Quantifying the value of another voice: a proposed simulation

Ben Klemens

November 17, 2002

Abstract

How many information providers do we need to get full information? In a perfectly rational world, where everyone is perfectly strategic, full information can be revealed by one—or at most two—information providers. Unfortunately, these results are technically fragile, and are contrary to the evidence we see in the real world. I instead answer the question with a simulation, describing a reporter who conveys facts via a story and a listener who is trying to determine the true facts based on that story. As one would expect, each reporter adds more information, but at a diminishing rate. One or two reporters is sufficient to guarantee full information revelation only in the case of simple situations. As the complexity of the question grows, the number of information providers needed expands.

1 Introduction

This paper presents a simulation of a listener who gathers information from a group of reporters. The primary assumptions of the model are that the reporters and the listener do not act strategically, that facts can be broken down into verifiable yes/no questions, that reporters never lie but may fail to reveal facts which disagree with their own biases, and that reporters gather and provide information in a similar manner. The output of the model is a measure of the vaule of an additional reporter, and a measure of the number of reporters one would need to ensure full information revelation with certainty.

1.1 Theory

I will first discuss the theoretical approaches to the model, notably those proposed by Milgrom & Roberts [5]. They depend on the standard assumptions of game theory: that reporters predict how listeners will respond and act accordingly, and listeners are adept at filtering the information they receive from reporters.

The listener in all of these models processes information in a very direct way: her sole goal is to gather all the facts, and such issues as context or source are not relevant. Although mathematically simple, this is a fatally inaccurate characterization of how people actually think. If the methods used by listeners are not simply a direct catalogging of facts, then information providers will be aware of this and will change their methods of presenting information accordingly.

As evidence of this, I cite some textbooks for law students on trial strategy. Lawyers and reporters are in a simliar position: they have the job of presenting facts, but have strategies at their disposal that may allow them to slant the information toward their own point of view. The legal setting is clearer and better studied, and is therefore better suited for the discussion here, but the conclusions carry over directly to all information providers.

The law textbooks make the point that a lawyer should not list facts, but should tell a story. Facts are merely support for a plausible description of what happened. The same is clearly applicable to what it is that reporters do: they gather facts and bundle them into a story.

1.2 Simulation

This insight from the legal textbooks provides motivation for the simulation which comprises the remainder of the paper. The simulation captures the idea the facts can be organized into stories, and that reporters have certain stories they prefer to tell. After reporters filter information, we can ask: will a listener who is interested only in gathering all of the facts be able to do so?

The simulation sidesteps a number of major problems that writing down a full theoretical model would encounter—notably, the fact that the utility function of true listeners is almost impossible to clearly describe, and is not monotonically increasing in the quantity of information. This will be discussed further below. It also allows calculation of how many reporters would actually be necessary to guarantee that a listener who wishes to gather full information can do so. I conclude that one or two reporters would be sufficient only for the case of relatively simple situations. For complex questions such as the evaluation of a political candidate, I find that many more reporters are necessary.

2 Why game theory fails us

The game-theoretic approach to this question would be to assume that reporters and readers are simultaneously maximizing their utility subject to certain constraints, given information about the other actors. As will be shown below, the results are that it takes only one biased information provider to arrive at full information, but these results are fragile. Notably, they do not provide a sequential or a trembling-hand equilibrium. So if actors hope implement an equilibrium that is sufficiently robust that it is sequentially rational and immune to trembles—both reasonable assumptions—then the game theoric literature offers no predictions.

2.1 Crawford & Sobel's model

The most popular model of information provision by one advocate, upon which a number of other models have been built, is that of Crawford & Sobel[2]. Their model differs from those here, including the simulation below, because the reporters of their model do not reveal their bias. They provide true-but-vague information, with the intent of obscuring both their bias and the true value of the information the decisionmaker hopes to learn. Those who argue that the bias of a media source is impossible to determine will find that this model is applicable, and that reporters will provide partial information, as in the simulation I present below. Those who feel that the bias of a media source is detectable will find that Crawford & Sobel's model is not applicable to the question here; the models which follow are all models with advocates of known bias.

2.2 Milgrom & Roberts's one-advocate game

Milgrom & Roberts [5] have a simple model of information revelation by an interested party. They arrive at full information revelation with only one reporter, when both reporter and decisionmaker are strategic. But on a technical level, their equilibrium fails to meet a few basic refinements; or on an intuitive level, a lot could go wrong if someone were to attempt to implement this in a real-world situation. The basic problem is that the reporter may insist on providing no information at all, and it is unclear what the decisionmaker should do in such a case.

