Human subjects perform a computationally wide range of tasks from only local, networked interactions.

BY MICHAEL KEARNS

Experiments in Social Computation

SINCE 2005, WE have conducted an extensive series of behavioral experiments at the University of Pennsylvania on the ability of human subjects to solve challenging global tasks in social networks from only local, distributed interactions. In these experiments, dozens of subjects simultaneously gather in a laboratory of networked workstations, and are given financial incentives to resolve "their" local piece of some collective problem, which is specified via individual incentives and may involve aspects of coordination, competition, and strategy. The underlying network structures mediating the interaction are unknown to the subjects, and are often chosen from well-studied stochastic models for social network formation. The tasks examined have been drawn from a wide variety of sources, including computer science and complexity theory, game theory and economics, and sociology. They include problems as diverse as graph coloring, networked trading, and biased voting. This article surveys these experiments and their findings.

Our experiments are inherently interdisciplinary, and draw their formulations and motivations from a number of distinct fields. Here, I mention some of these related areas and the questions they have led us to focus upon.

▸ **Computer science.** Within computer science there is current interest in the field's intersection with economics (in the form of algorithmic game theory and mechanism design[22]), including on the topic of strategic interaction in networks, of which our experiments are a behavioral instance. Within the broader technology community, there is also rising interest in the phenomenon of crowdsourcing,[26] citizen science,[18] and related areas, which have yielded impressive "point solutions," but which remains poorly understood in general. What kinds of computational problems can populations of human subjects (perhaps aided by traditional machine resources) solve in a distributed manner from relatively local information and interaction? Does complexity theory or some variant of it provide any guidance? Our experiments have deliberately examined a wide range of problems with varying computational difficulty and strategic properties. In particular, almost all the tasks we have examined entail much more interdependence between user actions than most crowdsourcing efforts to date.

▸ **Behavioral economics and game theory.** Many of our experiments have

» **key insights**

■ **Groups of human subjects are able to solve challenging collective tasks that require considerably more interdependence than most fielded crowdsourcing systems exhibit.**

■ **In its current form, computational complexity is a poor predictor of the outcome of our experiments. Equilibrium concepts from economics are more appropriate in some instances.**

■ **The possibility of Web-scale versions of our experiments is intriguing, but they will present their own special challenges of subject recruitment, retention, and management.**

an underlying game-theoretic or economic model, and all are conducted via monetary incentives at the level of individual subjects. They can thus be viewed as experiments in behavioral economics,[1] but taking place in (artificial) social networks, an area of growing interest but with little prior experimental literature. In some cases we can make detailed comparisons between behavior and equilibrium predictions, and find systematic (and therefore potentially rectifiable) differences, such as networked instances of phenomena like inequality aversion.

▸ **Network science.** Network Science is itself an interdisciplinary and emerging area[9,25] that seeks to document "universal" structural properties of social and other large-scale networks, and ask how they might form and influence network formation and dynamics. Our experiments can be viewed as extending this line of questioning into a laboratory setting with human subjects, and examining the ways in which network structure influences human behavior, strategies, and performance.

▸ **Computational social science.** While our experimental designs have often emphasized collective problem solving, it is an inescapable fact that individual human subjects make up the collective, and individual decision-making, strategies, and personalities influence the outcomes. What are these influences, and in what ways do they matter? In many of our experiments there are natural and quantifiable notions of traits like stubbornness, stability, and cooperation whose variation across subjects can be measured and correlated with collective behavior and performance, and in turn used to develop simple computational models of individual behavior for predictive and explanatory purposes.

This article surveys our experiments and results to date, emphasizing overall collective performance, behavioral phenomena arising repeatedly across different tasks, task- and network-specific findings that are particularly striking, and the overall methodology and analyses employed. It is worth noting at the outset that one of the greatest challenges posed by this line of work has been the enormous size of the design space: each experimental session involves the selection of a collective prob-

lem, a set of network structures, their decomposition into local interactions and subject incentives, and values for many other design variables. Early on we were faced with a choice between breadth and depth—that is, designing experiments to try to populate many points in this space, or picking very specific types of problems and networks, and examining these more deeply over the years. Since the overarching goal of the project has been to explore the broad themes and questions here, and to develop early pieces of a behavioral science of human computation in networked settings, we have opted for breadth, making direct comparisons between some of our experiments difficult. Clearly much more work is needed for a comprehensive picture to emerge.

In the remainder of this article, I describe the methodology of our experiments, including the system and its GUIs, human subject methodology, and session design. I then summarize our experiments to date and remark on findings that are common to all or most of the different tasks and highlight more specific experimental results on a task-by-task basis.

Experimental Methodology
All of the experiments discussed here were held over a roughly six-year period, in a series of approximately two-hour sessions in the same laboratory of workstations at the University of Pennsylvania. The experiments used an extensive software, network and visualization platform we have developed for this line of research, and which has been used by colleagues at other institutions as well. In all experiments the number of simultaneous subjects was approximately 36, and almost all of the subjects were drawn from Penn undergraduates taking a survey course on the science of social networks.[12] Each experimental session was preceded by a training and demonstration period in which the task, financial incentives, and GUI were explained, and a practice game was held. Sessions were closely proctored to make sure subjects were attending to their workstation and understood the rules and GUI; under no circumstances was advice on strategy provided. Physical partitions were erected around workstations to ensure subjects could only see their own GUI.

> While our experimental designs have often emphasized collective problem solving, it is an inescapable fact that individual human subjects make up the collective, and individual decision-making, strategies, and personalities influence the outcomes.

No communication or interaction of any kind outside that provided by the system was permitted. The system tabulated the total financial compensation earned by each subject throughout a session, and subjects were paid by check at a later date following the session. Compensation was strictly limited to the actual earnings of each individual subject according to their own play and the rules of the particular task or game; there was no compensation for mere participation. Following a session, subjects were given an exit survey in which they were asked to describe any strategies they employed and behaviors they observed during the experiments.

