The Wisdom of Reluctant Crowds

Christian Wagner
Department of Information Systems
City University of Hong Kong
iscw@cityu.edu.hk

Sesia Zhao
USTC-CityU Joint Research Center
sesiazj@mail.ustc.edu.cn

Christoph Schneider
Department of Information Systems
City University of Hong Kong
christoph.schneider@cityu.edu.hk

Huaping Chen
School of Management
University of Science and Technology of China
hpchen@ustc.edu.cn

Abstract

Estimating is difficult. This is true whether the task requires forecasting uncertain future events, or whether the estimation task is complex in itself and based on insufficient information. As a result, even perceived experts frequently estimate poorly. Surprisingly, recent research suggests that groups of non-experts can outperform individual experts, given certain conditions are met. The resulting capability has been described as collective intelligence or "wisdom of crowds". Yet crowds (and individuals) do not necessarily like to make guesses, whether because it is cognitively hard, or emotionally undesirable. If crowd members prefer not to estimate, but instead seek to transfer this responsibility to others, are they able to identify good surrogates? We empirically tested these two aspects of collective intelligence. First, we explored whether collective intelligence was able to produce estimates that are significantly better than those of individuals, and second whether perceived experts as surrogate estimators were able to perform the task equally well. Our findings demonstrate good estimation ability for the crowd as well as its surrogates. We discuss implications for scenarios where estimates involve both beliefs and preferences, and where collective estimates thus have to be negotiated. Resulting requirements for information systems are outlined.

1. Introduction

Making risky decisions and predicting[1] unknown events are two activities which people cannot avoid in their lives. Yet estimating is difficult. This is true whether the task requires forecasting uncertain future events, or whether the estimation task is complex in itself and based on insufficient information. Consequently people are not good at performing estimation tasks [1], frequently exhibiting biases and making errors [2, 3]. For example, physicians faced with the difficult question "Doc, how long do I have left to live" systematically demonstrate an optimistic bias [4]. Research has uncovered that systematic, non-pathological biases in cognitive processes, which are neither dependent on intelligence, nor on education, are distributed equally in the population [5]. Thus experts are as prone to biases and errors which may influence their prediction and decision making behavior as are non-experts. Hence, the overarching objective of our research is to find out in how far these biases and errors can be overcome by relying on collectives, rather than on skilled individuals.

1.1. Ability of crowds to guess well

While individual experts and non-experts are not good in estimating, collectives apparently are. This interesting insight can be drawn for instance from TV game shows such as "Who Wants to Be a Millionaire" (with its ask-the-audience feature) or prediction games such as Yahoo's "College Pickem", where the crowd year after year rivals or beats expert punters in predicting football game outcomes. Collective intelligence has come to widespread attention through Surowiecki's influential book The Wisdom of Crowds [6], which describes both the condition under which collective intelligence manifests itself and which illustrates, through numerous scenarios, the effectiveness of crowd wisdom. In line with our overarching objective, our first research goal therefore became the validation of this suggested ability of crowds to predict well.

1.2. Reluctance to predict

People don't like to predict. This has several reasons, namely own perceived capabilities, risk propensity, and outcome related preferences. Given their general lack of ability, people's usual attitude is reluctance to forecast future outcome probabilities

[1] We use the terms predicting, forecasting, estimating, and guessing interchangeably in this article.

IEEE computer society

[7]. Furthermore, forecasting activity cannot be separated from uncertainty and risk [8]. When placed in scenarios where predictions have to be made involving "ambiguity", people exhibit behaviors that indicate ambiguity aversion [9] and reluctance to assign numerical probabilities to uncertain events [10]. Finally, people are reluctant to guess when the outcome is undesirable, possibly because they do not want to dwell on the impact of the undesirable outcome, or because they worry about self-fulfilling predictions, e.g., [11]. As result of this multiplicity of reasons, people will be reluctant to guess, will seek information from others, or will abdicate the responsibility to make an estimate to others. Thus, our second research goal was to assess the value of the choice made by crowds in abdicating their estimation responsibility to others.

1.3. Crowd designees – is the reluctant crowd wise enough to pick good surrogates?

Hence, when we ask others to make assessments for us, is this a wise decision? According to crowd wisdom, crowd choices on who is best in predicting for them should be (collectively) intelligent. Crowd members should know whom to choose as surrogate experts, and these experts, as long as they have enough diversity among them, should be performing well.

