Linguistic Harbingers of Betrayal: A Case Study of Online Strategy Games

Abstract

Interpersonal relations are in constant flux; close friendships often dissolve into enmity. We examine linguistic cues augur transitions from friendly bonds into hostile relations through conversations from online strategy games. We look at betrayal from two perspectives in a classification framework: we characterize the difference between friendships that break and friendships that last, and we examine the temporal evolution of impending betrayal.

Our study reveals linguistic cues of a fragile friendship. We show that subtle signs of an imminent betrayal are encoded in changes of conversational patterns, when at least one participant is unaware of the relationship’s fate. In particular, we find that lasting friendships exhibit a form of social symmetry that also manifests through language. Imbalances of certain attributes, such as positivity or argumentativeness, should be taken as warning signs.

1 Introduction

The emerging field of computational social science attempts to explain social interactions through data. However, social interactions are complicated, as we rarely see all of the data that define the relationship between friends or enemies. Instead, thought experiments like the prisoner’s dilemma (Axelrod and Dion, 1988) are used to explain behavior. Two prisoners—denied communication—must decide whether to cooperate with each other or defect. Such simple and elegant tools helped understand many real world scenarios: pricing products, advertising, athletes doping. Despite its power, the prisoner’s dilemma remains woefully unrealistic. Cooperation and betrayal do not happen in a cell cut off from the rest of the world. Instead, real interactions are mediated by communication: promises are made, then broken, and met with recriminations.

In contrast, we study a game called Diplomacy (Sharp, 1978), where alliances and betrayals are orchestrated primarily through language. Diplomacy, like the prisoner’s dilemma, is a repeated game where players choose to either cooperate or betray other players. Diplomacy is so engaging that it is played around the world, including over the Internet. Players talk to each other, convincing other players to work with them or betray other players. Here is an abridged example of an exchange between two allies in a Diplomacy game:

Hi Baron, Congratulations on getting France to move [there]. I don’t know if you orchestrated this, but if so it was brilliant! Looks like France was suspicious of our co-operation. I hope my builds show I am committed to our relationship. Best regards, Tsar Nicholas

Dear Tsar Nicholas, I deduced that French [moving there] was likely from his other behaviour so far, but did not directly encourage it except by continued words of support. [Long, detailed message with future plans and analysis.] Thank you kindly, Baron Luftkopf

Hi Germany, Thanks for your long press (long messages typically mean players are serious about alliance!) Your suggestions seem good. I’ll leave it up to you, get back to me if you like saying which you will do. Tsar Nicholas

“The Tsar” is very polite and positive, and comments on how “the Baron’s” attention to detail is a good sign of their long-lasting friendship. The two have been good allies for the better part of the game, and immediately after the exchange, the Baron follows through and helps the Tsar. Immediately after, however, the Tsar backstabs the Baron. The intention to do so was so well concealed, that the Baron did not see it coming; otherwise he would have taken advantage first.

Such scenarios suggest a research challenge. Any hint of an intention to betray, hidden in language, would be valuable: the betrayer would unknowingly be giving away their intent without the victim...
noticing. Therefore, detecting the situation computationally would mean outperforming good human players.

After briefly describing the game (Section 2), we focus on how the structure of the game provides convenient, reliable indicators of whether pairs of participants are friends or foes (Section 3). Given these labels, we build computational methods that predict whether friendships will end in betrayal (Section 4) and—if so—when the betrayal will happen (Section 5). We find evidence of subtle but consistent patterns that can foretell the dissolution of a friendship. For instance, pairs where one player uses significantly more positive sentiment are more likely to end with that player betraying the other. People who put increasingly more pressure on a relationship by making more requests and making more plans attract betrayal.

While our focus is on a single popular game, these methods help reveal dynamics present in other social interactions (Section 7).

2 Communication and Conflict in Diplomacy

Figure 1: The full Diplomacy board representing Europe circa 1914. The seven nations struggle to control the map.

A game of Diplomacy begins in 1914 with players casting themselves as the European powers at the eve of the first world war: England, Germany, France, Russia, Austria, Italy, and the Ottoman Empire. The goal of the game (like other war games such as Risk or Axis & Allies) is to capture all of the territories on the game board (Figure 1). The games are divided into years starting from 1914 and each year is divided into two seasons—Spring and Fall. Each season of the game consists of two phases which alternate: diplomacy—the players communicate with each other—and orders—the players submit their moves for the season.

