The Wisdom of Crowds: Methods of Human Judgement Aggregation

Aidan Lyon, Eric Pacuit

Philosophy, University of Maryland, College Park.

alyon@umd.edu, epacuit@umd.edu

Draft: July 24, 2014.

Abstract

Although the idea is an old one, there has been a recent boom in research into the Wisdom of Crowds, and this appears to be at least partly due to the now widespread availability of the Internet, and the advent of social media and Web 2.0 applications. In this paper, we start by laying out a simple conceptual framework for thinking about the Wisdom of the Crowds. We identify six core aspects that are part of any instance of the Wisdom of the Crowds. One of these aspects, called *aggregation*, is the main focus of this paper. An aggregation method is the method of bringing the many contributions of a crowd together into a collective output. We discuss three different types of aggregation methods: mathematical aggregation, group deliberation and prediction markets.

1 Introduction

The Wisdom of Crowds is all the rage in these heady Web 2.0 days. But the idea is an old one, and one that goes back to the philosophers of antiquity:

> "For the many, of whom each individual is but an ordinary person, when they meet together may very likely be better than the few good, if regarded not individually but collectively, just as a feast to which many contribute is better than a dinner provided out of a single purse."
>
> —– Aristotle, *Politics*, Book III, §XI.

The basic idea is also a simple and familiar one: two heads are often better than one, and more are even better. A classic example comes from a contest of some 800 people at a county fair in Plymouth, 1906. The contest was to guess the weight of an ox, slaughtered and dressed. Francis Galton found that the *average* of the crowd's guesses was within 1% of the true weight of the ox, despite huge errors in most of the individual guesses (see Galton 1907b, 1907a). Somehow, the crowd knew more as a collective than many of its individuals.

Although the idea is an old one, there has been a recent boom in research into the Wisdom of Crowds, and this appears to be at least partly due to the now widespread availability of the Internet, and the advent of social media and Web 2.0 applications. Never before has it been so easy to get a crowd and leverage their collective wisdom for some task. There are now many well-documented and contemporary examples of the so-called Wisdom of Crowds:[1]

- Amazon's product recommendations.
- Wikipedia and Intellipedia.
- Netflix's movie recommendation algorithm.
- Prediction markets.
- Online citizen science.
- Google's PageRank algorithm.

Discussions of these examples (and many others) can be found in Surowieki 2005, Page 2008, Nielsen 2011, and Landemore and Elster 2012. This paper will provide an overview of some of the theory

[1] We say "so-called", because examples of the Wisdom of Crowds often have little to do with the notion of wisdom that philosophers care about (see Andler 2012 for further discussion), and they often involve only a group of people—even just a handful—and not a *crowd* in the usual sense of the word. Nevertheless, we will stick with the words that seem to have stuck.

behind all of the examples. By thinking carefully about what they have in common and how they differ from each other, we can find new ways to make these applications better. Sometimes such research will simply result in better movie recommendation services, but sometimes it will have much more serious consequences. For example, there are now many Web 2.0 tools being designed to help track and predict the outbreaks of emerging infectious diseases (*cf.* Collier *et al.* 2006, Brownstein *et al.* 2009, Keller *et al.* 2009, Lyon *et al.* 2012, 2013) and even to diagnose rare diseases (e.g., Nuwer 2013). By developing a better understanding of the Wisdom of Crowds, we should be able to improve upon such tools, and thereby make better forecasts of disease outbreaks (among other things).

To begin, in section 2, we'll lay out a simple conceptual framework for thinking about the Wisdom of Crowds. We'll identify six core aspects that are part of any instance of the Wisdom of Crowds. One of these aspects is called *aggregation*, and this will be our primary focus for the remainder of the paper. An aggregation method is the method of bringing the many contributions of a crowd together into a collective output. In the example of the crowd at Plymouth guessing the ox's weight, the aggregation method was the *averaging* of the crowd's individual guesses. This, however, is not the only method of aggregation available. In sections 3, 4, and 5, we'll discuss three broad kinds of aggregation methods: mathematical aggregation, group deliberation, and prediction markets.

2 Thinking about the Wisdom of Crowds

A good way to start thinking systematically about the Wisdom of Crowds is to think about what you would do if you had a burning desire to use the Wisdom of Crowds to do something—because, say, it just seems like a fun thing to do.

The very first thing you have to do is decide what you want to achieve. Do you want to predict the outcome of an election? Recommend products to customers? Decide if someone is guilty of a crime? Write an academic paper? Solve a murder mystery? Predict disease outbreaks? Whatever it is you want to do, we will call this the desired *output* of your endeavour; it's what you want to get out of the Wisdom of Crowds. As we'll soon see, it's important to be clear about your desired output, because this can have a big impact on how you use your crowd.

Speaking of which: you need to get yourself a crowd. Perhaps you have one already, because you have some willing friends. Or perhaps you don't have any friends, but you have some cash to rent a crowd. Or perhaps you see a free crowd—e.g., there could be people on Twitter regularly tweeting information that you could use. We'll call the process of getting a crowd *recruitment*—even if you don't recruit anyone in the usual sense of the word. This recruitment process is very important, for there are many things to consider. Does your crowd need to consist of experts on some topic? Or can they just all be regular folk? How large does your crowd have to be?[2] Does your crowd have to be diverse? Will members of the crowd talk to each other? And so on. These are all important and complicated issues to deal with, and we'll put them aside for now; we simply flag them here because they are important.