The reminder of the article pursues both described research objectives. Next we outline the research background and foundations, followed by a description of our research framework. We then present results of the empirical analysis, followed by discussion and implications. The paper ends with conclusions and suggestions for future work.

2. Background and Foundations

2.1. Collectives

Collectives differ from traditional groups, not only in term of their size, but also with regard to their characteristics. Groups are frequently defined as "social aggregates that involve mutual awareness and potential mutual interaction. … are relatively small and relatively structured or organized" [12, p. 7] While their structure and cohesion gives groups an advantage in a number of tasks, these characteristics can also lead to reasoning biases such as group polarization [13] or representativeness fallacy [14]. Collectives, which are not 'normed' [15] and don't necessarily share the same attitudes may perform worse on tasks requiring integrated action, but in turn benefit from members' relative independence.

However, not all collectives are alike, and collectives do not necessarily make wise decisions. Numerous examples (such as stock market bubbles) exist that demonstrate situations in which the behavior of the crowds led to suboptimal outcomes. Surowiecki [6] argues that such instances of crowd madness are a product of a faulty system or decision-making environment, rather than failing wisdom of the crowds. For instance, stock market bubbles are not a product of aggregating the wisdom of crowds, but a result of herding behavior as people are following others' opinions and actions. Hence, Surowiecki proposed three requirements for collective wisdom to emerge, diversity, decentralization of opinion, and independence. Diversity of opinion refers to the availability of multiple viewpoints (ideally many), within the group, as each further viewpoint may help to explain the phenomenon better. Independence means that peoples' opinions are not determined by the opinions of those around them, which is typically the case in groups or teams. Decentralization requires that people can draw on their local and specific knowledge and make independent decision.

2.2. Requirements for performance

For crowds to be able to predict, several criteria have to be met. First, the variable to be estimated cannot be completely random. Asking a crowd (or individual) to predict the outcome of a coin flip is futile. Second, individuals have to have some reasoning capability and information. If individuals guess at random because they lack either information of reasoning ability, the collective outcome will carry no information. Individuals, however, do not have to know the exact right value. Most importantly is some ability to eliminate impossible values. To simply illustrate, if three individuals (I1-I3) need to determine the right answer among choices A-D, and each one is able to eliminate two choices, then a process of approval voting might yield: I1: A and B, I2: B and C, I3: B and D. In this case, A, C, and D each would receive one vote, while B would receive three votes, thus making it the favorite guess. Similarly, in guessing the number of candies in a jar, ascertaining the right number will not be as important as being able to exclude impossible ranges. To achieve this outcome, we need the crowd to eliminate as many impossible values as can be done. The crowd must therefore approach the problem with different mechanisms so as not to replicate the same elimination mechanism time and again. In other word, the crowd needs diversity.

2.3. Not the law of large numbers

Frequently it is assumed that the wisdom of crowds is merely a manifestation of the Law of Large Numbers, such that when a crowd estimates, the error terms of the extreme cases will cancel each other out and thus the mean estimate approximates a good guess. While one aspect of having a large crowd is the reduction of errors and noise, this is not the main effect the crowd has. In fact, as mentioned already, diversity is sought. The logic of collective intelligence is that different individuals will apply different "theories", or more appropriately heuristics, to the guessing task, the aggregate of which results in a highly precise estimate of the variable in question. While each "theory" would only be able to predict part of the variance in the observed outcome, the collection of theories brought together can explain much of the variance and lead to a highly precise result. Asked "will it rain tomorrow", one individual might observe the clouds to make a prediction, someone else may view the barometer, someone else may refer to the Farmer's Almanac e.g., [16], and so on. The aggregate precision and breadth of applicability of these approaches would expectedly be high. For this logic to apply, crowd members must use different heuristics, or else their value is diminished to the law of large numbers. This is the reason why diversity in the crowd is fundamentally important. A similar explanation, also assuming diversity, can be found in Condorcet's Jury Theorem [e.g., 18]. According to this theorem, independent voters, each of whom is able to identify the correct answer with more than 50% accuracy will in aggregate pick the correct one among alternatives. Given our decision situation, with a potentially infinite number of alternatives, combined beliefs and preferences, and no assumption of >50% accuracy, Condorcet's theorem is not directly applicable to the scenarios we are presenting, but could apply to a narrower set of problems and decision makers.