Within an individual experimental session, the overall collective task or problem was fixed or varied only slightly (for example, an entire session on graph coloring), while the underlying network structures mediating the interaction would vary considerably. Thus, the sessions were structured as a series of short (1 to 5 minutes) experiments, each with its own network structure but on the same task. This is the natural session format, since once the task and incentives are explained to the subjects, it is relatively easy for them to engage in a series of experiments on differing networks, whereas explaining a new task is time-consuming. Each experiment had a time limit imposed by the system, in order to ensure the subjects would not remain stuck indefinitely on any single experiment. In some sessions, there were also conditions for early termination of an experiment, typically when the instance was "solved" (for example, a proper coloring was found). A typical session thus produced between 50 and 100 short experiments.
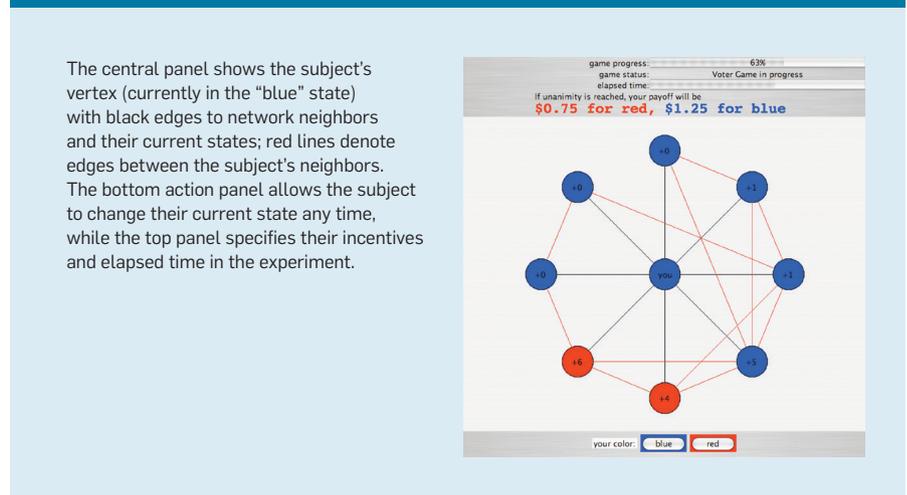
Within an individual experiment, the system randomly assigned subjects to one of the vertices in the network (thus there was neither persistence nor identifiability of network neighbors across experiments). Each subject's GUI (see Figure 1) showed them a local view of the current state of the network—usually a local fragment of the overall network in which the subject's vertex was in the middle and clearly labeled, as well as edges shown to the subject's network neighbors. Edges between a subject's neighbors were shown as well, but no more distant structure. The GUI also always clearly showed the incentives and current payoffs for each subject (which might vary from subject to subject within an experiment), as well the time remaining in the experiment. Typical incentives might pay subjects for being a different color than all their neighbors (graph coloring), the same color (consensus), or perhaps the same color but with different payoffs for different colors (biased voting). Other experiments involved financial scenarios, and the interface provided a mechanism for subjects to bargain or trade with their network neighbors. In general, GUIs always provided enough information for subjects to see the state of their neighbors' current play, and for them to determine their current (financial) best response.

Summary of Experiments

The accompanying table briefly summarizes the nature of the experiments conducted to date, describing the collective task, the network structures used, the individual incentives or mechanism employed, and some of the main findings that we detail below. Our first remark is on the diversity of these experiments along multiple dimensions. In terms of the

Figure 1. Sample screenshot of subject GUI for a biased-voting experiment; many other sessions involved similar GUIs.

The central panel shows the subject's vertex (currently in the "blue" state) with black edges to network neighbors and their current states; red lines denote edges between the subject's neighbors. The bottom action panel allows the subject to change their current state any time, while the top panel specifies their incentives and elapsed time in the experiment.

Summary of experiments to date. ER stands for Erdös-Renyi, PA for preferential attachment.

Task Description	Networks	Incentives/Mechanism	Sample Findings
graph coloring[17]	cycle+chords; PA	differ with neighbors	chords help; importance of information view
coloring and consensus[10]	clique chain w/rewiring	differ/agree with neighbors	opposite structure/task effects
networked trade[13]	ER; PA; structured; all bipartite	limit orders for trades for opposing good	comparison to equilibrium theory; networked inequality aversion
networked bargaining[3]	assorted	Nash bargain on each edge	behavioral price of obstinacy
independent set[15]	assorted	kings and pawns with side payments	side payments help; conflict and fairness
biased voting[14]	ER and PA between types; minority power	consensus with competing individual preferences	well-connected minority rules
network formation[16]	endogenous to the game	biased voting minus edge expenditures	poor collective performance

tasks, the computational complexity of the problems studied[a] varies from the trivial (biased voting and consensus, though this latter problem is difficult in standard models of distributed computation); to the tractable but challenging (networked trade, for which the closest corresponding algorithmic problem is the computation of market equilibria); to the likely intractable (graph coloring and independent set, both *NP*-hard). In terms of the networks, we have investigated standard generative models from the literature such as Erdös-Renyi, preferential attachment, and small worlds; highly structured networks whose design was chosen to highlight strategic tensions in the task and incentives; regular networks without obvious mechanisms to break symmetry; and

various other topologies. Figure 2 depicts visualizations of a sampling of network structures investigated. And finally, regarding the financial incentives, these have varied from cooperative (tasks where all players could simultaneously achieve their maximum payoff in the solution); to competitive (where higher payoffs for some players necessarily entail lower payoffs for others); to market-based trading and bargaining, where there are nontrivial networked equilibrium theories and predictions; and to settings where side payments were permitted.

