Behavioral experiments on biased voting in networks

Michael Kearns1, Stephen Judd, Jinsong Tan, and Jennifer Wortman

University of Pennsylvania, Department of Computer and Information Science, 3330 Walnut Street, Philadelphia, PA 19104

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved December 5, 2008 (received for review August 19, 2008)

Many distributed collective decision-making processes must balance diverse individual preferences with a desire for collective unity. We report here on an extensive session of behavioral experiments on biased voting in networks of individuals. In each of 81 experiments, 36 human subjects arranged in a virtual network were financially motivated to reach global consensus to one of two opposing choices. No payments were made unless the entire population reached a unanimous decision within 1 min, but different subjects were paid more for consensus to one choice or the other, and subjects could view only the current choices of their network neighbors, thus creating tensions between private incentives and preferences, global unity, and network structure. Along with analyses of how collective and individual performance vary with network structure and incentives generally, we find that there are well-studied network topologies in which the minority preference consistently wins globally; that the presence of “extremist” individuals, or the awareness of opposing incentives, reliably improve collective performance; and that certain behavioral characteristics of individual subjects, such as “stubbornness,” are strongly correlated with earnings.

behavioral game theory | collective decision making | network science

The tension between the expression of individual preferences and the desire for collective unity appears in decision-making and voting processes in politics, business, and many other arenas. Furthermore, such processes often take place in social or organizational networks, in which individuals are most influenced by, or aware of, the current views of their network neighbors.

The 2008 Democratic National Primary race offers a recent, if approximate, example of this phenomenon. On the one hand, individual voters held opposing and sometimes strong preferences that were apparently very nearly balanced across the population; however, there was a strong and explicit desire that once the winning candidate was identified, the entire party should unify behind that candidate (1). Obviously primary voters could be influenced by many global factors (such as polls and mainstream media) outside the scope of their individual social and organizational networks, but presumably for many voters these local influences still played an important and perhaps even dominant role.

Although there is now a significant literature on the diffusion of opinion in social networks (2–4), the topic is typically studied in the absence of any incentives toward collective unity. In many contagion-metaphor models, individuals are simply more or less susceptible to “catching” an opinion or fad from their neighbors, and are not directly cognizant of, or concerned with, the global state. In contrast, we are specifically interested in scenarios in which individual preferences are present but are subordinate to reaching a unanimous global consensus.

We report here on an extensive session of human-subject experiments meant to provide a simple abstraction of the key properties and tensions discussed above. In each experiment, 36 subjects each simultaneously sit at workstations and control the state of a single vertex in a 36-vertex network whose connectivity structure is determined exogenously and is unknown to the subjects. The state of a subject’s vertex is simply one of 2 colors (red or blue), and can be asynchronously updated as often as desired during the 1-min experiment. Subjects are able to view the current color choices of their immediate neighbors in the network at all times but otherwise have no global information on the current state of the network (aside from a crude and relatively uninformative “progress bar”; see Fig. 1). No communication between subjects outside the experimental platform is permitted.

In each experiment, each subject is given a financial incentive that varies across the network, and specifies both individual preferences and the demand for collective unity. For instance, one player might be paid $1.25 for blue consensus and $0.75 for red consensus, whereas another might be paid $0.50 for blue consensus and $1.50 for red consensus, thus creating distinct and competing preferences across individuals. However, payments for an experiment are made only if (red or blue) global unanimity is reached, so subjects must balance their preference for higher payoffs with their desire for any payoff at all. A screenshot for a particular subject in a typical experiment is shown in Fig. 1. We note that our experiments may also be viewed as a distributed, networked version of the classic “Battle of the Sexes” game, or as a networked coordination game (5).

We note that although our experimental framework deliberately omits global “broadcast” mechanisms for consensus (other than the aforementioned progress bar) that are common in many public electoral processes—such as media polls, “mainstream” media reports and analyses—many other real-world sources of both small- and large-scale influence can be modeled via network structure. For instance, individuals whose opinion reaches an inordinately large number of others (such as might be expected of some political bloggers) can be modeled by high-degree vertices. Cohesive or close-knit groups of like-minded individuals can be modeled by subsets of vertices with similar incentives and dense connectivity. Our experiments deliberately introduce such structures and others. We also remark that our demand for complete unanimity before any payoffs are made is an abstraction of most real decision-making and voting processes, where a sufficiently strong consensus is typically enough to yield the benefits of unity. Although we expect most of our findings would be robust to such weakening, we leave its investigation to future research.