To paraphrase the model here, the listener must decide on the value of some unobservable variable x, a nonnegative number. If she decides d, then her utility will be $-|d-x|$. The reporter knows the true value of x, but is biased and wants the listener to choose as large a value of d as possible. The reporter's report will consist of an interval $[a, b]$ of the nonnegative numbers, in which x lies. He must be honest, meaning that x is indeed within the reported interval, but the interval may be as large (and therefore uninformative) as the reporter chooses.

Milgrom & Roberts show that the decisionmaker can extract full information from the reporter with the following threat: 'if you report $[a, b]$, then I will assume that $x = a$.' Then the reporter will not want to report $a < x$, since the decisionmaker will choose a $d < x$; while presenting $a > x$ would mean that x is no longer in the reported interval, which is against the honesty rule. Therefore the decisionmaker will be able to extract full information from one information provider.

This one-dimensional model can easily be generalized to (non-negatively correlated) multiple dimensions, or even to probability densities, implying that it takes only one information provider to reach a full-information equilibrium in many cases.

However, this equilibrium is fragile for a number of reasons. It depends vitally on x being a nonnegative number. If x is a real number, then the reporter may give a maximally vague signal $(-\infty, \infty)$.[1] The assume-the-worst rule above forces the decisionmaker to choose $d = -\infty$, which gives her an impossibly bad payoff. Any real number would give her a better one. In technical terms, the game has no sequential equilibrium, since it requires the decisionmaker to shoot herself in the foot if the reporter doesn't do as expected.

The intuition carries over to an interval bounded below as well, but in this case it fails to be a trembling-hand equilibrium. The reporter can not report $a > x$ because of a hard constraint set in the rules, but does not report $a < x$ due to strategic considerations. Therefore, if the reporter were to commit an error, he would always commit an error towards a lower value of a than he should have reported. Bearing this in mind, the decisonmaker should decide $d = a + \epsilon$, where we can not specify the value of ϵ unless we go into great detail about the beliefs of both players. At the extreme of trembles, the reporter could forget to report entirely, and then the decisonmaker again has to decide whether to assume the worst and choose $d = 0$, or to choose some small but positive value.

These refinements are not technical details: they describe problems we would expect to have if we implemented this rule in reality. Since the assume-the-worst equilibrium fails to meet the criteria of these refinements, it is hard to believe that real people will implement it in a complex decisionmaking situation. The key problem is that it is hard to guess what the decisionmaker will do if the reporter offers no information, and the reporter can use that ambiguity to his advantage.

Another problem with this and any other model of information aggregation by listing facts is best discussed in the context of two advocates.

[1] Sticklers should note that this interval is closed with respect to the real numbers, and it is only notational custom that precludes writing it as $[-\infty, \infty]$.

2.3 Two opposing advocates

The argument about two advocates is usuall given in a legal setting: every piece of information benefits either the defense or the prosecution, so one or the other will always reveal any given fact. This idea is easy to formalize into a robust game theoretic model. However, it fails to match reality in any but the most rudimentary of situations.[2]

The purpose of this digression is to show that lawyers do not list facts, but tell stories, going against the prediction of any game theoretic model based on listeners are absorbing facts without consideration of context or source. The fact that a model of information aggregation using facts and not stories fails to describe a situation as orderly and methodical as a trial indicates that such a model will have no descriptive power at all for less formal situations like information dissemination via the media.

For a look at how lawyers argue, I found the two books on trials from USC Law's Fall 2002 reading list, by Mauet [4] and Murray [6]. Mauet also appears on what other reading lists I could find (those of law schools at NYU and the University of Chicago), and is in its sixth edition, implying that it has come recommended by many law professors in the past.

Mauet is clear on advising the law student to become a storyteller: "Trials involve much more than merely introducing a set of facts; those facts must be organized and presented as part of a memorable story." [p 27] Nor must this story be burdened by too much detail: "A theory of the case is a clear, simple story of 'what really happened' from your point of view [...] If you can not state your theory of the case in a minute or two, it needs more work." [p 24]

Murray [p 9] corroborates this testimony:

> The trial lawyer's task can be analogized to the role of raconteur at a social party or similar informal gathering. The storyteller entertains his audience by recreating in the consciousness of the listener an image of an event or episode from another time and place.

A key aspect of storytelling is deciding which facts will be used to tell the story. Both authors cited above advocate against throwing everything possible at the judge or jury: "Do not overprove your case", Mauet advises [p 516].

All of this applies to the media world, with more force. After all, a trial is a slow, deliberative process, where briefs may run into hundreds of pages. A newspaper article, colloquially referred to as a "story", has few column-inches to convey the state of the world the reporter perceives, and other media such as television have still tighter restrictions.