The next thing to do is decide how your crowd will contribute to your output. For example, if you want to determine someone's guilt or innocence, perhaps your crowd can contribute by giving their own judgements of guilt or innocence. You might then judge the person to be guilty if and only if everyone in your crowd judges the person to be guilty. However, maybe you need to be more nuanced: instead of your crowd giving outright "guilty" or "innocent" verdicts, perhaps you want to know how *confident* they are in their verdicts. If everyone judges the person to be guilty, but they are only 70% confident in their judgements, then you might be reluctant to, say, send the person to death row. When

[2] Psychologists have found that even just single person can function as a crowd of individuals (see e.g., Vul and Pashler 2008, Herzog and Hertwig 2009, Hourihan *et al.* 2010).

you've decided whether you want outright judgements or probabilities—or something else—we'll say that you've decided your *inputs*; you've decided what input the members of your crowd are going to have in your endeavour to achieve your desired output. Note that everyone needn't give the same kind of input. For example, you may want one half of your group to give product reviews, and the other half to rate the qualities of those reviews. Also note that the inputs needn't be of the same kind as your desired output—e.g., there are ways to turn probabilities (inputs) into an outright judgment (output); and there ways to turn outright judgements (inputs) into a final probability (output). We've mentioned a few kinds of inputs, but there are many others. To name just a few, inputs could be: votes, preferences, sentences, arguments, probability distributions, lines of computer code, quality ratings, translations, relevance rankings, or text transcriptions through services like reCAPTCHA (von Ahn *et al.* 2008).

One you've decided what kind of inputs you want to get out of your crowd, you have to work out how to *get them out* of your crowd. This is really important for all sorts of reasons. For example, some members of your crowd may have an incentive to lie to you (perhaps the person on trial seduced the wife of someone in your crowd). Or maybe their contributions are valuable to them, and so you need to pay for their contributions in some way. Maybe some of your members are shy while others are overbearing, and so you may need to make sure everyone has equal opportunity to make their contribution. We call this process of getting the inputs out of your crowd *elicitation*. Your method of elicitation can be crucial for getting the most out of your crowd. For example, psychologists have shown that how you ask for probability assignments from people can have a dramatic effect on how overconfident they they are (see e.g., Klayman *et al.* 1999).

Let's say you've decided what you want to do (output), got yourself a crowd (recruitment), worked out how they will contribute to your endeavour (inputs), and how you will get those contributions (elicitation). The next step is called *aggregation*: you need to convert the contributions of your crowd into your desired output. We've touched on an aggregation method already: judge the person on trial to be guilty if and only if everyone in your crowd judges them guilty. Another aggregation method is: judge them guilty if and only if a *majority* of your crowd judges them guilty. Yet another: judge the person guilty if and only if the average probability assigned to the guilty verdict by your crowd is above 90%. As you can probably tell, there are a lot of aggregation methods to choose from, and different aggregation methods will have different properties. Much of the rest of this paper is devoted to the topic of aggregation, so we'll leave further discussion of these matters to later sections.

There is one final aspect, and we call it *evaluation*, and it is how you evaluate the output of your endeavor. Sometimes evaluation will be straightforward. For example, if your crowd judged the person to be guilty, and they are in fact guilty, then your crowd got it *right*, and maybe that's all you care about. But you might also be concerned that your crowd will mistakenly judge an innocent person to be guilty, and that being wrong in this way (a false positive) is much worse than judging a guilty person to be innocent (a false negative). If so, you may have to decide how to balance these different kinds of error against each other. There are plenty of other standards of evaluation. If you're guessing the weight of an ox, you might want to *minimise the error* of your crowd's judgement. If you're forecasting the weather, you might want your announced "chances of rain" to be *well calibrated*.[3] If your crowd is writing encyclopedia articles, you might want the articles to have *few grammatical errors*, or to have few *factual inaccuracies*, or to have a *unified style*—or, probably, some combination of all of these virtues. How you choose to evaluate the output will have a big impact on your choices regarding the other five aspects we've identified. For example, some aggregation methods can be good at producing a collective judgement with the appropriate level of confidence (thus resulting in neither

[3] For example, it should rain on 90% of the days that your crowd says there is a 90% chance of rain.

over or underconfidence) but not very good at producing accurate judgements (*cf.*, Lyon *et al.* 2012).

To summarise, we've identified six core aspects to the Wisdom of Crowds:

1. The Output
2. The Recruitment
3. The Inputs
4. The Elicitation Method
5. The Aggregation Method
6. The Standard of Evaluation

There are two important qualifications that we now need to make. The first is that we presented these components as steps in a chronological process: decide what you want to, get your crowd, decide on your inputs, work out how to elicit them, work out how to aggregate them, and then work out how to evaluate them. However, it should be clear by now that there is no set chronological order to these aspects. For example, perhaps your most important criterion is that the output is a *fair* one. If so, this will put heavy constraints on the how you settle the other issues—e.g., your aggregation method may have to give equal weight to everyone's input, rather than unequal weight (*cf.* section 3). So instead of thinking of the aspects as steps in a chronological process, they should thought of as components of a reflective equilibrium.