2.1 Movement, Orders, and Battles

On the board, each player can operate a unit for each city they control. Each turn, these pieces have the option of moving to an adjacent territory. What makes Diplomacy unique is that all players submit their written (or electronic) orders; these orders are executed simultaneously; and there is no randomness (e.g., dice). Thus, the outcome of the game depends only on the communication, cooperation, and movements of players.

When two units end their turn in the same territory, it implies a battle. Who wins the battle is based only on numerical superiority (ties go to defenders). Instead of moving, a unit can support another unit; through intricate networks of support armies are created. The side with the largest army wins the battle.

The process of supporting a unit is thus critical for a successful offensive. Often, a lone player lacks the units to provide enough support and thus needs other players to help them. Because these orders (both movement and support) are machine readable, we have a clear indication of when players are working together (supporting each other) or working against each other (attacking each other); we will use this to define relationships between players (Section 3). However, coordinating these actions between players requires cooperation and diplomacy, the other phase of the game.

2.2 Communication

In the diplomacy phase of the game, players talk to each other. These conversations are either global or—more typically—one-on-one. Conversations include greetings, extra-game discussions (e.g., “did you see Game of Thrones?”), low-level tactics (“if you attack Armenia, I’ll support you”), high-level strategy (“we need to control central Europe”). These communication messages are the key elements of our study.

A recent episode of This American Life reveals insight into the tactics used by top players. A key

1 Instead of moving a unit, a player can have that unit support another unit. For example, if an English army in Belgium is attacking a Germany Army in Ruhr, a French army in Burgundy could support that attack rather than making an attack on its own. This is accomplished by the French player writing a move explicitly stating “I support England’s attack from Belgium to Ruhr”.

2 http://www.thisamericanlife.
component to succeeding in Diplomacy is to forge alliances and then break them at the right time. Thus, successful play depends on convincing competitors to help (support) and then to betray them.

Because of the centrality of language to Diplomacy, we can learn the rhetorical and social devices players use to build and break trust. Because this language is embedded in a game, it has convenient properties: similar situations are repeated, the goals are clear, and the machine-readable orders let us know who is working with whom and who the enemies are. In the next section, we describe some preliminary analyses of the Diplomacy data.

2.3 Preprocessing

We use games from two popular online platforms for playing Diplomacy.\(^3\) When playing online, one game season lasts about 9 days on average. After obtaining the data, we remove non-standard games caused by differences between the two platforms, as well as games that are still in progress. Moreover, in each game, we filtered out setup messages, regulatory messages to and from the administrator of the game and messages declaring the state of the game, leaving only messages between the players. This leaves us with 249 games with 145,000 total communications, 62,000 types and 12,700,000 tokens.

The dataset confirms that communication is an essential part of Diplomacy: half of the games have over 515 messages exchanged between the players, while the top quartile has over 750 messages per game. Also, non-trivial messages (with at least one sentence) tend to be complex: over half of them have at least 5 sentences, and the top quartile consists of messages with 8 or more sentences.

3 Will this Relationship Last?

In this section, we explore how interactions within the game of Diplomacy define the relationships between players. While most relationships between players are undefined (e.g., England and Turkey are in opposite corners of the map), specific interactions between players define whether they are friendly or hostile to each other. In the next sections, we show how the subtle linguistic patterns of in-game player conversations can reveal whether a friendship relation will turn hostile or not.

**Friendships and hostilities.** Alliances are a natural part of the game of Diplomacy. While the best outcome for a player is a solo victory against all other players, this is rare and difficult to achieve without any cooperation and assistance. Instead, the game’s structure encourages players to form long-term alliances. Allies sometimes often settle for (less prestigious) team victories, but these coalitions can also crumble as players seek a (more prestigious) solo victory for themselves. This game dynamic naturally leads to the formation of friendly and hostile dyads, which are relatively easy to identify in a post-hoc analysis of the game.