Despite this diversity, and the difficulties in making direct comparisons across sessions and experiments it engenders, there is one unmistakable commonality that has emerged across our six-year investigation: human subjects perform remarkably well at the collective level. While we have observed significant variability in performance across tasks, networks, and incentives, overall the populations have consistently exceeded our expectations. There is a natural and easy way of quantifying this performance: for any given short experiment, we of

course know the exact network used, and the incentives and their arrangement within the network, and thus can compute the maximum welfare solution for that particular experiment—that is, the state or arrangement of subject play that would generate the greatest collective payments to the subjects. For each experiment, our system has also recorded the actual payments made, which are by definition less than the maximum social welfare. We can thus sum up all of the actual payments made across all sessions and experiments, and divide it by the sum of all the maximum social welfare payments to arrive at a measure of the overall efficiency of the subject pools over the years.

The resulting figure across the lifetime of our project[b] is 0.88—thus, overall subjects have extracted close to 90% of the payments available to them in principle. In interpreting this figure it should be emphasized that it is an average taken over the particular ensemble of tasks and networks we have studied, which as mentioned before was chosen for its breadth and not in a globally systematic fashion. Clearly it is possible to craft behaviorally "hard" problems and networks.
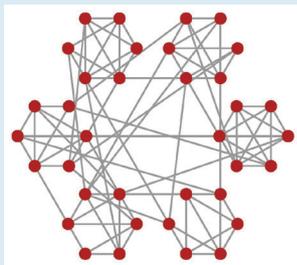
Nevertheless, their efficiency shows that subjects are capable of high performance on a wide variety of tasks and graph topologies.

Another phenomenon consistent across tasks has been the importance of network structure. For most tasks, we found there was a systematic and meaningful dependence of collective behavior on structure, and often an approximate ordering of difficulty of the network topologies could be inferred. Thus, simple cycles prove more difficult for coloring than preferential attachment networks,[17] denser networks result in higher social welfare in networked trading,[13] and so on. However, such dependences on structure are highly task-specific—which is perhaps not surprising for fixed heuristics or algorithms, but has not been documented behaviorally before. Indeed, in one set of experiments we isolated
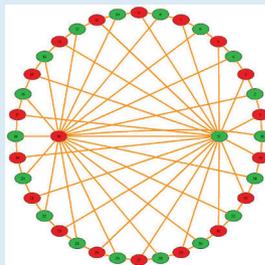
---

a Clearly computational complexity provides limited insight at best here, since it examines worst-case, centralized, asymptotic computation, all of which are violated in the experiments. But it remains the only comprehensive taxonomy of computational difficulty we have; perhaps these experiments call for a behavioral variant, much as behavioral game theory has provided for its parent field.

Figure 2. A small sampling of network structures in experiments.

(a) from consensus and independent-set experiments, a chain of six cliques of size 6, with a fraction of the internal clique edges "rewired" to random vertices, thus allowing interpolation between a highly "tribal" network and effectively random networks.

(b) from coloring experiments, an engineered structure with a cycle and two "leaders" in a two-colorable graph.

(c) from biased-voting experiments, a preferential attachment network with a minority of high-degree players preferring red.

(d) from many tasks, a sample Erdös-Renyi network.

---

b This excludes the most recent experiments in network formation, which are of a qualitatively different nature than the rest, and result in a rather surprising outcome discussed later.
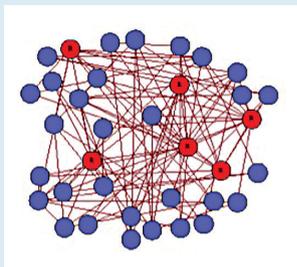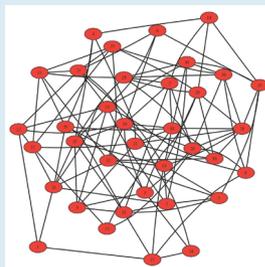
this phenomenon by showing that for two cognitively similar (but computationally different) problems, and for a particular generative model for networks, the effects of structure on collective behavioral performance is the opposite in the two tasks,[10] a finding discussed later in greater detail.

The third consistency we found across both tasks and networks was the emergence of individual subject "personalities" or behavioral traits. Our experimental platform is deliberately stylized, and effectively shoehorns the complexity of real human subjects into a highly constrained system, where language, emotion, and other natural forms of communication are eradicated, and all interactions must take place only via simple actions like selecting a color or offering a trade. While there are obvious drawbacks to this stylization in terms of realism, one benefit is that when we make a clear finding—such as the ability of a small but well-connected minority to systematically impose its preferences on the majority[14]—we have done so in a way that might identify the minimal network and task conditions for it to emerge.

Nevertheless, in our experiments we consistently find subjects differentiating and expressing themselves within the constraints of our system in ways that can be measured and compared. For instance, in many of our experiments there are natural notions of traits like stubbornness, stability, selfishness, patience, among others, that can be directly measured in the data, and the frequency of such behavior tallied for each subject. We often find the variation in such behaviors across a population indeed exceeds what can be expected by chance, and thus can be viewed as the personalities of human subjects peeking through our constraints. Harder to measure but still clearly present in almost every experiment we have conducted is the emergence of (sometimes complex) "signaling" mechanisms—it seems that when our system takes language away, the first thing subjects do is try to reintroduce it. From such behavioral traits arise many interesting questions, such as whether specific traits such as stubbornness are correlated with higher

> A theme running throughout our experiments is that intuitions about what networks might be easy or difficult can be strongly violated when considering a distributed human population using only local information.

payoffs (sometimes they are, other times not), and whether certain mixtures of subject personalities are necessary for effective collective performance (such as a mixture of stubborn and acquiescent individuals in coordination problems).

Highlights of Results
We now turn our attention to results at the level of specific tasks. For each task, I briefly outline any noteworthy details of the GUI or experimental setup, and then highlight some of the main findings.

Coloring and consensus. Our first set of experiments[17] explored the behavioral graph coloring task already alluded to—subjects were given financial incentives to be a different color than their network neighbors, saw only the colors of their local neighborhood, and were free to change their color at any time, choosing from a fixed set of colors whose size was the chromatic number of the underlying graph (thus demanding the subjects find an optimal coloring). It was in these initial experiments that we first found strong effects of network structure. For instance, while a simple two-colorable cycle proved surprisingly hard for the subjects—comparable to their difficulty with more complex and dense preferential attachment graphs—this difficulty was greatly eased by the addition of random chords to the cycle, which reduces diameter and increases edge density. But the preferential attachment networks had the smallest diameter and highest edge density, so these structural properties do not alone explain collective performance.

