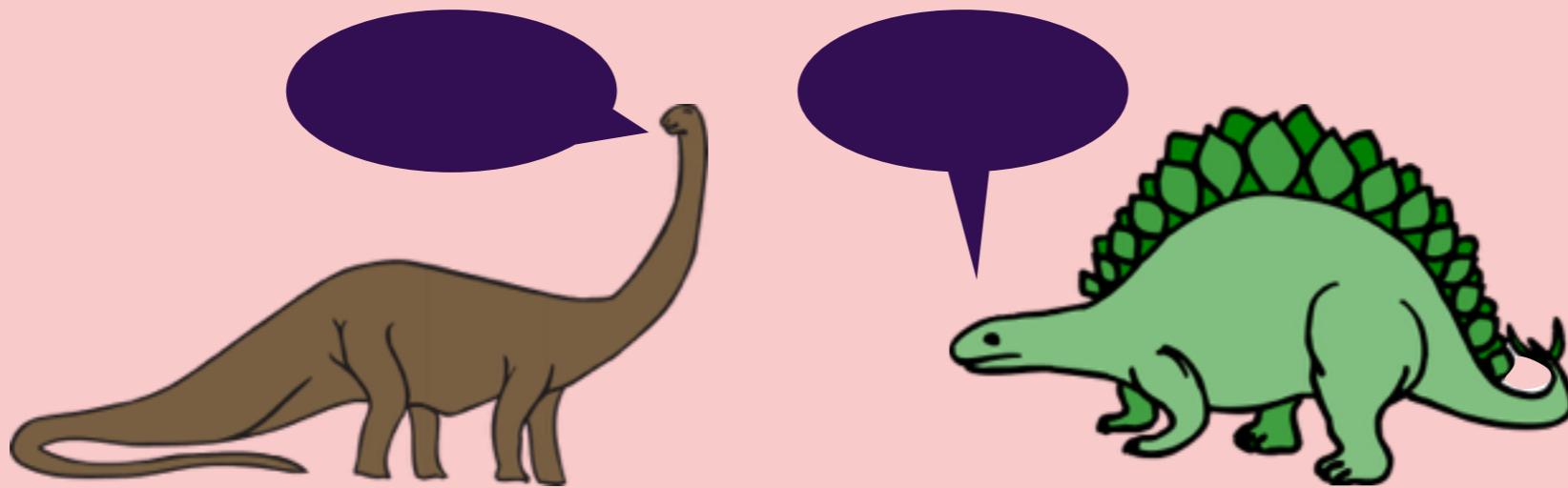
*Deceptive review spam*  
(Li, Huang, Yang & Zhu, 2011)  
(Ott, Choi, Cardie & Hancock, 2011)  
(Feng, Banerjee & Choi, 2012)  
...



*Deception in court cases*  
(Bachenko, Fitzpatrick & Schonwetter, 2008)  
(Fornaciari & Poesio, 2013)  
...



*Elicited deception in essays*  
(Newman, Pennebaker, Berry & Richards, 2003)  
(Mihalcea & Strapparava, 2009)  
(Pérez-Rosas & Mihalcea, 2014)  
...

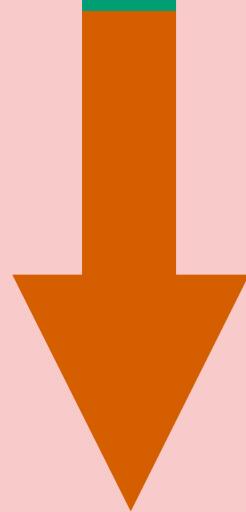
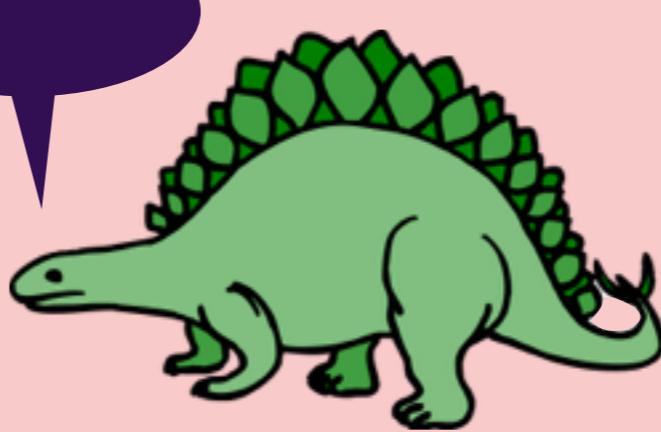
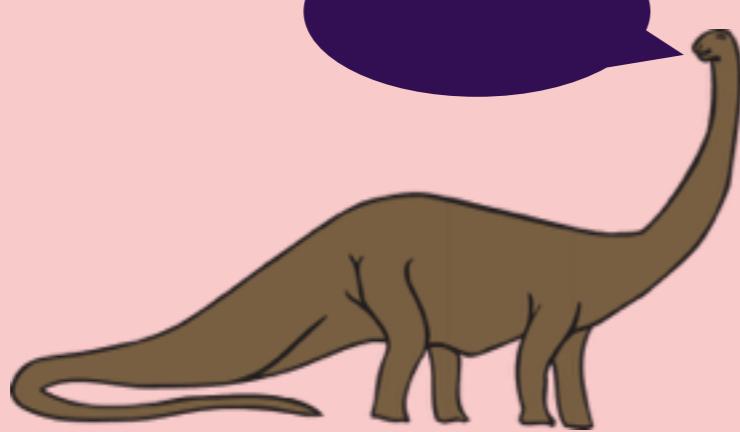
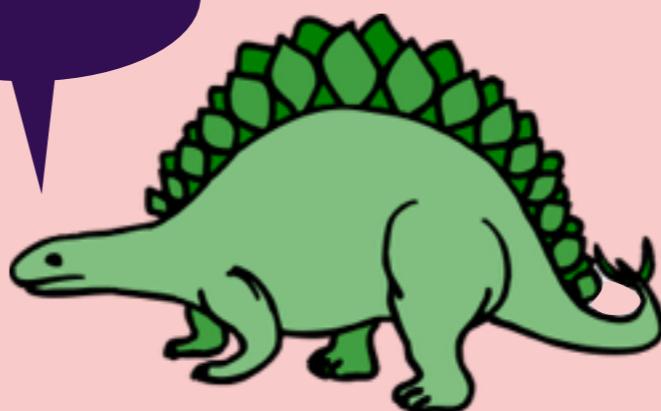
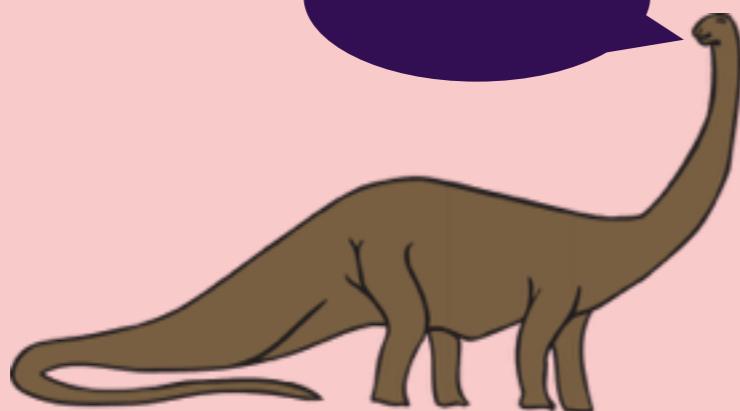
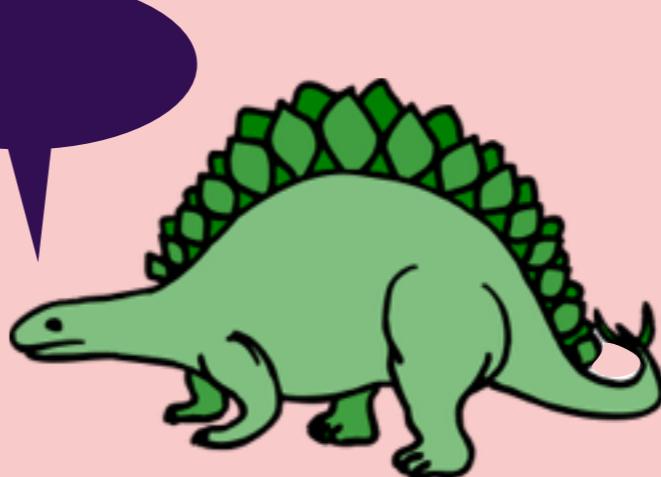
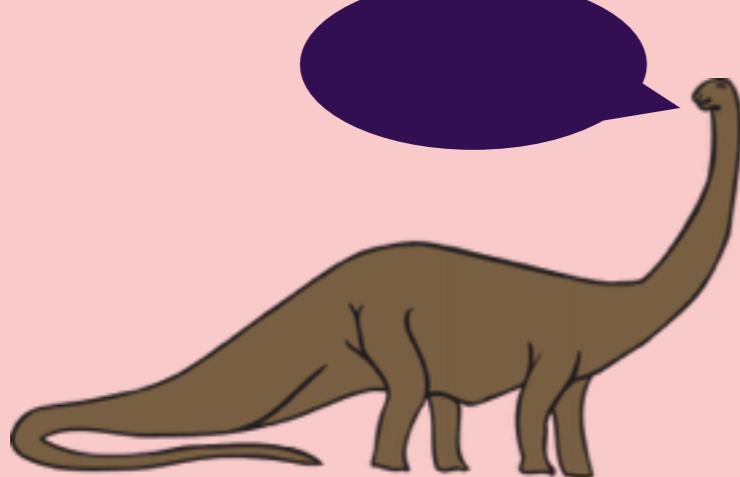
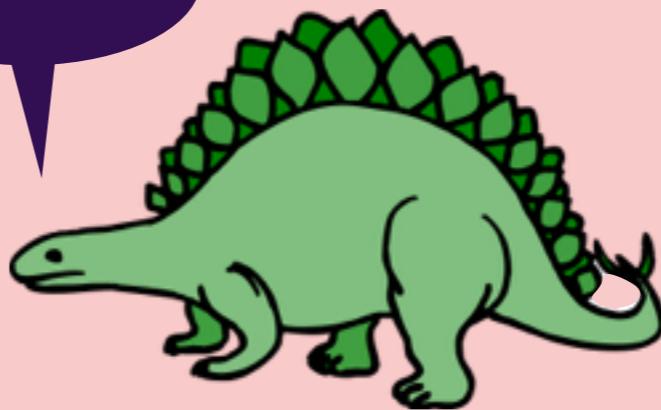
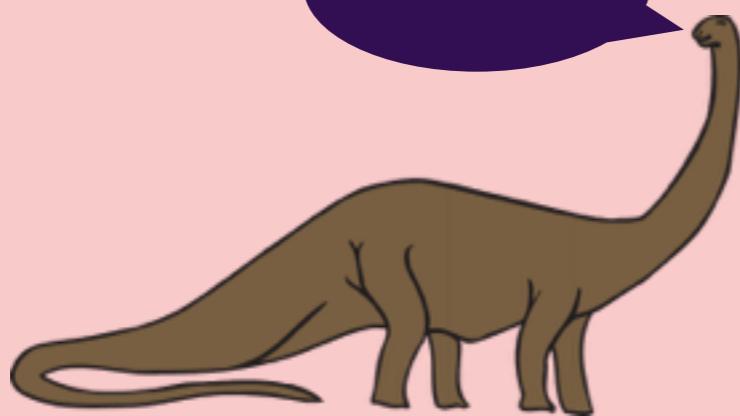


**Okay, how about this,  
can this be a betrayal?**

(Feldman and Happ, 2002)  
(Hancock, Curry, Goorha & Woodworth, 2011)

...

*t*

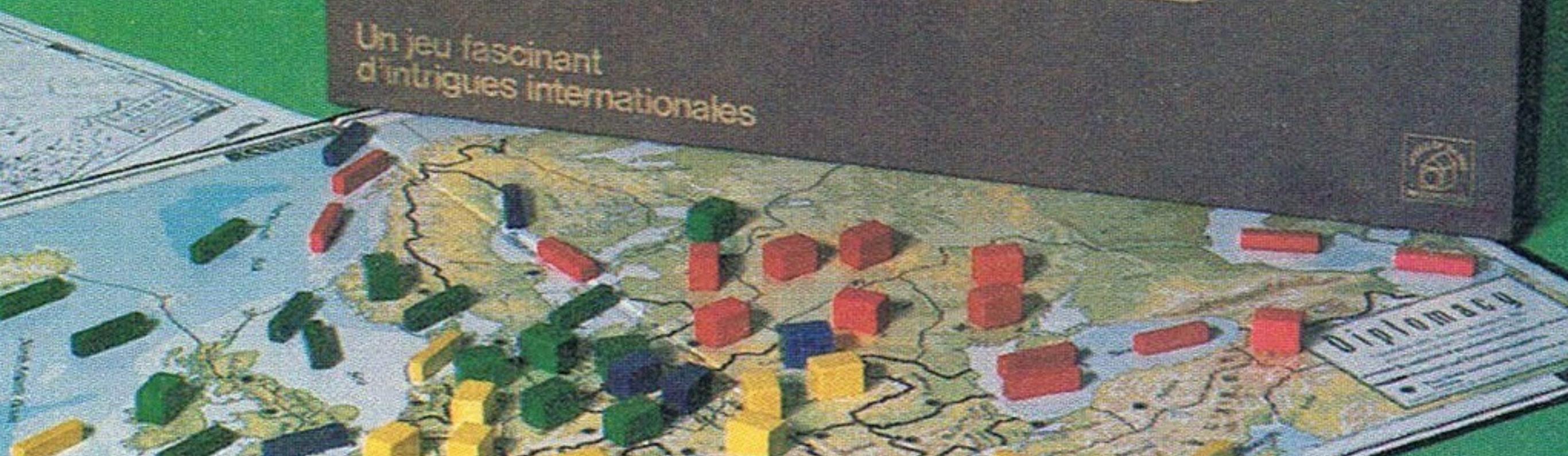


The exciting game  
of international intrigue

“The game that  
ruins friendships”

# Diplomacy

Un jeu fascinant  
d'intrigues internationales

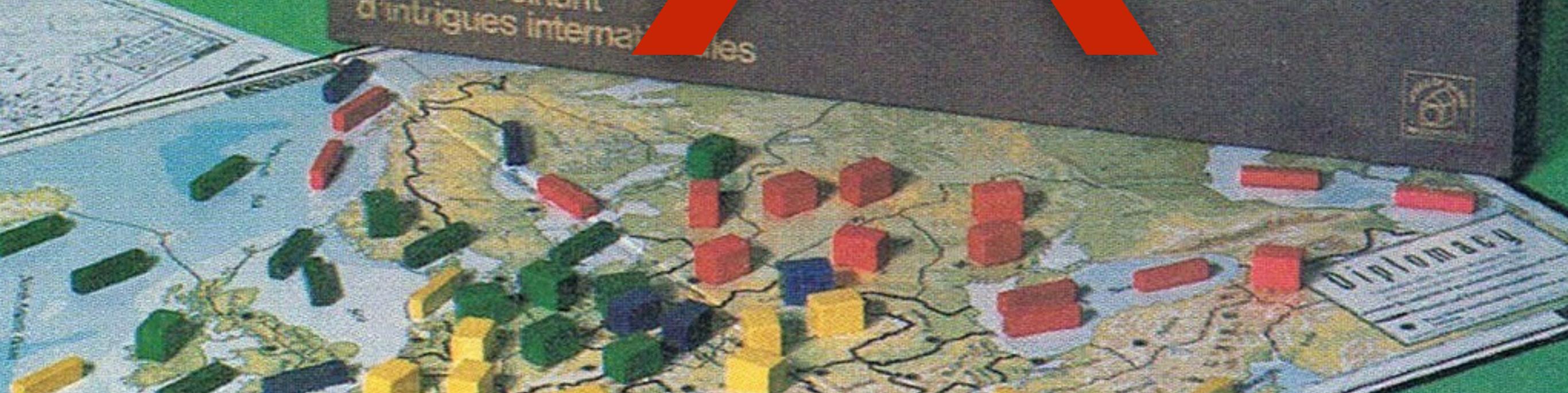


The exciting game  
of international intrigue

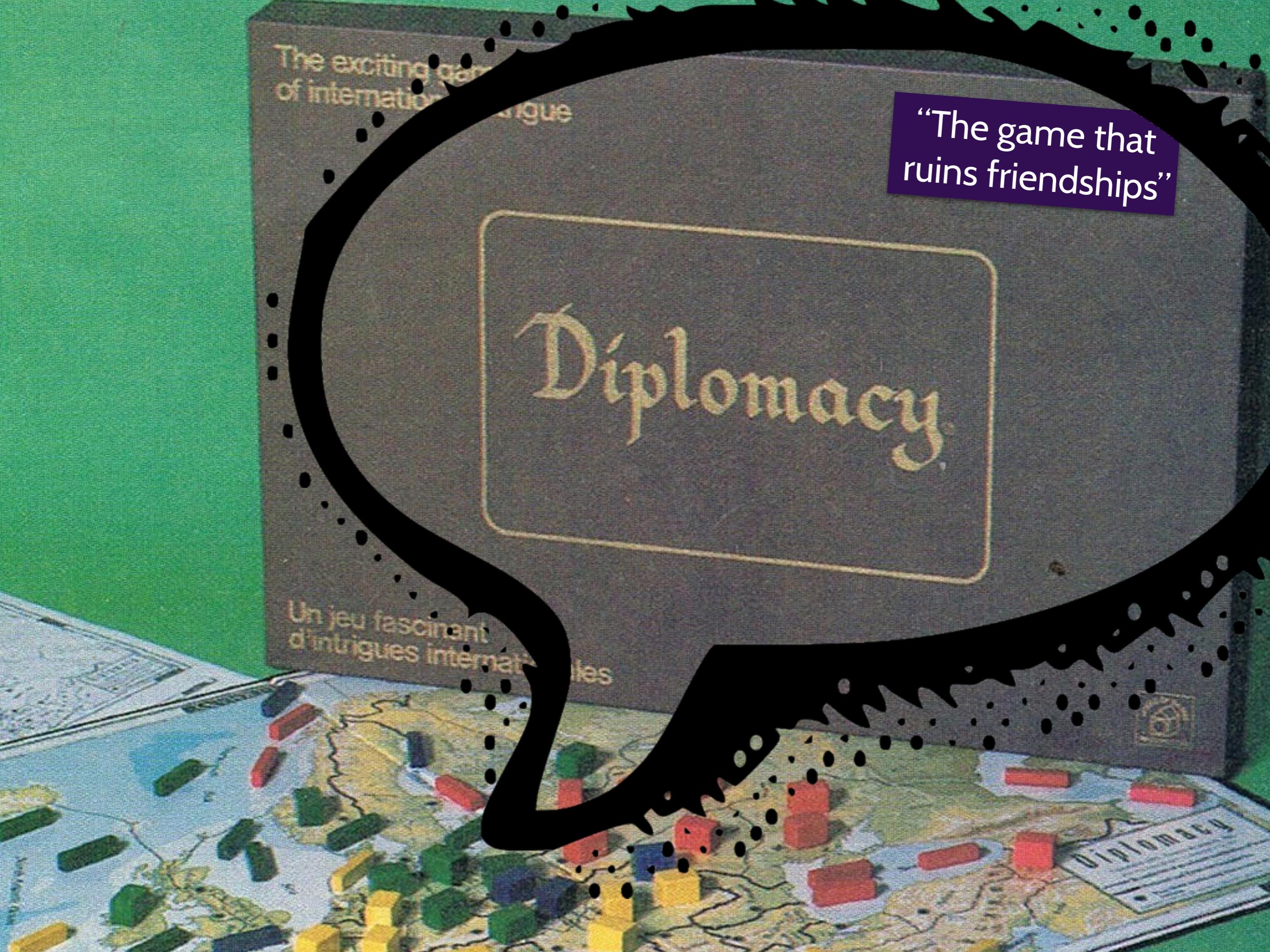
“The game that  
ruins friendships”