**Acts of friendship.** Diplomacy provides a support option for players to help each other: this game mechanism (discussed at large in Section 2) provides unequivocal evidence of friendship. When two players engage in a series of such friendly acts, we will say that the two have a friendship relation.

**Acts of hostility.** Unlike support, hostile actions are not explicitly marked in Diplomacy. We consider two players to be hostile if they get involved in any unambiguous belligerent action, such as invading one another’s territory, or if one supports an enemy of the other.\(^4\)

**Betrayal.** As in real life, friendships can be broken unilaterally: an individual can betray his friend by engaging in a hostile act towards them. Figure 2 shows two players who started out as friends (green) but became hostile (red) after a betrayal. Importantly, until the last act of friendship (\(t = 1\)), the victim is unaware that she will be betrayed (otherwise she would not engage in an act of friendship); also, the betrayer has no interest in signaling to her partner that a betrayal is happening.

This setting poses the following research challenge: are there linguistic cues that appear during the friendly conversations and portend upcoming betrayal? A positive answer would have two implications: the betrayer unknowingly hints at their future treachery, and the victim fails to notice. We will explore this question in the following sections.

**Relationstability.** Before venturing into the linguistic analysis of betrayals, we briefly explore the dynamics underlying these state transitions. We

\(^3\)More details on the data available after blind review.

\(^4\)In Diplomacy all game actions are simultaneous, and this can lead to ambiguous interpretation of the nature of a pair of user’s interactions. Our definition of hostility intentionally discards such ambiguous evidence. For instance, if two players attempt to move into the same unoccupied territory, this is not necessarily aggressive: allies sometimes use this tactic (“bouncing”) to ensure that a territory remains unoccupied.
<table>
<thead>
<tr>
<th>Event</th>
<th>Time</th>
<th>What happened</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>4</td>
<td>B supports V’s army in Vienna</td>
</tr>
<tr>
<td>$F_2$</td>
<td>3</td>
<td>V supports B’s attack from Warsaw to Silesia</td>
</tr>
<tr>
<td>$F_3$</td>
<td>3</td>
<td>B again supports V in Vienna</td>
</tr>
<tr>
<td>$F_4$</td>
<td>1</td>
<td>V supports B’s move from Venice to Tyrolia</td>
</tr>
<tr>
<td>$H_5$</td>
<td>0</td>
<td>B attacks V in Vienna</td>
</tr>
<tr>
<td>$H_6$</td>
<td>-1</td>
<td>V retaliates, attacking B in Warsaw</td>
</tr>
</tbody>
</table>

**Figure 2:** A friendship between Player B (eventual betrayer) and Player V (eventual victim) unravels. For the first four events, the players exchange friendly acts (in green). Eventually B’s unilateral hostile act betrays V’s trust, leading to hostility (in red). We define that dissolution as time ($t = 0$) and index backward from that betrayal (i.e., number of seasons before the betrayal).

Figure 3: A stable relationship should be symmetrical: friends will help each other while enemies will fight each other. A precarious friendship feels one-sided, while a conflict may turn to friendship through a magnanimous olive branch.

4 Language Foretelling Betrayal

In this section, we examine whether the messages exchanged between two Diplomacy players reveal whether a friendship will last or end in betrayal. Given the relationships defined in the previous section, we attempt to predict whether messages exchanged between players while ostensibly being friends belie a future betrayal.

This is a hard task, as at least one of the players has a clear incentive to predict betrayal: if your erstwhile friend will eventually betray you, it is in your interest to defect first.

To find betrayals, we must first find friendships. To be a true friendship, a relationship must be ongoing, established, and reciprocal. Thus, we focus on sequences of at least two friendly game interactions, lasting at least three seasons in game time, and containing friendly actions by both players. We also check that no more than five seasons pass between two interactions, since friendships can fade.

For some of these sequences, players never show hostility towards each other for the rest of the game. We take the longest such sequences to represent stable friendships.

Betrayals are friendships (as defined above) followed by at least two hostile acts. If the betrayal is mutual (i.e., both players start attacking each other in the same season), then we consider both directions.

For each example of betrayal, we find the most similar friendship that was never dissolved by betrayal. We compare friendships using two statistics: the length of the friendship and number of turns since the start of the game. We greedily select a friendly counterpart for each betrayal and verify that no significant difference in either of the two variables (Mann-Whitney $p > 0.3$).

