ECARES

How Trump Triumphed: Multi-candidate Primaries with Buffoons

Micael Castanheira
Université libre de Bruxelles

Steffen Huck
WZB and UCL

Johannes Leutgeb
WZB

October 2020

ECARES working paper 2020-45

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

How Trump Triumphed:
Multi-Candidate Primaries with Buffoons

Micael Castanheira
ULB and FNRS

Steffen Huck
WZB and UCL

Johannes Leutgeb
WZB

and

Andrew Schotter
NYU

October 13, 2020

Abstract

While people on all sides of the political spectrum were amazed that Donald Trump won the Republican nomination this paper demonstrates that Trump's victory was not a crazy event but rather the equilibrium outcome of a multi-candidate race where one candidate, the buffoon, is viewed as likely to self-destruct and hence unworthy of attack. We model such primaries as a truel (a three-way duel), solve for its equilibrium, and test its implications in the lab. We find that people recognize a buffoon when they see one and aim their attacks elsewhere with the unfortunate consequence that the buffoon has an enhanced probability of winning. This result is strongest amongst those subjects who demonstrate an ability to best respond suggesting that our results would only be stronger when this game is played by experts and for higher stakes.

Keywords: truel, political primaries, Trump

JEL classification: C72, C92, D72, D74

1 Introduction

People on all sides of the political spectrum were amazed that Donald Trump won the Republican nomination for president of the United States in 2016. This amazement turned to complete disbelief when he was thereafter elected president. The question then arises as to how this all happened. How did a political outsider with a flamboyant personality and a flamboyant past defeat a field of well funded political veterans some with outstanding pedigrees?

This is the question we ask in this paper and the answer we provide is simple. During the early stages of the Republican primaries all the contestants viewed Mr. Trump as a buffoon, or a candidate who, if left to his own devices, would soon implode or self destruct leaving his supporters to back another candidate, hopefully them. For a mainstream candidate the strategy was to leave “The Donald” alone and use precious airtime to attack the other mainstream candidates. Why waste scarce resources attacking an opponent who is likely to implode on his own? Why not save one’s powder for others?[1]

This logic obviously failed since by the time the Republican field realized that Trump was not going to implode, it was too late and he was the clear front runner upon whom all other candidates eventually aimed their fire. Ironically, as we will see in this paper, this outcome is in line with the equilibrium prediction of a simple model where candidates in a political contest have to choose who to attack. In the equilibrium it can happen that poor candidates who are likely to self destruct—candidates to whom we shall refer as buffoons—do end up with excellent chances of winning the contest.

The paper proposes a theoretical model and uses a laboratory experiment to test the validity of our results. We model the Republican primary as a truel, a three-way duel, between conventional candidates and buffoons (i.e., candidates

[1] This clearly worked in the case of Ben Carson who, while a front runner at some point, did destroy his own chances of gaining the nomination.

with positive implosion probabilities). Each candidate has one shot with given accuracy that he can fire at any candidate he wants and if there is only one person standing at the end, that person gets the nomination. If more candidates survive they share the survival benefits equally (perhaps in a runoff duel). While we do examine the general model where each candidate is characterized by a shooting accuracy and an implosion probability, we will be most interested in the case where only one candidate has a chance of imploding — the "buffoon." As we will show, in equilibrium the other candidates do not aim at the buffoon but rather at their more stable competitors. As a result, in equilibrium, the buffoon can achieve a comparatively high survival probability, especially, if his opponents are good shooters. The buffoon may even be the player with the highest survival probability and if he does not implode the nomination is very likely to be his.

Our experimental results offer support for our theory. In those treatments where buffoons exist they tend to be attacked far less than similar non-buffoon candidates in our other treatments. In other words, as our theory predicts, if we were to transform a candidate from a non-buffoon to a buffoon in a *ceteris paribus* manner by increasing her implosion probability, that candidate draws significantly less fire from the other candidates than her non-buffoon counterpart.

This result is nuanced, however. The theory seems to work best among a subset of subjects who prove themselves most adroit at best responding to the actions of computerized competitors in a second part of our experiment. Above-median subjects are more likely not to attack the buffoon when that is the equilibrium outcome than below-median subjects. We take this result as support for the external validity of our results in that, in the real world, we would expect political candidates to hire experts to advise them and the greater sophistication of these advisors only serves to enhance the predictions of the theory.

The remainder of the paper proceeds as follows. In Section 2 we will present some background of the Republican primaries and demonstrate that Trump was indeed perceived as a candidate likely to self destruct and hence not worthy of attack. In Section 3 we discuss some of the literature related to primary elections. In Section 4 we introduce our model and discuss its equilibrium properties. Section 5 introduces our experimental design and the parameters used in our experiment and also presents a set of hypotheses to be tested. Section 6 presents our results by first testing the hypotheses in the aggregate and then by grouping subjects by their ability. Finally, in Section 7 we offer some conclusions.

2 The Background

There were 17 candidates at the beginning of the 2016 Republican primaries, each vying for the nomination. Some, like Jeb Bush, were Republican aristocrats, while others were established politicians who were current or former U.S. Senators (Rubio, Cruz, Graham, Santorum, Paul) or Governors (Kasich, Walker, Christie, Jindahl, Huckabee, Pataki, Gilmore, Bush). Candidates like Fiorina and Trump came from business backgrounds, while there were two medical doctors (Carson and Paul who was also a Senator). None, of course, had the type of broad media exposure as did Trump whose reality show “The Apprentice” was a popular success.

From the start, possibly because of his name recognition, Trump rose to front runner status. Figure 1 shows the polling status of each candidate during the main primary season from 1 January 2016 to 5 May 2016 as reported by Real Clear Politics poll averages.[2] Obviously Trump had established himself as the clear front runner very early on.

Despite this front-runner status, the other candidates spent relatively little time or money trying to burst the Trump bubble. Using data from the Political

[2] Data scraped from `https://www.realclearpolitics.com/epolls/2016/president/us/2016_republican_presidential_nomination-3823.html` (last accessed 14 September 2020)

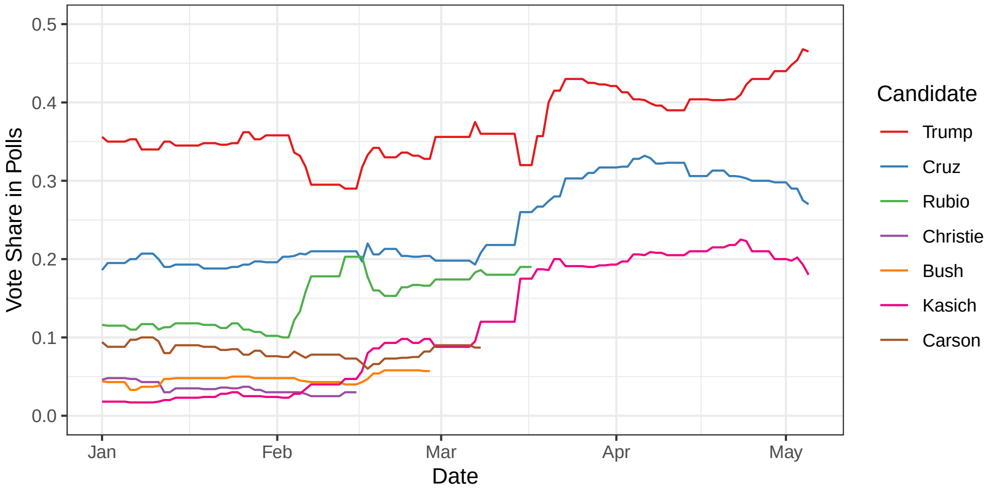


Figure 1: Average Polling Data From Real Clear Politics from 1 January 2016 to 5 May 2016

Note: In this graph we only show candidates who polled at more than 4% at any point between 1 January 2016 and 5 May 2016.

TV Ad Archive[3] (Internet-Archive, 2016), we try to back out each candidate's attack strategy across the primary campaign. We find that, in the first three primaries at least, Donald Trump was among the least attacked.

The Political TV Ad Archive is a freely accessible data base, which tracks airings of political ads in selected US TV markets. The database allows us to identify the sponsor and the candidates mentioned in the ad. First, we match TV markets to states and sponsors to candidates using publicly available information. The database has information on Donald Trump, Marco Rubio, Ted Cruz, Ben Carson, Jeb Bush, Carly Fiorina, Rand Paul, Chris Christie, Mike Huckabee, Jim Gilmore, George Pataki, Rick Santorum and John Kasich. Tables 8 and 9 in Appendix A1 show the matching the matching between TV markets and states, and between candidates and sponsors. Second, the database contains information on which candidates were mentioned in an ad. Under the assumption that the candidates will not have anything nice to say about one another we classify all ads that mention a candidate but have not been sponsored by

[3] https://politicaladarchive.org/ (last accessed 14 September 2020)

someone associated to that candidate's campaign as an attack ad. We use the database to track all airings of political attack ads 30 days before a Republican primary between January and May 2016. Furthermore, we restrict our analysis to markets in which there are at least 100 attack ads in the database and to attacks by and at candidates for which we record more than 100 attack ads in those markets in total. This leaves us with six candidates (Donald Trump, Ted Cruz, Marco Rubio, Chris Christie, Jeb Bush and John Kasich) in TV markets in 7 US states (Florida, Iowa, North Carolina, New Hampshire, Nevada, Ohio and South Carolina) with a total number of 16047 observations.

Figure 2 shows the number of attack ads directed at the various candidates 30 days before each primary. Throughout the first three primaries, despite Trump's front runner status, Marco Rubio drew the most fire from other candidates while Trump attracted little attention. It was not until the Florida primary when Rubio faced a do-or-die situation, that Trump finally drew the fire of others but by then it was starting to be too late.

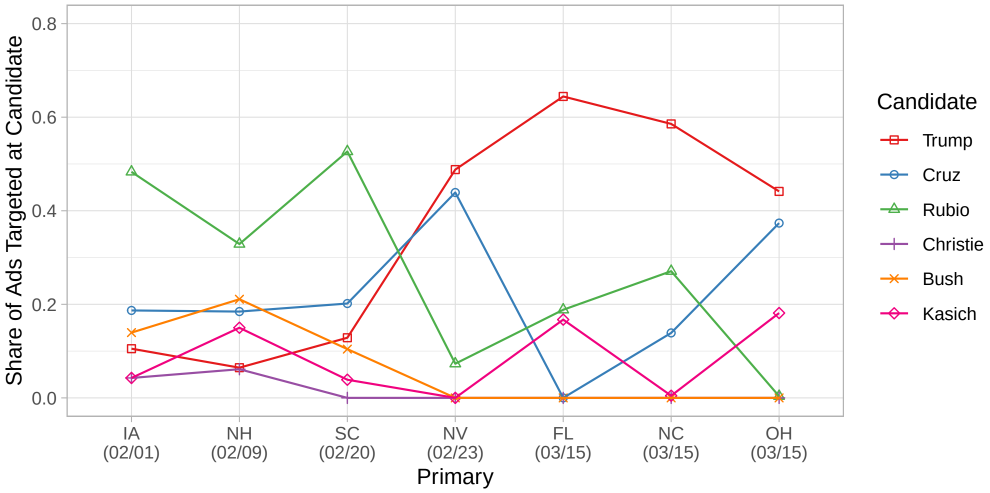


Figure 2: Attack Ads 30 Days Before Each Primary

There is consistent anecdotal evidence that this lack of attention to Trump was intentional and there are many quotes to be found that express the expectation that Trump's campaign would eventually implode. Famously, in an appearance on Fox Business on August 24, 2015, Anthony Scaramucci, who would later serve as Trump's White House Communication Director, called Trump a "political hack" who would "eventually implode."[4]

Interestingly, the idea of Trump's implosion being imminent persisted even after he had secured his party's nomination. For example, an article in the *New Statesman* in August 2016 by Woolf (2016) entitled "Is Donald Trump finally imploding? It is looking like this might finally be the end for the Republican party's ill-fated Trump experiment" summarizes the well considered view that Trump was a flash in the pan. He states:

> "Time and again, journalists and analysts expressed with great certainty that one cretinous gaffe or another would finally put Donald Trump's rickety, ridiculous, idiosyncratic campaign into a tailspin. He lied, and lied, and lied. His contempt for the truth, and for the constitution of the United States, was breathtaking – and matched only by his wilful ignorance of both. This was a bubble and, received wisdom held in the summer of 2015, it would burst soon enough."

Later in the piece, Woolf seems to see the light at the end of the tunnel. "This week, [the week commencing August 15, 2016] that longed-for tailspin seems to have finally arrived. Having goaded the news cycle, bare-facedly telling lie after madcap lie, Trump's quittance is, maybe, finally here."