2.4. Rationale of crowd wisdom

For our set of decision problems and relaxed assumptions concerning decision maker accuracy, we apply Page's [17] logic, rather than Condorcet's [18]. Page categorized cognitive diversity into four dimensions: diversity of perspectives (ways of representing situations and problems), diversity of interpretations (ways of categorizing or portioning perspectives), diversity of heuristics (ways of representing situations and problems) and diversity of predictive models (ways of inferring cause and effect). He formalizes the range of diversity in a prediction diversity (PD) variable, and further conceptualizes two other parameters of collective intelligence, collective error (CE), and (average) individual error (IE).

Prediction diversity is defined as the aggregate squared difference between individual guesses and the average guess. It reflects how far individuals, on average, veer from the group.

$$\text{PD} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ;$$

Prediction diversity is important as it demonstrates the diversity among decision makers, which is a commonly assumed requirement for collective intelligence [6, 17, 18].

Individual error aggregates the squared errors of all the participants. It thus captures the average accuracy of the individual guesses.

$$\text{IE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{true})^2 \ ;$$

Individual error (squared-average) is a benchmark to compare the performance of the collective.

Collective error, the squared error of the collective prediction, represents the difference between the correct answer x_{true} and the average guess $\bar{x}$ aggregated from all individuals.

$$\text{CE} = (\bar{x} - x_{true})^2$$

Collective error (squared), correspondingly, is the error that the collective makes in prediction. Page then formulates the *diversity prediction theorem*, which states:

Collective Error (CE) = Individual Error (IE) – Prediction Diversity (PD)

A low collective error signifies high overall collective intelligence and accuracy. The collective error, being composed of individual error and prediction diversity, pits these two parameters against each other. In other words, we can lower collective error by reducing individual error (raising individual expertise) or by raising prediction (group) diversity.

2.5. The problem: deterministic vs. heuristic vs. preferential: How good is the crowd?

Typically, questions can be defined along two dimensions: deterministic vs. non-deterministic and belief vs. preference. Deterministic questions can be answered with certainty, although they may require complex reasoning and information that is difficult to obtain. Non-deterministic ones have elements of chance in them, and in our case refer to the future.

Preference based questions solicit an answer that might reflect both the respondent's beliefs and preferences. The two dimensions with two values each result in four types of problems (questions) for decision makers. Type 1 questions require the estimate for a difficult but deterministic and belief-based problem. Type 2 questions require a belief-based answer for a probabilistic future event. Type 3 questions require an answer that is deterministic but incorporates both belief and preference. Type 4 questions require an answer to a non-deterministic problem which involves both belief-based and preferential decision making.

2.6. Hypotheses

Based on earlier reflections concerning the performance of crowds versus individuals along a range of tasks, we formulate three hypotheses for this research, referring to (1) the ability of crowds to outperform average individuals, (2) the ability of crowds to choose good surrogate experts for tasks involving beliefs estimates and task involving preferences, and (3) the ability of crowds to estimate very accurately in absolute terms.

Hypothesis 1: The crowd will show significant better prediction capabilities than individuals (relative crowd wisdom).

Hypothesis 2a: There will be no difference between estimates made by the crowd and its surrogates in non-deterministic questions.

Hypothesis 2b: There will be no difference between estimates made by crowds and its surrogates in preference based questions.

Hypothesis 3: The crowd will be able to make estimates which are not significantly different from true values (absolute crowd wisdom).

3. Research Methods

3.1. Operationalization

To test our hypotheses we posed several questions to two groups of graduate students. The first sample was a group of 19 students (G19), the second a group of 30 students (G30). To motivate participation, we offered monetary prizes. For each scenario, we asked subjects to make their own guess, and to identify three other individuals in the group (group members knew each other quite well) whom to ask so as to receive the best estimate. The three scenarios are illustrated in Table 1.