[2] The technical details of which of the model's assumptions need modification will not be discussed here. The reader is referred to my dissertation [3].

In advertising, we see pairs of opposing information providers failing to provide full information in any of a number of situations. The key problem is when opposing sides do not tell opposing stories. For example, Apple advertises the story that its products are stylish, while its competitors tell the story that their products are technically superior. What would happen to test results that show that an Apple computer is faster for certain tasks? Apple would not put the statistics in their neat, number-free advertisements because it is not a part of the story they wish to tell about style, while the competitors will certainly fail to mention these tests.

Without going into technical detail, this is the failure of the claim that two opposing advocates always reveal full information: since listeners process information in the framework of stories, and therefore information providers offer information in the framework of stories, facts which do not fit into either side's best story will not be revealed.

In summary, the theoretical literature predicts that if the bias of the information providers is known, then we will only need one or two information providers to extract full information. But these models are not descriptive of how information is transmitted in real-world situations. The one-advocate model is fragile to the assumptions of perfect individuals who precommit to their strategies; the two-advocate model depends on the assumption that the recipients of information simply gather facts without regard to their context, whereas information providers, like the experienced trial lawyers cited above, know that the story in which the facts are placed is essential to the processing of the facts.

3 The simulation

The motivation for the simulation's structure should be clear now: reporters have access to a great deal of information, but filter out that information that does not fit into the story they wish to tell. Listeners must then divine all the facts based on that subset which the reporter was willing to present. We can measure what percentage of facts they correctly divine given one reporter, two reporters, et cetera.

I will first present the model, then discuss the implicit assumptions made therein, and finally show the quantitative results.

3.1 Description

The space of events and filters The first component of this model, provided by Nature, is a set of events. An event is an ordered list of yes/no facts, constructed as follows: first, a center is randomly drawn. Then, a random draw determines which facts are true, with high probability the central fact is true, and decreasing likelihood that facts further out from the center are true. The procedure which does this is generate_biasedrow, listed in the appendix, Section 4.

Nature provides us with an extensive string of these events. I chose 200 events consisting of 20 yes/no facts each. From these events, today's state is randomly chosen.

Reporters have a filter of facts that they are willing to report. Filters look like events: a string of ones and zeros.

Figure 1 shows a set of filters which were drawn for a test run. Each filter is a row of twenty facts, each having exactly ten ones, and a center at the leftmost fact. Combining all filters would give full information about the lefttmost thirteen facts, and one fact near the right end of the spectrum.

1 1 1 1 1 1 1 1 1 0 0 0 1 0 0 0 0 0 0 0
1 1 1 1 1 0 1 0 1 1 1 1 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 0 0

Figure 1: A selection of filters. A selection of events would look the same, but be interpreted differently.

Define the *quality of a filter* to be the percentage of information that it reports. For example, all of the filters in Figure 1 reveal ten out of twenty facts, and therefore all have a quality of 0.5.

The only feature that the model as written imposes on the facts is that there is correlation between facts: if facts one and three are true, then fact number two is also likely to be true; if facts one and three are in a reporter's filter, then fact number two is also likely to be in his filter.

There are a number of interpretations that may be given to such ordered strings. One is that the space is the familiar left-right ideology space. An event with many ones on the left side of the range represents a state of the world that is amenable to a liberal ideology. As for the filter, the center represents the reporter's ideology, and the ones in the filter are the facts that he is willing to report—those facts that are consistent with his outlook. Notice that by randomly drawing the center of the reporter's belifs from anywhere in the sequence, we allow more than just the simple case where advocates are either at one extreme or the other. Some people will have a centrist bias.

The string may represent a content or category space. Returning to the legal metaphor, we can describe legal argument as a question of categorization: which legal precedenets and protocols best describe the situation at hand? [See, e.g., [8].] Some categorizations will lead to judgements for the defendant and some will favor the prosecution. Similarly for the stories told by reporters: some will categorize an election as a question of economic policy, while others will characterize it as a question of social policy.

The listener's interpretation The listener of this model is an automaton who receives information and calculates the most likely state of the world. As discussed above, this is not descriptive of real listeners. We know most voters are not well informed and put no effort into becoming informed, and we can see that media outlets are aware of this when they decide what to report on. But within this context, can that minority of individuals who do exert effort to become informed voters succeed in their goal? In order to answer this question, I assume that the listener is the statistically apt automaton described here.