In this section, we attempt to predict, given the conversation between two players in a game season, whether their friendship will end in betrayal or not.

A stable relationship should be symmetrical: friends will help each other while enemies will fight each other. A precarious friendship feels one-sided, while a conflict may turn to friendship through a magnanimous olive branch.

4.1 What Makes a Friendship

To find betrayals, we must first find friendships. To be a true friendship, a relationship must be ongoing, established, and reciprocal. Thus, we focus on sequences of at least two friendly game interactions, lasting at least three seasons in game time, and containing friendly actions by both players. We also check that no more than five seasons pass between two interactions, since friendships can fade.

For some of these sequences, players never show hostility towards each other for the rest of the game. We take the longest such sequences to represent stable friendships.

Betrayals are friendships (as defined above) followed by at least two hostile acts. If the betrayal is mutual (i.e., both players start attacking each other in the same season), then we consider both directions.

For each example of betrayal, we find the most similar friendship that was never dissolved by betrayal. We compare friendships using two statistics: the length of the friendship and number of turns since the start of the game. We greedily select a friendly counterpart for each betrayal and verify that no significant difference in either of the two variables (Mann-Whitney $p > 0.3$).

In this section, we attempt to predict, given the conversation between two players in a game season, whether their friendship will end in betrayal or not.
4.2 Linguistic Cues Predicting Betrayal

We explore linguistic correlates of social aspects that can indicate imbalance in relationships.

**Sentiment.** Changes in the sentiment expressed in language can reflect emotional responses, as well as the status of the relationship as a whole. We quantify the proportion of exchanged sentences that transmit positive, neutral and negative sentiment using the Stanford Sentiment Analyzer (Socher et al., 2013). Figure 4a suggests that increased positive sentiment in the messages sent by somebody can be a telltale sign that the person will betray (two-sample t-test, $p = 0.006$).

**Politeness.** How polite one is when requesting something can reveal differences in status and power. We measure the politeness of each message using the Stanford Politeness classifier (Danescu-Niculescu-Mizil et al., 2013). Figure 4b shows that temporal discourse markers referring to the future exhibit a double imbalance: the betrayer is less likely to use them (two-sample t-test $p = 0.07$), and the victim is more likely to use them (two-sample t-test $p = 0.11$). Combined, the difference is significantly below zero in friendships that will break (one-sample t-test, $p = 0.001$).

**Talkativeness.** We consider the amount of textual communication between the players, in each direction. To this end, we use the number of messages sent, the average number of sentences per message, and the average number of words per sentence. Abnormal communication patterns can indicate relationship breakdown. For example, friendships that dissolve are characterized by an imbalance in the number of messages exchanged between the two players (one-sample t-test $p = 0.0002$).

4.3 Classification

To test the combined power of the language markers we found, we turn to binary classification to discriminate between friendships that last and friendships that will break, based on the text from individual game seasons. We use the features described above, and summarized in Table 1. Motivated by the intuition of imbalance, we split the features into sent by the betrayer and received by the victim. (For friendships without betrayal, the direction is arbitrarily

---

5 We collapse the extreme positive and extreme negative examples into positive and negative, due to their uncommonness.

6 We remove the connectors that appear in over a third of the messages (and, for, but, if).

---

(Prasad et al., 2008). The markers belong to four coarse classes: comparison, contingency, expansive, and temporal\(^7\). For argumentation, we count the number of claim and premise markers as identified in the annotation guidelines from (Stab and Gurevych, 2014). We also measure the number of request sentences in each message, as identified by the heuristics from the Stanford Politeness classifier (Danescu-Niculescu-Mizil et al., 2013). Figure 4b shows that temporal discourse markers referring to the future exhibit a double imbalance: the betrayer is less likely to use them (two-sample t-test $p = 0.07$), and the victim is more likely to use them (two-sample t-test $p = 0.11$). Combined, the difference is significantly below zero in friendships that will break (one-sample t-test, $p = 0.001$).