Un jeu fascinant  
d'intrigues internationales



# “The game that ruins friendships”



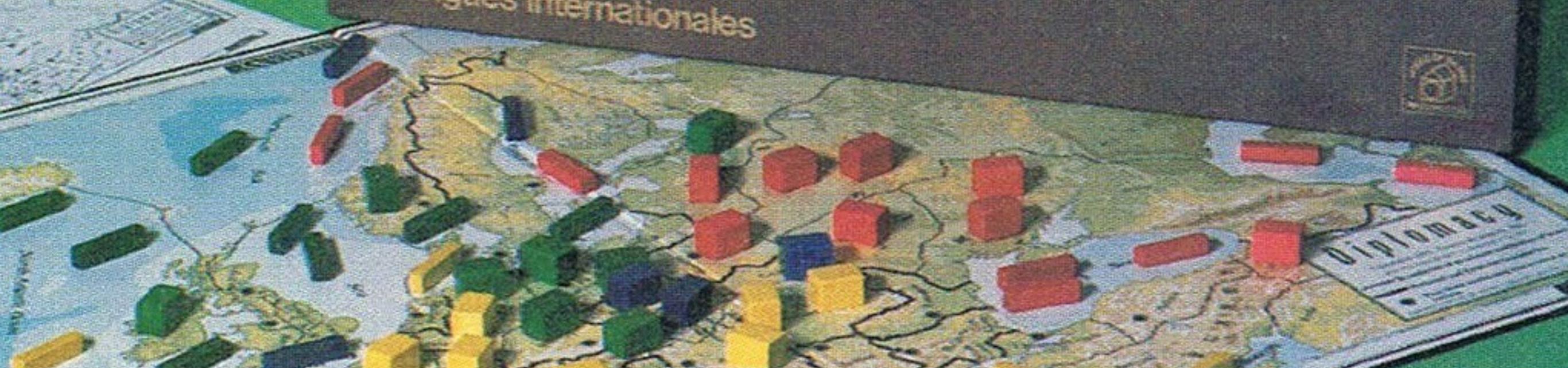
The exciting game  
of international intrigue

"The game that  
ruins friendships"

# Diplomacy

online!

Un jeu fascinant  
d'intrigues internationales



The exciting game  
of international intrigue

"The game that  
ruins friendships"

# Diplomacy

online!

249 games

~6 months/game

145k messages

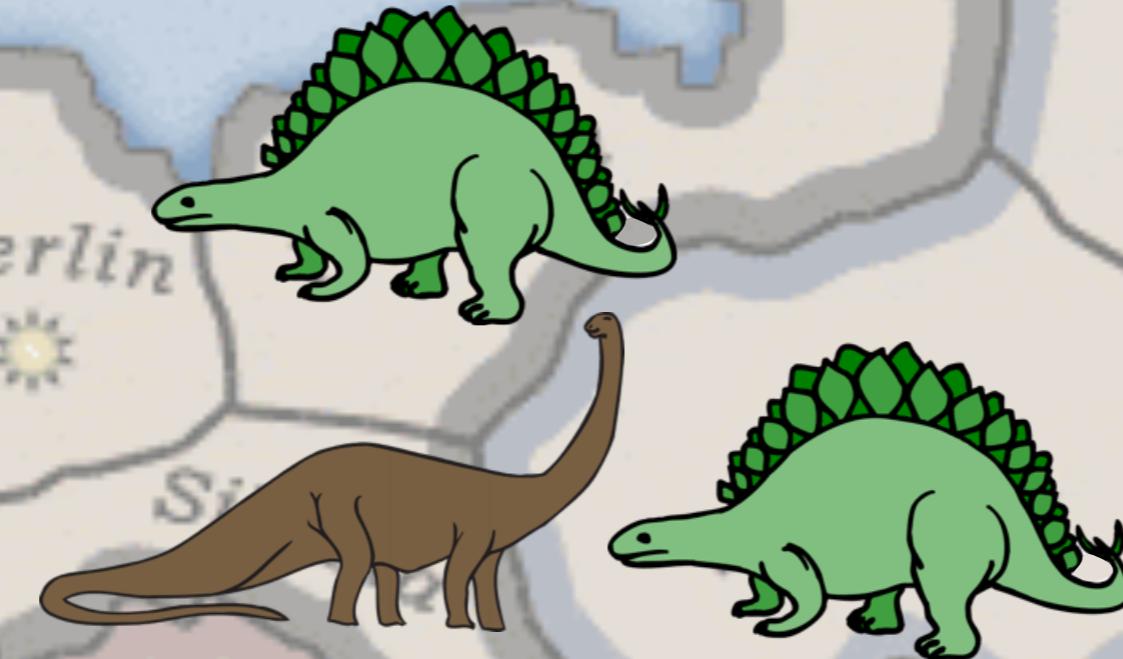
[diplom.org](http://diplom.org); [usak.asciiiking.com](http://usak.asciiiking.com)



Diplomacy  
by Allan B. Calhamer  
Copyright 1999, Avalon Hill  
Map by J. Fatula, III



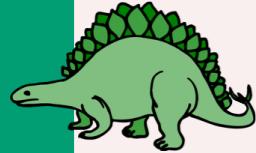








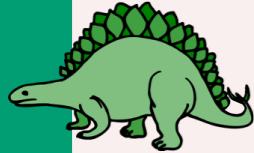
help?



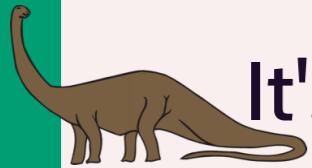
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.



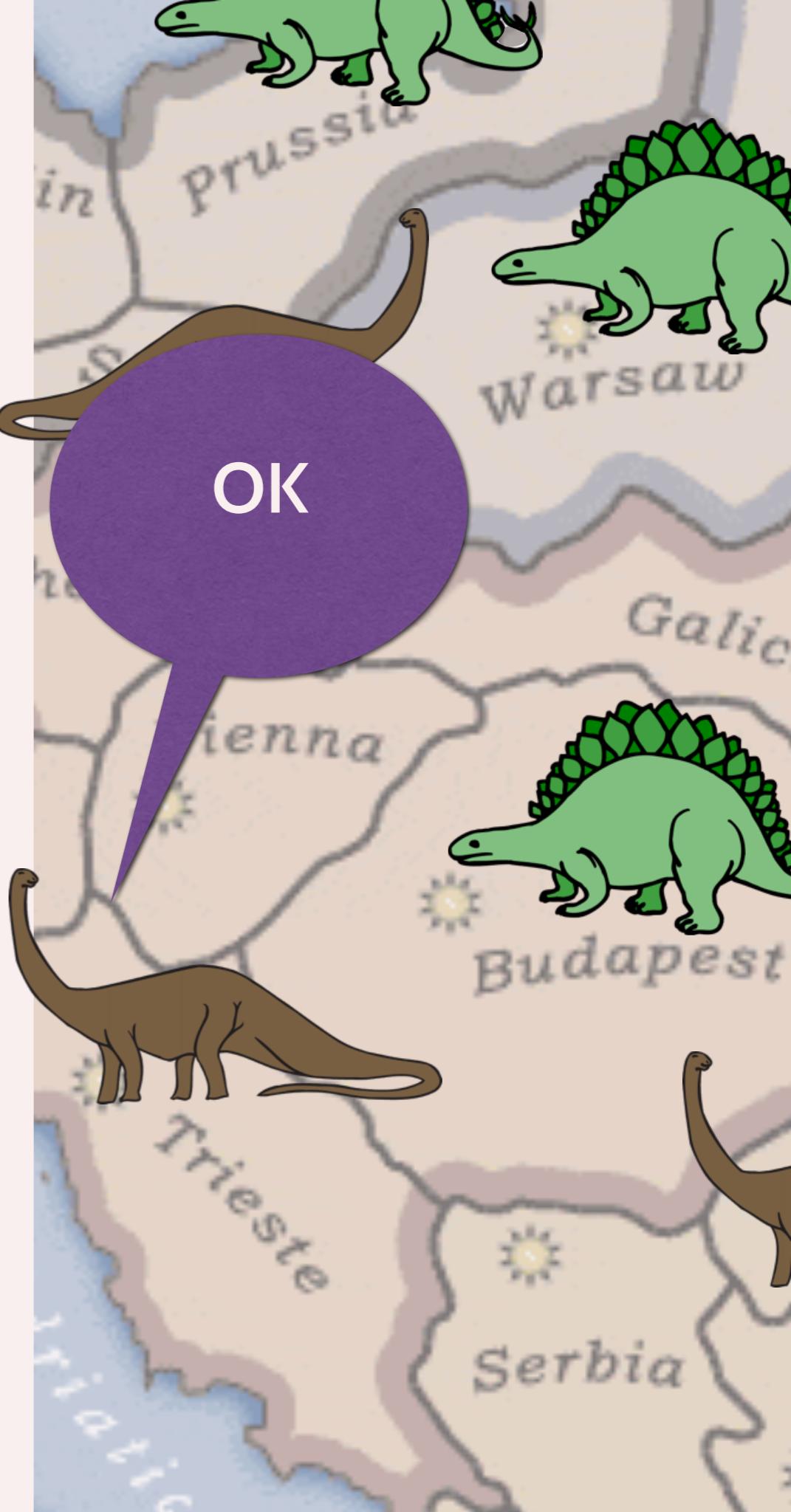
F



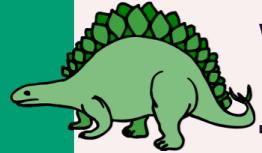
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.



It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

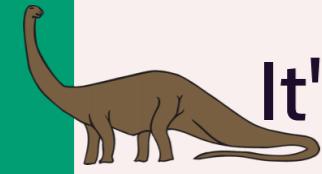


F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

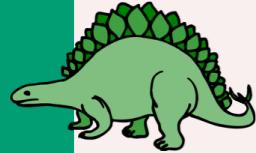
F



It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

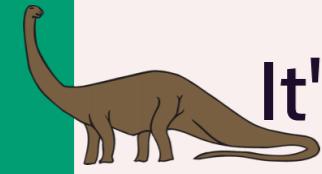


F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F



It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

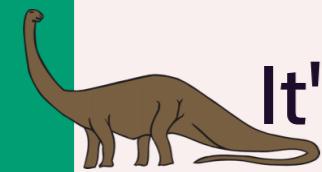


F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F



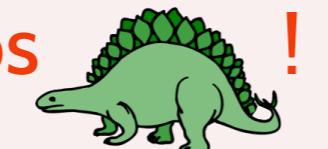
It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E

...



stabs



!

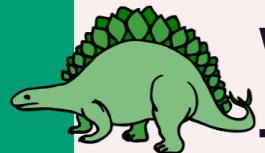


NOW STAND BACK,



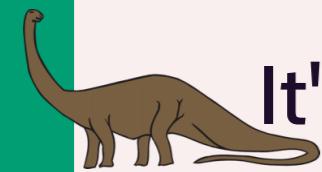
I GOTTA PRACTICE MY STABBIN'

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F



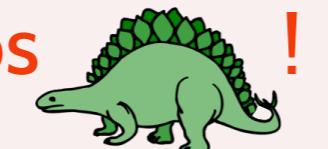
It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E

...



stabs



!



F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

E

...



stabs



!

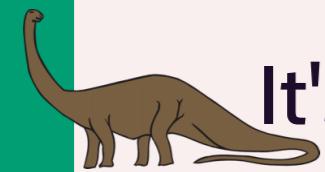


F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F



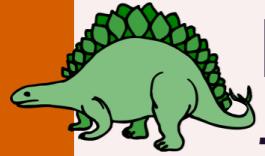
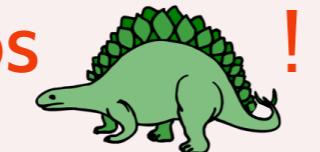
It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E

...



stabs



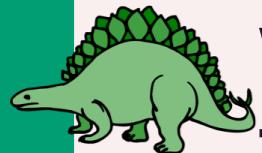
Not really sure what to say, except that I regret you did what you did.



A close-up photograph of a man's face, looking down with a shocked or surprised expression. He has short brown hair and is wearing a dark-colored shirt. In the foreground, two toy dinosaurs are positioned as if they are interacting. On the left, a blue and white Stegosaurus-like toy is held by a hand. On the right, a yellow and brown Tyrannosaurus-like toy is also held by a hand. The background is dark and out of focus.

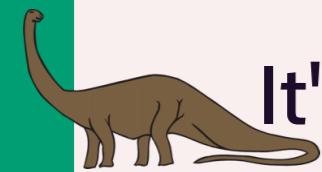
Curse your sudden  
but inevitable  
betrayal!

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there

F

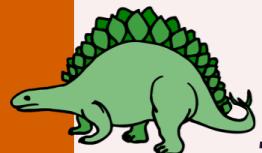


It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E



E



Not really sure what to say, except that I regret you did what you did.



# Identifying Betrayals

(So That We Can Analyze Their Language)

# Identifying Betrayals



Solid friendships:

established ( $\geq$  two supports),  
mutual ( $\geq$  one from each).

Betrayal:

friendship followed by  
 $\geq$  two attacks.

# Identifying Betrayals

250 such betrayals in our Diplomacy dataset.



# Matching Friendship

250 such betrayals in our Diplomacy dataset.

We find 250 matching friendships.

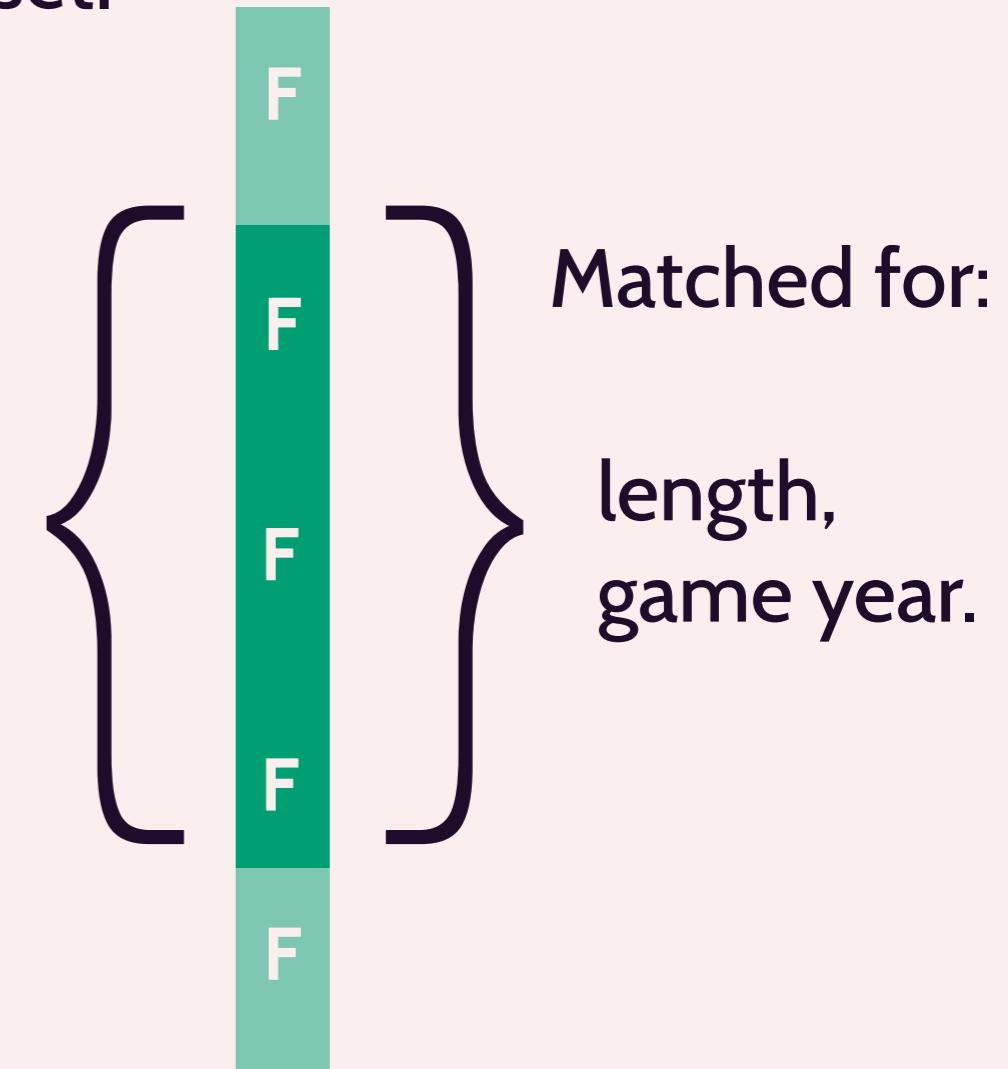
(9660 msgs., 59 words/msg.)



Solid friendships:

established ( $\geq$  two supports),  
mutual ( $\geq$  one from each).

Betrayal.