The experiments described here are part of an extensive and continuing series that have been conducted at the University of Pennsylvania since 2005, in which collective problem-solving from only local interactions in a network has been studied on a wide range of tasks, including graph coloring (6), trading of virtual goods (7), and several other problems. An overarching goal of this line of research is to establish the ways in which network structure and task type and difficulty interact to influence individual and collective behavior and performance.

Author contributions: M.K., S.J., J.T., and J.W. designed research; M.K., S.J., J.T., and J.W. performed research; M.K., S.J., J.T., and J.W. contributed new reagents/analytic tools; M.K., S.J., J.T., and J.W. analyzed data; and M.K. wrote the paper.

The authors declare no conflict of interest.

This article contains supporting information online at www.pnas.org/cgi/content/full/0808147106/DCSupplemental.

© 2009 by The National Academy of Sciences of the USA
The overarching goal of the Cohesion experiments was to systematically investigate how collective and individual performance and behavior varied with neighborhood diversity and the strength of preferences. Although it is perhaps most natural to hypothesize that increased inter-group connectivity should improve collective performance—this would be consistent with several mathematical network theories and metrics, including the aforementioned cohesion, and notions of expansion from the graph theory literature (9)—the degree of improvement, and how it might be influenced by the detailed structure (Erdos–Renyi vs. preferential attachment), the variability of individual human behavior, and so on, are difficult to predict.

In the Minority Power experiments, all networks were generated via preferential attachment (10). A minority of the vertices with the highest degrees (number of neighbors) were then assigned incentives preferring red global consensus to blue, whereas the remaining majority were assigned incentives preferring blue global consensus. The size of the chosen minority was varied (6, 9, or 14), as were the relative strengths of preferences. See Fig. 2 and the SI for further details.

The overarching goal of the Minority Power experiments was to systematically investigate the influence that a small but well-connected set of individuals could have on collective decision-making—in particular, to investigate whether such a group could reliably cause their preferred outcome to hold globally and unanimously.

For each of the different network structures in the Cohesion and Minority Power families, we ran experiments in which there were “strong symmetric,” “weak symmetric,” and “asymmetric” incentive structures. By “symmetric” we mean that the incentives of those players preferring blue and those preferring red were symmetrically opposed (such as $0.75/$1.25 for consensus to red/blue vs. $1.25/$0.75); by “weak” and “strong” we refer to the relative magnitudes of the preferred and non-preferred payments ($1.25 to $0.75 for weak, $1.50 to $0.50 for strong). In the asymmetric incentives experiments, the group preferring one color would be given strong incentives, whereas groups preferring the other color would be given weak incentives. We thus imposed scenarios in which 2
opposing groups “cared” equally but mildly about the global outcome, equally and strongly, or in which one group cared more than the other.

Thus, each of the 9 network structures was combined with weak symmetric, strong symmetric and asymmetric incentive schemes, yielding 27 distinct scenarios that were each executed in 3 trials, for a total of 81 experiments.

**Human Subject Methodology**

We now briefly remark on some further details of the experimental methodology and system. All experiments were held in a single session lasting several hours, and the participants were 36 University of Pennsylvania students enrolled in an undergraduate survey course on network science. Each of the 81 experiments had a fixed network and incentive structure, and the system assigned each of the 36 subjects randomly to one of the 36 network positions at the start of each experiment, thus assuring there was no systematic bias in the position of subjects in the networks. To prevent the establishment of social conventions that could trivialize the experiments (such as all subjects playing red for the remainder of the session following a successful global consensus to red), the system used a local randomization scheme on the colors, which might make what appeared red to one player appear blue to another. Each experiment had a 1-min limit for the population to reach a unanimous color choice; if they did so before then, the experiment ended and payments were tallied by the system. The session was closely proctored to ensure that no communication between subjects took place outside of the system, and physical partitions were erected around workstations to prevent inadvertent information leakage.

**Results**

**Collective Behavior.** Overall the subject population exhibited fairly strong collective performance. Of the 81 experiments, 55 ended in global consensus within 1 min (resulting in some payoff to all participants), with the mean completion time of the successful experiments being 43.9 s (standard deviation 9.6 s). We now proceed to describe more specific findings quantifying the impact of network structure, incentive schemes, and individual behavior.

Network structure influenced collective performance in a variety of notable ways. The Cohesion experiments were considerably harder for the subjects than the Minority Power experiments; only 31 of 54 of the former were solved compared with 24 of 27 of the latter (difference significant at $P < 0.001$). Furthermore, in all 24 of the successfully completed Minority Power experiments, the global consensus reached was in fact the preferred color of the well-connected minority. Together these results suggest that not only can an influentially positioned minority group reliably override the majority preference, but that such a group can in fact facilitate global unity.