Whether there was an objective high probability for Trump's self destruction or whether this was only perceived by his rivals is debatable and does not matter for the model that we will introduce below. Hanson (2016) argued in the National Review in March 2016 that it was naive perception:

> "Then there was the Republican establishment's assumption that the

[4] The clip can be found at `https://youtu.be/KZ0eqL2ZSWA` (last accessed 14 September 2020).

> supernova Trump would on its own burn out by last autumn. That was an odd expectation for a variety of reasons. It required ten debates and a winnowed-down field for other candidates finally to do to Trump what they had already done to one another. And by then the desperate level of invective needed to damage the Trump locomotive ensured that the attacker appeared as mean-spirited as Trump himself."

These quotes, which are just a few of many, all make the same point; Trump was not worth attacking. He would either implode or if not the best way to win the nomination would be to kill off one's conventional opponents, grab their support, and then face Trump one on one with the establishment lined up behind you. We know how well that worked.

3 Related Literature

The formal analysis of primary elections is extensive. Ware (2002) and Hirano and Snyder (2019) provide an outstanding historical and analytical account of primaries in the US. The first question is "why organize primaries?" Caillaud and Tirole (2002), Meirowitz (2005), Adams and Merrill (2008), Ansolabehere et al. (2010), and Hortala-Vallve and Mueller (2015) among others, argue that by stimulating intra-party competition, primaries improve the quality of the chosen candidate. Yet, primaries also feature a number of drawbacks. One is the risk of depressing the efforts of all contestants when the party trails behind (Castanheira et al., 2009) and Huck et al. (2001) identify conditions under which it pays off for a faction to send multiple contestants instead of a single one. Intra-party competition may also split (Ware, 1979) or polarize (Burden, 2001; Hirano et al., 2009) the party at the primary stage, and may require flip-flopping at the general election stage (Hummel, 2010; Agranov, 2016). All these effects hurt the party that runs primaries when facing an opponent with a pre-selected front-runner, e.g. an incumbent who can run for reelection.

A second set of questions relates to candidate behavior during a primary contest. The theoretical analysis of contests has been pioneered by Tullock (1980), Hirshleifer (1989), and Baron (1994), and is extensively surveyed by Corchón (2007) and Konrad (2009). Some noteworthy results are that contestants tend to invest more resources at attacking each other when the outcome is uncertain (Esteban and Ray, 1999; Herrera et al., 2014, 2016; Bouton et al., 2018); the equilibrium level of effort tends to be higher than with traditional incentives schemes, to the point that there can be more than full rent dissipation (Potters et al., 1998; Gradstein and Konrad, 1999), and experimental research shows that observed efforts can be even higher than theoretically predicted (Nalbantian and Schotter, 1997; Sheremeta, 2011).

In multi-candidate contests there can be positive effort to foster one's own campaign or negative effort to harm competitors. Positive effort can be seen as investing more into advertising one's own campaign (Congleton, 1986): it increases the player's chances of winning, without singling out a particular opponent. Negative effort can be seen as actively sabotaging the effort of a specific opponent (Konrad, 2000). In the case of primary elections, this translates into a so-called *negative advertising* strategy (see Lau and Rovner (2009) for an extensive survey). Negative advertising serves the purpose of reducing the support of the opponent, but sometimes at the cost of hurting the attacker's own base (Boyer et al., 2017).

Skaperdas and Grofman (1995) study both, and propose a model in which positive advertising attracts support from a pool of undecided voters, whereas negative advertising repels initially committed voters, both on the attacker's and the target's side. Hence, the candidate with the largest support has the most to lose from going negative and tends to use more positive campaign strategies in equilibrium. Conditional on going negative, the return of the attack is proportional to the initial base of the opponent. Consequently, attacks never target the opponent with the lowest support.

Our model differs in a number of ways from these established approaches. First, while the above papers model primaries as a contest, we propose that the nature of televised political debates is different: each contestant is given a fixed amount of time, which must largely be devoted to displaying the superiority of one's candidacy over the other contestants. Hence, the key strategic variable in a contest game, the amount of resources to be invested, does not feature. Rather, the main decision is *who to attack*.

Following that observation, we model the primary as a multi-lateral duel, focussing mainly on the three-player case, a *truel*. In a truel the competing players either decide simultaneously or sequentially who to shoot at with the empty set ("shooting into the air") being considered as an option in some models (Larson, 1948; Kilgour, 1971, 1975; Shubik, 1982). A typical result is that the best shooters get targeted the most, with the implication that under certain conditions bad shooters may not be wiped out by Darwinian forces (Archetti, 2012). We build on a simple simultaneous-move truel but introduce a key modification. Following the notion that a candidate's campaign may simply implode we allow this to happen in our model. Candidates have an exogenous probability to exit even if they are not attacked.

4 The Model

In this section and most of the paper we use a three-candidate static version of a truel, because such a version can easily be brought to the laboratory. Generalizations and extensions are discussed in Section 4.4. Most proofs are relegated to Appendix A2.

Consider a primary with three candidates $i \in \{X, Y, Z\}$. Each candidate has one chance to attack another candidate, say, in an ad or a statement in a debate (given the overwhelming majority of men in the above examples, we shall assign the pronoun "he" to a candidate). In line with the truel terminology and keeping open whether the attack is an ad or a statement we shall simply say that each candidate can fire one bullet at another candidate. Call the probability that candidate i hits his target the candidate's *precision* which we denote by $\pi_i \in \{\pi_X, \pi_Y, \pi_Z\}$. Note that a candidate's precision depends on his own identity only and is independent of the target's. If a targeted candidate is hit successfully, he is out of the race.

In addition to a candidate's accuracy in attacking opponents, each candidate also has a *probability of imploding* or self-destructing. The implosion probabilities are $\beta_i \in \{\beta_X, \beta_Y, \beta_Z\}$. If a candidate implodes, he is also out of the race. We order the candidates by increasing implosion probabilities: $\beta_X \leq \beta_Y \leq \beta_Z$. (In the experiment we will, in fact, focus on the case where there is only one candidate who has a strictly positive implosion probability.)

In our truel a candidate's innate strength is measured by, both, his precision as a shooter and the inverse of his implosion probability. Consequently, we think of a buffoon as a candidate who is weak on both of these dimensions – a sufficiently bad shooter (comparatively low π_i), prone to self-destruction (comparatively high β_i) – and we identify the formal condition in Section 4.2. In the model that we also test experimentally in the lab this condition boils down to $\pi_i < \beta_i$.

The game is simple. Each candidate aims his one attack at one of his opponents, that is, candidate i's strategy set is $\{X, Y, Z\} \setminus \{i\}$. Candidates are knocked out either if they are successfully hit by an attack or if they implode. If there is a single survivor, he gets a payoff of 1 (the nomination). Two survivors split the chance of being nominated, for a payoff of 1/2 each. Three survivors split it for a payoff of 1/3 each.[5]

4.1 Best Responses

One observation is crucial to derive the equilibrium of this game: as a candidate, you have no control over your own survival probability. It is entirely determined by your implosion probability and by who shoots at you, both of which you cannot influence. What you do have control over is your expected payoff *conditional* on survival. This, you can influence by aiming at the right opponent.

To identify best responses, we first compute the *counterfactual expected payoff*, Π_i^c, of a candidate i before he aims at either of his opponents. Let p_i be candidate i's exit probability given his opponents' choices, their precisions and i's own implosion probability. Let p_j^{-i} be candidate j's *counterfactual exit probability* that would result were candidate i not to attack anybody. These counterfactual exit probabilities depend on both other candidates' implosion probabilities and on whom they target. We obtain:

$$\begin{aligned}\Pi_i^c / \left(1 - p_i\right) = \quad & p_j^{-i} p_k^{-i} \times 1 + \left[p_j^{-i}\left(1 - p_k^{-i}\right) + \left(1 - p_j^{-i}\right) p_k^{-i}\right] \times \tfrac{1}{2} \\ & + \left(1 - p_j^{-i}\right)\left(1 - p_k^{-i}\right) \times \tfrac{1}{3},\end{aligned} \tag{1}$$

where the first term of the right-hand side is the probability that i remains alone in the race, and gets the nomination (value: 1). The second term is the probability that i and a single opponent remain in the race, multiplied by a split of the payoff in two. Finally, the third term is the probability that all remain in the race and split the payoff three ways. On the left-hand side we divided by i's own survival probability as the payoff only obtains conditional on survival.

If i chooses to attack opponent j, his expected payoff increases relative to

[5] If all candidates get eliminated, everyone receives a payoff of zero. In terms of a primary, think of this case as the party choosing an external candidate over any of the eliminated candidates.

the counterfactual expectation above by:

$$\begin{aligned}\Delta\Pi_i(j)/\left(1-p_i\right) &= \pi_i \times \left[\left(1-\tfrac{1}{2}\right)\left(1-p_j^{-i}\right)p_k^{-i}+...\right.\\ &\left. ...\left(\tfrac{1}{2}-\tfrac{1}{3}\right)\left(1-p_j^{-i}\right)\left(1-p_k^{-i}\right)\right], \qquad (2)\end{aligned}$$

where the right-hand side is the probability of successfully hitting j when either of two situations realize. The first summand is when the field of survivors is initially composed of i and j only. Then, successfully hitting j implies that i gets the nomination with probability 1 instead of 1/2. The second summand is when all three candidates initially survive. Then, successfully hitting j reduces the field of survivors from 3 to 2, and increases the probability of nomination from 1/3 to 1/2.

Candidate i will optimally attack candidate j if:

$$\begin{aligned}&\tfrac{1}{2}\left(1-p_j^{-i}\right)p_k^{-i}+\tfrac{1}{6}\left(1-p_j^{-i}\right)\left(1-p_k^{-i}\right)\geq ...\\ &...\geq \tfrac{1}{2}\left(1-p_k^{-i}\right)p_j^{-i}+\tfrac{1}{6}\left(1-p_k^{-i}\right)\left(1-p_j^{-i}\right), \qquad (3)\end{aligned}$$

which simplifies to $p_j^{-i} \leq p_k^{-i}$. In other words, a candidate simply shoots at the opponent whose counterfactual exit probability (absent an attack from i) is minimal. Notice that this result is independent of the payoffs that the player obtains for sharing survival with one or two others—both payoff increments, 1/2 and 1/6, could be replaced by any other positive value. Hence, the result holds independently of whether a successful attack drives an exit (as we assume here) or only marginally hurts the opponent's odds of winning. For the same reason, it is also independent of player i's risk attitude:

Lemma 1 *The best-response of player i is always to aim at the candidate with the highest survival probability in the absence of his own attack, that is, i aims at j iff* $p_j^{-i} \leq p_k^{-i}$. *This holds irrespective of player i's risk attitude.*

Lemma 1 is intuitive and already contains the core of the point we are making in this paper: candidates may be shielded from an attack if they are prone to implosion. Hence, increasing the implosion probability for a candidate has two countervailing effects. On the one hand, it hurts the candidate as he does become more likely to actually implode but on the other hand, he becomes a less attractive target for his opponents and the latter (indirect) effect can be strong enough such that nobody shoots at the implosion-prone candidate. Consequently, a player's expected equilibrium payoff can be locally increasing in his implosion probability. We will see this in detail below.

4.2 Equilibrium Analysis

Let us now turn to the search for pure-strategy equilibria in our truel game. An equilibrium will be a constellation of targets such that each candidate's target is a best response to the targets of his opponents. Since every player has to choose between two options, there are $2^3 = 8$ possible strategy combinations: (*i*) two circular cases: $X \to Y,\ Y \to Z,\ Z \to X$ and $X \to Z,\ Y \to X,\ Z \to Y$, where $i \to j$ means "i targets j"; and (*ii*) six cases in which two candidates target each other, and the third candidate shoots at one of these two. For instance: $X \to Y,\ Y \to Z,\ Z \to Y$.

W.l.o.g., let us order X, Y and Z from the lowest to the highest probability of implosion: $\beta_X \leq \beta_Y \leq \beta_Z$. The application of Lemma 1 immediately rules out four of the six non-circular candidate equilibria, namely those where X and Y are not attacked. The intuition is as follows: consider one of the other configuration where players $i \in \{X, Y\}$ and player Z attack each other while player $j \in \{X, Y\} \setminus \{i\}$ is unscathed. In such a configuration player i violates the best response condition from Lemma 1 as player Z must have a lower survival probability than player j prior to i's choice.

We are thus left with only four candidate equilibria: the two circular ones, labeled Cases 1 and 2 below, and two in which X and Y target each other, labeled Cases 3 and 4. These are depicted in Table 1, together with their associated best-response conditions.