The second important qualification is that these aspects can overlap with each other and that their borders are blurry. In fact, two of the main kinds of aggregation methods discussed in this paper—discussion groups and prediction markets—can also be thought of as elicitation methods. For example, a prediction market works by getting people to place bets with each other on whether some event will occur or not—e.g., whether Hilary Clinton will win the 2016 US Presidential Election. The "market price" of a prediction market is an aggregate of all of the individual bets, and, when interpreted as the probability of the event in question happening, can be highly effective in forecasting whether the event will happen. However, the market price is determined by the individual bets being made, and those bets can be used to infer the people's individual subjective probabilities of the event happening. So the prediction market both elicits and aggregates the human judgement inputs. Although the above six aspects overlap with each other, we believe they nevertheless provide a convenient conceptual framework for thinking about the Wisdom of Crowds.

All of the aspects are extremely important, but due to limitations on space, in this paper we will restrict our focus to the aggregation aspect of the Wisdom of Crowds. In fact, we will need to restrict our focus even further: we'll limit our discussion to aggregation methods that take only simple kinds of human judgements as input: votes, estimates, probabilities, etc., and these will always be *epistemic* judgements—that is, we won't the discuss the aggregation of inputs such as preferences, judgements of fairness, etc. And we won't discuss methods for aggregating more complex kinds of inputs, such as sentences to wikipedia articles, product reviews, text translations, contributions to legislation, etc.

The aggregation methods that we will focus on fall roughly into three broad categories: Mathematical Aggregation (section 3), Deliberation Methods (section 4), and Prediction Markets (section 5).

3 Mathematical Aggregation

Perhaps the most common aggregation method is averaging, specifically, *unweighted linear averaging*. Suppose there are N people in your crowd, and we number each individual, $i = 1, 2, 3, ..., N$. Let j_i be the elicited judgement of person i (e.g., the number of jelly beans in a jar). The unweighted linear average of your crowd's judgements is defined as:

$$\text{Unweighted Linear Average} = \frac{1}{N}\sum_{i=1}^{N} j_i$$

This simple method of averaging is considered by many as a standard benchmark, or gold standard of aggregation. For example, Armstrong 2001b recommends it as a good default option, especially if you don't know anything about the abilities of the individuals in the group. If you do have such information, you may want to use some kind of weighted average (see below).

Averaging has its drawbacks. It can make sense when the individual judgements are clustering around a central value, but it can have undesirable consequences when the distribution of judgements takes on another shape. For example, consider the following hypothetical estimates of the effect that Obama's economic policies will have on US GDP. Average growth in GDP for the next decade will be:

(i) $-0.1\%, 0.1\%, 0.2\%, -0.3\%, 0.1\%, 0.3\%, -0.3\%, 0.2\%, -0.1\%, -0.1\%$

(ii) $-19.1\%, 5.1\%, 5.2\%, 4.7\%, -20.5\%, 5.4\%, 4.7\%, 4.6\%, 4.8\%, 5.1\%$

In both cases, the average of the estimates is 0%, but the distributions of the guesses differ in an important way. The first set of estimates cluster around 0%, but the second tend to cluster more around 5% than they do around 0%. The only reason why the average of the second set of estimates is 0% is because of the two extremely negative estimates. In the first case, the individuals could agree to 0% as the collective judgement as a compromise—perhaps because 0% is so close to each individual estimate. However, in the second set, no one believes that the effect will be about 0%, so to take 0% as the collective judgement seems like a rather odd thing to do. For this sort of reason, a better strategy may be to take the *mode* of the estimates. In this way, the mode can be a more democratic aggregation method than the average. The mode is just one statistical property of the distribution of guesses we could use as an alternative to the mean. Other options include the median, the mean with outliers removed, the geometric mean, the maximum entropy expectation, and so on. In short, any of the tools of statistics can be used to construct a more sophisticated aggregation method.

Another way to move beyond simple averaging is to use a *weighted* average. A weighted average gives more weight to some of the estimates over than others. Using the same notation as before, but where w_i is the weight given to judgement j_i, the weighted linear average of the crowd's judgements is defined as:

$$\text{Weighted Linear Average} = \frac{1}{N}\sum_{i=1}^{N} w_i j_i$$

Using a weighted average can make sense when, say, you know some members of your crowd are more reliable than others. For example, if you know from past experience that Ann is twice as good at guessing the number of jelly beans in a jar than Bob is, then you might want to take the average of their guesses, but give twice as much weight to Ann's guess than to Bob's. This is a variant of a method known as *Cooke's method* (*cf.* Cooke 1991). The core idea is that you should use the past performance of the members of your crowd to determine how much weight you should give to their current judgements (see Clemen 2008 for a study of the method's performance). This is not the only way to use a weighted average. It may make sense to weight the judgements by how confident the individuals are in their judgements. If someone is not very confident in their judgement, then perhaps their judgement shouldn't contribute much to the collective judgement. Various results in pyschology suggest that, at least in some cases, confidence in a judgement correlates with the accuracy of that judgement (e.g., Koriat 2012).