Returning to the sequence of events, the reporter reports those true facts that fit into his story. The listener can divine that if a fact falls into the reporter's filter, but is not reported, then it must be the case that that fact is false. Therefore, the listener has full information about those facts that fall within the reporter's informational filter.

The listener now knows the story that the reporter wants to tell, in the form of the filter he is using to filter the facts, those facts which fall within the filter, and the database of past events. She knows those facts which fell within the filter with certainty, since the reporter can not lie, but needs to do some calculation to guess at the facts which the reporter does not report.

The calculation is simplified by the fact that we have the database of events from which the true state was drawn. The listener can remove all of the events which are not consistent with the facts reported, giving her a subdatabase of events. From that subdatabase, she can calculate the likelihood that each unreported fact is true by simply looking at its frequency in the subdatabase. If an event happens with greater than 50% likelihood in the subdatabase, then she believes that this fact is true. This is a useful divination method because a fact has some correlation to adjacent facts.

We now have the true state of the world and the facts that the listener believes, based on reports and calculations. We can compare the two and determine what percentage of facts she was able to divine correctly.

3.2 Adding reporters

Having determined the score, we can add another reporter. I assume that reporters share outlook to a certain extent. The same data is easy to gather for all reporters, who all learned data-gathering methods from one of a few journalism schools. Socioeconomically, reporters are generally in the same category, which may also lead to similar opinions (as Wanderer [9] found with regard to critics, for example). If a story is 'compelling' then it will be repeated by several media sources, meaning that they will all tell the same basic story. This subject is discussed further in section 3.4, below.

Returning to the model, I simulate the correlation of reporters by assuming that within a test run, their filters all have at least one point in common.[3] Since the reports are random, different filters will still allow different information to pass through. As shown by Figure 1, where each filter has the same quality and center, this method captures the idea that different media often report similar stories, but at the same time are not identical.

After adding another reporter, the listener now has facts which fit into two filters. She may repeat the guessing procedure with the facts which fit into neither filter to arrive at a new likelihood for nonreported facts and arrive at new beliefs. We can again calculate the percentage of correct beliefs.

The procedure may be repeated with more reporters.

3.3 Calibration

We need only ascertain two numbers to calibrate the model: the mean and variance of the size of the filter. The model is based on asking a series of yes/no questions, and looking at how many of these questions are answered by a given news source. Such a situation is easy to implement: all we need is an issue for which one would look to the media for information, and a series of questions which the news sources may or may not answer. Here, I will remain vague about the details of the study—which topics are of interest, which media are to be studied—since the answer to such questions depends on the scope of the specific study. Having decided on these questions, it is a simple matter to check a given source and verify which of the questions on the list have been answered by the source.

This data will provide us with descriptive statistics about the filter: the mean and variance of the number of questions answered. This data is all one needs to calibrate the model here, and the model can then be run to give a more general idea of the marginal value of each additional medium, in direct and indirect information provision.

This paper is interested in the questions that a voter would need answered to be an informed voter. Although a newspaper article may easily answer 100% of the questions regarding an isolated piece of news, any one source's coverage of a more abstract issue, such as the president's performance over the last four years or the potential impact of a proposed change in law, will be decidedly incomplete. I will report on the full range of reporter quality from zero to 100% and leave the reader to decide what is appropriate for a given situation.

[3] Another method would be to assume that all reporters have a center within a small range. The results are basically equivalent; for the sake of not overproving my case, I chose to report a table of statistics for only one method. The `get_good_filters` procedure shows this alternate method in commented-out code.

3.4 Discussion

In this section, I will list some of the assumptions of the model and the considerations which engendered them.

Information is from a fixed, closed space. This assumption is a necessary simplification for the structure of the model, although it is clearly untrue in reality. The state of the world in the model is drawn from a fixed pool of states, but news—the word is the plural of 'new'—is about new states.

The listener is a Bayesian information gatherer. This means that listeners have no biases, heuristics, or cognitive problems in calculating the probability of non-reported events. Assuming a closed space of events saves us the trouble of calculating prior conditional probabilities and all of the usual Bayesian machinery, which is implicit to, and consistent with, the framework here.

The listener in this model has one and only one interest: to collect accurate answers to her questions. This is not representative of real listeners, who want to hear stories for the sake of entertainment as well as information's sake, who is uncomfortable with opinions which clash with her own and therefore generally listens to sources which have the same bias, and who processes facts differently depending on their context. But the intent here is to ask: if the listener were the ideal information gatherer living in a world of less-than-ideal reporters, under what conditions will she be able to gather full information?