**Talkativeness.** We consider the amount of textual communication between the players, in each direction. To this end, we use the number of messages sent, the average number of sentences per message, and the average number of words per sentence. Abnormal communication patterns can indicate relationship breakdown. For example, friendships that dissolve are characterized by an imbalance in the number of messages exchanged between the two players (one-sample t-test $p = 0.0002$).
We use logistic regression after univariate feature selection. We choose the best setting for the model parameters after 10-fold cross validation, ensuring that instances from the same game are never found in both train and validation folds.

The best model achieves a Matthews correlation coefficient of 0.1 with a 95% bootstrapped lower bound of 0.05, showing predictive correlation significantly better than chance. The chosen features and their coefficients are reported in Table 2.

On top of the observations we previously made, the feature ranking reveals that writing more sentences per message is more common when one will betray. This goes against the intuition expressed by “the Tsar”, an experienced player, in the example from Section 1, that longer messages indicate good alliances. Discourse features prove relevant: more complex discourse indicates a lower likelihood of betrayal, with one important exception: One’s use of more words referring to the future make it more likely for one to be betrayed (Figure 4b).

Overall, the linguistic features capture a consistent signal that characterizes people’s language when they are planning to betray: be wary of your partner showing too much positivity, writing longer messages, and not using enough structured discourse and argumentation. Avoid putting pressure on the relationship by planning too much about the future and making too many claims.

## 5 Imminent Relationship Disruptions

The results from Section 4 suggest that language cues can be subtle signs of future relationship disruption. However, in real life, people are aware that most relationships eventually end, but keeping them alive can be beneficial. In Diplomacy, this is aggravated by the common knowledge that every player should prefer to win alone. We still encounter many long-lasting alliances, which suggests an alternative approach: assuming that a relationship will be disrupted, how soon can one expect a betrayal?

We investigate whether linguistic cues are indicative of imminent change in the relationship.
Figure 5: Evolution of features in friendships that end in betrayal, as the breakdown approaches. Error bars mark bootstrapped standard errors.

5.1 Setting
We consider only the subset of disrupted relationship sequences used in Section 4. We look at individual game seasons, and label each season with its distance from the end of the friendship. In this setting, the risk of incorrectly labeling older messages from a long friendship decreases, so we remove the limit used for the previous task. Instead, we want to prevent short alliances of circumstance from distorting the features from close to betrayal. We enforce this by keeping only friendships lasting at least four seasons.

5.2 Classification and Discussion
We consider the same predictors described in Table 1. We train a classifier to discriminate between the season preceding the last friendly interaction from all the older seasons. The best model achieves a Matthews correlation coefficient of 0.16 with a 95% bootstrapped lower bound of 0.06. The selected features, displayed in Table 3, reflect some of the effects identified in Section 4, such as the importance of positive sentiment and temporal discourse markers pertaining to the future. Betrayers have a tendency to use more positive sentiment exactly before the pair’s final friendly season (Figure 5a). A partner using increasingly future-looking discourse connectors is at higher risk of being betrayed (Figure 5b). This can be explained by the pressure that talking about the future can put on a relationship. A similar reasoning applies for making many requests.

We also find that decrease in a partner’s politeness is indicative of their imminent betrayal. When plotting the change in politeness over time (Figure 5c) there is a complete reversal in the politeness imbalance of the pair. This explains why politeness is not as effective a feature in detecting long-term betrayal. The behavior has two intuitive explanations. On one hand, if the betrayer has planned the act in advance, politeness can be a strategy for deception. On the other hand, if the betrayer receives impolite requests, the value of the relationship can decrease, which can determine betrayal. We observe a similar dynamic for the average number of sentences per message sent by the betrayer. The feature is selected in both cases, but with opposite signs: more complex messages suggest that betrayal will happen, but not right away.

Studying changes in language use as betrayal approaches adds extra insight to our results, by uncovering subtle changes that cannot be seen when looking at an entire friendship on average. Immediate betrayal can be forewarned by your partner suddenly becoming less polite while becoming more positive. But betrayal is more likely to happen against people who have themselves become less polite and then trying to redeem and match the higher politeness of their partner.