	Beliefs	**Preferences**
Deterministic	Jelly Bean Quantity	Best Movie
Non-deterministic	Future Temperature	---

Table 1: Estimation task dimensions

The three scenarios were as follows: The first scenario asked subjects to guess the number of jelly beans in a container shown to them (Type 1). This is a widely reported and popular crowd wisdom problem (e.g., [6, 19]), resembling other Type 1 problems, ranging from identifying mineral deposits to solving scientific problems. The second scenario asked them to guess the temperature in their local region one week in advance (Type 2). This is a scenario similar to those found in prediction markets, where subjects trade predictions like securities in a market [20]. The outcome for which predictions are traded is exactly known eventually, but not at the time the predictions is made. Typical prediction markets forecast the outcomes of football games, or stock market results. The third task asked them to assess the quality of a famous movie, known to all subjects (Type 3). This is a popular problem to be judged by crowds, such as in the influential website rottentomatoes.com [21]. We left out the highly speculative non-deterministic—preference scenario (Type 4). All of the presented scenarios could potentially be addressed by a single expert (e.g., a meteorologist or movie critic). And yet, collective intelligence mechanisms have emerged for problems such as these, specifically to draw on the insights of many [6].

3.2. Determining intelligence

A crowd might be argued to be intelligent if its collective error is lower than the average individual error. However, as per the diversity prediction theorem, this is true for any crowd which demonstrates at least some prediction diversity. Thus, to confirm intelligence, we sought a higher yardstick. We defined two measures of intelligence, one based on expert-novice performance differences, the other based on statistical reasoning.

3.2.1. Expert-novice differences. The literature on expertise has repeatedly, although not entirely consistently, found evidence of experts outperforming non-experts (often called "novices"). Vessey [22] for instance found in a study of computer programmers that, on average, experts completed a software debugging task, requiring diagnosis and planning, in 55.8% of the time it took non-experts. In other words, experts were almost twice (1.79x) as

fast. Lee et al. [23] similarly found that performance gained through expertise led to a reduction in task completion time to 57.8% (1.73x) and effectiveness increase of 64% for a DSS user group. Lee et al. even found a 212.5% effectiveness increase for a user group without DSS. Thus, with expertise raising productivity by a factor of 1.79 (time) and up to 3.13 (effectiveness), we chose to define an expert performance threshold accordingly. To evaluate expert-novice differences, we use a "collective intelligence quotient," i.e., the ratio of individual error (IE) to collective error (CE). For a collective estimate to be considered intelligent, we required IE/CE > 10; with IE and CE being squared terms, this represented a performance ratio of $\sqrt{10} = 3.16$.

3.2.2. Proximity of collective guess to the true value. We also considered a statistical definition of collective intelligence which captured the distance between true value and collective guess. We would consider a collective guess to be good if the true value falls within a +/-5% confidence interval around the collective guess. This formulation does not exactly represent the nature of collective intelligence, as it considers absolute quality of the guess, not the comparison of a collective versus an individual guess. For all we know, a collective could beat any individual expert, but still be relatively far off the true value for any particular estimate. For example, the crowd in Yahoo's 2008 College Pickem was correct only 77.1% (216 out of 280 guesses), but this was the best performance, a tie with expert punter Olin Buchanan (http://rivals.yahoo.com/ncaa/football/pickem?w=16).

4. Results

4.1. Crowd guesses

The results of our analysis are captured in Table 2. The table shows both groups (G19 and G30) as well as the results of perceived experts (G19E, G30E). In terms of performance relative to that of the average individual (Hypothesis 1), the crowd performed well. In all but one case (Future Temperature, G30) the ratio of IE/CE was considerably higher than 10, our threshold for collective intelligence.

Perceived experts performed as well as the collectives they represented, if not better (Hypothesis 2). All perceived expert collective guesses met our threshold IE/CE > 10, even when the collective itself had been below the threshold (Future Temperature, G30). Thus, the perceived experts did as well, or arguably better than the collectives they represented. To further test Hypothesis 2a, we conducted independent samples *t*-tests. The results revealed that the differences between the guesses of the collectives and the surrogate experts were non-significant with G19 vs. G19E: $t(62)=.06$, $p=.955$ and G30 vs. G30E: $t(97)=1.30$, $p=.196$. However, whereas the +/-5% confidence intervals around the sample means overlapped in the case of G19 vs. G19E (power=.90), there was no overlap in the case of G30 and G30E; thus, Hypothesis 2a was partially supported.