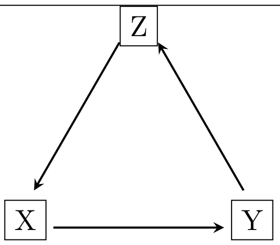
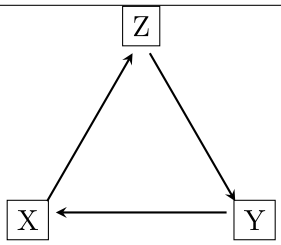
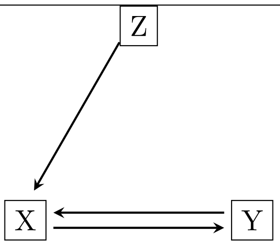
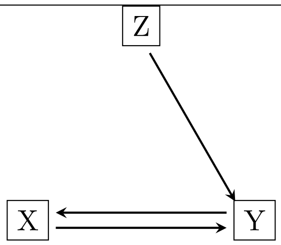
 $BR_X : \frac{\beta_Y - \beta_Z}{1-\beta_Z} < \pi_Y$ (always true) $BR_Y : \frac{\beta_Z - \beta_X}{1-\beta_X} < \pi_Z$ $BR_Z : \frac{\beta_X - \beta_Y}{1-\beta_Y} < \pi_X$ (always true) CASE 1	 $BR_X : \frac{\beta_Y - \beta_Z}{1-\beta_Z} < \pi_Z$ $BR_Y : \frac{\beta_X - \beta_Z}{1-\beta_Z} < \pi_X$ (always true) $BR_Z : \frac{\beta_Y - \beta_X}{1-\beta_X} < \pi_Y$ CASE 2
 $BR_X : \beta_Y < \beta_Z$ (always true) $BR_Y : \frac{\beta_Z - \beta_X}{1-\beta_X} > \pi_Z$ $BR_Z : \frac{\beta_Y - \beta_X}{1-\beta_X} + \frac{(1-\beta_Y)\pi_X}{1-\beta_X} > \pi_Y$ CASE 3	 $BR_X : \frac{\beta_Z - \beta_Y}{1-\beta_Y} > \pi_Z$ $BR_Y : \beta_X < \beta_Z$ (always true) $BR_Z : \frac{\beta_Y - \beta_X}{1-\beta_X} + \frac{(1-\beta_Y)\pi_X}{1-\beta_X} < \pi_Y$ CASE 4

Table 1: Equilibrium Conditions
Note: Arrows denote attacks. $X \to Y$ means "X attacks Y"

We are, of course, particularly interested in Cases 3 and 4 where Z, the candidate with the maximal implosion probability, is not attacked. As can be seen in Table 1, there is another necessary condition for this configuration:

$$\pi_Z < \max\left\{\frac{\beta_Z - \beta_X}{1-\beta_X}, \frac{\beta_Z - \beta_Y}{1-\beta_Y}\right\}. \tag{4}$$

In other words, to be left unattacked, the self-imploding candidate must also be a sufficiently bad shooter in comparison with his implosion probability.[6] If

[6] Note that only Z may remain unscathed, even if X and Y were also buffoons. Also, the result would remain identical if there was uncertainty about the exact values of β_Z: the latter would be replaced by its expected value in (4).

condition (4) holds we shall call candidate Z a *buffoon* and we shall refer to the equilibria where the buffoon remains unattacked as *buffoon equilibria.*

From now on, up until Section 4.4, we shall focus on the setup that we bring to the lab: $\beta_X = \beta_Y = 0$, that is, there is only one candidate prone to self-destruction. Notice that, in this setup, condition (4) boils down to:

$$\pi_Z < \beta_Z.$$

That is, player Z is a buffoon if his implosion probability is bigger than his precision. Combining the BR conditions from each of the four Cases in Table 1 yields our first proposition (all proofs are in Appendix A2):

Proposition 1 *For $\beta_X = \beta_Y = 0 < \beta_Z$, circular and buffoon equilibria are mutually exclusive.*

For $\pi_Z > \beta_Z$, the two circular equilibria coexist.

For $\pi_Z < \beta_Z$, there is a unique, buffoon, equilibrium:

- *with $\pi_X > \pi_Y$, the equilibrium is $X \to Y,\ Y \to X,\ Z \to X$,*
- *with $\pi_X < \pi_Y$, the equilibrium is $X \to Y,\ Y \to X,\ Z \to Y$.*

Figure 3 summarizes the proposition graphically. As stated in the proposition, the structure of the equilibrium depends on whether or not player Z is a buffoon. If he is not, that is, if he can at least shoot reasonably well, there are just the two circular equilibria that also arise when nobody else may implode. The reason is straightforward in light of Lemma 1: since Z shoots at either X or Y, that candidate's counterfactual survival probability is $(1 - \pi_Z)$, which is lower than Z's. Hence, Z must be shot at in equilibrium.

Conversely, if Z is a buffoon, then there exists only a buffoon equilibrium and which one emerges depends simply on whether X or Y is the better shot. If it is candidate X, the buffoon will shoot at X, if it is Y, he will shoot at Y. Notice

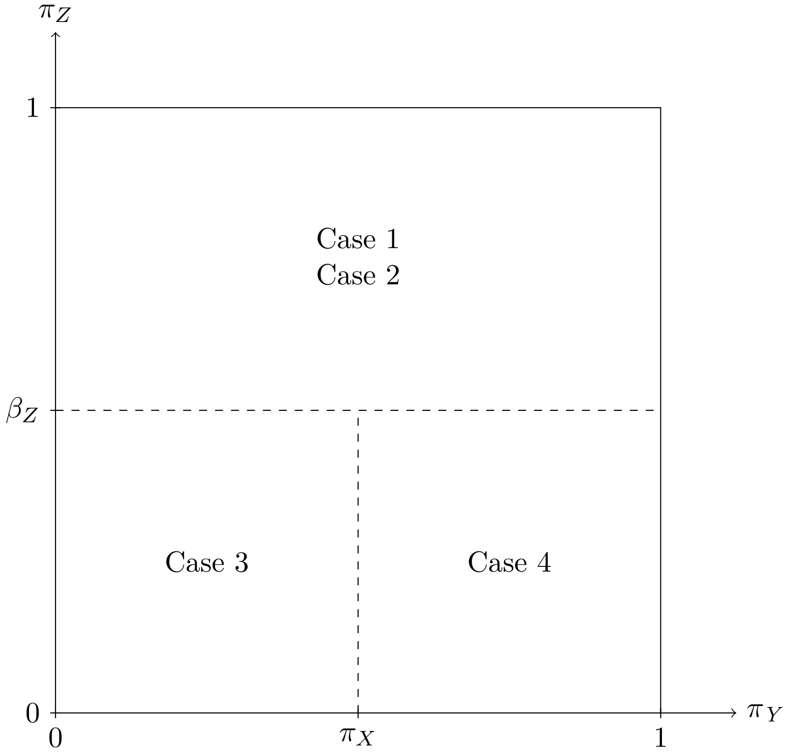


Figure 3: Equilibria when $\beta_X = \beta_Y = 0$

also that the truel, once Z becomes a buffoon, is dominance solvable. Given that $\beta_Z > \pi_Z$ shooting X becomes a dominant action for Y and vice versa. After that first round of elimination, Z's decision is determined by dominance in the second round.

In light of Proposition 1 it is informative to look back at Figures 1 and 2 in Section 2 and try to connect its predictions to what happened in the primaries.

Trump was ahead in the polls from the very beginning but despite his lead he was initially not attacked. The punchline of our paper is to point out that this was not an anomaly but an equilibrium outcome. Even as a front runner he was not to attacked because of the common belief that he would implode. Attacks on him would be wasted money. This perception started to change at the end of February when it dawned on campaign advisers that Trump was here to stay and was not going to implode, at which point we see a steady rise in attacks going into the Nevada and Florida primaries. Interestingly, this shift happens before Trump's support in opinion polls started to take off. If anything, the only movement in the polls was increasing support for Rubio and Kasich. Based on the model, our interpretation is that campaign professionals realized earlier than many that Trump's implosion probability was not as high as initially thought. Around the end of February Trump's perceived implosion probability became sufficiently low, i.e., $\beta < \pi$, and he started to draw fire form other candidates but, as we know with hindsight, that was too late.

4.3 Comparative statics

We can now examine how different types of players perform in truels. We are still focusing on the case with $\beta_X = \beta_Y = 0$ in this section.

Corollary 1 *Z's survival probability is generically decreasing in β_Z, except at $\beta_Z = \pi_Z$, where it displays a discontinuous upward jump.*

The most important part of this corollary is, of course, in the exception: the moment candidate Z turns into a buffoon, his survival probability jumps upwards as the players move from a circular equilibrium to a buffoon equilibrium, and Z no longer gets shot at. Notice that this jump can be big: when both players X and Y are good shooters (with precisions above Z's implosion probability) candidate Z's survival probability will reach a global maximum for an interior implosion probability.

This is illustrated in Figure 4 which, for a specific set of parameters, also plots the other players' survival probabilities as a function of β_Z and we assume coordination on the $X \to Y,\ Y \to Z,\ Z \to X$ equilibrium for low values of β_Z. At $\beta_Z = \pi_Z$, Z's probability of survival jumps from $(1 - \pi_Y)(1 - \beta_Z) \approx 0.36$ to $(1 - \beta_Z) = 0.65$. Notice also the dramatic fall in Y's survival probability once he gets targeted by two players instead of one. The same exercise can, of course, be

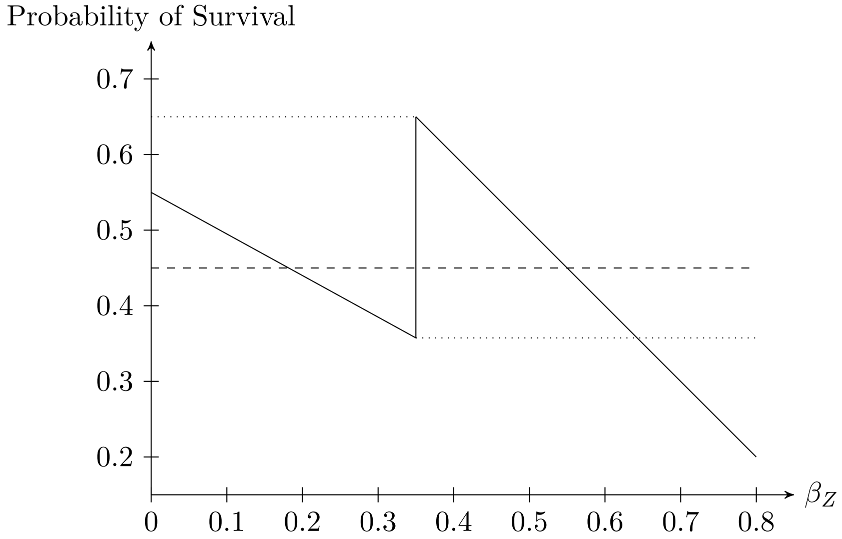


Figure 4: Buffoon Equilibria: Survival Probabilities as a Function of β_Z
Survival probabilities of X (dotted), Y (dashed) and Z (solid) as β_Z varies from 0 to 0.8, for $\pi_X = .55,\ \pi_Y = .45,$ and $\pi_Z = .35$.

done varying Z's precision π_Z. Then, we would see that his survival probability jumps upwards when his precision drops below his implosion probability.

It is also interesting to think about variations in the precision of one of the other players when we are in a buffoon equilibrium. As an illustration, we plot the survival probabilities for all three players as a function of player Y's precision in Figure 5. The buffoon's survival probability is, of course, independent of Y's precision as nobody shoots at Z. Candidate X's survival falls almost everywhere in Y's precision but jumps up discretely once Y's precision exceeds his own—simply because Z now changes his target from X to Y. Consequently, Y's own survival probability drops. X and Y essentially swap places at this point.