A theme that runs throughout our experiments is that intuitions about what networks might be easy or difficult can be strongly violated when considering a distributed human population using only local information. The challenge of finding simple explanations of such structural results is highlighted by the fact that a natural distributed, randomized heuristic for coloring—namely, not changing colors if there is no current conflict with neighbors, changing to a color resolving a local conflict if one exists, and picking a random color if conflict is unavoidable—produced an ordering of

the difficulty of the networks that was approximately the reverse of that for the subjects.

These first experiments were also the only ones in which we investigated the effects of global information views on performance. In a subset of the experiments, subjects actually saw the current state of the entire network (again with their own vertex in the network clearly indicated), not just the colors of their neighbors. Not surprisingly, this global view led to dramatically improved performance in a simple cycle, where the symmetric structure of the network and the optimal solution become immediately apparent. But strikingly, in preferential attachment networks, global views led to considerable *degradation* in collective performance—perhaps an instance of "information overload," or simply causing subjects to be distracted from attending to their local piece of the global problem.

In a later session,[10] we ran experiments on both coloring and *consensus* (where subjects were given financial incentives to be the same color as their neighbors, chosen from a fixed menu of nine colors), on the same set of underlying networks. Despite the vastly different (centralized) computational complexity of these problems—coloring being *NP*-hard, consensus trivial—the two tasks are cognitively very similar and easy for subjects to switch between: coloring is a problem of social differentiation, consensus one of social coordination.

In these experiments, the networks were drawn from a parametric family that begins with six cliques of size six loosely connected in a chain. A rewiring parameter q determines the fraction of internal clique edges that are replaced with random "long distance" edges, thus allowing interpolation between a highly clustered, "tribal" network, and the Erdös-Renyi random graph model; see Figure 2(a) for an example. The primary finding here was that the effect on collective performance of varying the rewiring parameter is systematic and *opposite* for the two problems— consensus performance benefits from more rewiring, coloring performance suffers. This effect can be qualitatively captured by simple distributed heuristics, but this does not diminish the striking behavioral phenomenon (see Figure 3). The result suggests that efforts to examine purely structural properties of social and organizational networks, without careful consideration of how structure interacts with the task(s) carried out in those networks, may provide only limited insights on collective behavior.

In addition to such systematic, statistically quantifiable results, our experiments often provide interesting opportunities to visualize collective and individual behavior in more anecdotal fashion. Figure 4 shows the actual play during one of the consensus experiments on a network with only a small amount of rewiring, thus largely preserving the tribal clique structure. Each row corresponds to one of the 36 players, and the horizontal axis represents elapsed time in the experiment. The horizontal bars then show the actual color choice by the player at that moment. The first six rows correspond to the players in the first (partially rewired) clique, the next six to the second clique, and so on. The underlying

Figure 3. Average time to global solution for coloring and consensus experiments (solid lines) as a function of edge rewiring in a clique-chain network, and simulation times (dashed lines) on the same networks for distributed heuristics. The parametric structure has the opposite effect on the two problems.
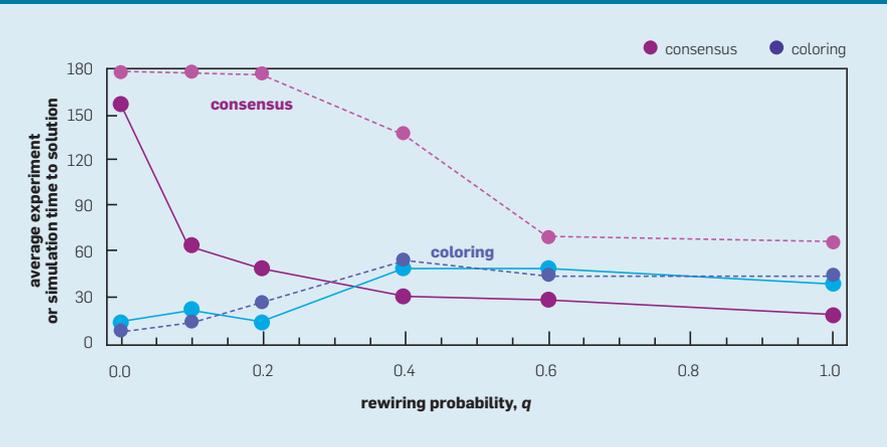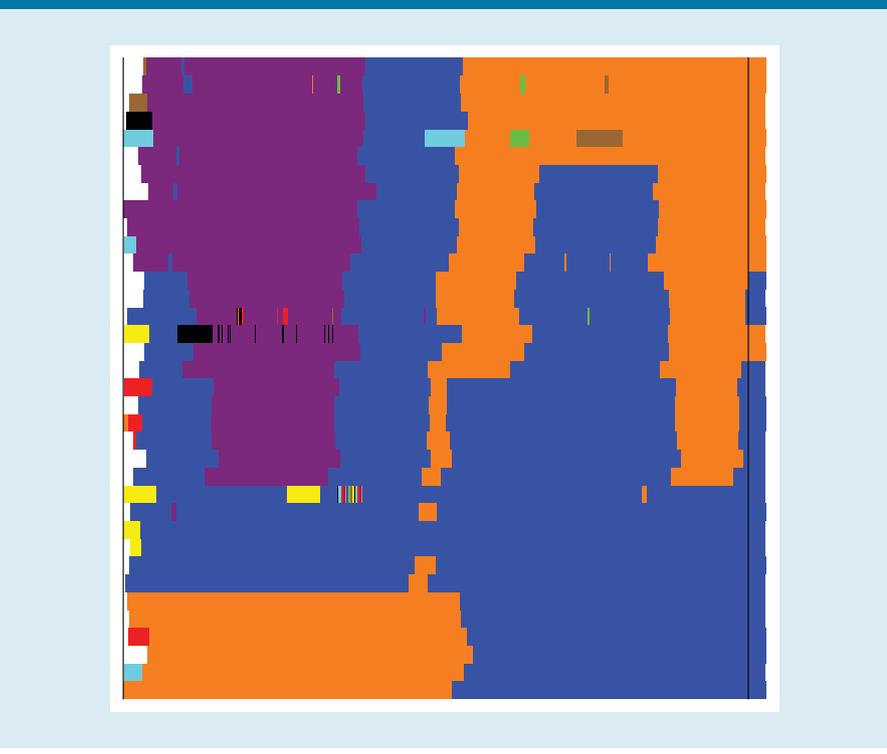