A more complicated way to take a weighted average is to elicit degrees of *peer respect* along with the judgements (thus making the inputs slightly more complex). Suppose you find yourself in a group of people who all give judgements about some issue, but you think some members of the group are experts on the issue at hand and others are not. You would probably be unhappy with any collective judgement that gave equal weight to everyone—you'd prefer a collective judgement that gave more weight to the experts than to the fools. Similarly, everyone else will feel the same way—although they may have different opinions as to who are the experts. For any individual k, if they respect each person i to degree w_{ki}, it looks like they should average as follows:

$$\text{Respect Weighted Average} = \frac{1}{N}\sum_{i=1}^{N} w_{ki} j_i$$

(where the w_{ki} are all between 0 and 1, and for each fixed i, the w_{ki} sum to 1; so there is no need for a normalisation term). This aggregation method will produce a new judgement j_i' for each person i. Lehrer and Wagner 1981 argue that there is nothing special about these new judgements, and so if they vary, then everyone should now average again, using the new judgements and original weights of respect. Lehrer and Wagner prove that if everyone continues to average in this way, they will reach a group *consensus*: all of the averaged judgements will approach a unique consensus judgement j_c. Lehrer and Wagner argue that this consensus judgement has a number of virtues—both pragmatic and epistemic. One potential drawback to this method of aggregation, however, is that people's judgements of each other's level of expertise do not track the accuracies of their judgements. Burgman *et al.* 2011 found that such ratings of expertise were poor guides to judgement accuracy. There can also be practical difficulties in getting people to rate each other's expertise—especially if those ratings are to be made public (*cf.* Regan *et al.* 2006). For an extensive discussion of the Lehrer–Wagner consensus model, see Loewer and Laddaga 1985 and the other papers in the same special issue of *Synthese*.

So far, we have only discussed examples where the inputs and outputs are quantity estimates and so can be represented with real numbers. If the inputs are are not like this, we have to choose a different kind of aggregation method. Another common sort of judgement are outright judgements of the form "guilty"/"innocent", "yea"/"nay", "black"/"white", and so on. Such judgements cannot be averaged, but there are, nonetheless, ways to aggregate them. Perhaps the most natural and common is what is known as the *majority rule*: the collective judgement is "guilty" ("innocent") if and only if more than 50% of the individual judgements are "guilty" ("innocent"). A famous theorem, known as the Condorcet 1785 jury theorem (rediscovered by Black 1963), shows that as you add more and more people to the crowd and aggregate their judgements using the marjority rule, then if each person has a greater than 50% chance of being right, and if they make their judgements independently of one another, then the probability that the collective judgement is correct will approach certainty. The theorem requires that the people in the crowd make their judgements independently of each other, which is a somewhat implausible of real life situations. However, Ladha 1992 generalised the theorem to allow for there to be some dependencies between the crowd's judgements. And there have now been a number of other generalisations of the theorem to make its application to real life situations more plausible. List and Goodin 2002 generalised the theorem to cover other aggregation methods and Grofman *et al.* 1983 generalised the theorem to allow for people who don't have a greater than 50% chance of judging correctly.

Things get tricky if the inputs and outputs are more complex than single all-or-nothing judgements. Suppose that instead of simply judging whether someone is guilty, we want our crowd to provide some reasoning for this judgement. For example, suppose that G means the person is *guilty*, N means they were *nearby* when the crime was committed, and $N \to G$ means that *if they were nearby, then they are guilty*. Now suppose we have 30 people in our crowd, and they make the following judgements on the three propositions, N, $N \to G$, and G:

	N	$N \rightarrow G$	G
10 people say	True	True	True
Another 10 people say	True	False	False
The remaining 10 people say	False	True	False
So, the greater-than-50% majority rule says	True	True	False

If the collective judgement is defined using the greater-than-50% majority rule, then the collective judgement on the three propositions will be *logically inconsistent*,[4] even though every individual in the group is perfectly consistent. This paradox has come to be known as the *doctrinal paradox*, and it has generated a large literature (see e.g., List 2012, Dietrich 2012, and Cariani 2011). This sort of inconsistency result shows that your choice of inputs and outputs can be incredibly important. Keep them simple, and you can get a result like the Condorcet Jury Theorem which says your crowd will probably do good things. But make the inputs and outputs a little more complex, and all of a sudden your crowd can be logically inconsistent. (Note that if the inputs are only judgements on N and $N \rightarrow G$, and the output is simply a judgement on G, then there is no inconsistency.)

In the above discussion, the collective output is evaluated in terms of its accuracy. However, as we explained in section 2, there are other standards of evaluation. May 1952, for example, identified four *procedural* constraints, which he called *Universal Domain*, *Anonymity*, *Neutrality*, and *Positive Responsiveness*. These are (arguably) plausible procedural constraints that an aggregation method should satisfy (with outright judgements as inputs and outputs). For example, Neutrality requires that if everyone changes their judgement, then the collective judgement should change accordingly. May proved that the majority rule is the only aggregation method that satisfies all four constraints. For further discussion of these issues see e.g., Maskin 1995, Woeginger 2003, and Asan and Sanver 2002.