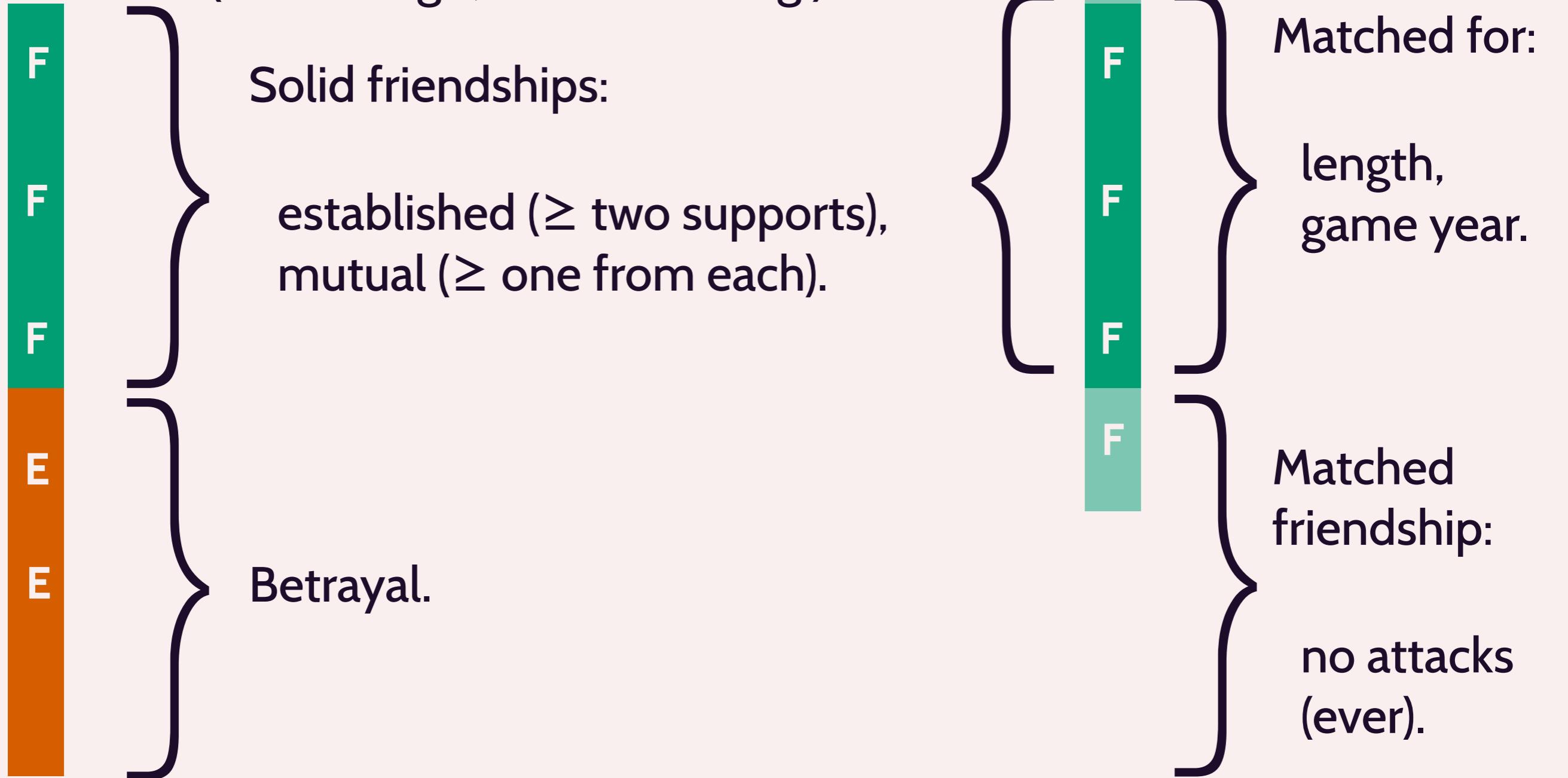
Matched for:  
length,  
game year.

# Matching Friendship

250 such betrayals in our Diplomacy dataset.

We find 250 matching friendships.

(9660 msgs., 59 words/msg.)

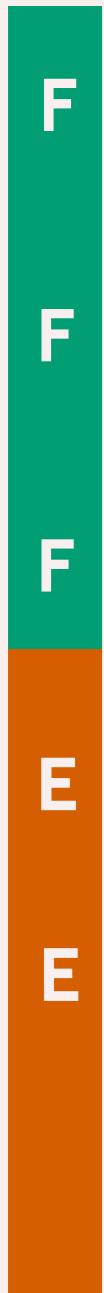


# Matching Friendship

250 such betrayals in our Diplomacy dataset.

We find 250 matching friendships.

(9660 msgs., 59 words/msg.)



# Matching Friendship

250 such betrayals in our Diplomacy dataset.

We find 250 matching friendships.

(9660 msgs., 59 words/msg.)

F

F

F

E

E

F

F

F

F

F

# Matching Friendship

250 such betrayals in our Diplomacy dataset.

We find 250 matching friendships.

(9660 msgs., 59 words/msg.)

F

Linguistic signs of betrayal  
while they act as friends?

F

The betrayers actively hide it.  
The victims didn't see it coming.

F

E

E

F

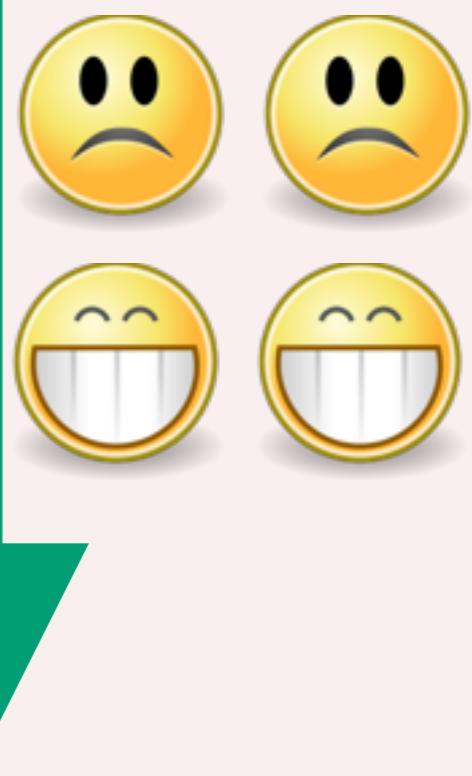
F

F

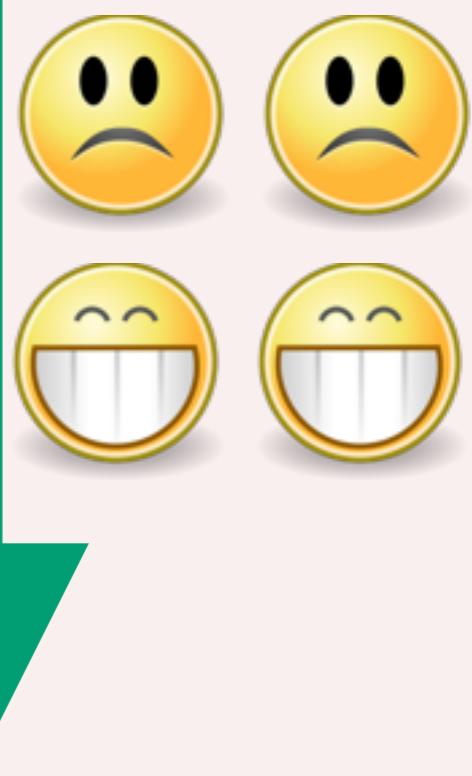
F

F

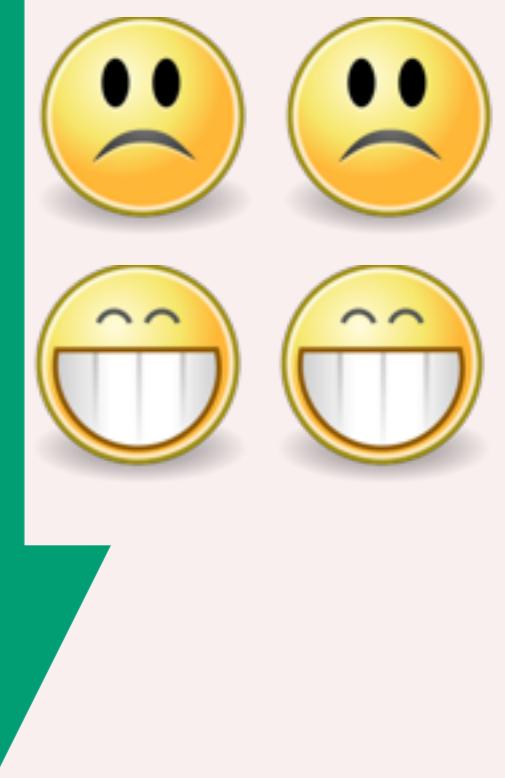
# Insight: Conversational Balance



# Insight: Conversational Balance

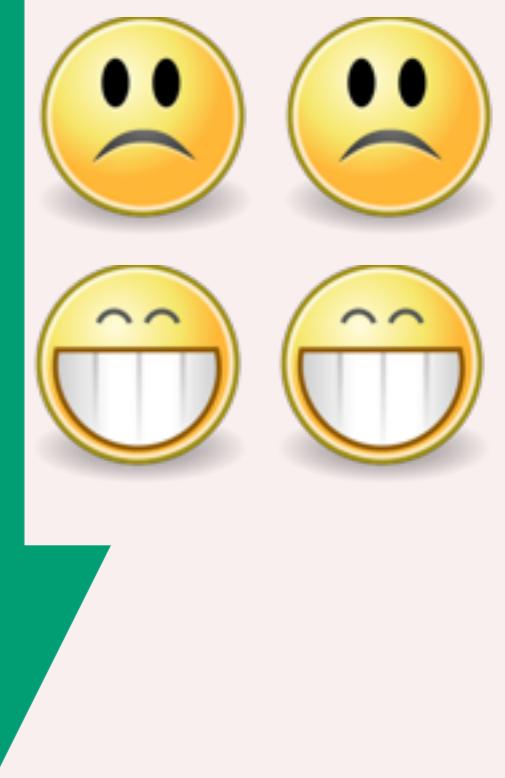


# Insight: Conversational Balance



- Stable marriages are balanced (Gottman, 1993).
- So are effective pair programming teams (Jung, Chong & Leifer, 2012).

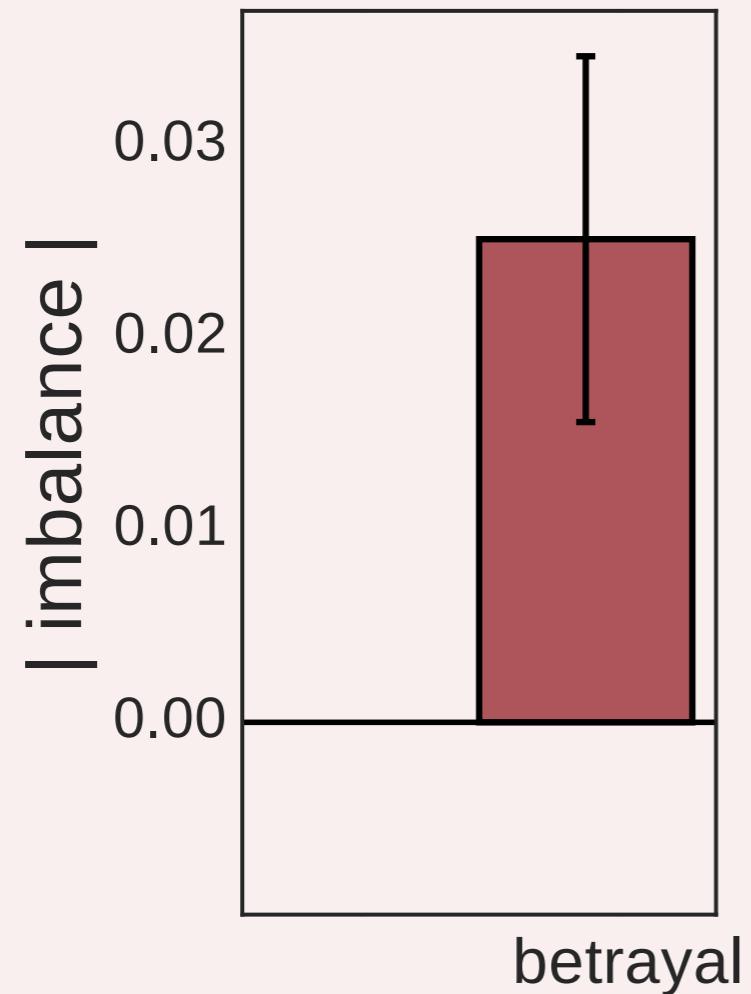
# Insight: Conversational Balance



- Stable marriages are balanced (Gottman, 1993).
- So are effective pair programming teams (Jung, Chong & Leifer, 2012).
- Can we apply this to linguistic conversational features?

# (Im)balance: Sentiment

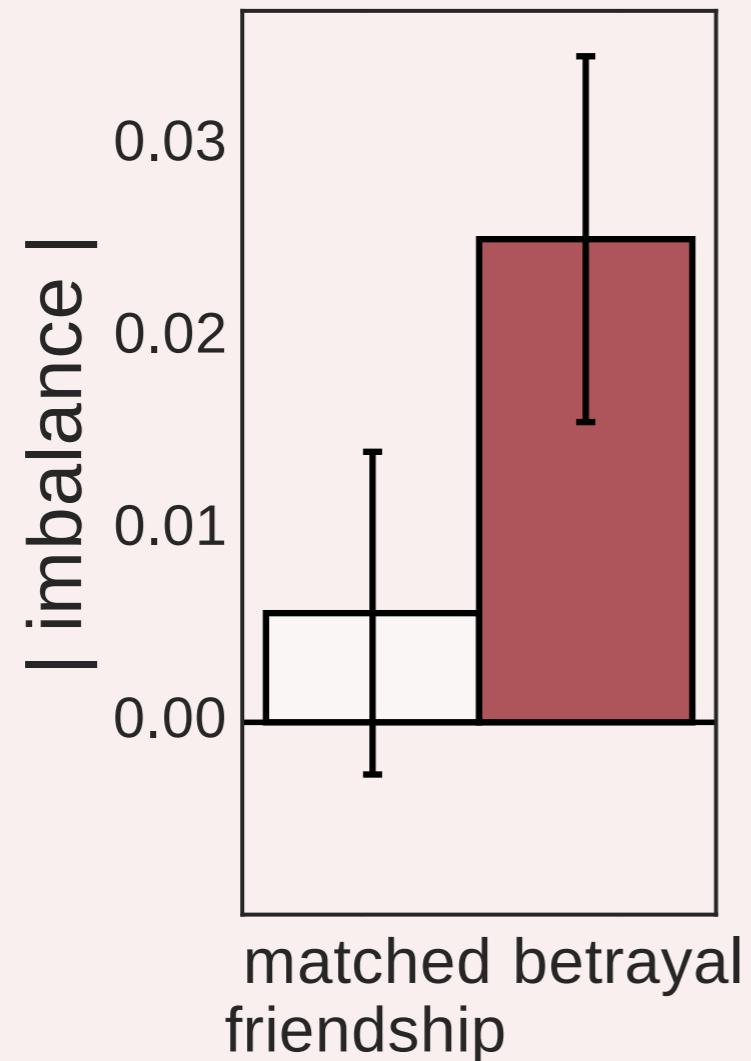
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Proportion of sentences showing positive sentiment.)  
(Error bars show standard error.)

# (Im)balance: Sentiment

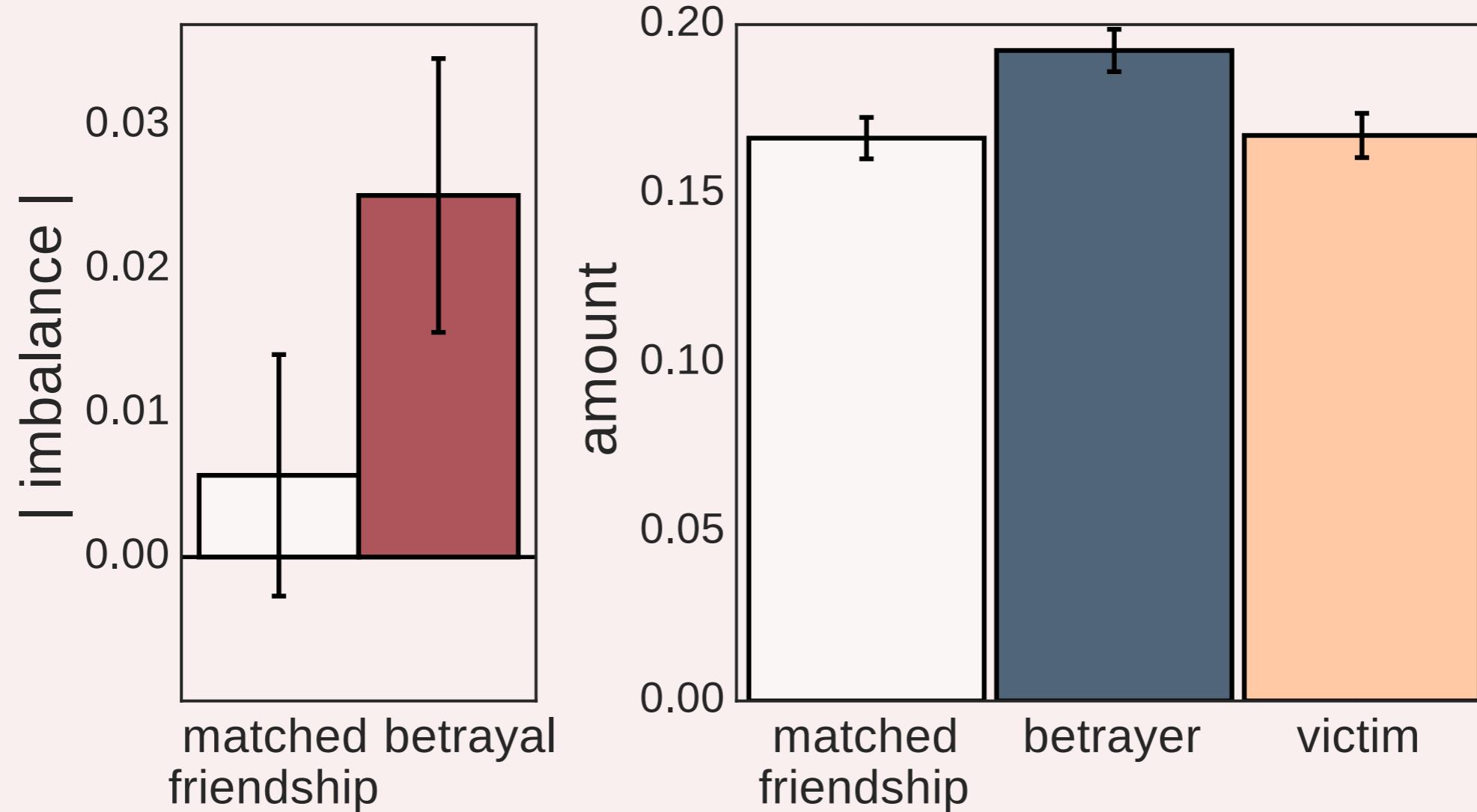
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Proportion of sentences showing positive sentiment.)  
(Error bars show standard error.)