The reporter's utility function is not explicitly described. We can presume that a reporter has some outlook that he wishes to share with the world, and at the same time wants to maximize readership, which means that he wants to present a sufficiently descriptive view of the world. How these two factors balance out is well beyond the scope of a study such as this. Assuming a constant filter is on par with assuming that the reporter has already gone through the mental process of balancing out his biases and abilities with the demands of the public.

If we want to design a model of a reporter who maximizes his utility by selectively providing information, we would first need to define the benefits or detriments of providing a given level of information. But since we lack a clear definition of what the typical consumer is looking for, it is impossible to draw a relationship between information provision and readership (though we do know that relation would not be monotonic). Therefore, I chose to abstract away from questions of what the reporter chooses to do, instead defining a set filter for the reporter which aggregates all of these decisions.

The filter does capture the fact that some facts are easy to gather while others are costly. If information is so costly that the reporter does not deem it worth gathering, then that information has been filtered out by that reporter. Since different reporters have different comparative advantages in information gathering, different reporters will have cost-based filters which all include the easy-to-gather points but which each have only a handful of the diffigult-to-gather points.

The listener has full information about reporters. The model could be modified so that the listeners do not have the reporter's filter before them. Instead, they would need to have prior beliefs about the filter the reporter could choose, out of the space of possible filters. It is clear that such a generalization would reduce the amount of information gleaned from any given number of reporters, since it requires more guesswork by the listener, and it is clear that the assumption of full knowledge of a reporter's filter is not realistic except on the most general level. But generalizing to priors over the space of filters adds a great deal of complication and requires many more assumptions about the listener and the world around her, but would not necessarily provide additional clarity.

Actors are nonstrategic. As above, it is impossible to write down a utility function and set of equations for maximization for the reporter; this also precludes game-theoretic maximizations. Although the sole goal of having a listener in this simulation is to determine the percentage of information that the listener receives, this is not necessarily descriptive of listeners in the real world. Therefore, since the true goal of the model is to realistically portray what real reporters tell their actual listeners, it is anathemic to out purposes to model a set of reporters who respond to our toy listener whose sole interest is information gathering.

Also, the assumption that listeners have full information about the reporters' filters and that those filters are constant means that reporters are unable to lie to their listeners: listeners may simply implement the assume-the-worst rule from Milgrom & Roberts's paper.

Reporters' filters are correlated. This assumption was discussed above. Since the filters are based on a number of motivations, this assumption merits discussion in a number of different directions.

The cost of gathering a certain fact tends to be the same for most reporters. For example, a press release or public statement is easy to obtain, while reducing a lengthy congressional hearing to a few useful facts may require days of effort. We can expect all reporters to report the cheap facts, while each reporter will be able to report only some subset of the more expensive facts.

Noelle-Neumann[7] presented the oft-cited idea that the choice of story is an information cascade in the style of Bikhchandani, Hirshleifer, and Welch[1]. If one media outlet presents a story and receives a good response, then other outlets can run a similar story with little risk that there will be no interest.

This effect is often true regardless of fundamental ideology. For example, if a conservative politician commits an error on a newsworthy scale, then liberal reporters will gladly report on it. Conservative reporters may then be obliged to respond with some mention of the error, perhaps adding additional facts which put the politician in a better light.

Unlike prosecution and defense, reporters tend not to be extremely polarized. Reporting with a centrist outlook is a risk minimizing strategy both for repoterts and the media which hire them. Extremists tend to be relegated to niche markets, where few readers will hear from them.

3.5 Results

The final output of the simulation is a function mapping number of reporters and quality of reporters to the quantity of information presented. Figure 2 shows a representative slice of the function for a fixed quality of reporter (around 0.7). It is the mean of a few hundred tests, in which the procedure listed above is run and the amount of information gathered at each stage is recorded. On average, having one reporter gives most but not all information, and each additional reporter provides more information, but with generally diminishing returns. As the number of reporters increases, the average amount of information their collective filters allow will approach, but never reach, one.

3.6 How many reporters do we need to guarantee full information?

Figure 3 shows the result of 200 independent hypothesis tests. The question each percentage answers is: If we have n reporters of a given quality, what is the likelihood that the amount of information we gather will be indistinguishable (in a statistical sense) from full information?

Figure 2 showed the mean percentage of information presented; that information is not present in the table here. For example, if we have twenty facts to determine and one reporter of 95% quality, then we will know 19 facts with certainty and one with about 50% likelihood. But as good as this information is, it is no guarantee of full information, and at the 95% or 99% confidence levels, we can reject the hypothesis that this setup will guarantee us full information. But we fail to reject the hypothesis that two reporters of 95% quality will not give full information at any significance level grater than the 1-.427 = 57.3% confidence level. Reversing the double negatives, we can say that if we wish to be 95% certain that reporters of 95% quality will provide us with full information, then we need at least two reporters.