Much more could be said on the topic of mathematical aggregation, and we have only discussed simple kinds of aggregation methods on fairly simple kinds of inputs and outputs. For further discussion see Armstrong 2001a, List and Pettit 2002, Grofman *et al.* 1983, and Pacuit 2012. We now turn to another way in which judgements can be aggregated: through group deliberation.

4 Deliberation Groups

The aggregation method that most readers will have had direct experience with is a deliberation group: the "crowd" meets to discuss the problem at hand, and after a period of discussion, they arrive at a collective judgement.[5] The group discussion can be structured or unstructured. In an ideal situation, the discussion will elicit from each member of the group not only their judgements, but also their reasons, arguments and evidence that back up these judgements. Through discussion and debate, the group can sort through all of the evidence and arguments leading to a more informed solution.

A common criticism of unstructured group discussion is that it *enhances* cognitive errors rather than mitigates them. In addition, there are many social phenomena that hinder a group's ability to reach a correct judgement, even if, in principle, the group has all the pieces needed to solve the problem. We note the following three issues. *Bias against the minority*: There is a tendency for groups to ignore isolated, minority or lower-status members. *Anchoring effect*: There is a tendency to rely too heavily, or "anchor", a judgement on one piece of information (for example, the first announced judgement,

[4] This is because N and $N \rightarrow G$ entail G, so if the former two propositions are true, the latter has to be true.

[5] As we noted in section 2, not all deliberation groups are instances of judgement aggregation. For example, the crowd could simply meet to share information and then still give different individual judgements, which could then be aggregated using one of the methods described in sections 2 or 5.

the judgement of the most senior person in the group, or the judgement of the loudest person in the group). *Common knowledge effect*: Information held by all members of the group has more influence on the final decision than information held by only a few members of the group (see Gigone and Hastie 1993). See Sunstein 2011 for a discussion of other problems with group deliberation.

Despite its many flaws, unstructured deliberation can be fruitful in certain circumstances. For instance, the unstructured discussion in the comments section of the polymath blog led to a new proof of the Hales-Jewett Theorem (see Polymath 2012). Other examples that may benefit from unstructured debate and discussion include writing a novel or finding the correct wording of a piece of legislation. Indeed, group brainstorming sessions are often used to generate new ideas and creative solutions to a variety of problems. However, some research shows that interacting brainstorming groups come up with fewer new ideas than does aggregating the collective ideas from a group of non-interacting individuals (Diehl *et al.* 1987). The social dynamics of the group also often interferes with the group's ability to achieve its intended goal. Therefore, it is important to develop methods to keep the group focused on the task at hand (e.g., see Gerber 2009 and Bao *et al.* 2010 for methods aimed at improving brainstorming sessions).

One way to diminish the effect of the psychological phenomena mentioned above is to *structure* the deliberation. A method that has been widely used is the *Delphi method* (Linstone and Turoff 1975). This actually refers to a whole range of methods. What is common among the different implementations is that the members of the group provide their initial judgement *before any discussion takes place*, then there are a number of rounds in which the group members can discuss and revise their judgements.

After the group members give their initial judgments to the moderator, there are a number of ways to proceed. A sample session may run as follows: The moderator shows everyone in the group the initial judgements (making public the judgement of each member of the group). Members of the group are encouraged to discuss their reasons for their initial judgements. After a round of discussion, each person in the group is asked (either privately or publicly) if they want to revise their initial judgement. The second round judgements are then given to the moderator who aggregates them using one of the methods from Section 3. There are many ways to vary the group interactions: (i) The initial judgements are kept anonymous. (ii) Members of the group are asked to judge how confident they are in their judgements. (iii) Rather than taking part in an unstructured discussion, the members of the group are given time to do their own research in light of receiving each other's judgements. (iv) Members of the group are asked to judge how confident they are in another (randomly selected) person in the group's judgements. (v) The process continues until consensus is achieved (or some large subgroup achieves consensus). There is mounting evidence that structuring group deliberation in this way leads to more accurate predictions (Armstrong 2006, 2011).

Sometimes no amount of discussion will lead the group to a consensus opinion. This means that group deliberation may only be a partial solution to an aggregation problem, and consequently, the moderator may have to use an additional aggregation method to form the final group judgement (e.g., the moderator might average the final estimates). However, one must be careful with how these aggregation methods are combined, for it is possible for group deliberation to improve the individual judgements in a group, while making the collective judgement worse. For example, suppose that there are 10 people estimating a parameter whose true value is 40 with the following initial estimates: 15, 18, 20, 22, 30, 45, 50, 55, 60, and 61. Using an unweighted average, the group estimate is 37.6. If the new estimates after the discussion period are: 16, 25, 21, 23, 31, 41, 41, 40, 41, and 45, then each individual improved their estimate. However, the average of these estimates is 32.4, and so the collective judgement (understood as the average) is worse after discussion. Nevertheless, there is data to show that discussion, in an appropriately structured deliberation group, can improve the group estimate—see e.g., Burgman *et al.* 2011.