For both the jelly bean quantity and the future temperature guesses (beliefs), the prediction diversity of the surrogate experts was lower than that of the full group they represented. For the guesses about movie quality, there was no "right" result. The assessment of the quality of a particular movie was, of course, highly value driven. We purposely did not ask for value free assessments such as box office success. We conducted further independent-samples *t*-tests to evaluate whether the surrogate experts differed from the crowds on such tasks (Hypothesis 2b). The results show that there were no significant differences, with G19 vs. G19E: $t(63)=.67$, $p=.503$ and G30 vs. G30E: $t(97)=.608$, $p=.545$. However, using our more stringent quality assessment, for both, the +/-5% confidence intervals around the means did not overlap; thus Hypothesis 2b was not supported.

Group	x_{true}	$\bar{x}$	s.d.	CE	IE	PD	IE/ CE
Jelly Bean Quantity							
G19	99	96.47	38.54	6.38	1413	1407	220.8
G30	46	44.37	15.62	2.66	238.4	235.7	89.6
Future Temperature							
G19	13	13.89	3.48	0.8	12.26	11.46	15.3
G19E	13	13.84	3.13	0.71	10.3	9.60	27.2
G30	29	29.89	2.46	0.8	6.66	5.86	*8.3*
G30E	29	29.28	2.03	0.08	4.15	4.07	51.9
Best Movie							
G19	---	3.32	0.89	---	---	0.74	---
G19E	---	3.15	0.89	---	---	0.78	---
G30	---	3.27	1.05	---	---	1.06	---
G30E	---	3.41	1.06	---	---	1.10	---

Table 2: Group results

Related to Hypothesis 3, the collectives did arguably quite well in terms of absolute values, guessing temperatures correctly within one degree (Celsius) and jelly bean quantities within 1.63 and 2.53 beans. A conventional *t*-test showed no significant difference between the sample means and the true values (although the differences approached significance for G19E and G30 guessing the future temperature, see Table 3).

Group	x_{true}	$\bar{x}$	s.d.	t	p
Jelly Bean Quantity					
G19	99	96.47	38.54	0.286	0.778
G30	46	44.37	15.62	0.573	0.571
Future Temperature					
G19	13	13.89	3.48	1.121	0.277
G19E	13	13.84	3.13	1.800	0.077
G30	29	29.89	2.46	1.987	0.056
G30E	29	29.28	2.03	1.130	0.264

Table 3: Absolute accuracy of collectives

Although the conventional t-test did not "prove" the estimates to be different, a more stringent test, could not "prove" that they were same. In none of our scenarios, the true value fell into a +/-5% confidence interval around the sample mean. Using this criterion, Hypothesis 3 was not supported.

5. Discussion and Implications

5.1. Result interpretation

The results confirmed our expectations in most respects. First, our relatively small "crowds" of 19 and 30 subjects demonstrated collective intelligence, performing in all but one case substantially better than the average individual, with an expert-like performance differential of at least 3.91 (√15.3),thus confirming our expectation. Second, surrogate groups of perceived experts performed equally as well as the collectives themselves. While their absolute guesses were (except for one case) not significantly different from those of the collectives they represented, the perceived experts "collective intelligence quotient" IE/CE was higher than that of the crowds they represented, thus confirming our expectation that perceived experts would perform as well. Apparently our crowds were as good at guessing their surrogates as they were at guessing values in our exercises. Also reassuring was that surrogate experts resembled the crowds quite closely on a preference assessment task, where the quality of a movie was to be assessed. Unlike individual critics in the media, whose preferences frequently do not match that of the general public, our perceived experts matched collective preferences very closely. Finally, in absolute terms, both "crowds" of 19 and 30 subjects did quite well, guessing temperatures correctly within one degree (Celsius) and jelly bean quantities within 1.63 and 2.53 beans. However, using our stringent criterion, the guesses were significantly different from the true values.

5.2. Bigger picture: collective knowledge creation

The results of our study show that collective guessing can be beneficial for at least two key reasons. First, the accuracy of collective guesses is superior individuals' guesses. Second, people commonly attempt to eschew making guesses, and surrogate guessing by perceived experts relieves them from this task. We see both these principles exercised in recent efforts to harness the wisdom of many through social software, ranging from collective tagging to prediction markets and crowdsourcing [24].