We can state in a second corollary under which conditions the buffoon is the candidate with the highest survival probability. Labeling X as the candidate with the highest precision ($\pi_X > \pi_Y$):

Corollary 2 *In a buffoon equilibrium (i.e. for $\beta_Z > \pi_Z$), candidate Z has the highest survival probability iff $\beta_Z < \min\{\pi_X, \pi_Y + (1 - \pi_Y)\pi_Z\}$.*

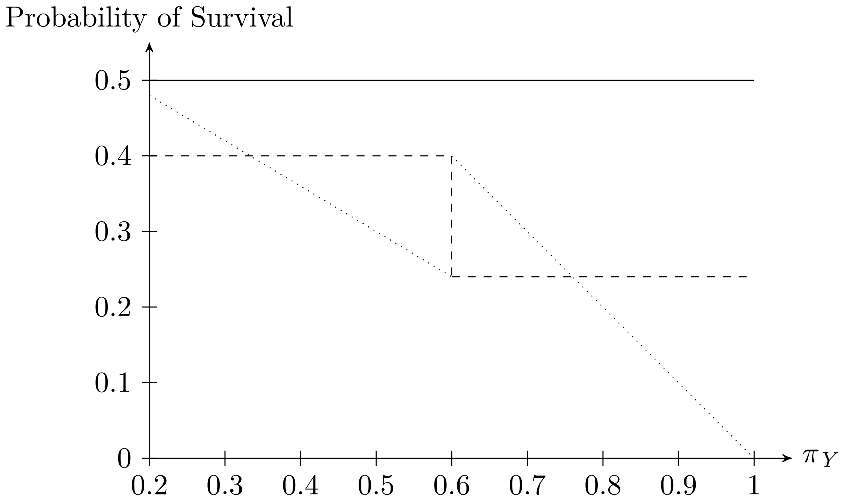


Figure 5: Buffoon Equilibria: Survival Probabilities as a Function of π_Y
Survival probabilities of X (dotted), Y (dashed) and Z (solid) as π_Y varies from 0.2 to 1, when $\pi_X = 0.6$, $\pi_Z = 0.4$, $\beta_Z = 0.5$.

A simple sufficient condition would be $\pi_Z < \beta_Z < \min\{\pi_X, \pi_Y\}$ which encapsulates the irony of our equilibrium results. In particular, when the buffoon competes against proper pro shooters, he emerges as the most likely candidate to win the contest. We believe that this encapsulates the core of what happened at the 2016 Republican primaries. Trump did not triumph despite competing against some excellent other candidates. He triumphed *because* he was a buffoon playing against real pros.

4.4 Extensions: mixed strategies and generalized setup

Before we move on to the experiment there are at least four open questions that require some attention. First we have, so far, focused on pure-strategy equilibria and we have to ask what role mixed-strategy equilibria might play. Second, we need some reassurance that our results do extend to a setup in which multiple candidates may implode. Third, actual primaries are held in multiple rounds: how do such dynamics modify best responses? Finally, we need to verify that our results are not an artefact that only arises with three players.

Mixed strategy equilibria

For a (fully) *mixed-strategy equilibrium* we need:

$$p_j^{-i} = p_h^{-i} \quad \forall i \text{ and } j \neq h. \tag{5}$$

That is, from player i's perspective the other two players are equally likely to survive in the counterfactual scenario where i himself shoots in the air. In this case, he is indifferent between his two targets and may mix.

For the case in which all implosion probabilities are zero, that is, for the standard truel, we obtain:

$$q_{ij} = \frac{\pi_i + \pi_j - \pi_h}{2\pi_i}, \tag{6}$$

where q_{ij} denotes the probability with which player i targets player j.

The existence of a fully mixed equilibrium requires relatively balanced shooting precisions. The precision of the best shooter must not be larger than the sum of the precisions of the other two players. Suppose this condition is met (as it will be in our experiment) and we introduce the possibility to self-destruct for player Z. How will that change the mixed-strategy equilibrium? Consider all players sticking to the equilibrium mixtures above while β_Z becomes positive. Clearly, to maintain the indifference of player X either player Y must decrease q_{YZ} or player Z must increase q_{ZY}. Similarly, to maintain the indifference of player Y either player X must decrease q_{XZ} or player Z must increase q_{ZX}. Notice that Z cannot do his part in both cases and also note that the Z indifference condition has not changed which implies that, as before, we must have:

$$q_{XY}\pi_X = q_{YX}\pi_Y. \tag{7}$$

Hence, q_{XY} and q_{YX} can only change keeping their proportion fixed. The above reasoning rules out that they both decrease as otherwise, both, q_{ZX} and q_{ZY} would have to increase as well, which is impossible as they sum to 1. Hence, as a consequence of Z starting to implode, X and Y must shoot with higher probability on each other. In other words, as Z's implosion probability increases, a fully mixed equilibrium necessarily moves closer to one of the buffoon equilibria.

The general description of fully mixed or degenerate mixed strategy equilibria is trivial but tedious and omitted for space reasons. But we shall derive them further below for the parameter settings that we implement in the experiment.

Generalized implosion probabilities

In this section, we extend the analysis to the case of all candidates facing a risk of implosion: $0 < \beta_X < \beta_Y < \beta_Z$. The intuition becomes less straightforward than in the simple setup. The equilibrium structure is summarized in Figure 6 below but the proof is relegated to Appendix A2. As illustrated, seven regions of relevance emerge, instead of three in Figure 3. Note that the three initial regions remain: they are the *Case 1, Case 2* box at the top right of the figure, and the two boxes at the bottom (*Case 3* and *Case 4*).

Going from $\beta_X = \beta_Y = 0$ to $\beta_Y > \beta_X > 0$ thus adds four new regions, three of which emerge for intermediate values of π_Z, namely when condition (4) is satisfied ($\pi_Z < \max\left\{\frac{\beta_Z-\beta_X}{1-\beta_X}, \frac{\beta_Z-\beta_Y}{1-\beta_Y}\right\}$), but π_Z is larger than the minimum of these two fractions ($\pi_Z > \min\left\{\frac{\beta_Z-\beta_X}{1-\beta_X}, \frac{\beta_Z-\beta_Y}{1-\beta_Y}\right\}$). In this new intermediate zone, the necessary condition for a buffoon equilibrium is satisfied but the equilibrium need not be dominance solvable. The intuition for the proof is as follows: consider first π_Y close to 1, in which case a buffoon equilibrium cannot exist, even though Z is a buffoon. Why? First, note that the candidate targeted by Y is almost certain to exit. Hence, one of the other candidate's best response will be to aim at the opponent not targeted by Y. Consider for instance the situation in which Y shoots at X. Then, Z's best response is to shoot at Y, and X's to

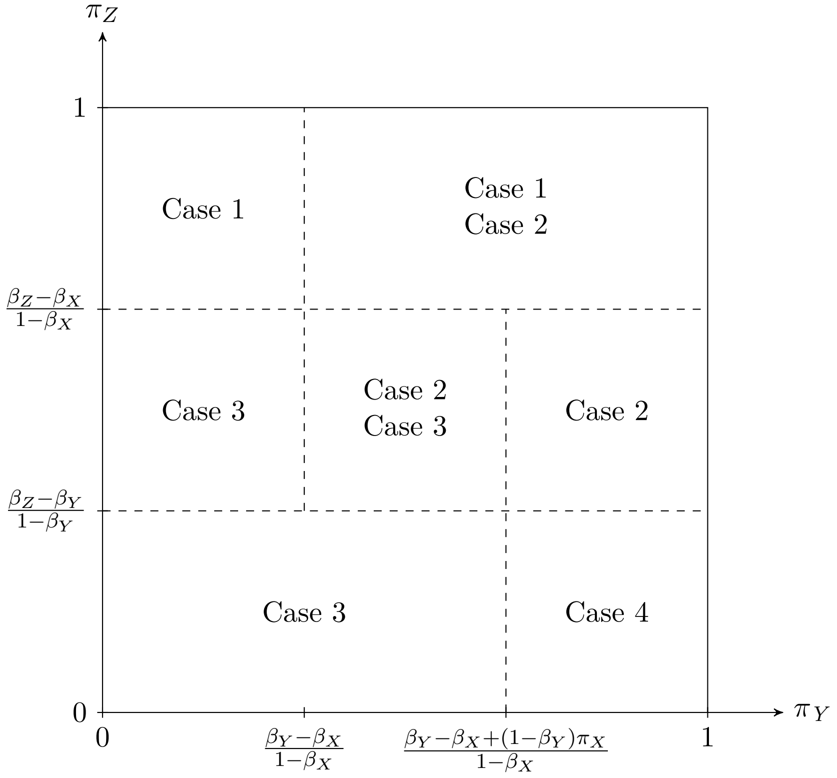

π_Z
1
$\frac{\beta_Z-\beta_X}{1-\beta_X}$
$\frac{\beta_Z-\beta_Y}{1-\beta_Y}$
0
Case 1
Case 1
Case 2
Case 3
Case 2
Case 3
Case 2
Case 3
Case 4
0
$\frac{\beta_Y-\beta_X}{1-\beta_X}$
$\frac{\beta_Y-\beta_X+(1-\beta_Y)\pi_X}{1-\beta_X}$
1
π_Y

Figure 6: Equilibria with non-zero β_X and β_Y

target Z: there is no buffoon equilibrium. By contrast, when π_Y is close to 0, the equilibrium is dominance solvable, and there is a buffoon equilibrium. In between these extremes, i.e. for intermediate values of π_Y, Z's best response is contingent on Y's action, giving rise to the coexistence of one circular and one buffoon equilibrium. Last, a new region appears at the top left, for π_Z sufficiently close to 1 and π_Y sufficiently close to 0. Only one of the circular equilibria survives in this case.

Introducing Dynamics

The Republican primaries described in Section 2 have a clear dynamic component; they are a multistage game. While characterizing the equilibria of such a dynamic game is beyond the scope of this paper, we show here how Lemma 1 generalizes to a two-period game. At time 1, the three contestants simultaneously choose a target, and nature selects which of them make it to the second primary. There are thus four possible outcomes of the first primary: if the three contestants survive the first round, the second primary is a repeat of the first one (this subgame is the same as in base model). If only two contestants do, the second primary is a simple duel – with two such duels possible. If only one contestant does, he wins the nomination outright.

In Appendix A2, we show that a candidate j gets less attacked in the first primary than in the second if and only if his implosion probability is high enough compared to his precision. The exact condition is:

$$\frac{1-\pi_k}{1-\pi_j} < \frac{2-(1-\beta_j)(1-\pi_i)}{2-(1-\beta_k)(1-\pi_i)}.$$

For equal precisions, the condition boils down to $\beta_j > \beta_k$. That is, our results in the static game get reinforced in the early stage of a two-period game.

One element is very intuitive: each candidate prefers to run against opponents with a high probability of implosion and a low precision (the later increases your own survival probability). Perhaps less straightforward, the effect of these two characteristics is different early and late in the race: late in the race, you lose the possibility to select your opponents. As in Lemma 1: your best response then only depends on your opponents' counterfactual exit probabilities. Early in the race, instead, you can manipulate against whom you'll be competing in the future: then, also their precision matters. Eliminating the best shooters early makes subsequent primaries less dangerous.

Larger number of players

Finally, let us consider what happens with *more than three players.* As shown in Appendix A2 generalizing Lemma 1 is straightforward: as discussed already, the results of Lemma 1 do not depend on the exact value of the realization payoffs

(1/3, 1/2, and 1). When comparing targeting opponent X vs. Y, these values can be replaced by the expected value of the prize being divided among any number of competitors. Two-by-two comparisons between potential targets yield the same conclusion: you must target the opponent with the lowest counterfactual exit probability. Iterating two-by-two comparisons proves the result.

Generalizing Proposition 1 is trickier: with N players, there are $(N-1)^N$ possible combinations. We thus focus on identifying a sufficient condition for the existence of a buffoon equilibrium. Consider a population of N candidates, with n_X "professionals", whose implosion probability is lower than their precision, and $n_Z := N - n_X$ buffoons: $\forall z_i \in \mathbb{Z} = \{z_1, ..., z_{n_Z}\}$, $\pi_{z_i} < \beta_{z_i}$. Moreover, $\min_{z_i} \beta_{z_i} > \max_{x_i} \beta_{x_i}$ and $\max_{z_i} \pi_{z_i} < \min_{x_i} \pi_{x_i}$ (see Appendix A2). That is, buffoons are uniformly worse candidates. Finally, $n_Z \leq n_X$: there are fewer buffoons than pros. In that setup, the sufficient condition for a buffoon equilibrium to exist is similar to what identifies the bottom-left region of Figure 6:

Proposition 2 *With n_X "professionals" and n_Z buffoons, a sufficient condition for the existence of an equilibrium in which no buffoon gets targeted is:*

$$\max_{x \in \mathbb{X}} \pi_x \leq \min_{x \in \mathbb{X}, z \in \mathbb{Z}} \frac{\beta_z - \beta_x}{1 - \beta_x}.$$

As detailed in the proof, this is actually a sufficient condition to satisfy another sufficient condition. But it has the advantage of being easy to interpret: with $n_Z \leq n_X$, all the pros get targeted by one other pro, and some of them get *also* targeted by one buffoon. When the precision of all opponents does not exceed the threshold identified in the proposition, the counterfactual exit probability of at least one of the fine candidates must be larger than that of a buffoon who hasn't been targeted by anyone. Note that such an equilibrium is no longer dominance solvable: if all candidates were targeting a single opponent, the latter's counterfactual exit probability must converge to zero as population size increases.