For a much more detailed overview of deliberative groups and collective group judgements in general, see Fidler *et al.* MS.

5 Prediction Markets

Recently, there has been quite a lot of interest in the use of *prediction markets* as a method for aggregating individuals opinions about future events. Suppose that we are interested in whether an event will take place at some time in the future (for example, will Hillary Clinton run for president in 2016?). Rather than gathering experts to discuss their opinions about this event, the approach we highlight in this section is to create a market in which individuals trade contracts whose payoffs are tied to the future event. For instance, suppose that there is an option that pays $10 if Hillary runs for president in 2016 and $0 otherwise. Ignoring any transaction costs, if an investor pays $7 for the event "Hillary Clinton will run for president in 2016", then she earns $3 dollars if Clinton runs and loses $7 if she does not run. Under standard decision-theoretic assumptions (such as that investors are *risk-neutral*), investors should be willing to pay up to a price that equals their estimated probability that an event will happen. The market price, or equilibrium price, is the value such that if an investor were willing to sell below the price, the other investors would buy the stock driving the price back up (similarly for anyone willing to buy above the market price). The market price has been interpreted as the aggregate probability of the investors (Manski 2006, Wolfers and Zitzewitz 2006a) and has been shown to be remarkably accurate in predicting events (Arrow *et al.* 2008 and Rothschild 2009).

Prediction markets have many advantages over deliberation as a method for aggregating individual judgements (see also Sunstein 2011 for a discussion of this). The primary advantage is that prediction markets provide the right incentives for a diverse population to disclose the information that they privately hold. Furthermore, the economic incentive in a market encourages traders to search for the best available information. Moreover, even if the investors are unsophisticated or not well-informed, the *efficient market hypothesis* states that markets are good aggregators of information (see Lo 2007 for an overview). See Wolfers and Zitzewitz 2006b and Arrow *et al.* 2008 for an an extensive discussion of predication markets, including an overview of the experimental evidence and case studies that demonstrate the benefits of using markets to predict future events.

Markets work well when there is a large and diverse group and each person is likely to get different types of information. This suggests that implementing a prediction market may not always be feasible. There are two central problems that can make a prediction market infeasible. The first is that there must be a large enough group of people that are interested and engaged with the market. The second is that in order to use a prediction market, you must be interested in predicting whether or not some event will happen at some specific moment in the future. This is important since It must be perfectly clear which bets to payoff. In addition to problems of feasibility, prediction markets face a number of other challenges.

The economic incentives provided by a prediction market do a good job mitigating many of the biases that infect group deliberation discussed in the previous section. Still, prediction markets are influenced by investor biases. The most well-known is the *favorite long-shot bias*. This bias causes investors to undervalue events with probabilities close to 1. Similarly, investors tend to over-value events that have probabilities close to 0. This type of bias is well-documented and can have an effect on the market price (Thaler and Ziemba 1988).

Recent work has questioned the relationship between the market price and the distribution of beliefs of the investors in a prediction market. Othman and Sandholm 2010 study the behavior of simple agents that sequentially interact with the market. They show that by varying the order of participation in a market, the market price can converge to an arbitrary value (see Frongillo *et al.* 2012 for a generalization of this result).

A final challenge for the use of prediction markets as an aggregation method is the possibility of manipulation. Since most prediction markets have a relatively low volume, it would be relatively inexpensive for an investor to take losses in order to affect the market price. An example of this type of manipulation was recently observed in the Intrade market to predict the outcome of the 2012 election. According to a Washington Post article (Plumer 2012), a few months before election day, there was a huge swing towards Romney in the market which appears to have been driven by someone spending about $17,800 to push up Romneys chances of winning. The surge only lasted about six minutes before other traders brought the price back down. It is still unclear whether this was a manipulation by an investor attempting to sway perceptions of the race or simply an example of a trader who made an expensive trade. There is evidence that attempts to manipulate prediction markets tend to fail (Hanson *et al.* 2006). In fact, Hanson and Oprea 2007 offer a model in which attempts to manipulate the market actually *increases* the accuracy of the market price (see also Chen *et al.* 2010 for a general study of manipulation in prediction markets).

We conclude this short section with a few brief comments about some computational aspects of prediction markets. Virtually all the prediction markets currently in use restrict trade to "simple" events that can all be explicitly listed and monitored. So, for example, bets are made on events of the form "horse A will win" rather than more complex events such as "horse A will beat horse B which will beat horse C", "horse A will win and horse B will come in third" or "horse A will win if horse B comes in second". Initial research shows that allowing individuals to trade on more complex and/or conditional events has significant advantages (see BEREA CHAPTER), but it raises many difficult computational challenges. For example, it is no longer feasible to explicitly list all the possible bets—e.g., in a horse race with 10 horses, there are $10! = 3,628,800$ many different possible permutations that would need to be listed. Therefore, it is important to develop *combinatorial betting mechanisms* that allow investors to succinctly express their bets. Another computational challenge is that allowing bets on more complex events makes it much more difficult to match buyers with sellers. In general, it may be necessary to look beyond bilateral trades and consider complex multilateral trades. There is much more to say about the computational aspects of prediction markets. See Pennock and Sami 2007 and Chen and Pennock 2010 for a discussion of these issues and further references to the relevant literature.