# (Im)balance: Sentiment

Imbalance:  $f(\text{betrayer}) - f(\text{victim})$

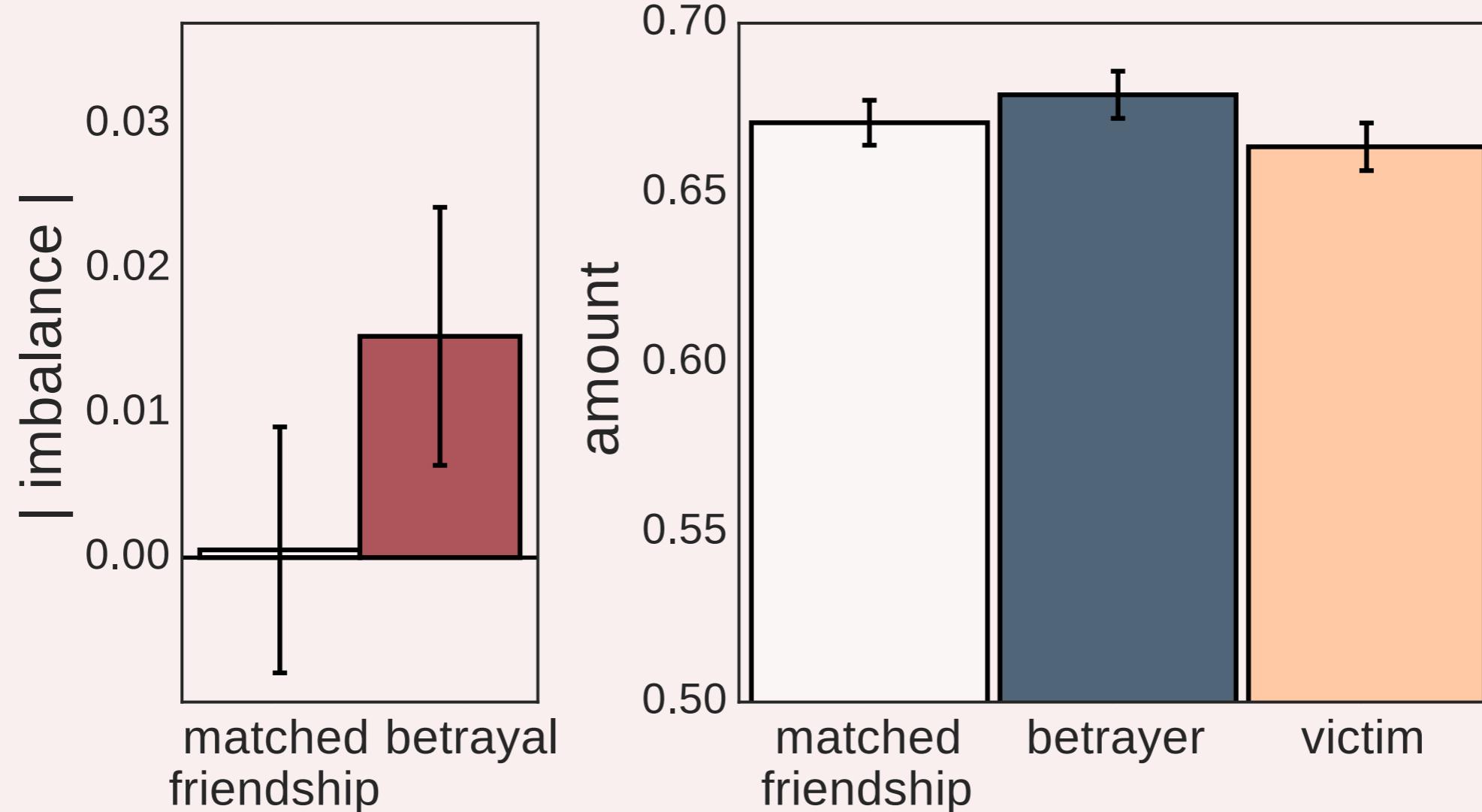


(Proportion of sentences showing positive sentiment.)

(Error bars show standard error.)

# (Im)balance: Politeness

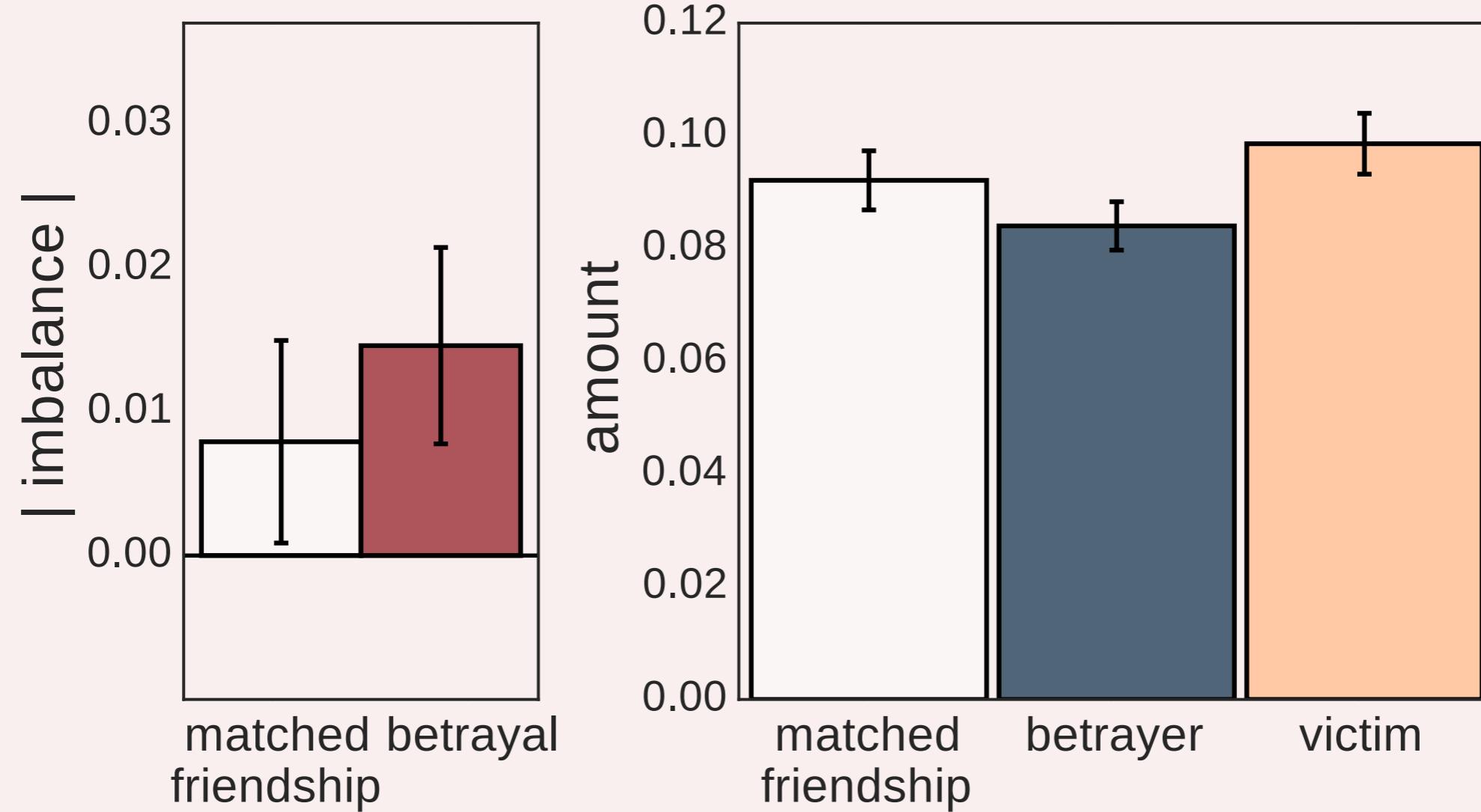
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Average 0-1 politeness score of requests: <http://politeness.mpi-sws.org>)  
(Error bars show standard error.)

# (Im)balance: Future Planning

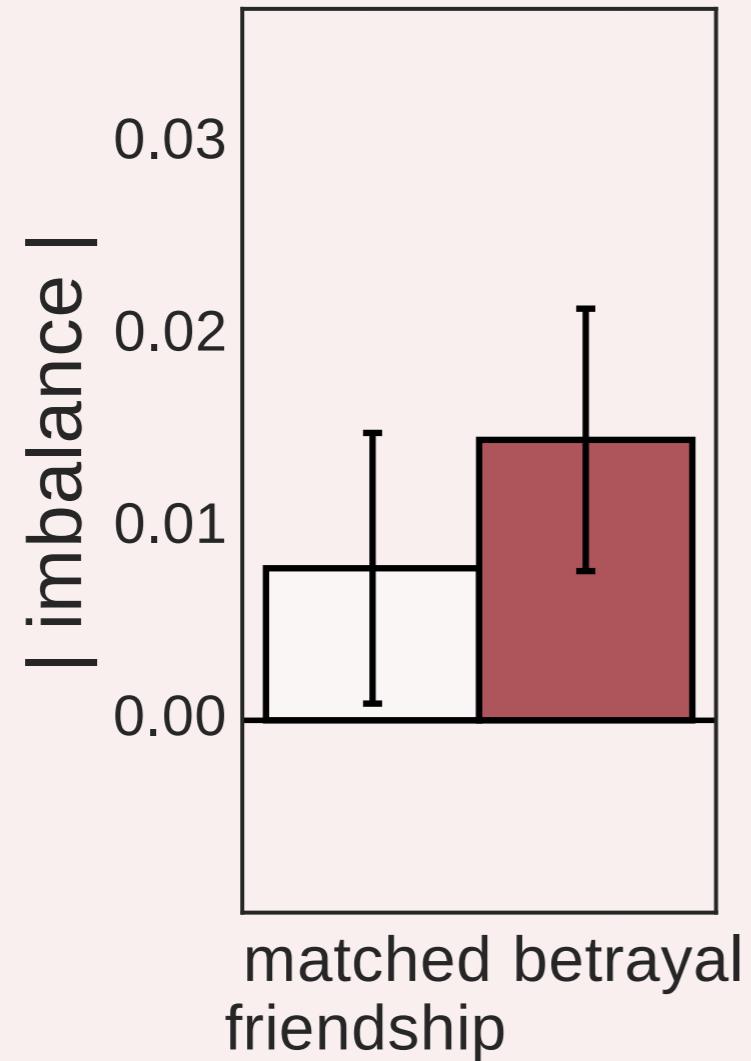
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



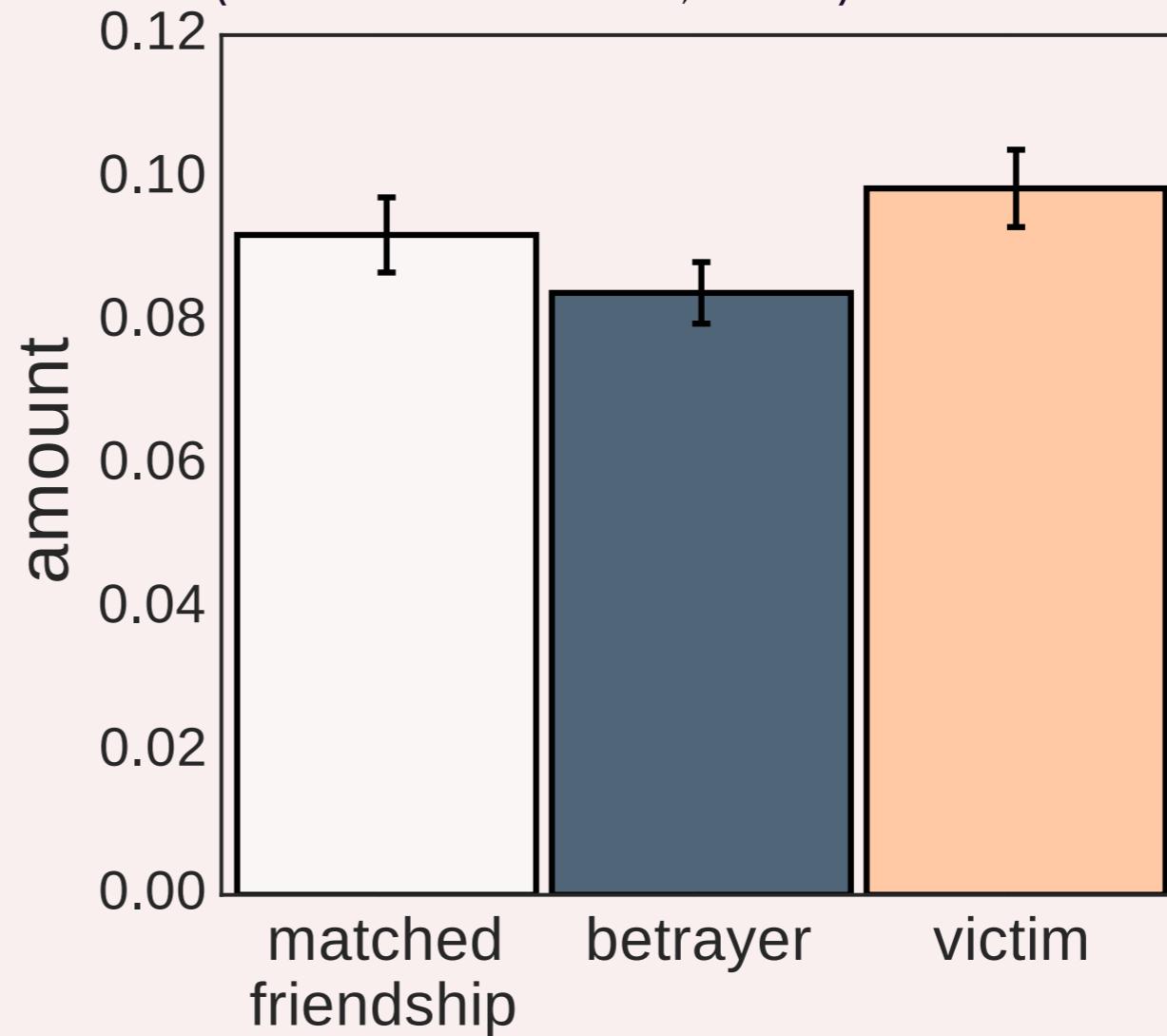
(Average number of planning connectors per message, e.g. “next”, “after”)  
(Error bars show standard error.)

# (Im)balance: Future Planning

Imbalance:  $f(\text{ betrayer}) - f(\text{ victim})$



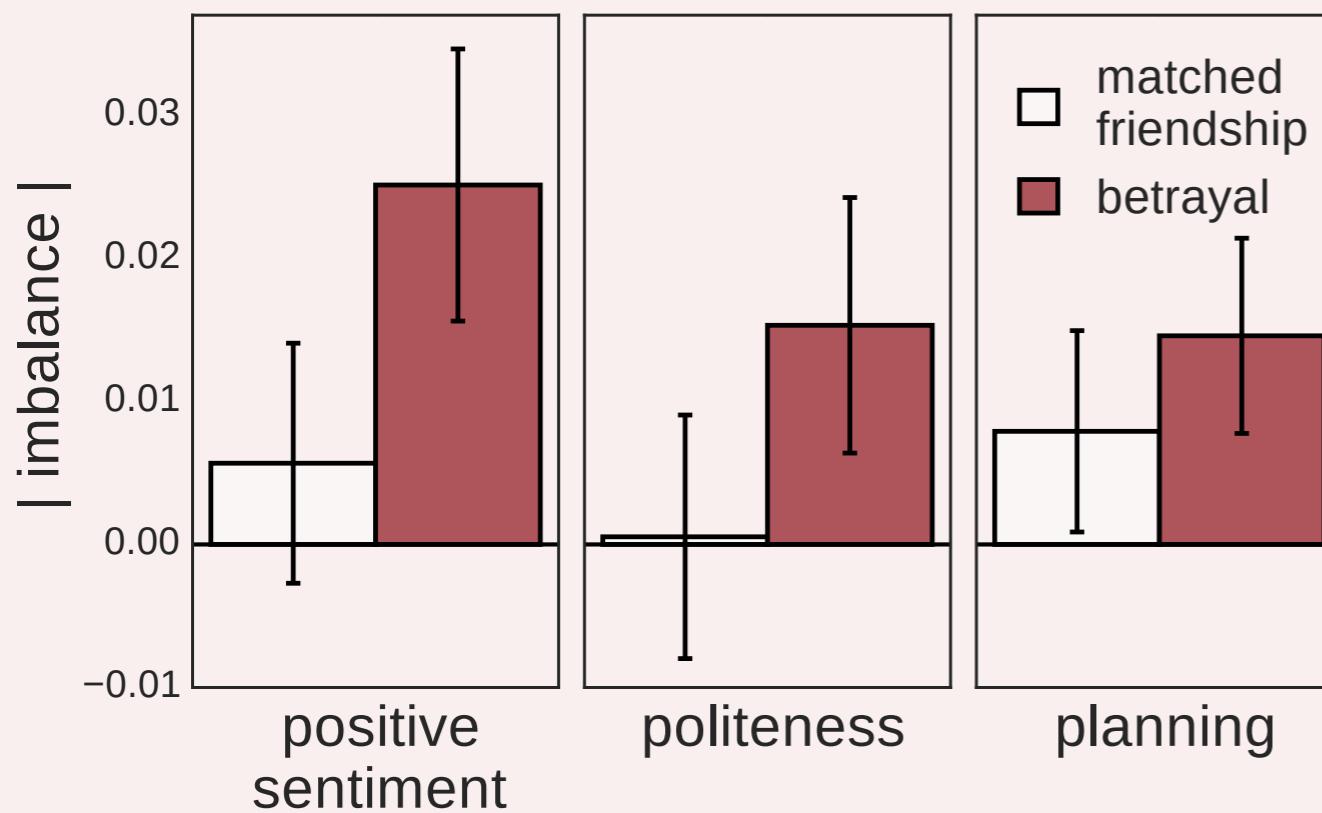
Demand-Withdraw pattern pre-divorce.  
(Gottman & Levenson, 2000)



(Average number of planning connectors per message, e.g. “next”, “after”)  
(Error bars show standard error.)

# Conversational (Im)balance

Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

Friendships that break  
exhibit imbalance  
through language cues.