5 Experiment

5.1 Experimental Setup

The experiment was run at the WZB-TU laboratory at Technical University Berlin in January and February 2020. In total 6 sessions took place for which 129 subjects were recruited from the laboratory's subject pool using ORSEE (Greiner, 2015). Subjects were students from a wide range of fields who had been studying for 3.6 semesters and were 22.2 years old on average. 68% of subjects were male, 30% were female and 2% were other. The program was written in oTree (Chen et al., 2016). The experiment lasted for one hour and subjects earned, on average, 13.45 Euro plus a show-up fee of 7 Euro.

The experiment had two parts. Subjects first received the instructions for Part 1 of the experiment. They were made aware that there would be a second part, but that Part 2 was independent from their choices in Part 1. After Part 1 concluded subjects received instructions for Part 2 without any feedback about outcomes and payoffs from Part 1. The experiment was run in German. An English translation of the instructions is provided in Appendix A3. Feedback for the results of both parts was only provided after Part 2.

In Part 1 subjects played the game against random human opponents. In Part 2 of the experiment we removed the strategic uncertainty of playing against human opponents and subjects played eight games against computerized opponents with a transparent strategy instead. We randomized the order in which the treatments were presented to subjects in both parts. Subjects finished the experiment by filling out a form and, finally, got paid privately in cash.

5.2 Experimental design and hypotheses testing

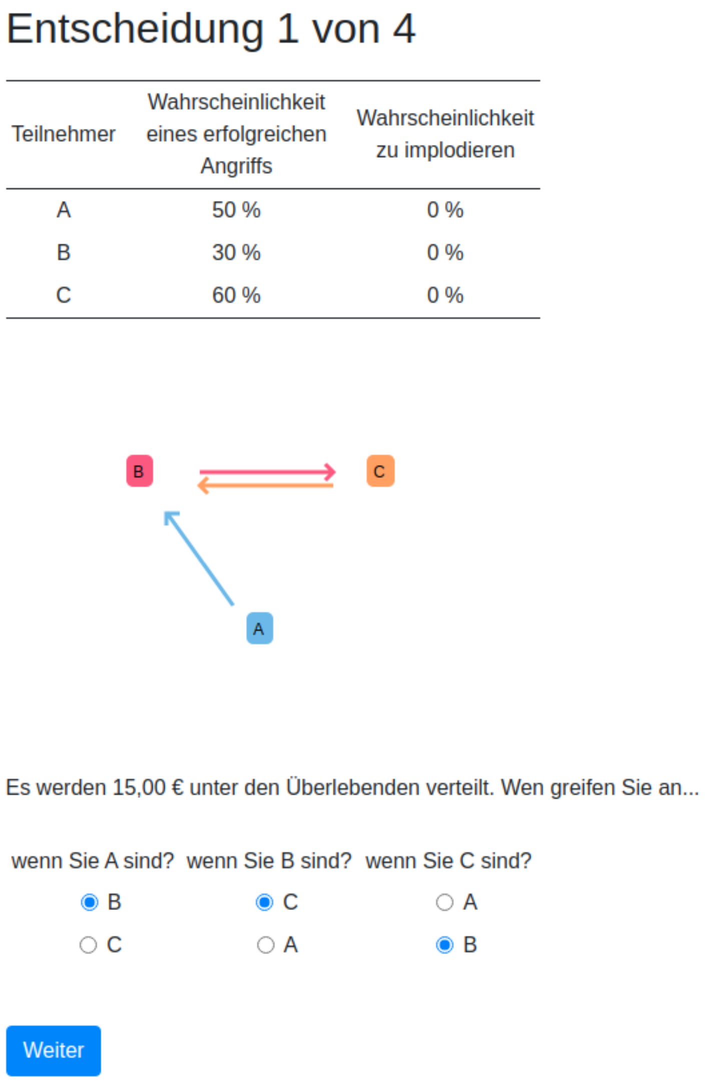
5.2.1 Part 1

In part 1 subjects played a truel with players $\{X, Y, Z\}$ against other human subjects and we asked subjects to make choices for all three roles.[7] There were four treatments with different parameter sets, shown in Table 2. Subjects faced each of these four treatments in random order. For each treatment, subjects had to choose targets for all three player roles. Subjects could make the decisions for all three roles in any order they wanted and they could go back and forth and revise their choices. Only when they were happy with all three decisions would they submit their strategies. The interface is shown in Figure 7.

There was no feedback after any of these choice triplets. Instead, payoffs were determined at the end of the experiment. One parameter set was chosen at random and subjects were grouped into sets of three with their roles $\{X, Y, Z\}$ also being randomly chosen. The choices that they specified for relevant treatment and role were then implemented and the random variables realized, i.e., the random realizations that determined whether the players would hit their targets and whether they would implode. There was a 15 Euro prize per group that was split equally among the survivors of the implemented truel. Finally, subjects were informed which parameter set was chosen, which role they were assigned to, what the choices of their opponents were, and whether they and their opponents survived. Subjects received this feedback only after completing Part 2. By not providing feedback until the end of the experiment we avoid possible confounds from learning and maximize the number of independent observations.

We chose four parameter settings to put the theory through its paces—and making only one parameter change at the time. In treatment Baseline all implosion probabilities are zero. The players are asymmetric in their attack

[7] In the experiment the players were labelled $\{A, B, C\}$, but in the analysis we will continue to label them $\{X, Y, Z\}$.



Entscheidung 1 von 4

Teilnehmer	Wahrscheinlichkeit eines erfolgreichen Angriffs	Wahrscheinlichkeit zu implodieren
A	50 %	0 %
B	30 %	0 %
C	60 %	0 %

Es werden 15,00 € unter den Überlebenden verteilt. Wen greifen Sie an...

wenn Sie A sind?	wenn Sie B sind?	wenn Sie C sind?
● B	● C	○ A
○ C	○ A	● B

Figure 7: Interface in Part 1
In the experiment, players were labelled $\{A, B, C\}$ instead of $\{X, Y, Z\}$. Subjects see a table summarizing the parameters on top. In the center they see a graph visualizing their attack choices that automatically updates when they change their choices in using the radio buttons for each role at the bottom. The text in the screenshot reads "The survivors split 15 Euro. You are A. Who do you attack... if you are A (B or C)."

probabilities but there are only circular pure-strategy equilibria in which everybody gets attacked. In addition, there is a fully mixed-strategy equilibrium with $q_{XY} = 1/5$, $q_{YZ} = 2/3$, and $q_{ZX} = 2/3$.

In Circular we increase Z's implosion probability slightly to 0.25. Since Z's precision is still greater than his implosion probability the pure strategy equilib-

Treatment	π_X	π_Y	π_Z	β_X	β_Y	β_Z
Baseline	0.5	0.3	0.6	0	0	0
Circular	0.5	0.3	0.6	0	0	0.25
BuffoonZX	0.5	0.3	0.6	0	0	0.7
BuffoonZY	0.5	0.8	0.6	0	0	0.7

Table 2: Treatment Parameters, Part 1

ria remain the same as in Baseline. Player Z is not yet a buffoon. However, in line with the prediction that we developed above, the probabilities with which X and Y attack Z in the mixed strategy equilibrium fall and they do so substantially. In fact, the mixed-strategy equilibium becomes degenerate as Y completely ceases to attack Z. The equilibrium mixtures are $q_{XY} = 3/5$, $q_{YZ} = 0$, and $q_{ZX} = 7/12$.

Finally, in treatments BuffoonZX and BuffoonZY we increase Z's implosion probability β_Z to 0.7 so that he becomes a buffoon who should be ignored by X and Y, who should instead attack each other. There is only one equilibrium in both treatments and this equilibrium is in pure strategies. In treatment BuffoonZX X's precision of 0.5 is higher than Y's precision of 0.3, so Z attacks X, and the unique equilibrium is $X \rightarrow Y$, $Y \rightarrow X$, $Z \rightarrow X$. In treatment BuffoonZY we increase Y's attack probability to 0.8, so Z now attacks Y and the unique equilibrium is $X \rightarrow Y$, $Y \rightarrow X$, $Z \rightarrow Y$.

We can now turn to developing some hypotheses for Part 1 of the experiment. The most fundamental prediction of the theory is that not attacking player Z is a dominant strategy for both player X and player Y, when player Z is a buffoon. In contrast, in both Baseline and Circular, every equilibrium has positive probability for Z being attacked. A weak version of this is summarized in:

Hypothesis 1 (Dominance) *Compared to Baseline and Circular, player Z does get attacked less often by X and Y in BuffoonZX and BuffoonZY.*

Examining the mixed-strategy equilibria we can also make predictions for differences in X's and Y's attack rates between Baseline and Circular, and Circular and BuffoonZX.

Hypothesis 2 (Mixed Nash) *The frequency with which Z gets attacked falls from Baseline to Circular and it falls further from Circular to BuffoonZX.*

Our final hypothesis deals with player Z who is predicted to shoot player X in BuffoonZX and player Y in BuffoonZY. This does not result from outright dominance but iterated elimination of dominated strategies or equilibrium reasoning.

Hypothesis 3 (Pure Nash) *Z attacks Y more in BuffoonZY than in BuffoonZX.*

5.2.2 Part 2

Part 2 of the experiment is designed to investigate behavior at the individual level more closely and see how subjects respond to the behavior of their opponents. Subjects are assigned the role of player X and they play against two computer players who are already committed to an action. The prize is again 15 Euro, but all money that the computer players earn is retained by the experimenter and not paid to any of the subjects. Subjects make eight choices in a random order without any feedback. This transforms the game into a series of decision problems such that the indeterminacy of what constitutes rational behavior that characterizes much of Part 1 disappears. In Part 1, both multiplicity and off-equilibrium beliefs may justify different actions in eight out of the twelve decisions. Only in BuffoonZX and BuffoonZY are the best responses of players X and Y belief-independent.

Of course, we could have opted to elicit beliefs in Part 1 but that would have required an additional 24 decisions and the usual measurement issues. So instead we decided to fix beliefs in a second part of the experiment by letting subjects play against computer players who play fully transparent strategies. An added benefit of this approach is that it also eliminates any potential confounds from social preferences.

		Computer		Prediction
Set	β_Z	q_{YZ}	q_{ZX}	q_{XY}
1	0.25	0	0	0
2	0.25	0	1	1
3	0.25	1	0	0
4	0.25	1	1	1
5	0.7	0	0	1
6	0.7	0	1	1
7	0.7	1	0	1
8	0.7	1	1	1

Table 3: Part 2 Treatment Parameters and Predictions
Note: Subjects play as player X. q_{ij} denotes the probability with which player i targets player j. Column Prediction shows the rational theory's prediction. $\pi_X = 0.5$, $\pi_Y = 0.3$, $\pi_Z = 0.6$ and $\beta_X = \beta_Y = 0$ remain constant.

We decided to re-use the parameters of the Circular and BuffoonZX treatments from Part 1. In all parameter sets subjects are assigned the role of player X, and they encounter all possible combinations of actions of computer players Y and Z. We vary only β_Z and the computers' actions, and keep $\pi_X = 0.5$, $\pi_Y = 0.3$, $\pi_Z = 0.6$ and $\beta_X = \beta_Y = 0$ constant. This generates eight decision problems in an otherwise stable environment. For each of these, expected utility maximization makes a unique prediction (for arbitrary monotone von Neumann Morgenstern utility functions, see the discussion of Lemma 1 above). Table 3 summarizes the parameters and predictions for each decision.

6 Results

Let us start by investigating the hypotheses listed in Section 5: Table 4 shows the average attack rates in Part 1 of our experiment for all four treatments.[8] As Hypothesis 1 suggests, we expect that, in our experiment, player Z will get attacked less often by X and Y in BuffoonZX and BuffoonZY than in the Baseline and Circular treatments. This is because only in those treatments are the implosion rates sufficiently high as to warrant subjects ignoring player Z when they are in the roles of X or Y. This hypothesis is supported in our data.

As we see in Table 4, the rate with which X targets Z or Y targets Z drops substantially when Z becomes a buffoon, in line with Hypothesis 1. To test this hypothesis we count how often a subject decides to attack Z when making decisions for roles X and Y, separately for the treatments when Z is a buffoon and when he is not, constructing a measure with values between 0 and 4. Since we observe the same subjects in both treatment groups, we employ a one-sided Wilcoxon matched pairs test. On average, subjects direct 2.55 attacks at Z in the Baseline in Circular treatments and 1.56 attacks in BuffoonZX and BuffoonZY treatments. This difference is highly significant ($p < 0.001$).[9] As a consequence, the probability that Z is not attacked by X and Y increases from a minimum of 0.11 in Baseline to a maximum of 0.42 in BuffoonZY treatment.