6 Conclusion

There is a growing literature focused on the Wisdom of Crowds spanning many disciplines such as philosophy, computer science, management science, social psychology and social choice theory. And it can be difficult to pin down exactly what the class of phenomena is that is loosely called *the* "Wisdom of Crowds" by these diverse research communities.

In section 2, we outlined out a simple conceptual framework for thinking systematically about the Wisdom of Crowds. We did this by taking the perspective of someone interested in using the Wisdom of Crowds to solve a problem. For example, suppose that you are interested in finding the answer to some question (e.g., Is the defendant guilty?), a prediction about a future event (e.g., Will Hillary Clinton run for president in 2016?), or an estimation of some parameter (e.g., How many jelly beans are in the jar?).

Once you identify the group of people that will make up your crowd, you must decide how to leverage the "wisdom" of the crowd to solve your problem. This involves eliciting useful information from each member of the group and deciding how to aggregate this information. There are many different methods that can be used to aggregate the judgements of a group of people. We discussed three broad categories of aggregation: Mathematical Aggregation, Group Deliberation and Prediction Markets.

All of the aggregation methods we discussed in this paper accept *human* judgments as inputs, and we primarily focused on *epistemic* judgements that were simple in form—e.g., they are about the true value of a parameter, or the answer to a yes/no question. Moving beyond this limited focus would allow us to examine a wider variety of examples of the Wisdom of Crowds. (See, for example, Nielsen 2011 for a discussion of collective judgements in situations where there are no objective facts against which the judgements can be evaluated.)

Clearly, there is much more to say about the Wisdom of Crowds. It is certainly going to take a diverse group of researchers to fully understand this phenomena.

References

Amrstrong, J. S. (2006). Should the forecasting process eliminate face-to-face meetings? *The International Journal of Applied Forecasting 5*, 3–8.

Andler, D. (2012). What has Collective Wisdom to do with Wisdom? In C. U. Press (Ed.), *Collective Wisdom*, pp. 72–84.

Armstrong, J. (2001a). *Principles of Forecasting*. Kluwer Academic Publishers.

Armstrong, S. (2001b). Combining Forecasts. In S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Arrow, K. J., R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz (2008). The promise of prediciton markets. *Science 320*, 877 – 878.

Asan, G. and R. Sanver (2002). Another characterization of majority rule. *Economic Letters 75*(3), 409 – 413.

Bao, P., E. Gerber, D. Gergle, and D. Hoffman (2010). Momentum: getting and staying on topic during a brainstorm. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1233 – 1236.

Black, D. (1963). *The theory of committees and elections*. Springer.

Brownstein, J. S., C. C. Freifeld, and L. C. Madoff (2009). Digital disease detection —harnessing the web for public health surveillance. *New England Journal of Medicine 360*(21), 2153–2157.

Burgman, M., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy (2011). Expert Status and Performance. *PLoS One 6*(7), e22998.

Cariani, F. (2011). Judgment aggregation. *Philosophy Compass 6*(1), 22 – 32.

Chen, Y., S. Dimitrov, R. Sami, D. Reeves, D. Pennock, R. Hanson, L. Fortnow, and R. Gonen (2010). Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica 58*(4), 930–969.

Chen, Y. and D. Pennock (2010). Designing markets for prediction. *AI Magazine 31*(4), 42 – 52.

Clemen, R. T. (2008). Comment on cooke's classical method. *Reliability Engineering & System Safety 93*(5), 760–765.

Collier, N., A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul (2006). A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges. *Language resources and evaluation 40*(3), 405–413.

Condorcet, M. (1785). *Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix.* Paris: l'Imprimerie Royale. Reprint. New York: Chelsea, 1972.

Cooke, R. M. (1991). Experts in uncertainty: opinion and subjective probability in science.

Diehl, M. and W. Stroebe (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology 53*(3), 497–509.

Dietrich, F. (2012). Judgment aggregation and the discursive dilemma. In *Encyclopedia of Philosophy and the Social Sciences*. Sage Publishers.

Fidler, F., B. Wintle, and N. Thomason (MS). Groups Making Wise Judgements. (Unpublished manuscript).

Frongillo, R., N. Della Penna, and M. Reid (2012). Interpreting prediction markets: a stochastic approach. Manuscript.

Galton, F. (1907a). Letters to the Editor: The Ballot-Box. *Nature 75*, 900–1.

Galton, F. (1907b). Vox Populi. *Nature 75*, 450–1.

Gerber, E. (2009). Using improvisation to enhance the effectiveness of brainstorming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 97 – 104.

Gigone, D. and R. Hastie (1993). The common knowledge effect: Informaiton sharing and group judgments. *Journal of Personality and Soical Psychology 65*(5), 959–974.

Graefe, A. and J. S. Armstrong (2011). Comparing face-to-face meetings, nominal groups, delphi and prediction markets on an estimation task. *International Journal of Forecasting 27*(1), 183 – 195.