Figure 4. Visualization of a consensus experiment with low rewiring parameter, showing collective and individual behaviors, and effects of underlying clique structure.

network structure manifests itself visually in the tendency for these groups of six to change colors approximately simultaneously. As was typical, after an initial diversity of colors, the population quickly settles down to just two or three, and nearly converges to blue before a trickle of orange propagates through the network and takes firm hold; at some point the majority is orange, but this wanes again until the experiment ends in deadlock. Acts of individual signaling (such as toggling between colors) and (apparent) irrationality or experimentation (playing a color not present anywhere else in the network) can also be observed.

Networked trading and bargaining. Our experiments on trading and bargaining differ from the others in that they are accompanied by nontrivial equilibrium theories that generalize certain classical microeconomic models to the networked setting.[4,11] In the networked trading experiments,[4] there were two virtual goods available for trade—call them milk and wheat—and two types of players: those that start with an endowment of milk, but whose payoff is proportional only to how much wheat they obtain via trade; and those that start with wheat but only value milk. All networks were bipartite between the two types of players, and trade was permitted only with network neighbors; players endowed with milk could only trade for wheat and vice-versa, so there were no "resale" or arbitrage opportunities. All endowments were fully divisible and equal, so the only asymmetries are due to network position. The system GUI allowed players to broadcast to their neighbors a proposed rate of exchange[c] of their endowment good for the other good in the form of a traditional limit order in financial markets, and to see the counter offers made by their neighbors; any time the rates of two neighboring limit orders crossed, an irrevocable trade was booked for both parties.

For the one-shot, simultaneous trade version of this model, there is a detailed equilibrium theory that precisely predicts the wealth of every player based on their position in the network;[11] in brief, the richest and poorest players at equilibrium are determined by finding the subset of vertices whose neighbor set yields the greatest contraction,[d] and this can be applied recursively to compute all equilibrium wealths. An implication is that the only bipartite networks in which there will not be variation in player wealths at equilibrium are those that contain perfect matchings. One of the primary goals of the experiment was to test this equilibrium theory behaviorally, particularly because equilibrium wealths are not determined by local structure alone, and thus might be challenging for human subjects to discover from only local interactions; even the best known centralized algorithm for computing equilibrium uses linear programming as a subroutine.[5] We again examined a wide variety of network structures, including several where equilibrium predictions have considerable variation in player wealth.
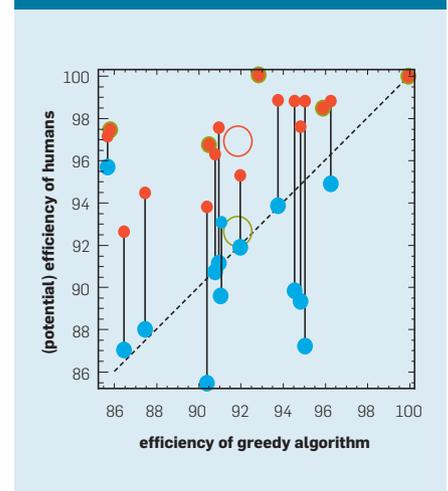
There were a number of notable findings regarding the comparison of subject behavior to the equilibrium theory. In particular, across all experiments and networks, there was strong *negative* correlation between the equilibrium predicted variation of wealth across players, and the collective earnings of the human subjects—even though there was strong *positive* correlation between equilibrium wealth variation and behavioral wealth variation. In other words, the greater the variation of wealth predicted by equilibrium, the greater the actual variation in behavioral wealth, but the more money that was left on the table by the subjects. This apparent distaste for unequal allocation of payoffs was confirmed by our best-fit model for player payoffs, which turned out to be a mixture of the equilibrium wealth distribution and the uniform distribution in approximately a (3/4; 1/4) weighting. Thus the equilibrium theory is definitely relevant, but is improved by tilting it toward greater equality. This

can be viewed as a networked instance of inequality aversion, a bias that has been noted repeatedly in the behavioral game theory literature.[1]

Our experiments on networked bargaining[3] have a similarly financial flavor, and are also accompanied by an equilibrium theory.[4] In these experiments, each edge in the network represents a separate instance of Nash's bargaining game:[21] if by the end of the experiment, the two subjects on each end of an edge can agree on how to split $2, they each receive their negotiated share (otherwise they receive nothing for this edge). Subjects were thus simultaneously bargaining independently with multiple neighbors for multiple payoffs. Network effects can arise due to the fact that different players have different degrees and thus varying numbers of deals, thus affecting their "outside options" regarding any particular deal. In many experiments, the system also enforced limits on the number of deals a player could close; these limits were less than the player's degree, incentivizing subjects to shop around for the best deals in their neighborhood. The system provided a GUI that let players make and see separate counter offers with each of their neighbors.