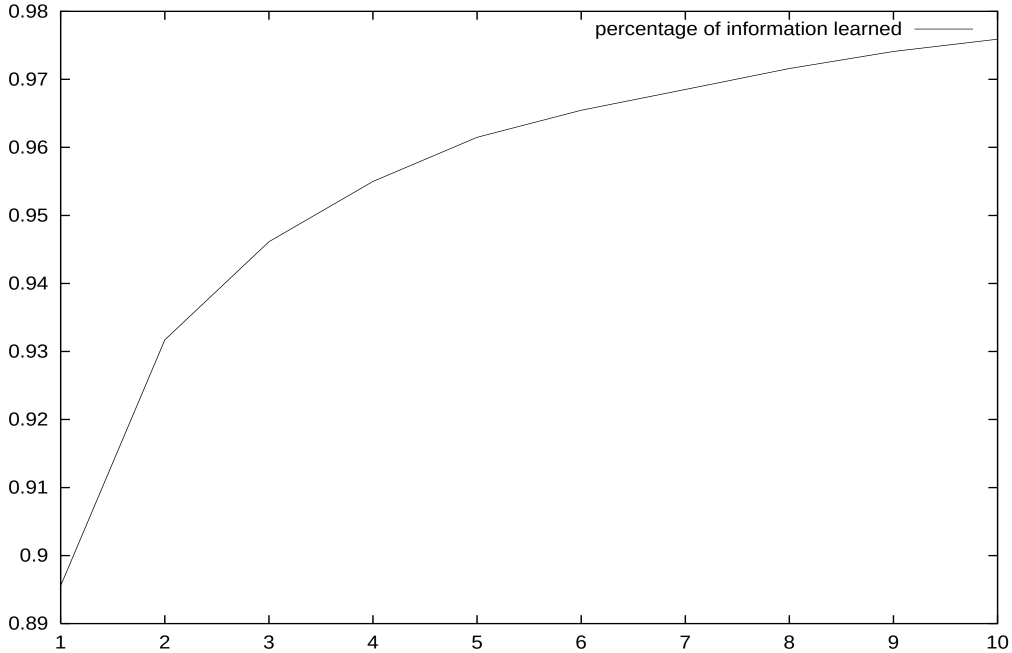

Figure 2: The percentage of correct information given n reporters

For visual clarity, I have marked the smallest reporter/quality pairs in each row or column which are indistinguishable from full information with 95% confidence. The reader should also note that I omitted the case where reporters give zero information and where all reporters give full information, since their outcome is obvious.

Using a significance level of 95%, the data shows a cutoff at a quality of about 40%. If reports include 40% of the information or more, as would be the case in a simple news story, then only one to four sources is necessary to ensure full information with significance. If reports include less than 40% of the information, as would be the case of a complex political question, six or more reporters would be necessary.

4 Appendix: Code

This is code written in Octave, a package freely available from www.octave.org. This code should also work in the MATLAB package.

The general procedure is as described in the text above, but there are a few tricks of implementation that the reader of the code should note. Most of these are based on the fact that producing normally distributed random numbers on the scale needed here is a slow process. Therefore, the `filters` table is produced beforehand, and is saved for later use. The time

	1	2	3	4	5	6	7	8	9	10
0.05	0.00000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.10	0.00000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.15	0.00000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.20	0.00000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004
0.25	0.00000	0.000	0.000	0.000	0.000	0.001	0.006	0.013	0.021	0.032
0.30	0.00000	0.000	0.000	0.000	0.003	0.010	0.026	0.046	0.070	0.097
0.35	0.00000	0.000	0.000	0.004	0.018	0.052	0.084	0.112	0.150	0.175
0.40	0.00000	0.000	0.002	0.024	0.071	0.124	0.169	0.218	0.245	0.265
0.45	0.00000	0.000	0.016	0.072	0.135	0.211	0.250	0.288	0.320	0.347
0.50	0.00000	0.000	0.024	0.101	0.180	0.223	0.294	0.330	0.374	0.405
0.55	0.00000	0.001	0.086	0.219	0.314	0.390	0.433	0.469	0.497	0.524
0.60	0.00000	0.003	0.062	0.202	0.310	0.371	0.424	0.450	0.474	0.491
0.65	0.00000	0.014	0.122	0.253	0.360	0.424	0.459	0.494	0.535	0.556
0.70	0.00000	0.019	0.200	0.393	0.465	0.512	0.547	0.575	0.579	0.599
0.75	0.00000	0.025	0.217	0.344	0.460	0.538	0.552	0.592	0.632	0.652
0.80	0.00000	0.047	0.254	0.456	0.607	0.669	0.692	0.728	0.750	0.768
0.85	0.00012	0.110	0.369	0.573	0.657	0.689	0.725	0.766	0.795	0.805
0.90	0.00009	0.167	0.454	0.617	0.752	0.795	1.000	1.000	1.000	1.000
0.95	0.00225	0.427	0.675	0.766	0.795	0.816	1.000	1.000	1.000	1.000