Grofman, B., G. Owen, and S. L. Feld (1983). Thirteen Theorems in Search of the Truth. *Theory and Decision 15*(3), 261–278.

Hanson, R. and R. Oprea (2007). A manipulator can aid prediction market accuracy.

Hanson, R., R. Oprea, and D. Porter (2006). Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior and Organization*.

Herzog, S. M. and R. Hertwig (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychological Science 20*(2), 231–237.

Hourihan, K. L. and A. S. Benjamin (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition 36*(4), 1068.

Keller, M., M. Blench, H. Tolentino, C. Freifeld, K. Mandl, A. Mawudeku, G. Eysenbach, and J. Brownstein (2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. *Emerging Infectious Diseases 15*(5), 689.

Klayman, J., J. Soll, C. González-Vallejo, and S. Barlas (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes 79*(3), 216–247.

Koriat, A. (2012). When are two heads better than one and why? *Science 336*, 360–2.

Ladha, K. K. (1992). The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, 617–634.

Landemore, H. and J. Elster (2012). Collective wisdom: Principes and mechanisms.

Lehrer, K. and C. Wagner (1981). *Rational consensus in science and society: A philosophical and mathematical study*, Volume 24. D. Reidel.

Linstone, H. A. and M. Turoff (1975). *The Delphi Method: Techniques and Applications*. Addison-Wesley.

List, C. (2012). The theory of judgment aggregation: An introductory review. *Synthese 187*(1), 179 – 207.

List, C. and R. E. Goodin (2002). Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy 9*(3), 277–306.

List, C. and P. Pettit (2002). Aggregating sets of judgments: An impossibility result. *Economics and Philosophy 18*(01), 89–110.

Lo, A. (2007). Efficient market hypothesis. In *The New Palgrave: A Dictionary of Economics, Second Edition*.

Loewer, B. and R. Laddaga (1985). Destroying the consensus. *Synthese 62*(1), 79–95.

Lyon, A., F. Fidler, and M. Burgman (2012). Judgement swapping and aggregation. In *2012 AAAI Fall Symposium Series*.

Lyon, A., G. Grossel, M. Nunn, and M. Burgman (2013). Using Internet Intelligence to Manage Biosecurity Risks: A Case Study for Aquatic Animal Health. *Diversity and Distributions*.

Lyon, A., M. Nunn, G. Grossel, and M. Burgman (2012). Comparison of web-based biosecurity intelligence systems: Biocaster, epispider and healthmap. *Transboundary and Emerging Diseases 59*(3), 223–232.

Manski, C. (2006). Interpreting the predictions of prediction markets. *Economic Letters 91*(3), 425 – 429.

Maskin, E. (1995). Majority rule, social welfare functions and game forms. In *Choice, Welfare and Development: A Festschrift in Honour of Amartya K. Sen*, pp. 100 – 109. Oxford University Press.

May, K. (1952). A set of independent necessary and sufficient conditions for simply majority decision. *Econometrica 20*(4), 680 – 684.

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press.

Nuwer, R. (2013). Software could make rare diseases easier to spot. *New Scientist 218*(2913), 21.

Othman, A. and T. Sandholm (2010). When do markets with simple agents fail? In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1 - Volume 1*, AAMAS '10, pp. 865–872.

Pacuit, E. (2012). Voting methods. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012 ed.).

Page, S. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)*. Princeton University Press.

Pennock, D. M. and R. Sami (2007). Computational aspects of prediction markets. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani (Eds.), *Algorithmic Game Theory*. Cambridge University Press.

Plumer, B. (2012). How to swing the prediction markets and boost Mitt Romney's fortunes.

Polymath, D. H. J. (2012). A new proof of the density Hales-Jewett theorem. *Annals of Mathematics 175*(3), 1283 – 1327.

Regan, H. M., M. Colyvan, and L. Markovchick-Nicholls (2006). A formal model for consensus and negotiation in environmental management. *Journal of Environmental Management 80*(2), 167–176.

Rothschild, D. (2009). Forecasting elections: Comparing prediction markets, polls and their biases. *Public Opinion Quarterly 73*(5), 895 – 916.

Sunstein, C. (2011). Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Habermas. In *Social Epistemology: Essential Readings*, pp. 314 – 337.

Surowiecki, J. (2005). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday.

Thaler, R. and W. Ziemba. (1988). Anomalies: Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives 2*(2), 161 – 174.

Von Ahn, L., B. Maurer, C. McMillen, D. Abraham, and M. Blum (2008). recaptcha: Human-based character recognition via web security measures. *Science 321*(5895), 1465–1468.

Vul, E. and H. Pashler (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science 19*(7), 645–647.

Woeginger, G. (2003). A new characterizaiton of the majority rule. *Economic Letters 81*(1), 89 – 94.

Wolfers, J. and E. Zitzewitz (2006a). Interpreting prediciton market prices as probabilities. Technical report, NBER Working Paper 12200.

Wolfers, J. and E. Zitzewitz (2006b). Prediciton markets. *Journal of Economic Perspectives*.