Perhaps the most interesting finding regarded the comparison between subject performance and a simple
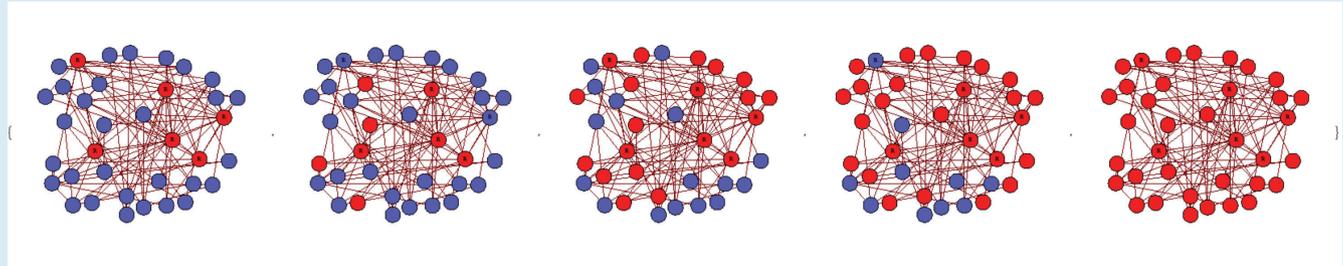
---

c As per the theoretical model, players were not able to offer different rates to different neighbors; thus conceptually prices label vertices, not edges.

d For instance, a set of 10 milk players who collectively have only three neighboring wheat players on the other side of the bipartite network has a contraction of 10/3.

Figure 7. Series of snapshots of global state in a minority power biased voting experiment, showing an instance in which a minority player (upper left vertex R) acquiesces at various times though eventually wins out.

greedy algorithm for approximating the maximum social welfare solution, summarized in Figure 5. This centralized greedy algorithm simply selects random edges in the network on which to close bargains, subject to any deal limits in the experiment, until no further deals could be closed without violating some deal limit. The social welfare obtained (which does not require specifying how the edge deals are split between the two players) is then simply $2 times the number of closed deals, as it is for the behavioral experiments as well.
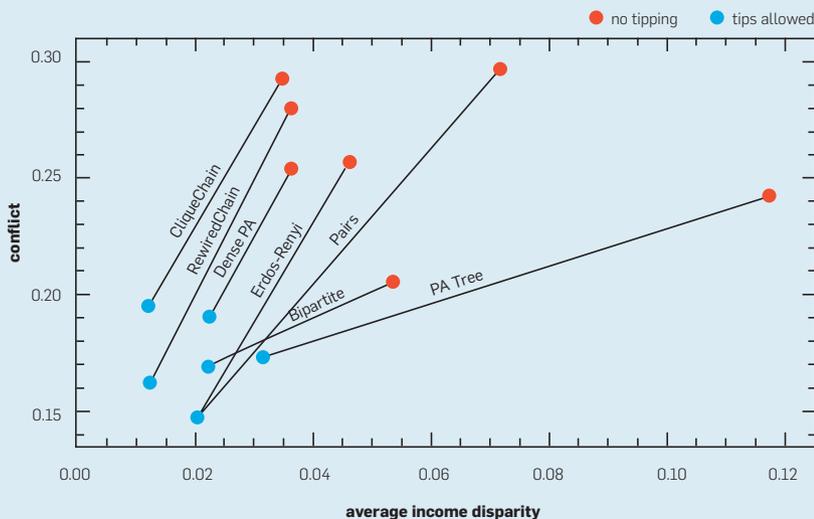
The blue dots in Figure 5 each represent averages over several trials of one of the network topologies examined (thus each dot corresponds to a different topological family). The x value shows the social welfare of the

greedy algorithm as a percentage of the maximum social welfare (optimal) solution, while the y value shows the same measure for the human subjects. Averaged over all topologies, both humans and greedy perform rather well—roughly 92% of optimal (blue open circle). However, while the greedy solutions are maximal and thus cannot be locally improved, much of the inefficiency of the subjects can be attributed to what we might call the *Price of Obstinacy*: at the end of many experiments, there were a number of deals that still could have been closed given the deal limits on the two endpoints, but on which the two human subjects had not been able to agree to a split. If we simply apply the greedy algorithm to the final state of each behavioral experiment, and greed-

ily close as many remaining deals as possible, the *potential* performance of the subjects on each topology, absent obstinacy, rises to the orange dot connected to the corresponding blue dot in the figure. This hypothetical subject performance is now well above the performance of pure greedy (all orange points above the diagonal now), and the average across topologies is close to 97% of optimal (orange open circle). In other words, the human subjects are consistently finding better underlying solutions than those obtained by simply running greedy on the initial graph, but are failing to realize those better solutions due to unclosed deals. While humans may show aversion to inequality of payoffs, they can also be stubborn to the point of significant lost payoffs.

Independent set. Another set of experiments required subjects to declare their vertex to be either a "king" or a "pawn" at each moment, with the following resulting payoffs: any player who is the only one that has declared kingship in his neighborhood enjoys the highest possible rate of pay; but if one or more of their neighbors are also kings, the player receives nothing. On the other hand, pawns receive an intermediate rate of pay regardless of the states of their neighbors. It is easily seen that the Nash equilibria of the one-shot, simultaneous move version of this game are the maximal independent sets (corresponding to the kings) of the graph, while the maximum social welfare state is the largest independent set, whose centralized computation is *NP*-hard. Because we were concerned that computing payoffs based on only the final state of the gain

Figure 6. From independent-set experiments: Average income disparity between neighbors (x-axis) vs. average time neighbors are conflicting kings (y-axis), both with (blue) and without (orange) side payments. Grouped by network structure. The side payments uniformly reduced conflict and disparity.

would lead to an uninteresting global "chicken" strategy (all players declaring king until the final seconds of the experiment, with some players then "blinking" and switching to pawn), in these experiments payoffs accrued continuously according to the pro-rated time players spent in each of the three possible states (pawn, king with no conflicting neighbors, conflicting kings).

Every experiment was run under two conditions—one just as described above, and another in which the GUI included an additional element: in the case that the player was the lone king in their neighborhood, and thus enjoying the highest rate of pay, a slider bar permitted them to specify a fraction of their earnings in that state to be shared equally among all their neighbors (whose pawn status allows the king's high payoff). These "tips" or side payments could range from 0% to 100% in increments of 10%, and could be adjusted at any time. Note that in some cases, depending on network structure, some vertices might be able to obtain a higher rate of pay by being a pawn receiving side payments from many neighboring kings than by being the lone king in their neighborhood.