<table>
<thead>
<tr>
<th>Direction</th>
<th>Feature</th>
<th>weight (e-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent</td>
<td>Positive sentiment</td>
<td>+4.5</td>
</tr>
<tr>
<td>Received</td>
<td>Temporal (future)</td>
<td>+4.2</td>
</tr>
<tr>
<td>Received</td>
<td>Requests</td>
<td>+3.5</td>
</tr>
<tr>
<td>Sent</td>
<td>Claims</td>
<td>−6.0</td>
</tr>
<tr>
<td>Sent</td>
<td>Politeness</td>
<td>−5.1</td>
</tr>
<tr>
<td>Sent</td>
<td>Sentences</td>
<td>−3.4</td>
</tr>
<tr>
<td>Sent</td>
<td>Subjectivity</td>
<td>−2.6</td>
</tr>
<tr>
<td>Received</td>
<td>Messages</td>
<td>−1.8</td>
</tr>
</tbody>
</table>

Table 3: Features for recognizing imminent betrayal. Features with positive weights suggest that the exchange will immediately be followed by the sender’s betrayal.
6 Stabbing Prediction

In this section, we will look at the task of predicting whether a player would betray the other, after making certain promises. During the diplomacy phase of each season, players try to selectively convince others to form alliances. As a part of an alliance, a player may support another player’s move. However, in order to promote itself into a stronger position, a player may choose to trick the other into thinking that it would provide the support, while it may eventually not do so. This is referred to as “stabbing” and in this task, we try to uncover whether linguistic cues provide clues of whether a player is going to stab the other. We specifically refrain ourselves from looking at game-specific features, such as position of players in the game, more advantageous moves for a player and so on, and focus only on features derived from the communication messages between the players. It may be noted that this type of betrayal is different from the previously discussed ones as in this case, we look at short term betrayal — a player promises support for a particular move and we predict whether the player would keep the promise — as opposed to long term betrayal — predicting whether a player will break an alliance in the future.

Setting

The task consists of classifying pairs of players into the positive (stabbing) and the negative (assisting) class. More specifically, an assist happens when one player provides support to the other, and a stab occurs when a player promises another to provide support for a particular move (which is ordered to take place in the order phase), but the player fails to deliver its promise. Each instance of the negative (assist) class has all the messages in a season between a supporter-supported player pair for that season. If there are multiple supports by two different players for the same move, these are treated as separate instances, as the messages exchanged between them may be different. Moreover, if same pair of supporter-supported player occurs in different seasons, then they are separate instances as well, by definition. Each instance in the positive (stab) class consists all the messages between the player asking for support, and the player who promised the support, but did not make the corresponding supporting move. These cases are more difficult to pinpoint, as there is no explicit statement saying whether someone stabbed the other. For finding such instances, we use regular expressions to search the messages between the players for support of each move that was made by either of the player. If such message exists, but there is no corresponding supporting move by the player, then we consider these cases to be stabbing cases. All messages in the season between the two players comprise of a single instance in the positive class. Here as well, if support for the same move was discussed with two different players and both of them do not support the move, then both the instances are counted separately as negative instances, by the same intuition as before.

All the messages between the players are divided into “to” and “from” messages, with respect to the player asking for support in each case. In all cases, all private and group messages involving both the players are considered.

Classification and Discussion

7 Related work not already discussed

Romantic relations. Linguistic correlates of initiation (Ranganath et al., 2009) and stability (Slatcher and Pennebaker, 2006; Gottman and Levenson, 2000; Ireland et al., 2011). While this body of work mostly focuses on indications of initial fit, we focus on evolving relations. Furthermore, we consider the directed problem of betrayal.

Success and performance. (Jung et al., 2012) shows that balance is important

Conflict.

8 Conclusions

While in this paper we use one game to develop our methodology, the framework developed here can be extended to be applied to a wide range of social interaction. For example, social dynamics in collaborative settings can bear striking similarities to those present in wargames. For example, in Wikipedia edit wars—where attacks correspond to edit reverts—are common on issues relating to politics, religion, history and nationality, among others (Kittur et al., 2007). As in Diplomacy, Wikipedia editors form alliances, argue and negotiate about possible compromises.

References

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Douglas W. Oard, and Philip
Resnik. 2011. Believe me: We can do this! In The AAAI 2011 workshop on Computational Models of Natural Argument.