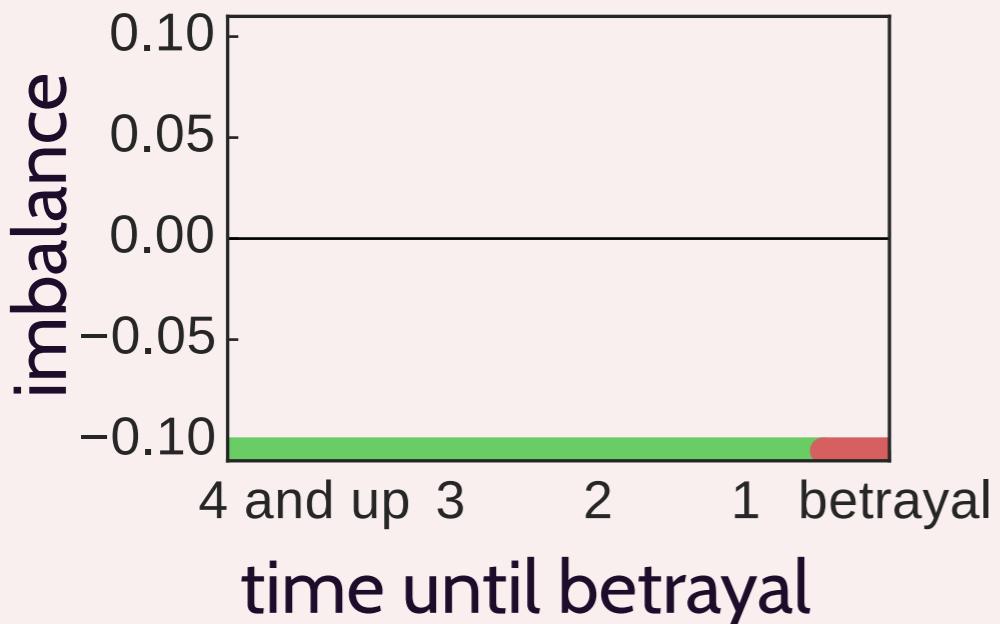
**Are backstabbing  
friendships doomed  
from the start?**

**Are backstabbing  
friendships doomed  
from the start?**

**Or do the dynamics change over time?**

# (Im)balance Over Time

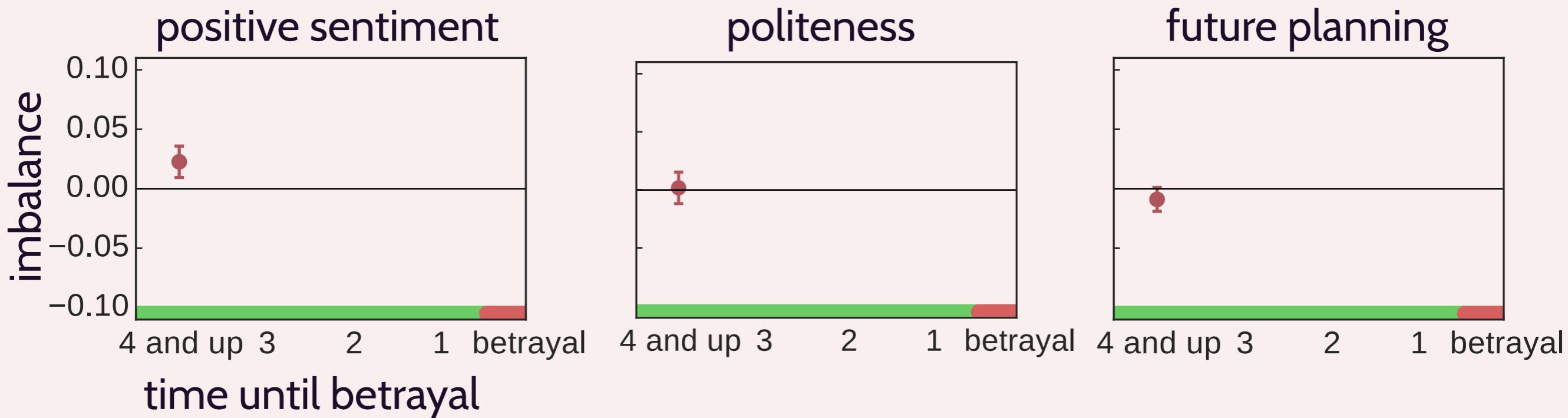
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$ . Looking only at betrayals.



(Error bars show standard error.)

# (Im)balance Over Time

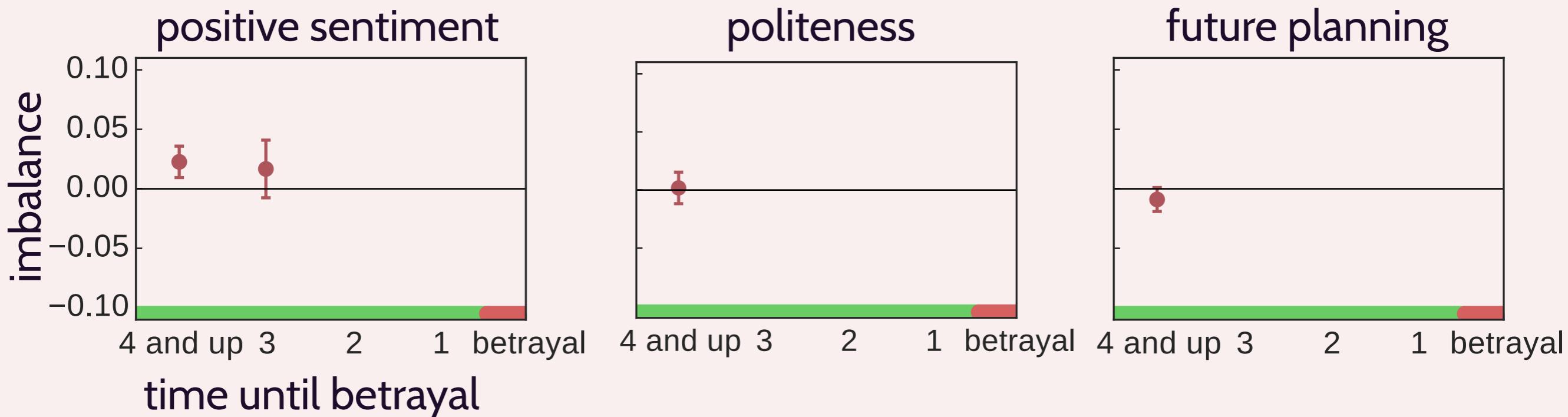
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

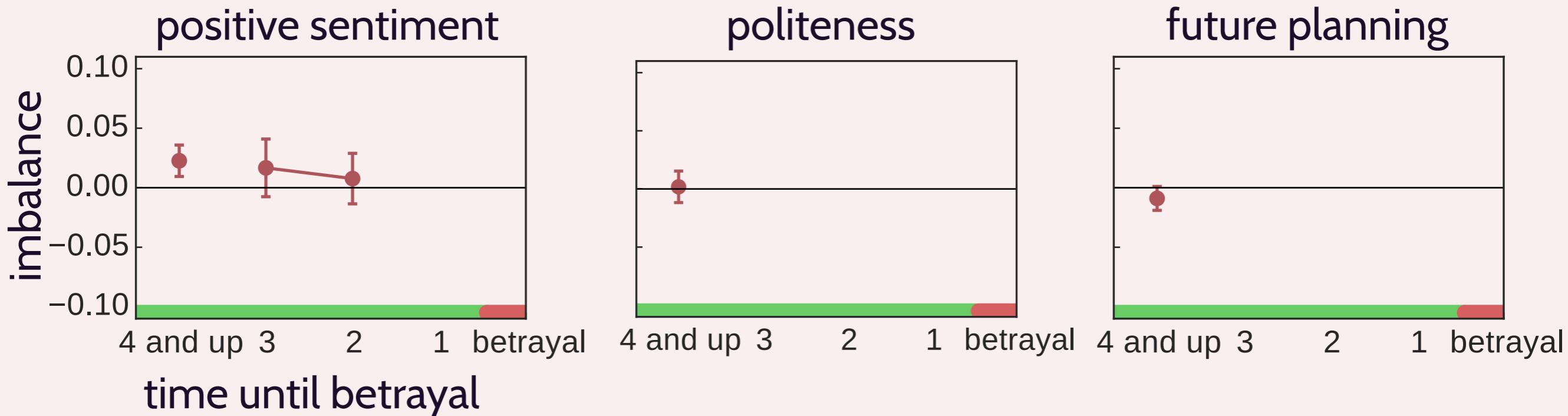
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

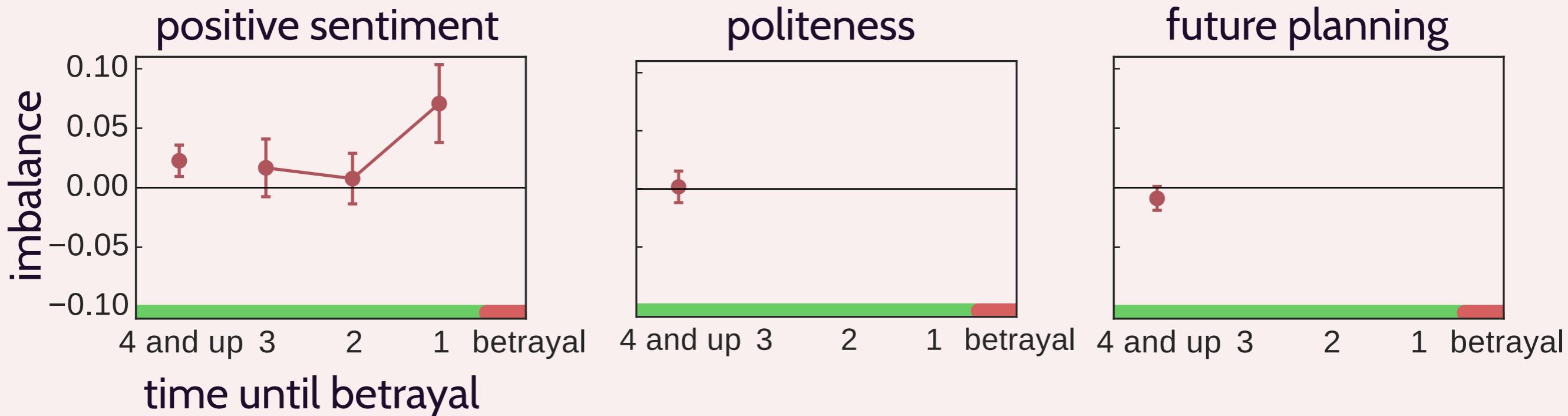
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

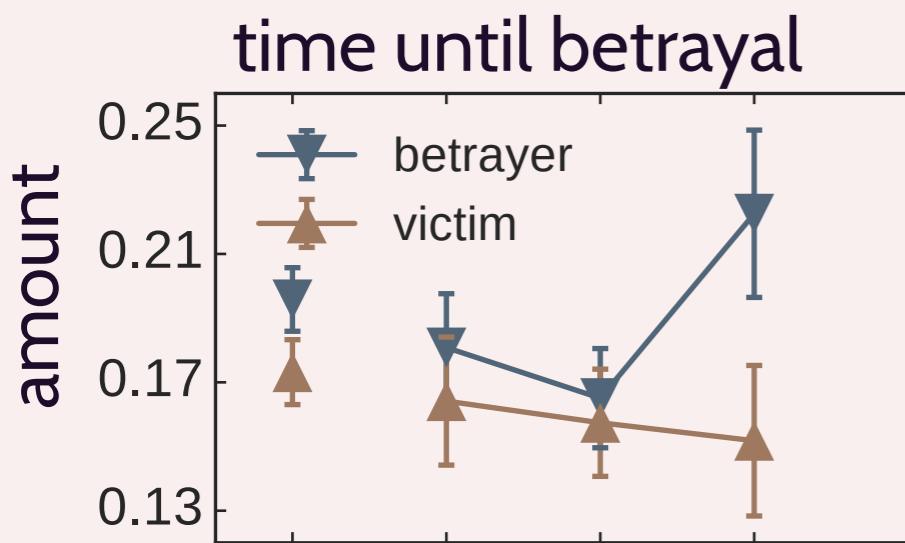
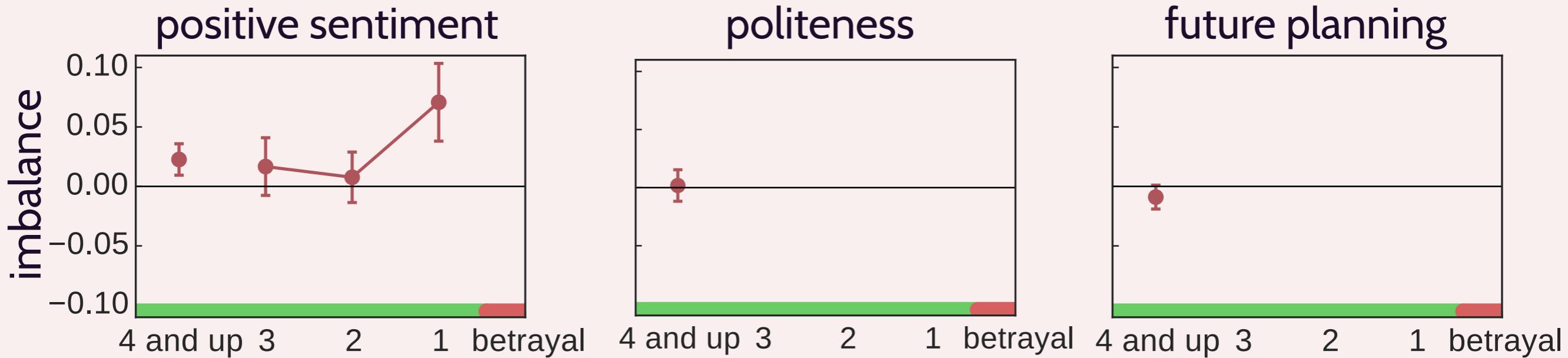
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

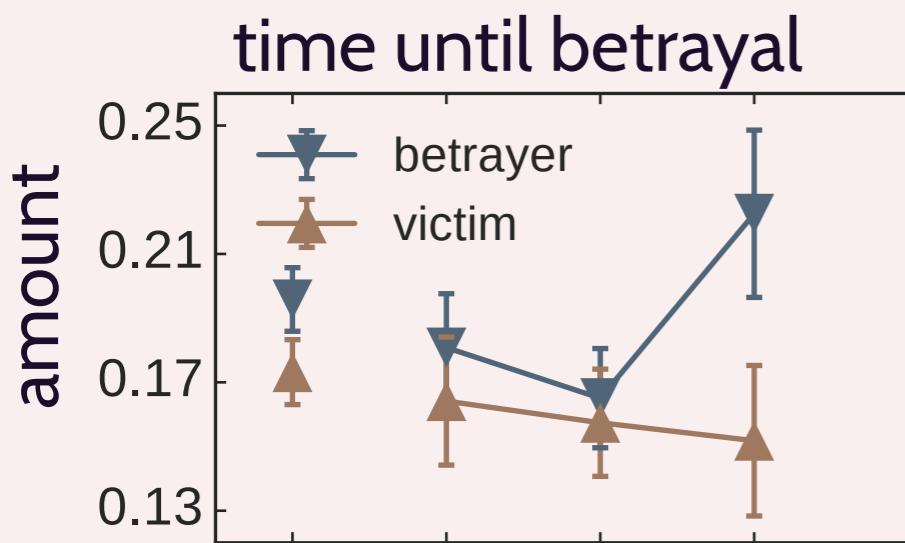
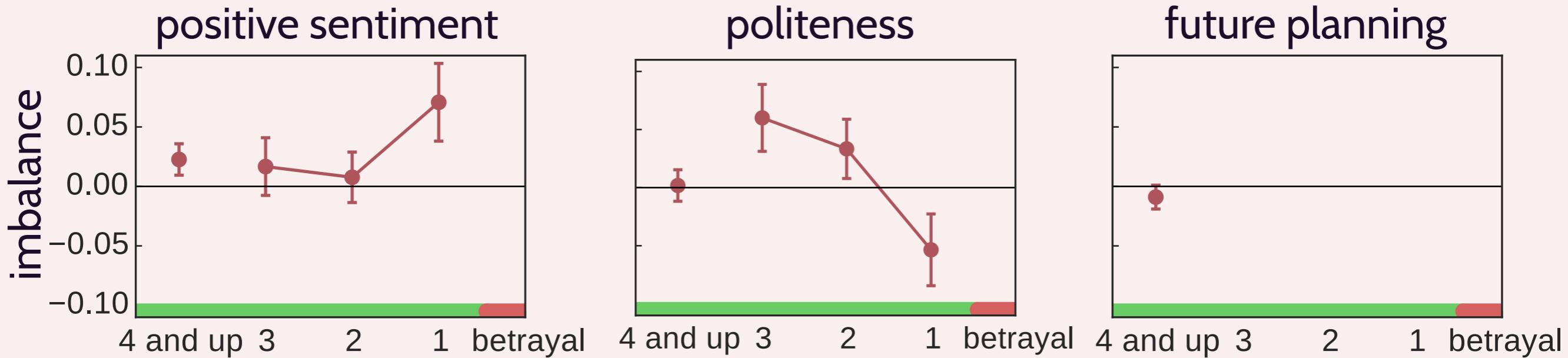
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