Hypothesis 2 predicts that the rate at which Z gets attacked by X and Y decreases from the Baseline to the Circular treatment and decreases further from the Circular to the BuffoonZX treatment. To test this hypothesis, we again count how often subjects attack Z when in roles X and Y, but this time separately for each treatment and test for differences between the Baseline and Circular

[8] Table 7 in the appendix shows the average attack rates that resulted from subjects' first choices only. The results in both tables are very similar: there are no order effects affecting subject behavior. Hence, we pool our observations and perform within-subjects tests.

[9] Alternatively, we can test for differences in the attack rates of X and Y between Baseline/Circular and BuffoonZX/BuffoonZY using McNemar's χ^2 test. All eight individual comparisons are again significant ($p < 0.05$).

Treatment	q_{XY}	q_{YZ}	q_{ZX}	Prob(Z safe)
Baseline	0.34	0.67	0.69	0.11
Circular	0.32	0.53	0.78	0.15
BuffoonZX	0.55	0.41	0.74	0.32
BuffoonZY	0.67	0.36	0.32	0.42

Table 4: Average Attack Rates in Part 1
Note: q_{ij} denotes the probability with which player i targets player j. Column Prob(Z safe) shows the resulting probability that Z is attacked by neither X nor Y.

treatments and for differences between Circular and BuffoonZX treatments using the same one-sided Wilcoxon matched pairs test. In the Baseline, subjects direct an average of 1.33 attacks at Z, which falls to 1.22 in the Circular treatment ($p = 0.094$) and to 0.86 in BuffonZX treatment ($p < 0.001$).[10]

Finally, to test Hypothesis 3 we use McNemar's χ^2 test to test for differences between the in the attack rates q_{ZX} between the BuffoonZX and BuffoonZY treatments, which does drop dramatically from 0.74 to 0.32 ($p < 0.001$).

We summarize the findings so far in the following

Summary 1 *We find support for all three hypotheses: Player Z does get attacked less often when he is a buffoon (H1 Dominance). Indeed, the frequency of attacks on Z falls as we move from Baseline to Circular and it falls further as me move from Circular to BuffoonZX (H2 Mixed Nash). Finally, the buffoon attacks Y more when he is predicted to (H3 Pure Nash).*

While our results bear out the qualitative predictions of our hypotheses, the point predictions are still off. In the BuffoonZX and BuffoonZY treatments it is a dominant strategy for X and Y to ignore Z but, even in the best case of the BuffoonZY treatment, about one third of attacks by X and Y are still directed at Z and hence violate dominance.

[10] If we test for differences in the attack rates by role using McNemar's χ^2 test separately, there is no evidence of a significant difference in attack rates between Baseline and Circular for role X ($p = 0.728$) but for role Y there is ($p = 0.028$). Comparing Circular and BuffoonZX, there is significant evidence of a drop in attack rates for both roles X ($p < 0.001$) and Y ($p = 0.010$).

	q_{XY}		Money at Stake
Set	Predicted	Observed	(in EUR)
1	0	0.30	0.92
2	1	0.39	0.26
3	0	0.38	0.47
4	1	0.44	0.71
5	1	0.68	0.26
6	1	0.63	0.73
7	1	0.74	0.71
8	1	0.65	1.18

Table 5: Average Attack Rates in Part 2
Note: Subjects play as player X. q_{ij} denotes the probability with which player i targets player j. Column Money at Stake shows the amount of money a subject is expected to lose if they do not follow the prediction and the parameter set is chosen for payment.

Since we do not give subjects the opportunity to learn from feedback it is likely that some strategically less savvy subjects will make mistakes. In an actual political primary, however, we might expect the participants and their advisors to be sophisticated experienced players. To investigate the impact of subject sophistication on behavior in Part 1 of our experiment we exploit Part 2 where we evaluate the ability of our subjects to best respond to the actions of their opponents and use a median split to classify our subjects as relatively more or less sophisticated. We ask whether subjects who perform better in that series of decision problems are less likely to violate dominance in Part 1.

As a reminder, in Part 2 subjects assumed the role of Player X and played against a pair of computerize opponents (Y and Z) who were assigned a specific action that we made observable to the subject. Since the actions of their opponents were fixed, the choice made by our human subjects revealed how good she was in best responding to the predetermined actions of her opponent. The better she was able to do so the more sophisticated we viewed her.

Table 5 summarizes the predicted and observed attack rates in each choice set together with the expected amount of money at stake if the parameter set was chosen for payment. As we see in the table, the observed attack rates correlate positively with the predicted attacks but subjects still make a sizeable number of errors. We weigh these errors by the total amount of money that subjects leave on the table conditional on the parameter set being chosen for payment.

Figure 8 plots a histogram of how much money subjects leave on the table in total. As we see, around 10% of subjects perfectly follow the predictions of the theory and leave no money on the table. The rest of the subjects leave a nonzero amount of money on the table, with a median of 1.71 Euro and a maximum of 5.26 Euro. Making costly errors in this series of relatively simple decision problem is akin to a violation of dominance. Using this money-on-the-table metric we split our subject sample in two at the median of money left and compare the behavior of the two resulting groups in Part 1. There are 66 subjects below or at the median performance and 63 subjects above the median.

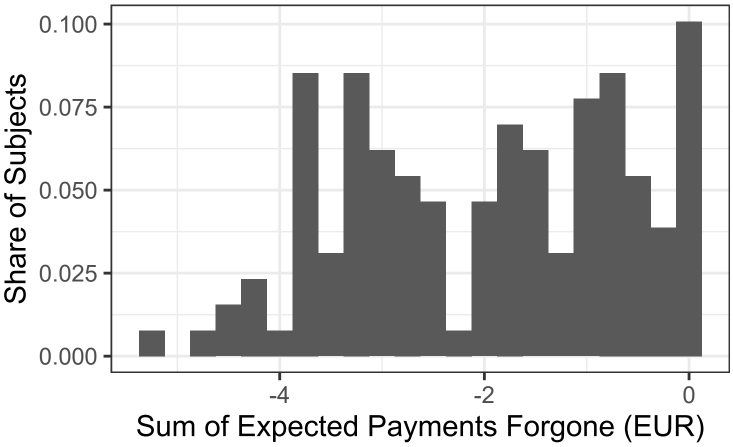


Figure 8: Histogram of Total Expected Cost of Errors

Using these two categories of subjects, we now look to see if their revealed sophistication in Part 2 of the experiment had consequences for their behavior in Part 1.

Table 6 shows the average attack rates in Part 1 split by below or above median performance in Part 2. While in the Baseline and Circular treatments the attack rates are rather similar there are stark differences between the groups

Treatment	q_{XY}	q_{YZ}	q_{ZX}	Prob(Z safe)
Below Median				
Baseline	0.35	0.62	0.64	0.13
Circular	0.26	0.65	0.67	0.09
BuffoonZX	0.36	0.59	0.68	0.15
BuffoonZY	0.53	0.53	0.36	0.25
Above Median				
Baseline	0.33	0.73	0.70	0.09
Circular	0.38	0.41	0.80	0.22
BuffoonZX	0.75	0.22	0.81	0.58
BuffoonZY	0.81	0.19	0.27	0.66

Table 6: Average Attack Rates in Part 1, Split by Performance in Part 2
Note: q_{ij} denotes the probability with which player i targets player j. Column Prob(Z safe) shows the resulting probability that Z is attacked by neither X nor Y.

in the BuffoonZX and BuffoonZY treatments. In the BuffoonZX treatment the below median group mainly attacks Z in both roles X and Y, and in BuffoonZY they are close to 50/50. For the above median group the picture is rather different with a much smaller fraction of subjects attacking Z when doing so is a dominated strategy. Since these comparisons are between subjects, we use two-sided Mann-Whitney-U tests to compare the attack rates between the groups. In the treatments without buffoons subjects in the below-median group direct an average number of 2.67 attacks at Z, while in the above-median group it is an average of 2.43 attacks ($p = 0.152$). In contrast, in the treatments with buffoons subjects in the below median group still directed an average number of 2.23 attacks on Z, while for those in the above-median group this number falls to 0.86 ($p < 0.001$).[11] Consequently, the probability that X and Y both avoid attacking Z in the buffoon treatments does not exceed 0.25 in the below-median group but reaches 0.66 in the above-median group.

[11] All comparisons between groups in the buffoon treatments hold when we use Pearson's χ^2 test to test for a difference in proportions in the attack rates of X and Y individually ($p < 0.01$).

7 Conclusion

In this paper we have investigated a phenomenon in political contests where it is possible for a candidate who appears on the surface to have no chance of being viable (a buffoon) to avoid being attacked by his opponents to the extent that he or she actually wins a primary election. We believe that this logic captures an important element of the Trump victory in the Republican primary of 2016.

We model political competition as a truel (a three-way duel) and derive the fact that, in the equilibrium of these contests, candidates with a high enough probability of self destruction may avoid being attacked. In the light of this, the Trump victory appears much less an accident but rather emerges as a plausible equilibrium outcome of a contest with rational players.

When we take this model to the lab we find evidence that supports the buffoon equilibrium prediction, i.e., that a candidate with a high enough implosion probability does get attacked less frequently. Our results appear strong as we do not give our subjects the opportunity to learn from feedback. In a primary election contest most candidates themselves have substantial experience in political campaigning and their advisors are usually seasoned veterans of many campaigns well trained in understanding the strategic intricacies of such contests. In the laboratory, we can see that subjects who make fewer costly errors in a series of decision tasks also follow the theory more closely. We expect that the effects for experienced players should only be larger. As we have shown Donald Trump was largely spared from attacks in the early stages of the 2016 Republican primaries. For the casual observer this looks like an amazing blunder from his opponents. Our study supports the opposite view: there were solid reasons to ignore the buffoon who may indeed be the most likely contender to be elected.

References

Adams, J. and Merrill, S. (2008). Candidate and party strategies in two-stage elections beginning with a primary. *American Journal of Political Science*, 52(2):344–359.

Agranov, M. (2016). Flip-Flopping, Primary Visibility, and the Selection of Candidates. *American Economic Journal: Microeconomics*, 8(2):61–85.

Ansolabehere, S., Hansen, J. M., Hirano, S., and Snyder, J. M. (2010). More Democracy: The Direct Primary and Competition in U.S. Elections. *Studies in American Political Development*, 24(2):190–205. Edition: 2010/08/10 Publisher: Cambridge University Press.

Archetti, M. (2012). Survival of the weakest in N-person duels and the maintenance of variation under constant selection. *Evolution*, 66(3):637–650.

Baron, D. P. (1994). Electoral Competition with Informed and Uninformed Voters. *American Political Science Review*, 88(1):33–47. Edition: 2013/09/02 Publisher: Cambridge University Press.

Bouton, L., Castanheira, M., and Drazen, A. (2018). A theory of small campaign contributions. Technical report, National Bureau of Economic Research.

Boyer, P. C., Konrad, K. A., and Roberson, B. (2017). Targeted campaign competition, loyal voters, and supermajorities. *Journal of Mathematical Economics*, 71:49–62.

Burden, B. C. (2001). The polarizing effects of congressional primaries. In Galderisi, P. F., Ezra, M., and Lyons, M., editors, *Congressional Primaries and the Politics of Representation*, pages 95–115. Rowman & Littlefield, Lanham, MD, United States.

Caillaud, B. and Tirole, J. (2002). Parties as Political Intermediaries. *The Quarterly Journal of Economics*, 117(4):1453–1489.

Castanheira, M., Crutzen, B. S. Y., and Sahuguet, N. (2009). Party Organization and Electoral Competition. *The Journal of Law, Economics, and Organization*, 26(2):212–242.

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Congleton, R. D. (1986). Rent-seeking aspects of political advertising. *Public Choice*, 49(3):249–263.

Corchón, L. C. (2007). The theory of contests: a survey. *Review of Economic Design*, 11(2):69–100.

Esteban, J. and Ray, D. (1999). Conflict and Distribution. *Journal of Economic Theory*, 87(2):379–415.

Gradstein, M. and Konrad, K. A. (1999). Orchestrating Rent Seeking Contests. *The Economic Journal*, 109(458):536–545. Publisher: John Wiley & Sons, Ltd.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Hanson, D. (2016). Republicans in Chaos. *National Review*.

Herrera, H., Morelli, M., and Nunnari, S. (2016). Turnout across democracies. *American Journal of Political Science*, 60(3):607–624.

Herrera, H., Morelli, M., and Palfrey, T. (2014). Turnout and power sharing. *The Economic Journal*, 124(574):F131–F162.

Hirano, S. and Snyder, J. M. (2019). *Primary elections in the United States*. Cambridge University Press.