One particularly surprising application has been the use of wikis, where anyone can freely add, edit, or modify anyone's content. It is a key tenet of wikis that a collective can, over time, produce better quality content than a single writer or a small group. A large and diverse group of people, coordinating their knowledge, and equipped with appropriate mechanisms to aggregate this knowledge, has the potential to create an outcome that is superior to any outcome that could be produced by an isolated individual [6]. This can be, as in the case of Wikipedia, a lengthy and overall high effort process of creating, editing, and reverting changes, until a commonly accepted version of the truth emerges. Yet the result can be a work product whose quality rivals or exceeds that of small expert teams [25],despite criticisms and corresponding controversy (e.g.,[26]).

A closer look of content creation in Wiki (such as Wikipedia or Wikitravel) demonstrates that the content creation process is often accompanied by meta-discussions about the content on so-called "talk pages". Editors use these talk pages to discuss issues related to the article and its creation process, for instance to seek feedback before making changes to the actual article. Even more frequently, talk pages are used after certain edits have been made to an article [27]. Especially when the edits or the topic of the article itself are not value-free and thus preference based and controversial, the resulting discussions can be lengthy and fierce. This demonstrates a collective truth finding process, which often takes place through discussion and persuasion, but sometimes resolves to formal methods [28], including voting.

5.3. Negotiating "truth" in collective knowledge creation

The extrapolation from our simple number guessing experiments to collective knowledge creation of non-quantitative knowledge raises an important question. When individual guesses cannot

simply be aggregated through averaging, how do collectives negotiate the process of knowledge creation, establishing a "truth" acceptable to the collective. A further review of the activities within Wikipedia provides some answers.

Kittur and Kraut [27] showed that almost 40% of edits made in Wikipedia related to consensus building, the creation of policies, or other activities not directly related to the writing of the actual articles. Thus, it can be argued that through the combination of article edits, talk page discussions, and (sometimes) the use of formal governance mechanisms, a "negotiated truth" emerges, that will eventually be accepted by the community.

The different ways of influencing the creation and modification of an article suggest that the level of community involvement—and thus the level of negotiation—can be characterized along a continuum, but with several distinct scenarios. On one extreme, by far the largest number of Wikipedia users never makes a single edit, and thus completely evades the collective intelligence and negotiation process. Essentially this group implicitly decides to accept those who edit as the perceived experts and allows them to be its surrogates. As we found in our experiment, this may not be an unreasonable decision. In another frequent scenario, individual editors freely make any edits they deem needed on relatively uncontested articles. Although these editors participate in the knowledge creation process, negotiation only takes place implicitly, through tacit acceptance of others' contributions in the same article, or acceptance of others' changes to one's own work without explicit discussion. When intended edits are more substantial, yet nevertheless belief oriented, contributors may attempt to seek others' input, and thus actively work towards the concept of collective intelligence. This defines a third scenario. An example can be found in the Wikipedia article on Kowloon (an urban area of Hong Kong). In an earlier version of that article, the origin of the word Kowloon was in question. One contributor stated the uncertain belief, in the talk pages, that the word Kowloon may be derived from a story of "nine dragons". Another Wikipedian confirmed this notion, also in the talk pages, whereupon a while later, the now supported belief moved from Wikipedia's talk pages to the actual content pages. Here a relatively small crowd of two collectively agreed, but gave others sufficient time to offer any opposing opinions. By acquiescence, the collective thus found its confirmed truth. With more vocal members of the collective involved, votes on the talk pages can take place and have been performed, although Wikipedia's wiki has no built-in voting technology.

Finally, the struggle for a negotiated truth becomes visible in the context of controversial issues, such as issues surrounding people's values or preferences. In such cases, editors often resort to using different influence tactics in order to persuade others of the merits of their arguments. In other words, participants self-regulate the creation of collective knowledge by drawing on multiple informal governance mechanisms (such as providing logical arguments or pointing to established guidelines or community norms), in an attempt to achieve commitment to suggested changes from other members of the collective [29]. Oftentimes, heated debates are conducted on the talk pages until a conflict is resolved. Typically, the use of formal governance mechanisms is only the last resort used in the article creation process. This clearly shows that as the collaborators' commitment to establishing their views grows, so does the need for negotiating a commonly accepted version of the truth. As the findings of this study reveal, a collective of surrogate experts can generate results that match those of the collectives, if the collective chooses its experts appropriately. In Wikipedia, for instance, this process is a very intensive one. Although anyone can be a content contributor, those with conflict resolution rights (administrators) are chosen by positive and negative votes based on meritocracy demonstrated frequently through thousands of prior contributions. One thus should expect that the resulting truth negotiation processes in Wikis such as Wikipedia do lead to high quality outcomes.