Within the Cohesion experiments, generating connectivity according to preferential attachment (20/27 solved) yielded better collective performance than generating it via Erdos–Renyi (11/27 solved; difference significant at $P < 0.015$). When combined with the high success rate of the preferential attachment Minority Power experiments (the difference between the 44/54 solved instances of all preferential attachment networks and the 11/27 solved Erdos–Renyi networks is significant at $P < 0.001$), this finding indicates that, for this class of consensus problems, preferential attachment connectivity may generally be easier for subjects than Erdos–Renyi connectivity, an interesting contrast to problems of social differentiation such as graph coloring (6), where preferential attachment networks appear to create behavioral difficulties.

Independent of the method for generating connectivity, Cohesion performance improved systematically as within-group connectivity was replaced by between-group connectivity, with the strongest performance coming from Cohesion networks in which most subjects might have a preferred color different from those of a majority of their neighbors. Across all Cohesion experiments, the success rate on the networks with the highest level of inter-group connectivity (14/18 solved) and the success rate when connectivity was either mainly intra-group or balanced (17/36 solved) are significantly different ($P < 0.03$). Thus, increased awareness of the presence of opposing preferences imposes social welfare. In terms of behavioral collective dynamics, it appears that this awareness leads to early “experimentation” with subjects’ nonpreferred colors, resulting in more rapid mixing of the population choices.

Across all network structures, asymmetric incentives yielded the strongest collective performance (the overall asymmetric success rate of 22/27 differs from the combined weak/strong symmetric success rate of 33/54 at $P < 0.05$), and, indeed, the extremist’s preferences were dominant, determining the consen-
sus outcome in 18 of the 22 successful asymmetric experiments. Strong symmetric incentives (14/27 successes) yielded worse performance than weak symmetric ones (19/27 successes). Thus, it appears most beneficial to have extremists present in a relatively indifferent population, and most harmful to have 2 opposing extremist groups.
The results on collective behavior described so far have focused on the final outcomes of experiments. The collective dynamics within individual experiments is also revealing, and shows notable effects of network structure. In Fig. 3 we provide visualizations of the collective dynamics in each of the 81 experiments, grouped by network structure and incentive scheme. As described in the caption, for each experiment there is a plot charting the progression of the number of players choosing the eventual consensus or majority color as a function of time within the experiment. Notable features include a ritual initial flurry of activity away from the minority preference in the Minority Power experiments, followed by an inevitable assertion of the minority influence over the population. There are also many instances in which a significant fraction of the experiment is spent quite far away from the eventual consensus choice, including near-total reversals of the collectively chosen color; see Fig. 3 and its caption for further details.

Although these visualizations of the dynamics are rich in detail, it is difficult to extract meaningful structural effects from them. In Fig. 4 we thus show the results of fitting simple 2-segment random walk models to the experimental dynamics within each family of experiments (fixed network and incentive structure). These models clearly show the effects of structure on collective dynamics. In terms of the rate of approach to the eventually favored color, Cohesion experiments with Erdos–Renyi connectivity tend to both begin and end slowly, whereas those with preferential attachment connectivity begin slowly but end more rapidly. Higher inter-group connectivity consistently increased late-game speed toward consensus. The Minority Power dynamics ended relatively fast, but early speed was heavily influenced by the size of their minorities.

**Individual Behavior.** It is natural to investigate the extent to which different human subjects exhibited distinct strategies or styles of play across the experimental session, and the degree to which such stylistic differences did or did not influence individual earnings. For any measure $M$ of individual subject behavior within an experiment (such as the number of color changes made by the subject), we can compute the 36 average values for $M$ obtained by taking the 81-game average for each subject, and compare these to the distribution of “random observer” averages, obtained by picking a random subject to observe in each experiment, and averaging the resulting 81 $M$ values. Because subjects were in fact randomly assigned their network positions and incentives at the start of each experiment, if the variance of the 36 actual subject averages significantly exceeds that of the random observer distribution (according to a standard variance test), we can conclude that subjects exhibited meaningful (greater than chance) variation on measure $M$. See Fig. 5.

Most noteworthy is the fact that when the measure is wealth, subjects did not exhibit meaningful variation—thus the disparity in average or total wealth across the session (which ranged from $46.50$ total earnings to $58.75$, with a mean of $52.76$ and standard deviation of $2.46$) is already well-explained by the random assignment of subjects to positions. However, this finding in no way precludes the possibility that subjects still display distinct “personalities,” nor that these differences might strongly correlate with final wealth. For instance, subject “stubbornness”—as measured by the amount of time a subject is playing their preferred color, but is the minority color in their neighborhood—varies meaningfully (Fig. 5) and is positively correlated with average wealth (correlation coefficient $\approx 0.43$, $P < 0.01$). Being stubborn at the outset of an experiment (during the first 9 s) shows even stronger correlation with wealth (correlation coefficient $\approx 0.55$, $P < 0.001$). The number of color changes made by subjects in the opening seconds of an experiment also varies significantly (Fig. 5) and is strongly negatively correlated with wealth ($-0.58$, $P < 0.001$). Together, these results suggest that stubborn and stable players set the tone of an experiment early.