The most striking finding was that, across a wide variety of network structures, the introduction of the side payments uniformly raised the collective payoffs or social welfare. Side payment rates were often generous, and averaged close to 20%. Furthermore, when side payments are introduced, both the average income disparity between neighboring players, and the amount of time they spend as conflicting kings, are considerably reduced, across all network structures examined (see Figure 6). This suggests that without side payments, subjects used conflict, which reduces the wealth of all players involved, to express perceived unfairness or inequality. The side payments reduce unfairness and consequently reduce conflict, thus facilitating coordination and raising the social welfare.

Biased voting. The biased voting experiments[14] shared with the earlier consensus experiments an incentive toward collective agreement and coordination, but with an important strategic twist. As in consensus, each

> ## The side payments reduce unfairness and consequently reduce conflict, thus facilitating coordination and raising the social welfare.

player had to simply select a color for their vertex, but now only between the two colors red and blue. If within the allotted time, the *entire population* converged unanimously to either red or blue, the experiment was halted and every player received some payoff. If this did not occur within the allotted time, every player received nothing for that experiment. Thus the incentives were now not at the individual level, but at the collective—players had to not only agree with their neighbors, but with the entire network, even though they were still given only local views and interactions.

The strategic twist was that different players were paid different amounts for convergence to the two colors within the same experiment. In particular, some players received a higher payoff for convergence to blue, while others received a higher payoff for convergence to red. Typical incentives might pay blue-preferring players $1.50 for blue convergence and only $0.50 for red, with red-preferring players receiving the reverse. Some experiments permitted asymmetries between higher and lower payoffs, thus incentivizing some players to "care" more about the color chosen by the population. These experiments thus set up a deliberate tension between competing individual preferences and the need for collective unity.

In the most dramatic set of experiments, networks were chosen according to preferential attachment—known to generate a small number of vertices with high degree—and the vast majority of players given incentives that paid more for convergence to blue. However, the minority of vertices preferring red was chosen to be the high-degree vertices. These experiments tested whether a small but well-connected minority could systematically impose its preferences on the majority, thus resulting in suboptimal social welfare.

The answer was resoundingly affirmative: in 27 such "minority power" experiments, 24 of them resulted in the subjects reaching a unanimous choice—in every case, the preferred choice of the well-connected minority. The finding is especially surprising when we remember that since everyone has only local views and information,

the powerful minority has no particular reason to believe they are powerful—in fact, their high degree ensured that at the start of each such experiment, they would see themselves surrounded by players choosing the opposing color. Indeed, the minority players would often acquiesce to the majority early in the experiment (see Figure 7, which shows a series of snapshots of actual play during an experiment). But the dynamics always eventually came to favor the minority choice.

A behavioral network formation game. Our most recent experiments[16] attempted to address what is perhaps the greatest of many artificialities in this line of research: the exogenous imposition of the social network structure mediating interactions. While corporations and other social entities of course often do impose organizational structure, it is natural to believe that in many circumstances, humans will organically construct the communication and interaction patterns required to solve a task efficiently—perhaps even circumventing any imposed hierarchy or structure. Given the aforementioned overall strong performance of our subjects across a wide variety of challenging tasks, even when network structures were complex and not directly optimized for the task, we were naturally interested in whether performance might improve even further if the subjects could collectively choose the networks themselves.

We thus ran among the first experiments in network formation games, on which there is an active theoretical literature.[8,24] We wanted to design such a game in which the formation of the network was not an end in itself, as it is in many of the theoretical works, but was in service of a collective task—which we again chose to be biased voting. The framework was thus as followed: the payoff functions for the players was exactly as described for biased voting, with all players wanting to reach unanimity, but having a preferred (higher payoff) color. Now, however, there were *no edges* in the network at the start of each experiment—every vertex was isolated, and players could thus see only their own color. Throughout the experiment, players could optionally and unilaterally *purchase*

These experiments thus tested whether a small but well-connected minority could systematically impose its preferences on the majority, resulting in suboptimal social welfare.

edges to other players, resulting in subsequent bilateral viewing of each other's colors for the two players; the GUI would adapt and grow each player's neighborhood view as edges were purchased. A player's edge purchases were deducted from any eventual payoffs from the biased voting task (subject to the constraints that net payoffs could never be negative).

Players were thus doing two things at once—building the network by purchasing edges, and choosing colors in the biased voting task. The GUI had an edge purchasing panel that showed players icons indicating the degrees and shortest-path distances of players they were not currently connected to, thus allowing them to choose to buy edges (for instance) to players that were far away in the current network and with high degree, perhaps in the hopes that such players would aggregate information from distant areas of the network; or (for instance) to low-degree vertices, perhaps in the hope of strongly influencing them. The formation game adds to the biased voting problem the tension that while the players must collectively build enough edges to facilitate global communication and coordination, individual players would of course prefer that others purchase the edges.

While there were many detailed findings, the overall results were surprising: the collective performance on this task was by far the worse we have seen in all of the experiments to date, and much worse than on the original, exogenous network, biased voting experiments. Across all experiments (that included some in which the subjects started not with the empty network, but with some "seed" edges that were provided for free), the fraction in which unanimity was reached (and thus players received nonzero payoffs) was only 41%—far below the aforementioned nearly 90% efficiency across all previous experiments. We were sufficiently surprised that we ran control experiments in which a subsequent set of subjects were once again given fixed, exogenously imposed networks—but this time, the "hard" networks created by the network formation subjects in cases where they failed to solve the biased voting task. This was

done to investigate the possibility that the formation subjects built good networks for the task, but either ran out of time to reach unanimity, or included subjects who behaved very stubbornly because they had significant edge expenditures and thus strongly held out for their preferred color.