Hirano, S., Snyder, J. M., and Ting, M. M. (2009). Distributive Politics with Primaries. *The Journal of Politics*, 71(4):1467–1480. Publisher: The University of Chicago Press.

Hirshleifer, J. (1989). Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. *Public Choice*, 63(2):101–112.

Hortala-Vallve, R. and Mueller, H. (2015). Primaries: the unifying force. *Public Choice*, 163(3-4):289–305.

Huck, S., Konrad, K. A., and Müller, W. (2001). Divisionalization in contests. *Economics Letters*, 70(1):89–93.

Hummel, P. (2010). Flip-flopping from primaries to general elections. *Journal of Public Economics*, 94(11):1020–1027.

Internet-Archive (2016). Political-TV-Ad-Archive.

Kilgour, D. M. (1971). The simultaneous truel. *International Journal of Game Theory*, 1(1):229–242.

Kilgour, D. M. (1975). The sequential truel. *International Journal of Game Theory*, 4(3):151–174.

Konrad, K. A. (2000). Sabotage in rent-seeking contests. *The Journal of Law, Economics, and Organization*, 16(1):155–165.

Konrad, K. A. (2009). *Strategy and dynamics in contests.* Oxford University Press, Oxford.

Larson, H. D. (1948). A dart game. *The American Mathematical Monthly*, 55:640–641.

Lau, R. R. and Rovner, I. B. (2009). Negative Campaigning. *Annual Review of Political Science*, 12(1):285–306. Publisher: Annual Reviews.

Meirowitz, A. (2005). Informational Party Primaries and Strategic Ambiguity. *Journal of Theoretical Politics*, 17(1):107–136. Publisher: SAGE Publications Ltd.

Nalbantian, H. R. and Schotter, A. (1997). Productivity Under Group Incentives: An Experimental Study. *The American Economic Review*, 87(3):314–341. Publisher: American Economic Association.

Potters, J., de Vries, C. G., and van Winden, F. (1998). An experimental examination of rational rent-seeking. *European Journal of Political Economy*, 14(4):783–800.

Sheremeta, R. M. (2011). Contest Design: An Experimental Investigation. *Economic Inquiry*, 49(2):573–590. Publisher: John Wiley & Sons, Ltd.

Shubik, M. (1982). *Game Theory in the Social Sciences: Concepts and Solutions.* MIT press.

Skaperdas, S. and Grofman, B. (1995). Modeling Negative Campaigning. *American Political Science Review*, 89(1):49–61. Edition: 2013/09/02 Publisher: Cambridge University Press.

Tullock, G. (1980). Efficient rent seeking. In Buchanan, J., Tollison, R., and Tullock, G., editors, *Toward a theory of the rent-seeking society*, pages 97–112. Texas A&M University Press, College Station.

Ware, A. (1979). 'divisive' primaries: The important questions. *British Journal of Political Science*, 9(3):381–384.

Ware, A. (2002). *The American Direct Primary: Party Institutionalization and Transformation in the North.* Cambridge University Press, Cambridge.

Woolf, N. (2016). Is Donald Trump finally imploding? *New Statesman.*

Appendices

A1 Additional Tables

Treatment	q_{XY}	q_{YZ}	q_{ZX}
Baseline	0.37	0.54	0.69
Circular	0.40	0.60	0.80
BuffoonZX	0.60	0.36	0.80
BuffoonZY	0.71	0.35	0.26

Table 7: Average Attack Rates, First Choice
Note: q_{ij} denotes the probability with which player i targets player j.

State	TV Market
Iowa (IA)	Ceder Rapids-Waterloo-Iowa City-Dublin, Iowa Des Moines-Ames, Iowa Sioux City, Iowa
New Hampshire (NH)	Boston, MA/Manchester, NH
South Carolina (SC)	Columbia, SC
Nevada (NV)	Reno, NV
North Carolina (NC)	Charlotte, NC Raleigh-Durham-Fayetteville, NC Norfolk-Portsmouth-Newport News, NC
Ohio (OH)	Cincinnati, OH
Florida (FL)	Tampa-St. Petersburg, FL Orlando-Daytona Beach-Melbourne, FL Miami-Fort Lauderdale, FL
Arizona (AZ)	Phoenix-Prescott, AZ
Wisconsin (WI)	Milwaukee, WI
Colorado (CO)	Denver, CO Colorado Springs-Pueblo, CO
New York (NY)	New York City, NY
Virginia (VA)	Roanoke-Lynchburg, VA
Pennsylvania (PA)	Philadelphia, PA
California (CA)	San Francisco-Oakland-San Jose, CA
Maryland (MD)	Washington, DC/Hagerstown, MD

Table 8: TV Markets and States Matching

Note: We drop data from the "Greenville-Spartanburg, SC/Asheville-Anderson, NC" TV market since the market spans two states whose primaries are only 25 days apart. Including them for either South Carolina or North Carolina does not change our results.

Candidate	Sponsors
Donald Trump	Donald J. Trump For President Great America PAC Rebuilding America Now Make America Number One
Marco Rubio	Marco Rubio for President Conservative Solutions Reclaim America
Ted Cruz	Cruz for President Keep the Promise Stand For Truth Trusted Leadership Courageous Conservatives
Ben Carson	Carson America 2016 Committee
Jeb Bush	Jeb 2016 Right to Rise USA
Carly Fiorina	Carly for President Carly for America Committee
Rand Paul	Rand Paul for President America's Liberty
Chris Christie	Chris Christie For President America Leads
Mike Huckabee	Huckabee For President Pursuing America's Greatness
Jim Gilmore	Gilmore For America
George Pataki	Pataki for President
Rick Santorum	Santorum For President 2016
John Kasich	Kasich for America New Day For America New Day Independent Media Committee

Table 9: Candidates and Ad Sponsors

A2 Proofs

Proof of Proposition 1

Case 1: $X \to Y, Y \to Z, Z \to X$.

Denote the incentive compatibility condition of some candidate j by IC_j. We have:

$$\begin{aligned} IC_X &: \; p_Y^{-X} = \beta_Y < \beta_Z + (1-\beta_Z)\,\pi_Y = p_Z^{-X} \\ IC_Y &: \; p_Z^{-Y} = \beta_Z < \beta_X + (1-\beta_X)\,\pi_Z = p_X^{-Y} \\ IC_Z &: \; p_X^{-Z} = \beta_X < \beta_Y + (1-\beta_Y)\,\pi_X = p_Y^{-Z} \end{aligned}$$

IC_X and IC_Z automatically follow from $\beta_X < \beta_Y < \beta_Z$. IC_Y instead requires that:

$$\pi_Z > \frac{\beta_Z - \beta_X}{1-\beta_X}. \tag{8}$$

Case 2: $X \to Z, Y \to X, Z \to Y$. We have:

$$\begin{aligned} IC_X &: \; p_Z^{-X} = \beta_Z < \beta_Y + (1-\beta_Y)\,\pi_Z = p_Y^{-X} \\ IC_Y &: \; p_X^{-Y} = \beta_X < \beta_Z + (1-\beta_Z)\,\pi_X = p_Z^{-Y} \\ IC_Z &: \; p_Y^{-Z} = \beta_Y < \beta_X + (1-\beta_X)\,\pi_Y = p_X^{-Z} \end{aligned}$$

IC_Y automatically follows from $\beta_X < \beta_Z$. IC_X and IC_Z instead require that:

$$\pi_Z > \frac{\beta_Z - \beta_Y}{1-\beta_Y} \text{ and } \pi_Y > \frac{\beta_Y - \beta_X}{1-\beta_X}. \tag{9}$$

Case 3: $X \to Y, Y \to X, Z \to X$. We have:

$$\begin{aligned} IC_X &: \; p_Y^{-X} = \beta_Y < \beta_Z = p_Z^{-X} \\ IC_Y &: \; p_X^{-Y} = \beta_X + (1-\beta_X)\,\pi_Z < \beta_Z = p_Z^{-Y} \\ IC_Z &: \; p_X^{-Z} = \beta_X + (1-\beta_X)\,\pi_Y < \beta_Y + (1-\beta_Y)\,\pi_X = p_Y^{-Z}. \end{aligned}$$

IC_X automatically follows from $\beta_Y < \beta_Z$. IC_Y and IC_Z instead require that:

$$\pi_Z < \frac{\beta_Z - \beta_X}{1 - \beta_X} \text{ and } \pi_Y < \frac{\beta_Y - \beta_X + (1 - \beta_Y)\,\pi_X}{1 - \beta_X}. \tag{10}$$

Note that the first of these conditions is the opposite of (8) , which means that Case 1 and Case 3 equilibria are mutually exclusive.

Case 4: $X \to Y, Y \to X, Z \to Y$. We have:

$$\begin{aligned} IC_X &: \; p_Y^{-X} = \beta_Y + (1 - \beta_Y)\,\pi_Z < \beta_Z = p_Z^{-X} \\ IC_Y &: \; p_X^{-Y} = \beta_X < \beta_Z = p_Z^{-Y} \\ IC_Z &: \; p_X^{-Z} = \beta_X + (1 - \beta_X)\,\pi_Y > \beta_Y + (1 - \beta_Y)\,\pi_X = p_Y^{-Z}. \end{aligned}$$

IC_Y automatically follows from $\beta_X < \beta_Z$. IC_X and IC_Z instead require:

$$\pi_Z < \frac{\beta_Z - \beta_Y}{1 - \beta_Y} \text{ and } \pi_Y > \frac{\beta_Y - \beta_X + (1 - \beta_Y)\,\pi_X}{1 - \beta_X},$$

where the first condition contradicts both (8) because $\frac{\beta_Z - \beta_Y}{1 - \beta_Y} < \frac{\beta_Z - \beta_X}{1 - \beta_X}$ and (9), and the second condition contradicts (10). Hence, Case 4 can never coexist with any of the other Cases.

This implies that Case 1 can only coexist with Case 2, and Case 2 can only coexist with Case 1 or Case 3. We depict the mapping from all possible parameter constellations to equilibrium existence in Figure 3 for the case $0 = \beta_X = \beta_Y < \beta_Z$ and in Figure 6 for the case $0 \leq \beta_X < \beta_Y < \beta_Z$.

Introducing Dynamics

Denote by $\Pi_i^*(j,k)$ the value for i of entering the second round with both j and k still in the contest. Conversely, $\Pi_i^*(j)$ and $\Pi_i^*(k)$ are the continuation values associated with i entering a duel, respectively against j and against k. Importantly, all three continuation values are strictly increasing in the remaining opponents' implosion probabilities.

Player i's counterfactual payoffs in the first round are a straightforward modification of (1). Denoting by $\Delta\Pi_i^1(j)$ the utility gain from targeting j in round 1, and comparing it against the value of targeting k, it emerges that:

$$\tfrac{\Delta\Pi_i^1(j)-\Delta\Pi_i^1(k)}{(1-p_i)\pi_i} = \left(p_k^{-i} - p_j^{-i}\right) + \left[\Pi_i^*(k)(1-p_k^{-i}) - \Pi_i^*(j)(1-p_j^{-i})\right]. \tag{11}$$

Note that this difference is strictly decreasing both in p_j^{-i} and in $\Pi_i^1(j)$, and hence in β_j. That is, the higher is j's implosion probability, the lower is i's incentive to shoot at j, which extends our previous results to the dynamic game.

Going further, we can compare (11) with (3), the difference obtained in the static game, which is $(p_k^{-i} - p_j^{-i})$. It thus appears that j is less targeted early in the race if the second term in (11), the one between square brackets, is negative. This happens if and only if:

$$\frac{1-\pi_k}{1-\pi_j} < \frac{2-(1-\beta_j)(1-\pi_i)}{2-(1-\beta_k)(1-\pi_i)},$$

which compares the opponents' precisions and implosion probabilities. For equal precisions, the condition boils down to $\beta_j > \beta_k$. For the general case, the condition is more easily satisfied the higher is β_j and less easily satisfied the larger is π_j.