Player stubbornness warrants further investigation, because it strikes at the heart of the tension that is a focal point of the experiments—by being stubborn, one might improve the chances of swaying the population toward one’s preferred color, but one also risks preventing global consensus being reached in time (and thus forgoing any payoff). It is clear that no subject was infinitely...
stubborn: The wealthiest player had their preferred color 28 times out of 55 successful games but acquiesced to group dynamics and accepted the lower payoff 27 times. All other players acquiesced more often—up to as many as 40 times out of 55. In the 26 games that failed to achieve unanimity, there were only 30 individual cases of players defying all of their neighbors as time expired, and only 5 games ended in failure due to players that defied all neighbors for more than the last 2 seconds of play. Only 3 individual players ever caused this kind of failure; one did it 3 times, but also acquiesced 38 out of 55 times and garnered relatively poor overall earnings. These facts combined with the aforementioned correlation of stubbornness with wealth suggest that successful players managed to be “tastefully” stubborn, and that overall behavior was quite acquiescent.

In addition to the raw experimental data, subjects were given an exit survey in which they were invited to comment on their own and others’ strategies, and these surveys provide a rich and often consistent source of insight into individual styles of play. Twenty-four subjects explicitly mentioned starting off by choosing the color that would give them the higher payoff upon consensus. Twenty-seven subjects mentioned either trying to signal others, or noticing others trying to signal; however, many also found this behavior annoying and said that it did not help. Twenty-one subjects noticed others being irrationally stubborn, or expressed suspicion that others were being irrationally stubborn. (Here we use the term “stubborn” in the informal way it was given in the surveys, as opposed to the formal measure discussed above.) Three subjects mentioned being stubborn themselves because they did not want small payoffs. Seven subjects mentioned using different strategies depending on whether their incentives were weak ($0.75 vs. $1.25) or strong ($0.50 vs. $1.50). Three subjects mentioned changing their behavior as the night progressed, 1 subject developed a strategy, or strong ($0.50 vs. $1.50). Three subjects mentioned either trying to signal others, or noticing others being irrationally stubborn, or expressed suspicion that others were being irrationally stubborn. (Here we use the term “stubborn” in the informal way it was given in the surveys, as opposed to the formal measure discussed above.) Three subjects mentioned being stubborn themselves because they did not want small payoffs. Seven subjects mentioned using different strategies depending on whether their incentives were weak ($0.75 vs. $1.25) or strong ($0.50 vs. $1.50). Three subjects mentioned changing their behavior as the night progressed, 1 subject developed a strategy, or strong ($0.50 vs. $1.50). Three subjects mentioned following the action choices of whose neighbors are currently playing c, plays c in the next time step with probability proportional to w(c)f(c) of whose neighbors are currently playing c, plays c in the next time step with probability proportional to w(c)f(c) (11). Such agents combine their preferences (as given by the values w(c)) with the current trend in their neighborhoods (the f(c)) to stochastically select their next color in a natural manner. If such agents are simulated using the same networks and incentives as in the 81 human subject experiments, and the number of simulation steps is capped (as it effectively is by the 1-min time limit of the human experiments), there is rather strong correlation (0.60, P < 0.001) between subject and simulation times to consensus.

Discussion
A number of further investigations are suggested by the findings summarized here. In particular, the variations in individual behavior and the apparently helpful presence of “extremists” raise the question of whether certain mixtures of behaviors and attitudes are required for optimal collective problem-solving. It would also be interesting to use the data from our experiments to develop richer statistical models of individual and population behavior, whose predictions in turn could be tested on further behavioral experiments.

Fig. 5. Illustration of the “random observer” method for detecting meaningful variation in subject behavior. (Left) Empirical cumulative distribution function (CDF) of total player wealth (blue), in which wealth (x axis) is plotted against the fraction of the 36 subjects earning at least that amount (y axis). It is very well-modeled by the theoretical expected CDF generated by choosing a random player’s wealth independently in each experiment (orange), so we may conclude that the variation in player wealth is explained by the random assignments to network position. In contrast, the CDFs of the number of color changes taken by each player in the first several seconds (Middle) and the total amount of “stubborn” time (Right) are poorly modeled by the random observer CDF, showing considerably greater variance in both cases. See text for details.