Figure 3: What is the likelihood that we will get full information from n reporters of the given quality?

savings from this trick is substantial, but (unlike with coffee) I did not find that reusing the filters was detrimental to later batches.

Both events and the reporter's filters are chosen from this database. Events are randomly chosen, while filters are randomly selected from those with a certain center and quality. The calibrations in the `generate_filters` procedure gives a table which is about half ones and half zeros, meaning that the radomly selected events will be 50-50 as well.

The Central Limit Theorem applies to the mean of a group of iid draws from some distribution. The loop beginning with `for k = 1:testrun` produces such an iid set of draws, whose mean is added to the `meanscore` table. The loop beginning with `for ct = 1:loops` produces a series of these means; that series is (asymptotically) normally distributed.

Finally, this code will be available at `http://avocado.caltech.edu/fcc`, along with a current version of this paper (probably until September 2003).

```
#! /usr/bin/octave -qf
n = 20;                 #number of facts
```

```
eventcount = 200;       #number of events
testrun = 20;           #iterations for one run
loops = 100;            #iterations of the whole model
rpts = 10;              #maximum number of reporters to listen to
correlate_filters = 1;  #Should we demand a central filter?

source("fns");          #see below.

#First, generate a sample of filters, noting the number of 1s.
#After generating them, they're kept in a file for reuse.
try      load -ascii -force "data/filters" filters;
catch    filters = generate_filters(eventcount*4,n);
end_try_catch

meanscore=[ ]; meandiff=[ ];
for ct = 1:loops
    score=zeros((n+1)*testrun,rpts+1);
  for m = 1:n
    ok_filters=get_good_filters(filters,m,n,correlate_filters);
    for k = 1:testrun
        rownumber=(m-1)*testrun + k;
        printf("%d ",rownumber + testrun*n*(ct-1));  #something to look at
        events = generate_events(eventcount,n,filters);
        state = vec(events(randomno(rows(events)),:))';
        mask=zeros(1,n);
        for rpt = 1:rpts
          new_mask = vec(ok_filters(randomno(rows(ok_filters)),1:n))';
           mask = any([mask ; new_mask]);     #combine masks
          if (sum(mask)==n) #everything is reported, consumers get a point
            score(rownumber,rpt)=1;
          else
            reported = state .* mask;        #reporter filters the state
            score(rownumber,rpt)= consumer_thinks(
                      whatcouldbe(events,reported,mask),mask,state,n);
          endif
        endfor
        score(rownumber,rpts+1)=mean(new_mask);
    endfor
  endfor
  meanscore = [meanscore ; summarize(score,rpts+1)];
  diff = [score(:,1:rpts)-[.5*ones((n+1)*testrun,1) score(:,1:(rpts-1))] ...
                                                    score(:,rpts+1)];
```

```
  meandiff = [meandiff; summarize(diff,rpts+1)];
  save -ascii "data/score" score
  save -ascii "data/diff" diff
  save -ascii "data/meanscore" meanscore
  save -ascii "data/meandiff" meandiff
endfor

#Summarize and output
finalscore=summarize(meanscore,rpts+1,'mean');
finaldiff=summarize(meandiff,rpts+1,'mean');
pstats = (1- finalscore) ./ summarize(meanscore,rpts+1,'stddev') ;
tstats = 1-normal_cdf(pstats,1,1);
tstats(:,columns(tstats))=finalscore(:,columns(finalscore));
save -ascii "data/tstats"  tstats;
save -ascii "data/finalscore"  finalscore
save -ascii "data/finaldiff"  finaldiff
```