5.4. Role of information technology

The fundamental principles of collective intelligence and the needs to create negotiated truths, define core requirements for supporting information technology, namely (1) enabling the creation of a diverse pool of beliefs, and (2) facilitating the aggregation of results. In addition, existing implementations of collective intelligence aggregators (including wikis) suggest (3) the creation of meta-knowledge about the group wisdom process.

5.4.1. Enabling the creation of a heterogeneous pool of beliefs. Surowiecki [6] proposed three requirements for collective wisdom to emerge: diversity, decentralization of opinion, and independence. Correspondingly, information technology must enable access to/by a group of individuals from a variety of backgrounds, and must enable them to voice their opinion separately and independently from each other. Why so? Access to personal knowledge from a range of sources is important. After all, any crowd member's "opinion"

is expression of the outcome of a theory applied to the situation at hand. A broader range of theories, in the end, will explain more of the variance in a phenomenon, and thus should lead to better results. Independence of ideas is equally important. Although the literature on prediction market research does not explain it, prior research on human judgment and biases [1] suggests that another individual's opinion can become an anchor, leading to only marginal adjustments in voicing one's opinion, thus reducing variance of opinions and also biasing judgments. In essence, although the collective in this case could potentially have access to a broad range of knowledge, it does not take advantage of this resource if independence is not maintained. Independence maintains the breadth of knowledge that has been created by diversity and decentralization. Consequently, to avoid process losses, diversity, decentralization, and independence must be monitored and potential disturbances must be identified.

5.4.2. Facilitating the aggregation of results. The aggregation of results is simple when information consists only of highly structured knowledge or information. As such, aggregating the price opinions of many in quantitative or categorical predictions is relatively easily automatable and straightforward. However, in situations where we need to aggregate relatively unstructured, non-quantitative information or knowledge, the complexity of aggregation increases. Wiki agreement is one such example. We already explained that Wikipedia's Mediawiki software (as well as several others) does not include a voting mechanism. Hence, aggregation remains a task of considerable difficulty, performed by only the minority of wiki users [24]. Effective and efficient aggregation of hard-to-structure crowd intelligence thus remains a key requirement for information technology.

5.4.3. Creation of Meta-Knowledge about the Group Process. Coordinating the collaboration of many is difficult. Supporting technology must deal with large group problems, and must do so with relatively little facilitation or "governance". Hence, the technology needs to signal its users information about the process, particularly this is vital to the process outcome. For example, if an unregistered user, and thus not a "qualified surrogate", is making significant changes to shared knowledge, other members of the collective may want to be alerted to ensure that this activity does not lower the quality of the collectively created knowledge asset. Alternatively, if the crowd that participates in a vote is too homogeneous, the system should inform the participants, so as to create opportunity for high prediction diversity. Likewise, if some crowd members demonstrate exceptional capabilities, the system could also signal this, thus helping in the process of identifying good surrogate experts.

6. Conclusions and Further Research

Our study confirms several notions of collective intelligence, and extends them to the use of surrogate perceived experts. It further extrapolates findings from the quantitative realm to collective knowledge creation and truth negotiation in non-quantitative collaborative tasks. We find that the answer to "how good are collectives" is that they are clearly better than individuals, and that they can be reasonably good in approaching true values (although not supported by statistics), as long as important conditions are met. We also learned that if collectives choose their surrogates appropriately (as they did in our study), those surrogates perform equally well, and possibly better, both on belief based and preference based tasks. This was reassuring in light of the bigger picture of collective knowledge creation, where we increasingly rely on knowledge created by the collective, through social media. It is further noteworthy that much of the collective knowledge we accept is generated by relatively few surrogates and generated uncontested in a process of silent agreement. Only at the edges of the collective knowledge creation process where contributors have different values and preferences do we see active discussion and even negotiation of the collective truth. At the same time, information technology used for collective knowledge creation today needs to offer more capabilities to match the requirements of collective intelligence, in diversity creation, knowledge aggregation, and process monitoring.