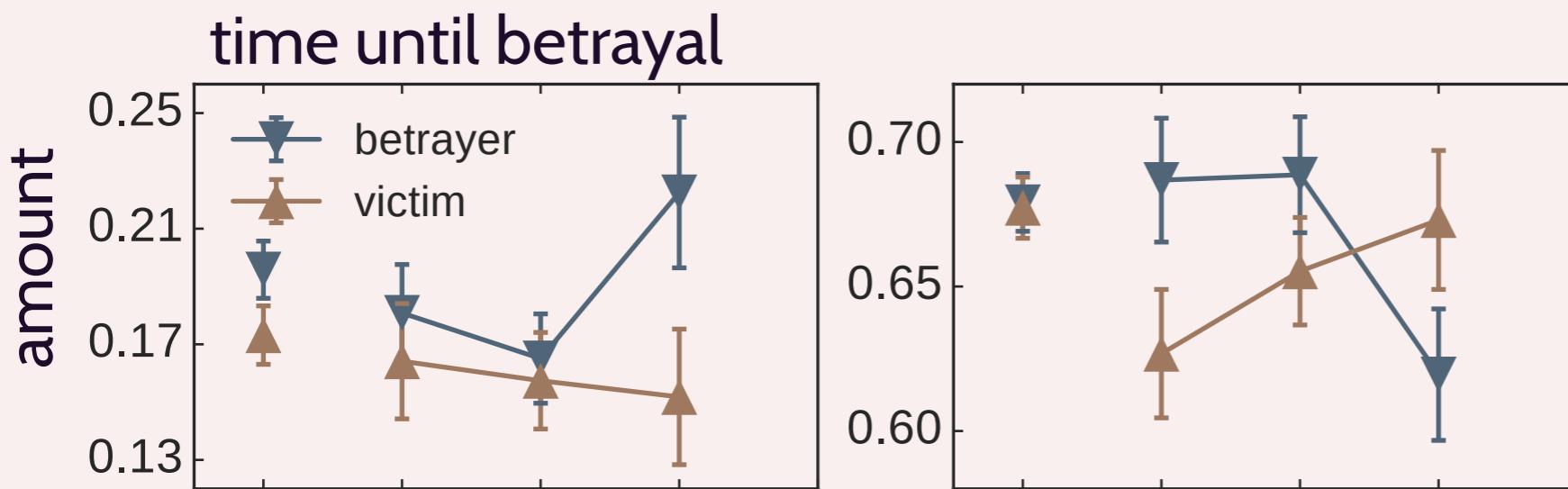
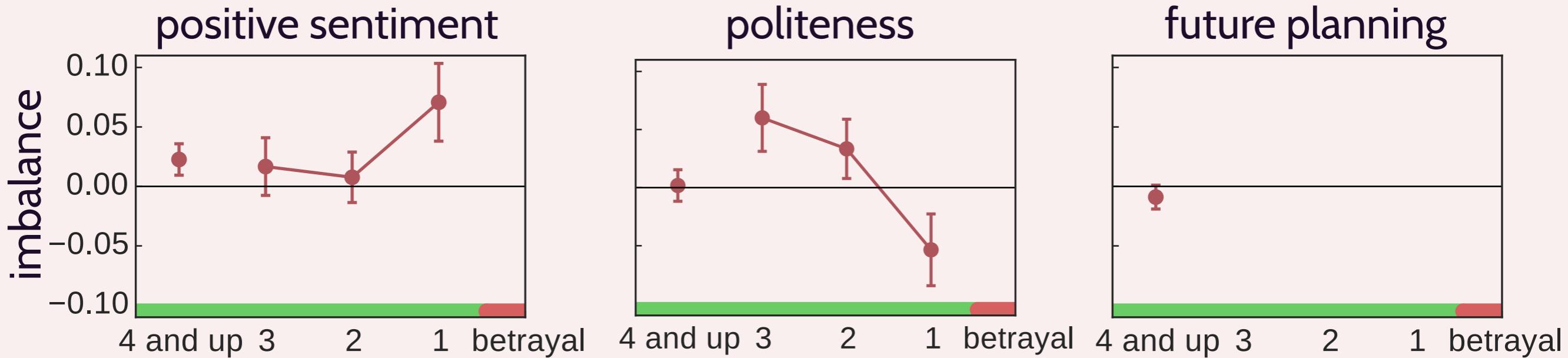
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# (Im)balance Over Time

Imbalance:  $f(\text{betrayer}) - f(\text{victim})$

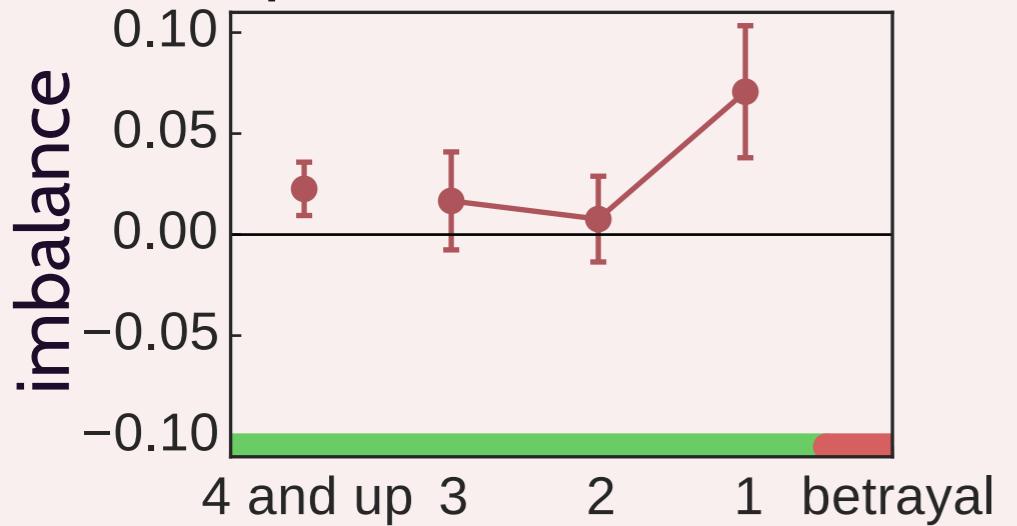


(Error bars show standard error.)

# (Im)balance Over Time

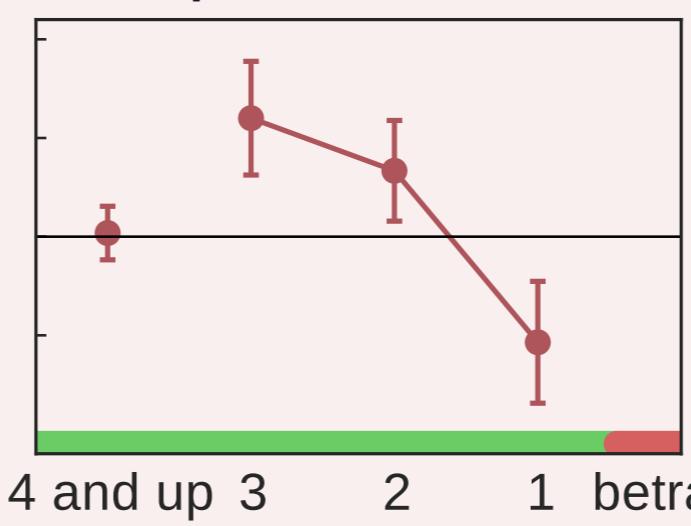
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$

positive sentiment

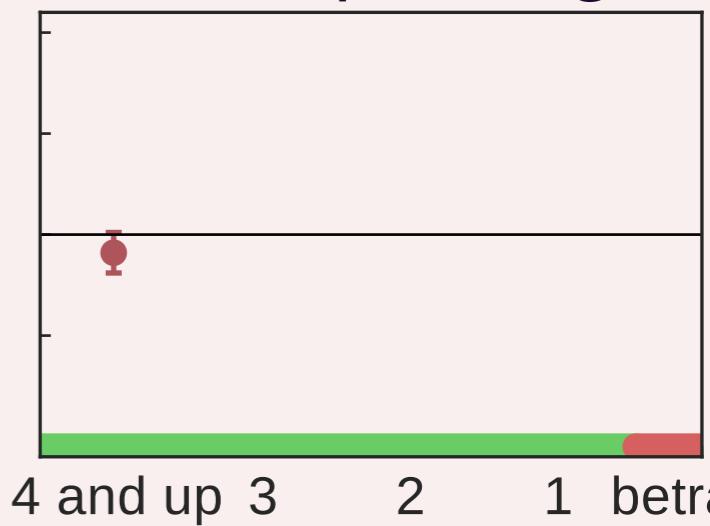


Demand-Withdraw pattern pre-divorce.  
(Gottman & Levenson, 2000)

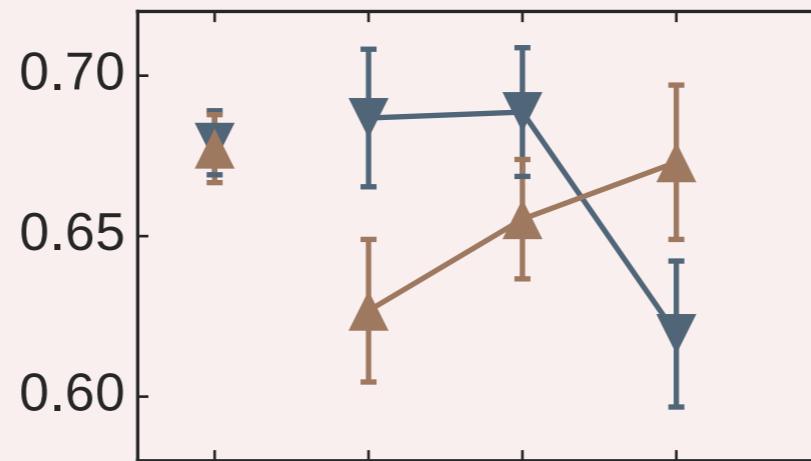
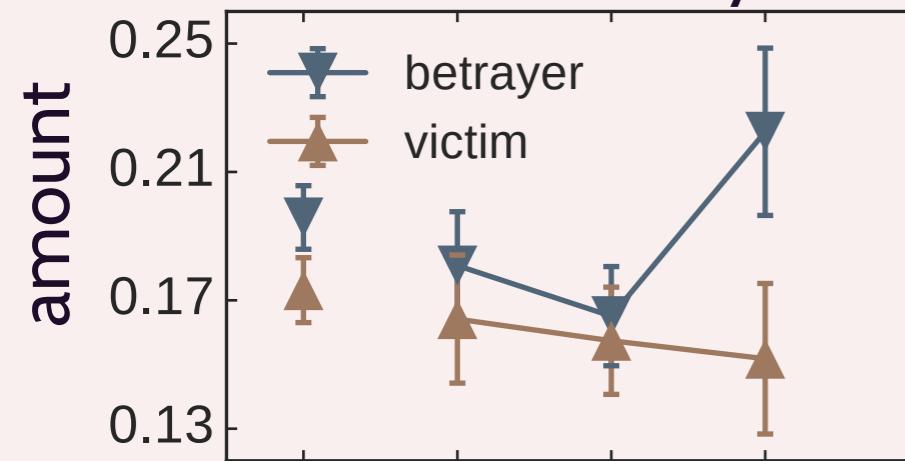
politeness



future planning



time until betrayal

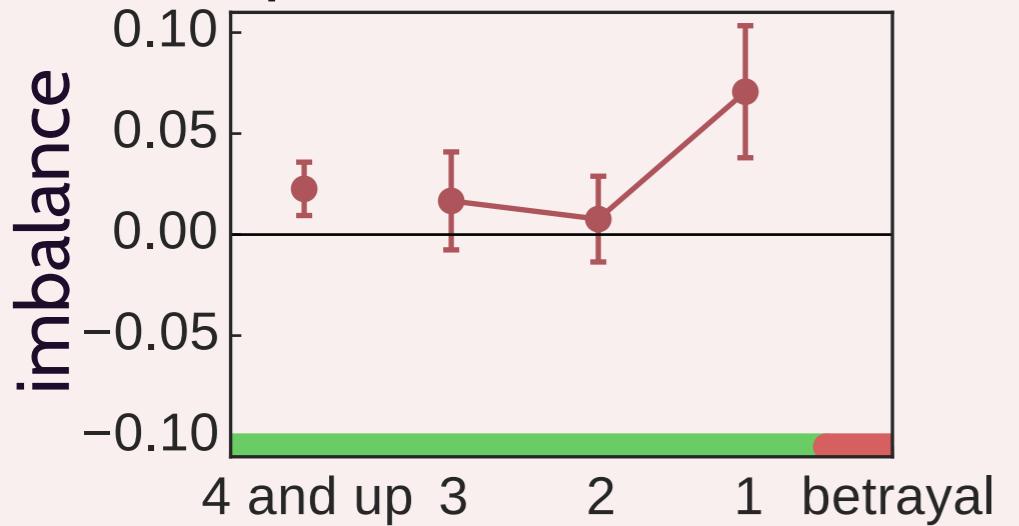


(Error bars show standard error.)

# (Im)balance Over Time

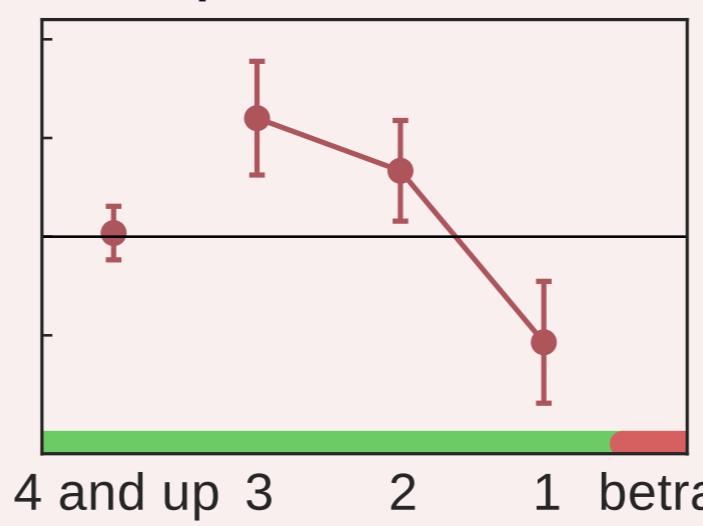
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$

positive sentiment

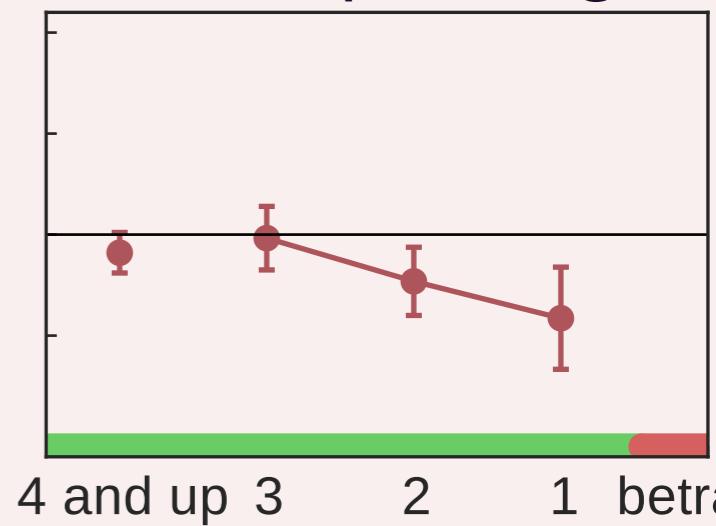


Demand-Withdraw pattern pre-divorce.  
(Gottman & Levenson, 2000)

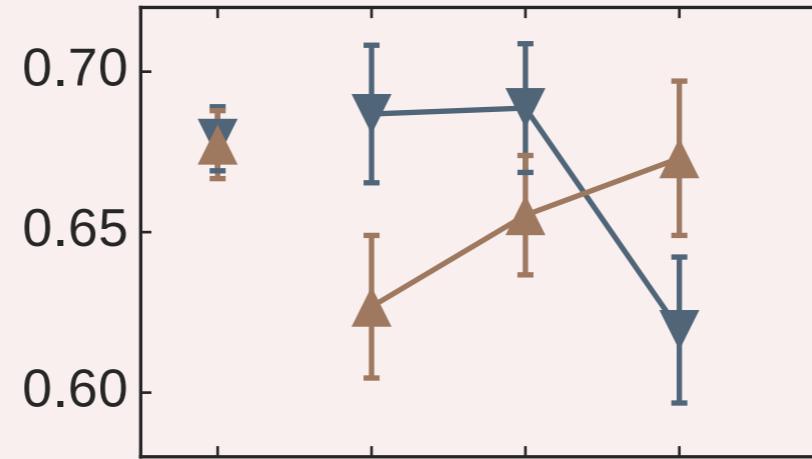
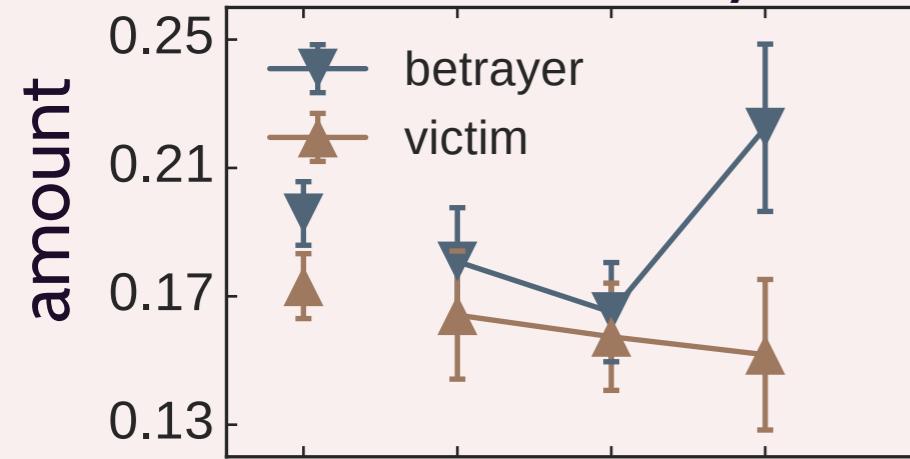
politeness



future planning



time until betrayal

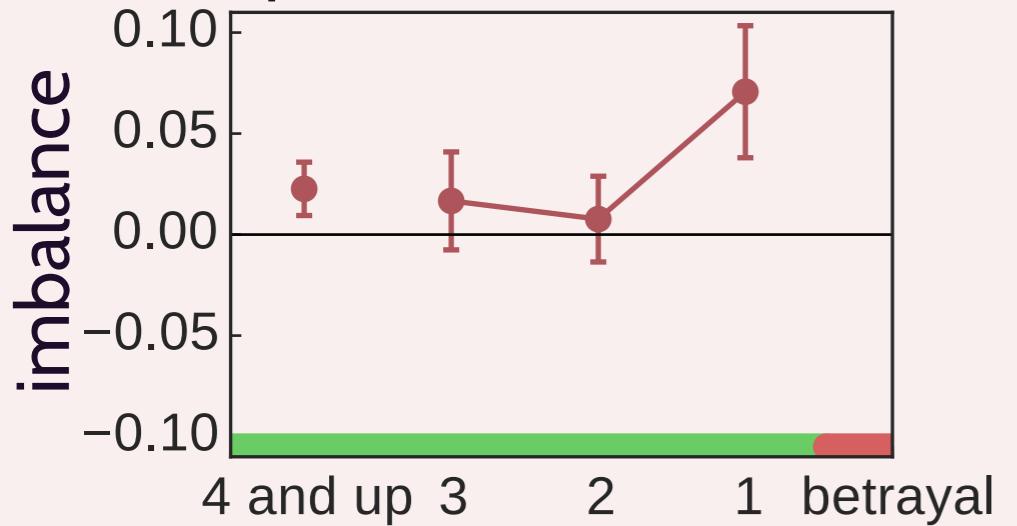


(Error bars show standard error.)