Here is the file `fns`, called from the main program above:

```
empty_list_elements_ok = 1;       #prevents a few error messages

function r = biased_randomno (bias,scale,shift)
#return a normally distributed binary random no
    r= (normal_rnd(((scale*shift)-bias),scale,1,1)>0);
endfunction

function pts = consumer_thinks(e,mask,state,n);
#find out what the listener knows; return her score
    if (rows(e)>1)
            cut_state=state(!mask);
            believed = round(sum(e)/rows(e));
                  pts=(sum(!xor(cut_state,believed))+sum(mask))/n;
    else #if there's only one option, consumers get a point
                 pts=1;
    endif
endfunction

function r = generate_biasedrow (bias,size,shift)
#create a single event
    r = zeros(size,1);
    for j = 1:size
```

```
            #r= [r biased_randomno(abs(j-bias),size,shift)];
            r(j)=biased_randomno(abs(j-bias),size*.5,shift);
    endfor
    r = r';
endfunction

function e= generate_events (qty, n, filters)
#generate the past events. They are correlated,but the center
#is random and I don't care how many facts are true for each event.
    e=zeros(qty,n);
    rr = rows(filters);
    for dummy = 1:(qty)
      e(dummy,1:n) = filters(randomno(rr),1:n);
          #generate_biasedrow(randomno(n),n,(randomno(20)-4)/10);
    endfor
endfunction

function filters= generate_filters (qty, n)
#create a file of sample filters from which to pull samples
    filters=zeros(20*qty,n+2);
    for i = 1:20
        shift=(i-4)/10
        for dummy = 1:(qty)
  center=randomno(n);
          indx=((i-1)*qty)+dummy;
          printf("%d ",indx);  #something to look at
          filters(indx,1:n) = generate_biasedrow(center,n,shift);
          filters(indx,n+2)=center;
        endfor
    endfor
filters(:,n+1)= (sum(filters')- filters(:,n+2)')';
save -ascii "data/filters" filters;
endfunction

function ok_filters=get_good_filters(filters,m,n,correlate_filters)
    ok_filters_prelim=filters(filters(:,n+1)==m,:);
    prevent_infinite_loop=0;
    do
      prevent_infinite_loop++;
      if (correlate_filters)
  t = randomno(n);
          ok_filters=ok_filters_prelim(ok_filters_prelim(:,t)==1,:);
```

```
  #goodindices=ok_filters_prelim(:,columns(filters))==t;
          #ok_filters=ok_filters_prelim(goodindices,:);
      else
          ok_filters=ok_filters_prelim;
      endif
    until (rows(ok_filters)>3 || prevent_infinite_loop>2000)
    if (prevent_infinite_loop>2000) ok_filters=ok_filters_prelim; endif
endfunction

function r = randomno (upto)
#return a uniform random number between 1 and upto
    r = ceil(rand() * upto);
endfunction

function avgtab = summarize(intab, indexcol,sum_type)
#find mean (default), 'var', or 'stddev' by index row
    if (nargin == 2) sum_type='mean'; endif
    indices = union(intab(:,indexcol),[]);
    avgtab = zeros(length(indices),columns(intab));
    for i = 1:length(indices)
        sum_me = intab(intab(:,indexcol)==indices(i),:);
switch lower(sum_type)
case 'var'
  avgtab(i,:) = var(sum_me);
  avgtab(i,indexcol) = indices(i);
case 'stddev'
  avgtab(i,:) = sqrt(var(sum_me));
  avgtab(i,indexcol) = indices(i);
otherwise
          #note that the last col of sum_me is constant,
          #so its mean is that constant
          avgtab(i,:) = mean(sum_me);
end
    endfor
endfunction

function e = whatcouldbe(events,reported,mask)
#remove any events inconsistent with what was reported
    chksum=sum(reported);
    realistic=[];
    for j = 1:rows(events)
            if (sum(reported .* events(j,:))==chksum);
```

```
            realistic = [realistic j];
            endif
    endfor
    #cut out events which don't fit the data and
    #anything that was accurately reported
    e = events(realistic,!mask);
endfunction
```

References

[1] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(51):992–1026, 1992.

[2] Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, November 1982.

[3] Ben Klemens. Information aggregation. Technical report, California Institute of Technology, 2002. http://avocado.caltech.edu/klemens/klemens_diss.pdf.

[4] Thomas A Mauet. *Trial Techniques.* Aspen Publishers, 6th edition, 2002.

[5] Paul Milgrom and John Roberts. Relying on the information of interested parties. *RAND Journal of Economics*, 17(1):18–32, Spring 1986.

[6] Peter Murray. *Basic Trial Advocacy.* Aspen Publishers, 1995.

[7] Elisabeth Noelle-Neumann. *The Spiral of Silence.* University of Chicago Press, 1993.

[8] Cass R Sunstein. *Legal Reasoning and Political Conflict.* Oxford University Press, 1996.

[9] Jules J Wanderer. In defense of popular taste: Film ratings among professional and lay audiences. *American Journal of Sociology*, 76(2):262–272, September 1970.