Performance on the control experiments was even worse. The surprising conclusion seems to be that despite the fact that subjects clearly understood the task, and were now given the opportunity to solve it not on an arbitrary network, but one collectively designed by the population in service of the task, they were unable to do so. One candidate for a structural property of the subject-built networks that might account for their difficulty in the biased voting task is *(betweenness) centrality*, a standard measure of a vertex's importance[e] in a network. Compared to the networks used in the original, exogenous-network biased voting experiments, the distribution (across vertices) of centrality in the subject-built networks is considerably more skewed.[16] This means that in the network formation experiments, there was effectively more reliance on a small number of high-centrality vertices or players, making performance less robust to stubbornness or other non-coordinating behaviors by these players. Indeed, there was moderately positive and highly significant correlation between centrality and earnings, indicating that players with high centrality tended to use their position for financial gain rather than global coordination and information aggregation.

Despite their demonstrated ability to solve a diverse range of computational problems on a diverse set of networks, human subjects seem poor at *building* networks, at least within the limited confines of our experiments so far. Further investigation of this phenomenon is clearly warranted.

Concluding Remarks

Despite their diversity, our experiments have established a number of rather consistent facts. At least in mod-

erate population sizes, human subjects can perform a computationally wide range of tasks from only local interaction. Network structure has strong but task-dependent effects. Notions of social fairness and inequality play important roles, despite the anonymity of our networked setting. Behavioral traits of individual subjects are revealed despite the highly simplified and stylized interactions; with language removed, subjects persistently try to invent signaling mechanisms.

There are a number of recent efforts related to the research described here. Some compelling new coloring experiments[7,20] have investigated the conditions under which increased connectivity improves performance. Our experimental approach has thus far aimed for breadth, but studies such as these are necessary to gain depth of understanding. We have also usually done only the most basic statistical analyses of our data, but others have begun to attempt more sophisticated models.[6]

Perhaps the greatest next frontier is to conduct similar experiments on the Web, where a necessary loss of control over subjects and the experimental environment may be compensated by orders of magnitude greater scale, both in population size and the number of experimental conditions investigated. Recent efforts using both the open web and Amazon's Mechanical Turk online labor market have started down this important path.[2,19,23]

Acknowledgments

Ⓒ

---

e   The betweenness centrality of vertex *v* is average, over all pairs of other vertices *u* and *w*, of the fraction of shortest paths between *u* and *w* in which *v* appears.

References
1. Camerer, C. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003.
2. Centola, D. The spread of behavior in an online social network experiment. *Science 329*, 5996 (2010), 1194–1197.
3. Chakraborty, T., Judd, J.S., Kearns, M., and Tan, J. A behavioral study of bargaining in social networks. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2010, 243–252.
4. Chakraborty, T., Kearns, M., and Khanna, S. Networked bargaining: Algorithms and structural results. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2009,159–168.
5. Devanur, N.R., Papadimitriou, C.H., Saberi, A., and Vazirani, V.V. Market equilibrium via a primal-dual algorithm for a convex program. *Journal of the ACM 55*, 5 (2008).
6. Duong, Q., Wellman, M.P., Singh, S., and Kearns, M. Learning and predicting dynamic behavior with graphical multiagent models. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (2012).
7. Enemark, D.P., McCubbins, M.D., Paturi, R., and Weller, N. Does more connectivity help groups to solve social problems? In *ACM Conference on Electronic Commerce*. ACM Press, New York, 2011, 21–26.
8. Jackson, M.O. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics: Networks, Clubs and Coalitions*, Cambridge University Press, 2005.
9. Jackson, M.O. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, 2010.
10. Judd, S., Kearns, M., and Vorobeychik, Y. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences 107*, 34 (2010), 14978–14982.
11. Kakade, S.M., Kearns, M.J., Ortiz, L.E., Pemantle, R., and Suri, S. Economic properties of social networks. *Neural Information Processing Systems* (2004).
12. Kearns, M. Networked Life. University of Pennsylvania Undergraduate Course; http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife
13. Kearns, M. and Judd, J.S. Behavioral experiments in networked trade. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2008, 150–159.
14. Kearns, M., Judd, S., Tan, J., and Wortman, J. Behavioral experiments in biased voting in networks. *Proceedings of the National Academy of Sciences 106*, 5 (2009), 1347–1352.
15. Kearns, M., Judd, S., and Vorobeychik, Y. Behavioral conflict and fairness in social networks. *Workshop on Internet and Network Economics* (2011).
16. Kearns, M., Judd, S., and Vorobeychik, Y. Behavioral experiments on a network formation game. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2012.
17. Kearns, M., Suri, S., and Montfort, N. An experimental study of the coloring problem on human subject networks. *Science 313* (2006), 824–827.
18. Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popovic, Z., Baker, D., and Foldit Players. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* (2011).
19. Mason, W. and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* (2011).
20. McCubbins, M.D., Paturi, R., and Weller, N. Connected coordination network structure and group coordination. *American Politics Research 37*, 5 (2009), 899–920.
21. Nash, J.F. The bargaining problem. *Econometrica 18*, 2 (1950), 155–162.
22. Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V.V., Eds. *Algorithmic Game Theory*. Cambridge University Press, 2007.
23. Suri, S. and Watts, D.J. Cooperation and contagion in Web-based, networked public goods experiments. *PLoS ONE 6*, 3 (2011).
24. Tardos, E. and Wexler, T. Network formation games and the potential function method. In *Algorithmic Game Theory*, Cambridge University Press, 2007, 487–513.
25. Watts, D.J. *Six Degrees: The Science of a Connected Age*. W.W. Norton & Company, 2003.
26. Wikipedia. Crowdsourcing; http://en.wikipedia.org/wiki/Crowdsourcing en.wikipedia.org/wiki/Crowdsourcing.

Michael Kearns (mkearns@cis.upenn.edu) is a professor in the Computer and Information Science Department of the University of Pennsylvania. His research interests include machine learning, social networks, algorithmic game theory, and computational finance.