# (Im)balance Over Time

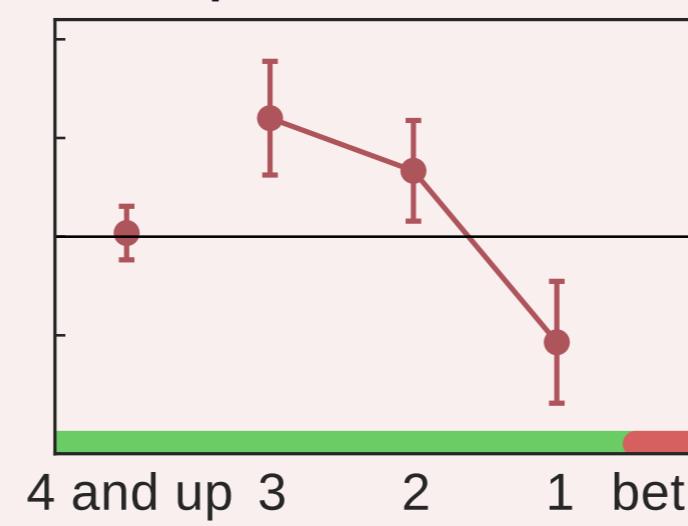
Imbalance:  $f(\text{betrayer}) - f(\text{victim})$

positive sentiment

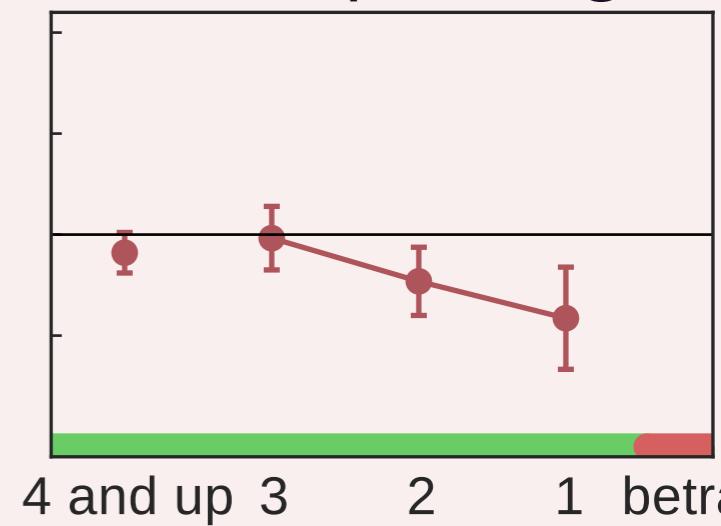


Demand-Withdraw pattern pre-divorce.  
(Gottman & Levenson, 2000)

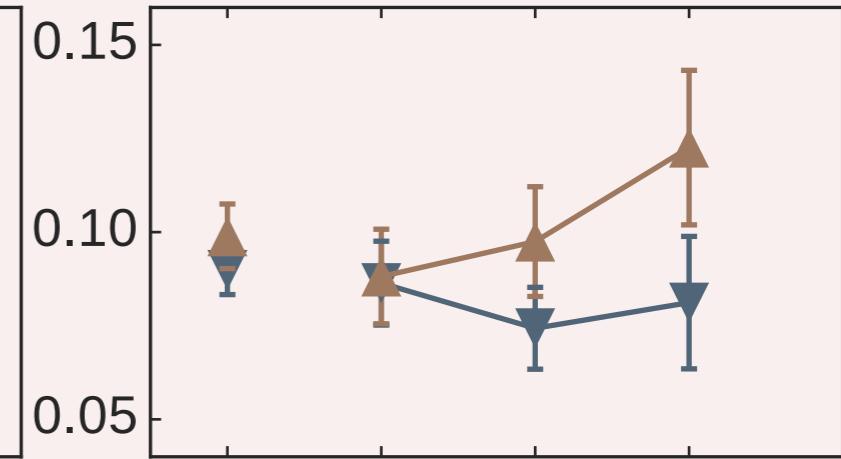
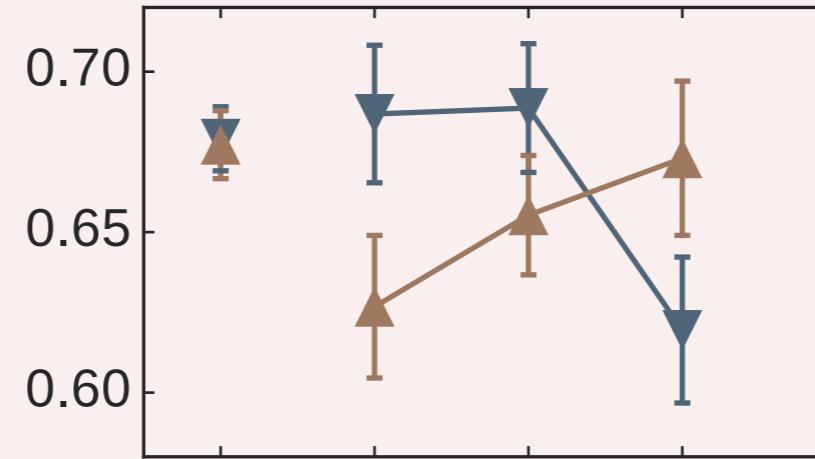
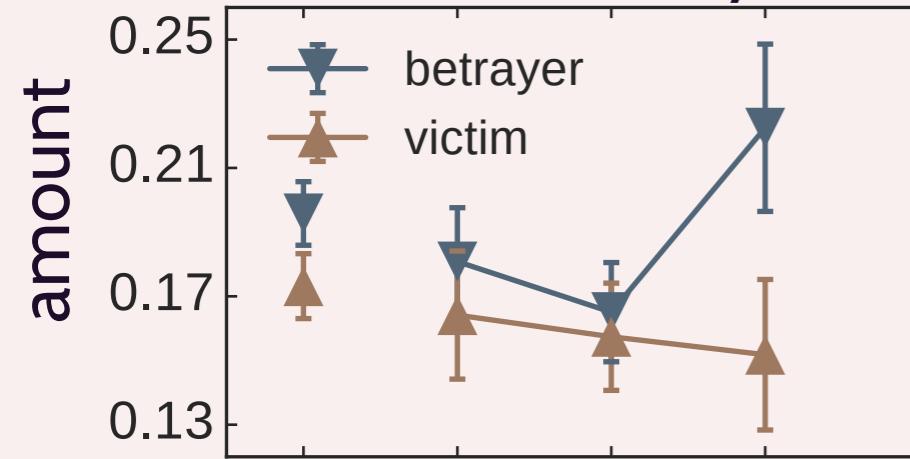
politeness



future planning



time until betrayal

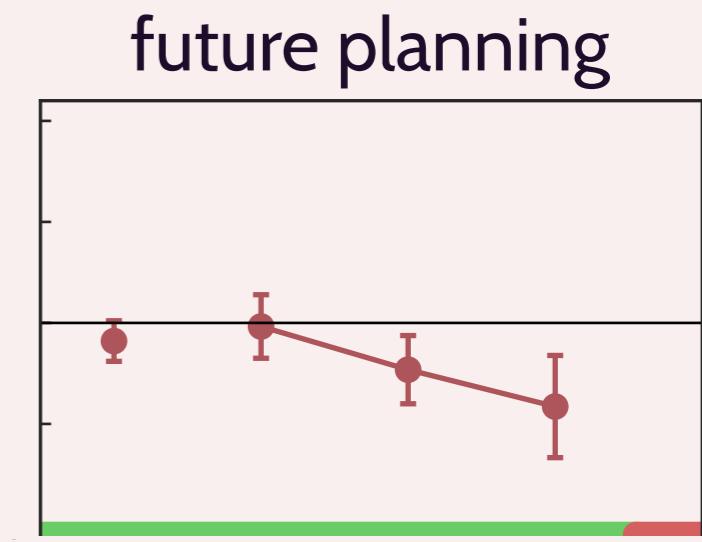
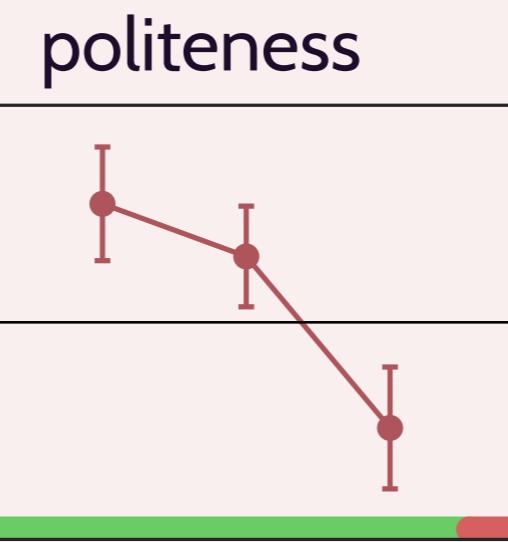
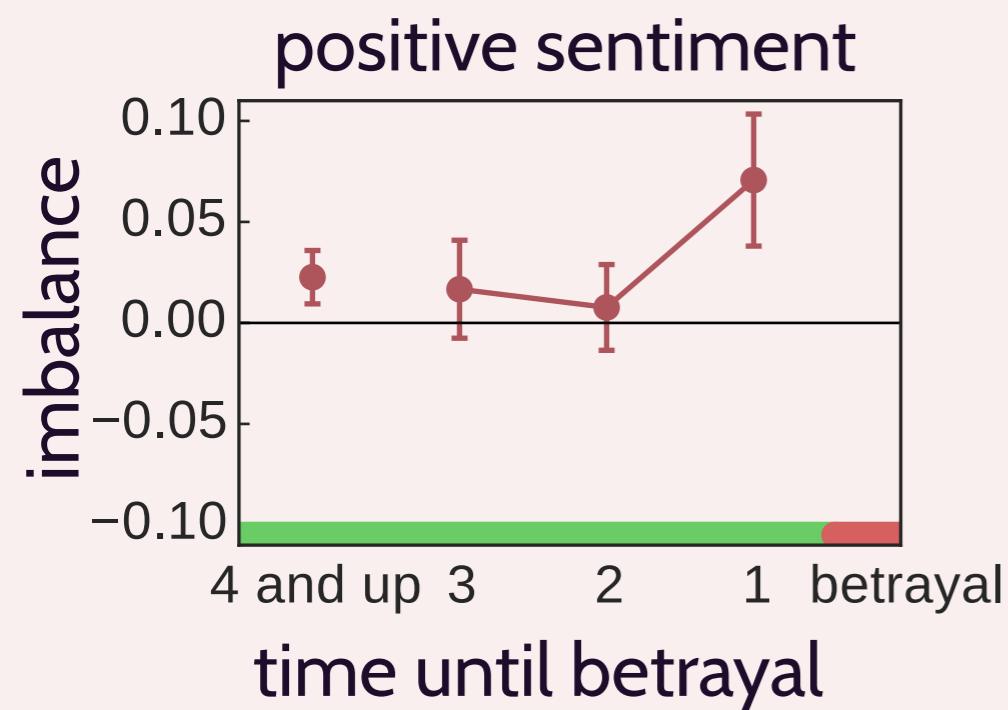


(Error bars show standard error.)

As betrayal  
draws nearer,  
balance is broken.

Attributes change  
at different rates.

Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



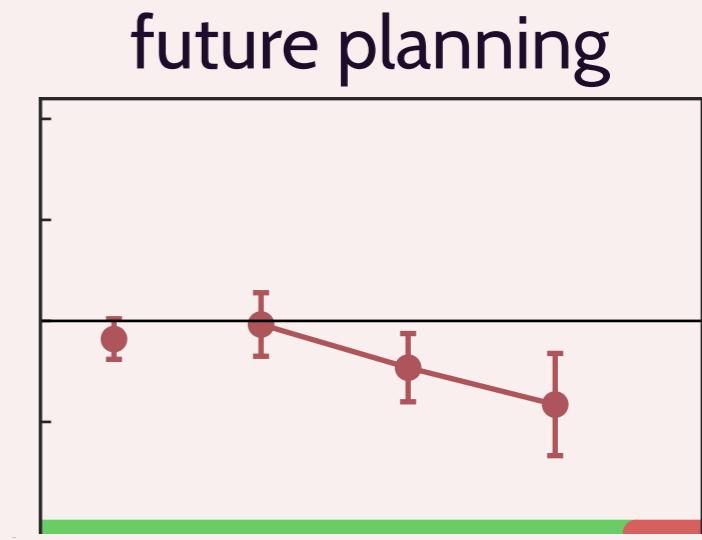
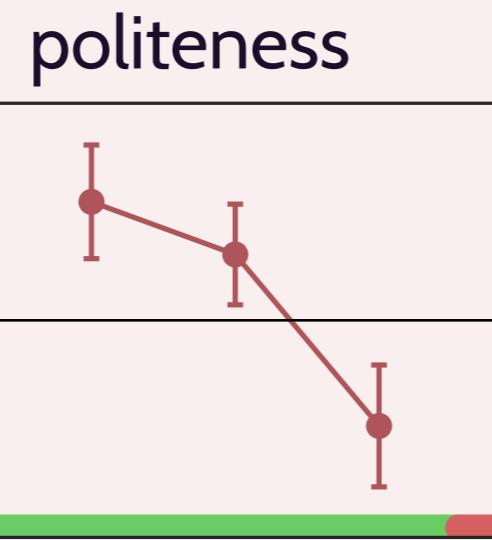
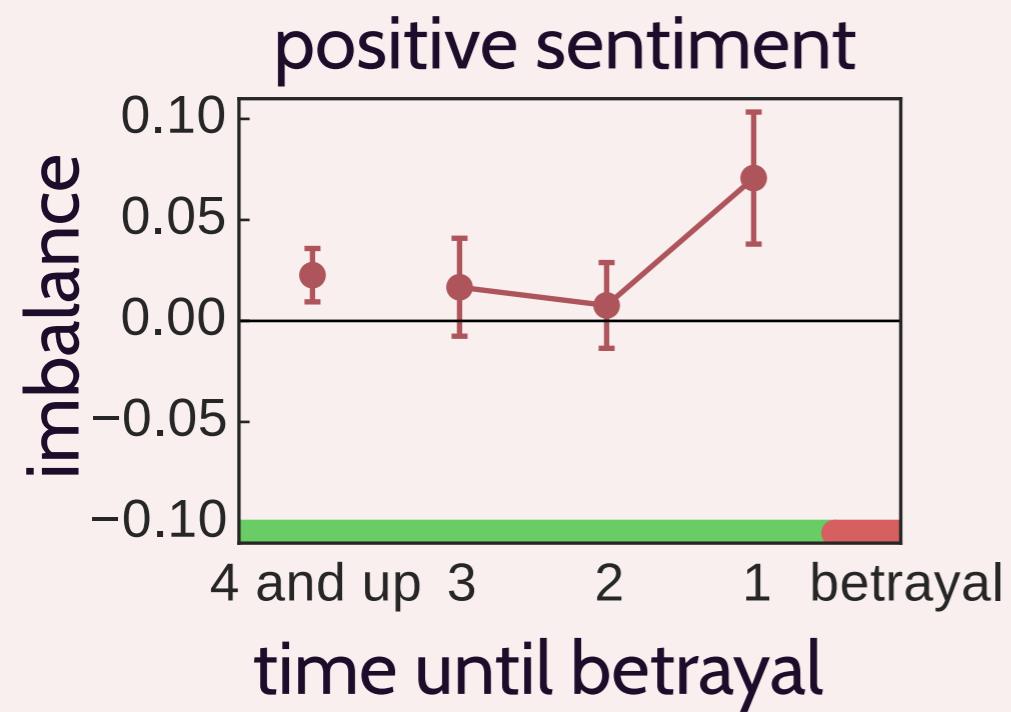
(Error bars show standard error.)

As betrayal  
draws nearer,  
balance is broken.

Attributes change  
at different rates.

Are these cues predictive?

Imbalance:  $f(\text{betrayer}) - f(\text{victim})$



(Error bars show standard error.)

# Predicting Betrayal



“Would it be ok with you if I took Denmark? I think I'm going to need it if I am going to hold France back.”



“Hi Germany, How about I give you back Denmark next year. This is because I probably won't get a centre this year and would rather not disband a unit.”

“I am supporting you into Sweden this turn if you want. If you want to be able to keep Sweden I suggest moving into Finland.  
Cheers, Harriet Jones, PM.”



“Thanks, I accept the support. I'll decide what I want to do with the army.”

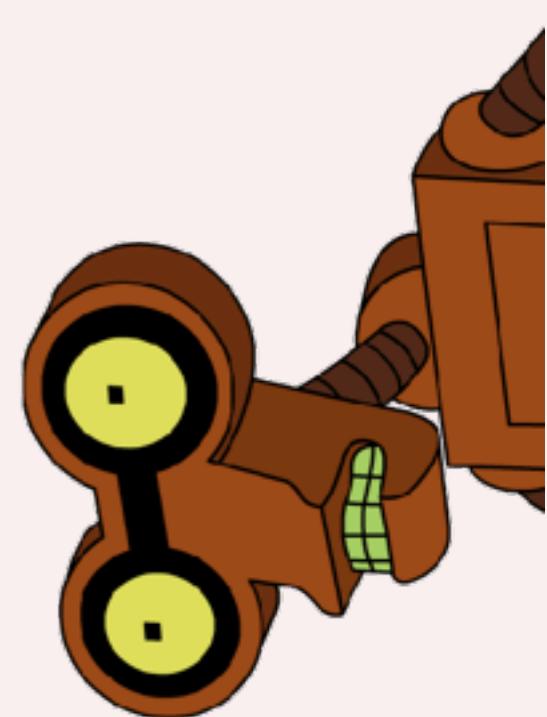
# Predicting Betrayal



“Would it be ok with you if I took Denmark? I think I'm going to need it if I am going to hold France back.”



“Hi Germany, How about I give you back Denmark next year. This is because I probably won't get a centre this year and would rather not disband a unit.”



Germany  
Stabs!

“Germany,  
Well that move was sour.  
This was a pity.  
Unfortunately now you  
have jumped out of the  
pan into the fire.”