Generalization of Lemma 1 to N+3 players

Consider a game with $N+3$ players (i, j, k and $N > 0$ others), and consider the perspective of player i. For a given action profile, define $\mathbb{K}$ as the set of players who aim at candidate k. k's probability of survival is then: $s_k := (1-\beta_k)\prod_{j\in\mathbb{K}}(1-\pi_j)$, and his probability of exit is $1 - s_k$. From the perspective of player i, k's counterfactual exit probability is:

$$p_k^{-i} := 1 - (1-\beta_k)\prod_{j\in\mathbb{K}\setminus i}(1-\pi_j).$$

To ease reading, in the developments below we denote p_k^{-i} by p_k. Define also $P := \prod_{l=1,..,N} p_l$, which is the probability that all the players $n_1, n_2, ..., N$ exit the race, conditional on i not aiming at anyone. We have:

Lemma 2 *Given an action profile, candidate i's best response is to aim at the other candidate j with the lowest counterfactual exit probability.*

Proof. First, we derive the value of aiming at some candidate j:

$$\begin{aligned}\frac{\Delta\Pi_i\left(j\right)}{\pi_i\left(1-p_i\right)} &= \left(1-p_j\right)p_k \overbrace{\left[\begin{array}{c}\left(1-\frac{1}{2}\right)P+\left(\frac{1}{2}-\frac{1}{3}\right)\sum_{l_1}\frac{\left(1-p_{l_1}\right)}{p_{l_1}}P+...\\ ...+\left(\frac{1}{3}-\frac{1}{4}\right)\sum_{l_1}\sum_{l_2\neq l_1}\frac{\left(1-p_{l_1}\right)\left(1-p_{l_2}\right)}{p_{l_1}p_{l_2}}P+...\end{array}\right]}^{\alpha}\\ &\quad+\left(1-p_j\right)\left(1-p_k\right)\underbrace{\left[\begin{array}{c}\left(\frac{1}{2}-\frac{1}{3}\right)P+\left(\frac{1}{3}-\frac{1}{4}\right)\sum_{l_1}\frac{\left(1-p_{l_1}\right)}{p_{l_1}}P+...\\ ...+\left(\frac{1}{4}-\frac{1}{5}\right)\sum_{l_1}\sum_{l_2\neq l_1}\frac{\left(1-p_{l_1}\right)\left(1-p_{l_2}\right)}{p_{l_1}p_{l_2}}+...\end{array}\right]}_{\gamma}.\end{aligned}$$

Conversely, value of aiming at k is:

$$\frac{\Delta\Pi_i\left(k\right)}{\pi_i\left(1-p_i\right)}=p_j\left(1-p_k\right)\ \alpha+\left(1-p_j\right)\left(1-p_k\right)\gamma$$

Comparing the two:

$$\frac{\Delta\Pi_i\left(j\right)-\Delta\Pi_i\left(k\right)}{\pi_i\left(1-p_i\right)}=\left[\left(1-p_j\right)p_k-p_j\left(1-p_k\right)\right]\times\alpha,$$

which is positive iff $p_j < p_k$. This comparison being valid for any pair of candidates j and k different from i, applying recursively proves the lemma. ■

Generalization of Proposition 1 to N players

Consider n_X “professional candidates” $x \in \mathbb{X} = \{1, 2, ..., n_X\}$ with:

$$\pi_x > \beta_x,$$

and n_Z candidates $z \in \mathbb{Z} = \{1, 2, ..., n_Z\}$ who are buffoons:

$$\pi_z < \beta_z.$$

We impose that: $\beta_z > \beta_x$ and $\pi_z < \pi_x$ for all $x \in \mathbb{X}$ and $z \in \mathbb{Z}$ that is, any buffoon has a probability of implosion strictly higher than any fine candidate. We also set $n_Z \leq n_X$: there are fewer buffoons than fine candidates.

Note that each candidate may a priori aim at any of the other $(n_X + n_Z - 1)$ candidates, implying a number of possible combinations equal to: $(n_X + n_Z - 1)^{n_X + n_Z}$.

There also are multiple combination of actions that may be outcome equivalent: consider for instance a slate of 4 professionals and one buffoon. Both of the following action profiles would characterize a buffoon equilibrium: first, consider a circular action profile among the pros, $x_1 \to x_2$, $x_2 \to x_3$, $x_3 \to x_4$, $x_4 \to x_1$, $z \to x_1$. Second, two duels among the pros: $x_1 \to x_2$, $x_2 \to x_1$, $x_3 \to x_4$, $x_4 \to x_3$, $z \to x_1$. Given the potentially large number of such profiles, we only look for a sufficient condition to ensure the existence of at least one buffoon equilibrium.

Proof of Proposition 2. We focus on the following action profile: first, let all the professional candidates pick their target in accordance with Lemma 2: none of the buffoons are targeted as long as $\max_{x\in\mathbb{X}} \beta_x < \min_{z\in\mathbb{Z}} \beta_z$. Next, let the buffoons pick their target in the same fashion. None of the buffoons get targeted if there are at least n_Z professionals such that:

$$1 - (1-\beta_x)(1-\pi_x) = \beta_x + \pi_x (1-\beta_x) \leq \min_{z\in\mathbb{Z}} \beta_z.$$

A sufficient condition for this to hold is:

$$\max_x \pi_x \leq \min_{x\in\mathbb{X}, z\in\mathbb{Z}} \frac{\beta_z - \beta_x}{1-\beta_x}.$$

By transitivity, no buffoon wants to aim at another buffoon if this condition is satisfied.

■

A3 Instructions

These instructions are translated from the original German instructions. The German instructions are available from the authors upon request. Subjects received the instructions for Part 2 only after finishing Part 1.

Part 1

Welcome to our experiment!

During the experiment you are not allowed to use electronic gadgets or to communicate with other participants. Please only use the programs and functions that were designed for the experiment. Please do not talk to other participants. If you have a question, please raise your hand. We will come to you and answer your question privately. Please do not ask questions aloud. If the question is relevant for all participants, we will repeat and answer it aloud. If you break these rules, we will have to exclude you from the experiment and from payment.

Part 1 of the experiment

This experiment has two independent parts. Your decisions in the first part will not influence the second part.

Short description

You take part in a tournament with three players. In total there are three players who we call A, B, and C. During the competition each player has the option to attack one other player. If an attack is successful, the attacked player is ELIMINATED and will not earn any money. You only earn money if at the end of the competition you are STILL STANDING. All players that are still standing will share 15 Euro between each other. Additionally, you will receive a participation fee of 5 Euro.

In this part of the experiment you will make decisions in 4 different situations. At the end of the experiment one of those decisions will be randomly drawn for payment.

Attacks

All players must decide whom to attack at the same time. That is, by the time you make your decision, you will not know who the other players will attack.

An attack is not always successful. When a player attacks, the computer rolls a die that determines whether his attack succeeds in eliminating the opponent. But players differ in their odds of successful attacks.

In addition, there is a probability that a player exits the competition and thus is ELIMINATED even if he was not attacked or if all attacks on him failed. If a player is eliminated in this way, we say that he IMPLODED.

All attacks are carried out at the same time. A player is still standing at the end of the tournament if and only if:

(1) All attacks on him failed.

(2) He did not implode.

This gives rise to three possible situations for every player:

(1) The player was not attacked by anyone. Then, he is STILL STANDING if he did not IMPLODE.

(2) The player was attacked by exactly one other player. Then, he is STILL STANDING if the attack failed AND he did not IMPLODE.

(3) The player was attacked by both other players. Then he is STILL STANDING only if both attacks failed AND he did not IMPLODE.

This means that you cannot influence your own likelihood to be STILL STANDING at the end of the competition. You can only influence the likelihood that the other players are STILL STANDING.

In the experiment we will present all the information in a table as follows. Imagine the following situation:

Player	Probability to attack successfully	Probability to implode
A	50%	0%
B	40%	0%
C	10%	30%

Note that in this situation two players, A and B, have a probability to implode of zero, while player C has a probability of 30% to implode. In this situation player C is at the same time the player with the lowest probability to successfully attack, while player A is the player with the highest probability to successfully attack.

Results

After the attacks, there are four possible scenarios that can arise:

(i) All players were ELIMINATED. In this case, all players receive 0 Euro.

(ii) Two players were ELIMINATED, one player is STILL STANDING. The players who were eliminated, receive 0 Euro. The player who has survived receives 15 Euro.

(iii) One player was ELIMINATED, two players are STILL STANDING. The player who was eliminated receives 0 Euro. The two players who are still standing, receive 7.50 Euro each.

(iv) All players are STILL STANDING. All players receive 5 Euro each.

Procedure of the experiment

In the experiment you will first make choices for 2 different situations. In each round we will ask you who you attack if you are player A, player B, or player C. You have to make exactly one decision for each role. It means that must to make a choice for the role of player A whether you want to attack B or C. For the role of player B you must decide whether you want to attack C or A. For the role of player C you must decide whether you want to attack A or B.

After every participant in the experiment has made all choices in all situations the computer will match you randomly and anonymously with two other participants in the room. Then the computer will randomly pick one of the situations and will assign you the role of player A, player B, or player C. The computer will then implement the choice that you made for this role in this situation. Simultaneously, the computer will implement the choices of the other two players, which they made for their roles. Afterwards, the computer will roll dice to determine if the players hit their targets, if they implode, and if they are still standing. The computer will use the probabilities which were defined in the chosen situation.

Note that you will not receive feedback of what other people are doing when you are making your choices! You will only learn more at the end of the experiment.

Payments

At the end of the experiment we will pay your earnings in cash. You will receive 5 Euro for your participation in the experiment plus your earning from Part 1 and Part 2.

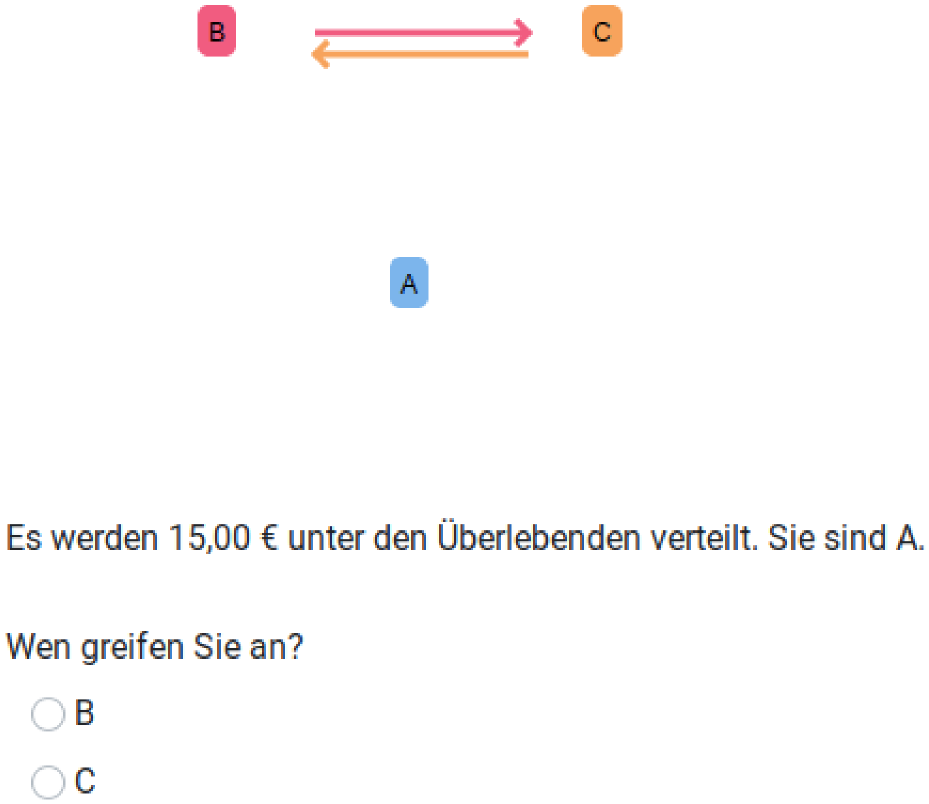
Part 2

Welcome to part 2 of the experiment

In this part of the experiment you do not interact with the other participants in the room but only with the computer. Like in part 1 you take part in a competition with three players.

The differences compared to Part 1:

- You do not have a random, but a specific role, which you will know when making your choice.
- The other two roles will be controlled by the computer.
- All earnings by players who are controlled by the computer will not be paid out to human participants.
- The players who are controlled by the computer will commit from the beginning who they will attack. You will know from the shown arrows who the computer will attack.



Es werden 15,00 € unter den Überlebenden verteilt. Sie sind A.

Wen greifen Sie an?

○ B

○ C

In this example the computer controls players B and C. B will attack C and C will attack B. You are player A and you have to decide whether to attack B or C.[12]

Apart from these differences the same rules like in Part 1 apply. You will see the same table containing all the information about attack and implosion probabilities. All players who were not eliminated again share 15 Euro between themselves. Any money that is earned by a computer-controlled player will not be paid out to human participants. You will make decisions for 8 situations. The computer will randomly pick one of these situations and will implement the choice that you made in this situation. Afterwards, the computer will roll dice to determine if the players hit their targets, if they implode, and if they are still standing. The computer will use the probabilities which were defined in the chosen situation.

[12] The text in the picture reads "The survivors split 15 Euro. You are A. Who do you attack?"

Payments

At the end of the experiment we will pay your earnings in cash. You will receive 5 Euro for your participation in the experiment plus your earning from Part 1 and Part 2.