# Predicting Betrayal

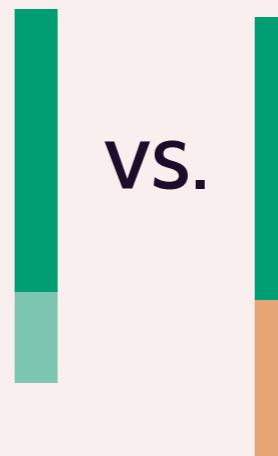
Toss in a few more features:

- Sentiment  
(Stanford Sentiment Analysis)
- Argumentation & discourse  
(Penn Discourse Treebank)  
(Stab & Gurevich, 2014)
- Politeness  
(<http://politeness.mpi-sws.org>)
- Subjectivity  
(Riloff & Wiebe, 2003)
- Talkativeness

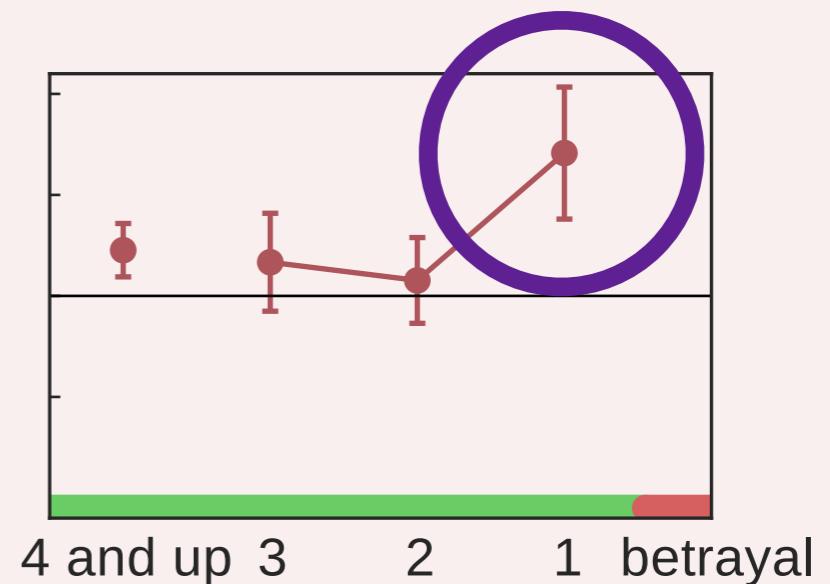
# Predicting Betrayal

Prediction tasks:

- Will this friendship break?



- Is betrayal imminent?



# Predicting Betrayal

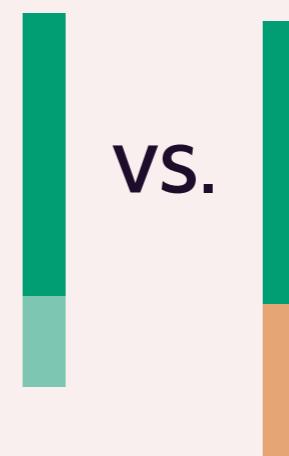
## Prediction tasks:

- Will this friendship break?

(1375 seasons, 48% *betrayals*)

Accuracy: (players: 52%)

MCC:<sup>\*</sup> (players: 0)

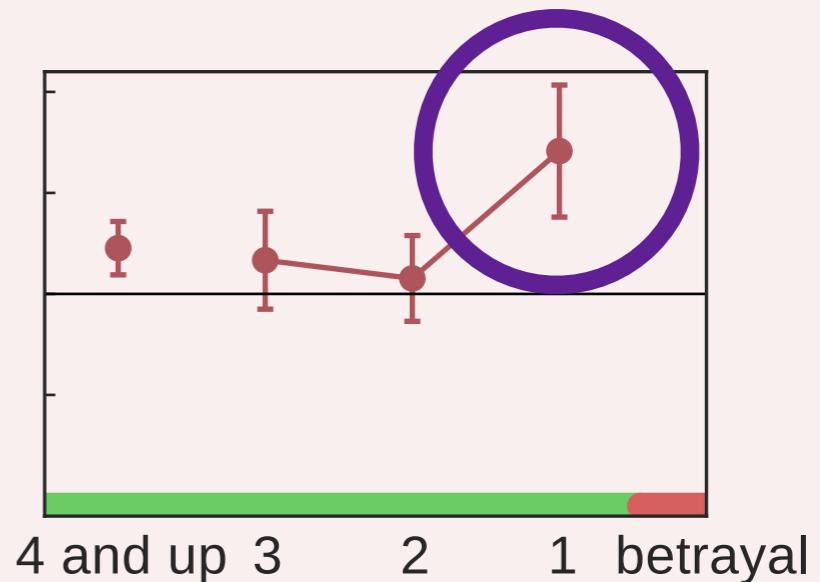


- Is betrayal imminent?

(663 seasons from betrayals,  
14% *immediately before betrayal*)

F<sub>1</sub>: (players: 0)

MCC:<sup>\*</sup> (players: 0)



<sup>\*</sup>Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.

# Predicting Betrayal

## Prediction tasks:

- Will this friendship break?

(1375 seasons, 48% *betrayals*)

Accuracy: (players: 52%) **57%**

MCC:<sup>\*</sup> (players: 0) **0.14**

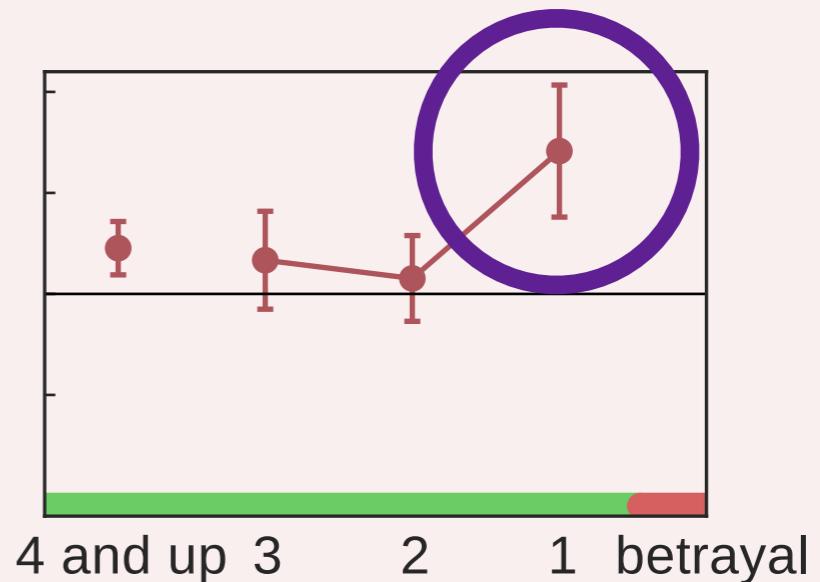


- Is betrayal imminent?

(663 seasons from betrayals,  
14% *immediately before betrayal*)

F<sub>1</sub>: (players: 0)

MCC:<sup>\*</sup> (players: 0)



<sup>\*</sup>Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.

# Predicting Betrayal

## Prediction tasks:

- Will this friendship break?

(1375 seasons, 48% *betrayals*)

Accuracy: (players: 52%) **57%**

MCC:<sup>\*</sup> (players: 0) **0.14**

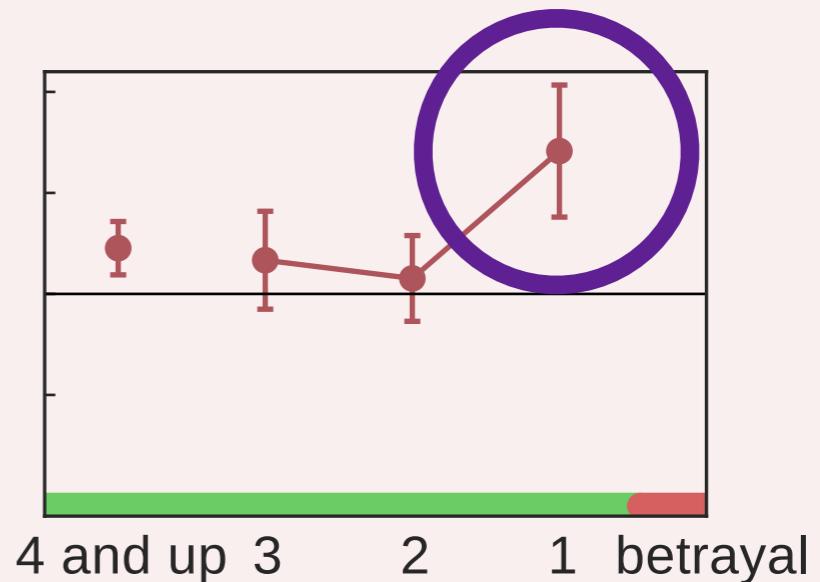


- Is betrayal imminent?

(663 seasons from betrayals,  
14% *immediately before betrayal*)

F<sub>1</sub>: (players: 0) **0.31**

MCC:<sup>\*</sup> (players: 0) **0.17**



<sup>\*</sup>Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.

# Predicting Betrayal

## Prediction tasks:

- Will this friendship break?

(1375 seasons, 48% *betrayals*)

Accuracy: (players: 52%) **57%**

MCC:<sup>\*</sup> (players: 0) **0.14**

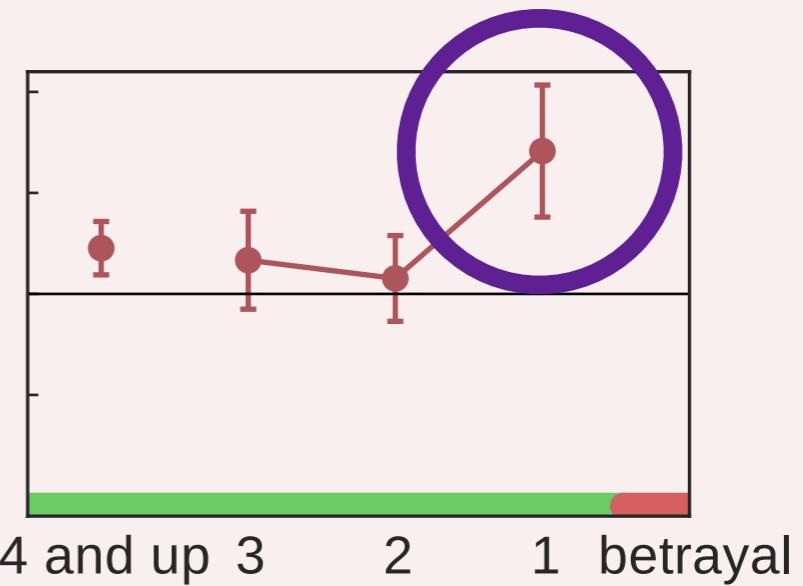


- Is betrayal imminent?

(663 seasons from betrayals,  
14% *immediately before betrayal*)

F<sub>1</sub>: (players: 0) **0.31**

MCC:<sup>\*</sup> (players: 0) **0.17**



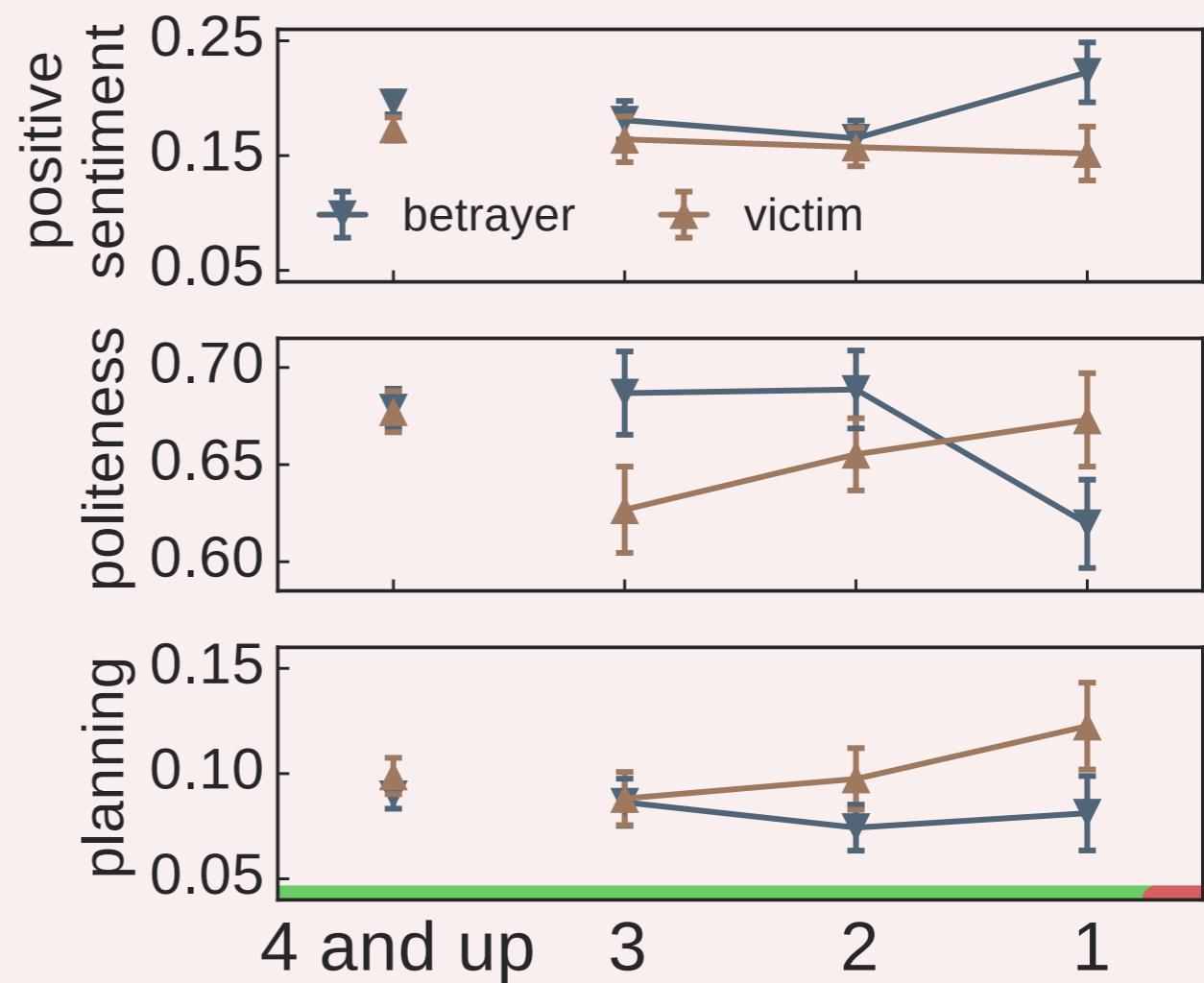
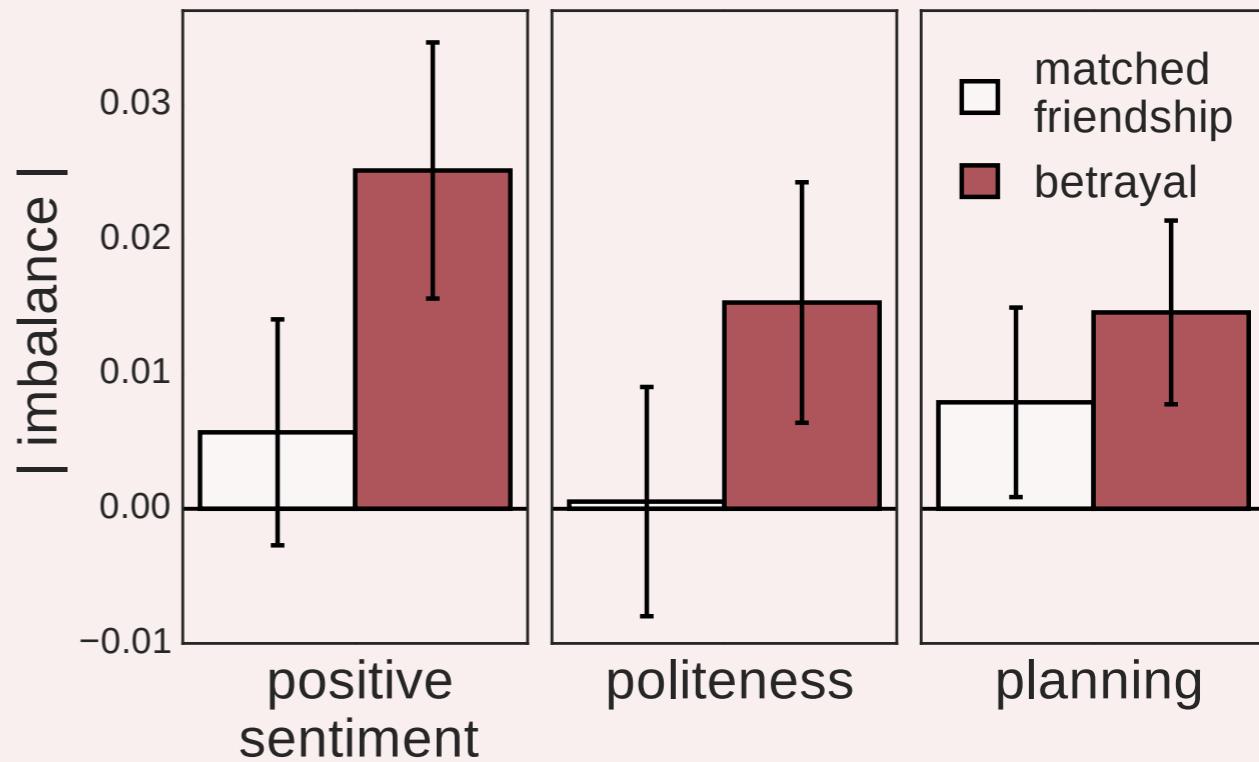
- Outperforming the players!

<sup>\*</sup>Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.

# The intention to betray can leak through words.

## Good friendships are balanced.

## Imbalance changes as betrayal draws near.



*extra slides*

# Feature Examples

Positive sentiment	I will still be trilled if you win this war.
Negative sentiment	It's not a great outcome, but still an OK one.
Neutral sentiment	Do you concur with my assumption?
Claim	I believe that E/F have discarded him.
Premise	I put italy out because I wanted to work with you.
Comparison	We can trade centers as much as we like.
Contingency	He did not, thus we are indeed in fine shape.
Expansion	Would you rather see A or B?
Temporal	i think he can still be effective while you take ROM.
Planning	HOL should fall next year, and then MUN after.
Subjectivity	I'm just curious what you think.
Politeness	I wonder if you shouldn't try to support Italy into MAR... What do you think?

# Selected Features

Will this friendship break?

Is betrayal imminent?

Sender	Positive feature	Sender	Negative feature	Sender	Positive feature	Sender	Negative feature
B	Positive sentiment	B	Expansion	V	Comparison	B	Claims
B	No. Sents	B	Comparison	V	Positive sentiment	B	Politeness
		B	Contingency	V	Contingency	B	Contingency
		B	No. Words	V	Planning	B	Subjectivity
		B	Planning	V	Requests	B	Expansion
		B	Negative sentiment	V	Expansion	B	No. Sentences
						B	Comparison