Our study has numerous limitations which create ample opportunity for further work. Specifically, we did not look into the process of how collectives pick their surrogates, even though the selection of appropriate surrogates is so important. Our measures of collective intelligence, specifically the "collective intelligence quotient" IE / CE also raise questions. Determining appropriate measures and thresholds for collective intelligence is important, as is their interpretation in light of the peculiarities of statistics. After all, with current measures, one guess can be "better" than another not because it is closer to the true value, but because its internal variance (diversity) is higher. Moreover, we used a very stringent criterion to determine if two guesses are equal, or whether a guess is equal to the true value. Using this criterion, both regular crowds and crowds

of perceived experts failed to be accurate; even though most people would for instance consider a temperature guess within 1° Celsius as quite accurate. A question is thus, what would be the best, or most useful, way to determine the quality of a guess? Further research will hopefully help us to understand the differences between the quality of the process and the quality of the guess itself.

7. Acknowledgement

This research was supported in part by GRF Grant 9041464, CityU Start-up Grant 7200147, and by the Centre for Applied Knowledge and Innovation Management.

8. References

[1] D. Kahneman and A. Tversky, "On the psychology of prediction," *psychologyical Review,* vol. 80, pp. 237-251, 1973.

[2] J. G. March, "Bounded Rationality, Ambiguity, and the Engineering of Choice " *The Bell Journal of Economics,* vol. 9, pp. 587-608, 1978.

[3] K. Oliven and T. A. Rietz, "Suckers are born but markets are made: individual rationality, arbitrage, and market efficiency on an electronic future market," *Management Science,* vol. 50, pp. 336-351, 2004.

[4] E. B. Lamont, *Formulating an accurate prognosis*: Prognosis in Advanced Cancer, P. Glare and N.A. Christakis, eds., 2008.

[5] S.-M. Ravahi, "Systematic biases in human cognition," in *IEEE International Conference on Systems*, 2002.

[6] J. Surowiecki, "*The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*," 2005.

[7] L. Carlsson, "Policy networks as collective action," *Policy Studies Journal,* vol. 28, pp. 502-520, 2000.

[8] M. Shahidehpour, H. Yamin, and Z. Li, *Market operations in electric power systems: forecasting, scheduling, and risk management*. New York, NY: Wiley, 2003.

[9] C. R. Fox and A. Tversky, "Ambiguity Aversion and Comparative Ignorance," *The Quarterly Journal of Economics,* vol. 110, pp. 585-603, 1995.

[10] C. Heath and A. Tversky, "Preference and belief: Ambiguity and competence in choice under uncertainty," *Journal of Risk and Uncertainty,* vol. 4, pp. 5-28, 1991.

[11] P. Glare and N. A. Christakis, *Overview: Advancing the clinical science of prognostication*: Prognosis in Advanced Cancer 2008.

[12] J. McGrath, "Groups: Interaction and Performance," p. 7, 1984.

[13] I. L. Janis, *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*: Boston: Houghton Mifflin, 1982.

[14] L.Argote, R. Devadas, and N. Melone, "The base-rate fallacy: Contrasting processes and outcomes of group and individual judgment," *Organizational Behavior and Human Decision Processes,* vol. 46, pp. 296-310, 1999.

[15] B. Tuckman, "Developmental sequence in small groups," *Psychological bulletin,* vol. 63, pp. 384-399, 1965.

[16] R. Thomas, *The (Old) Farmer's Almanack*. Boston: Boston: E.G.House, 1809.

[17] S. E. Page, "The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies," *Princeton University Press,* 2007.

[18] A. McLennan, "Consequences of the Condorcet Jury Theorem for Beneficial Information Aggregation by Rational Agents," *American Political Science Review,* vol. 92, pp. 413-418, 1998.

[19] M. J. Mauboussin, "Explaining the Wisdom of Crowds," *Legg Mason Capital Management,* 2007.

[20] J. Wolfers and E. Zitzewitz, "Prediction Markets," *Journal of Economic Perspectives* vol. 18, pp. 107-126, 2004.

[21] T. King, "Does film criticism affect box office earnings? Evidence from movies released in the U.S. in 2003," *Journal of Cultural Economics,* vol. 31, pp. 171-186, 2007.

[22] I. Vessey, "Expertise in Debugging Computer Programs: An Analysis of the Content of Verbal Protocols," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-16, pp. 621-637, 1986.

[23] Z. Lee, C. Wagner, and H. K. Shin, "The effect of decision support system expertise on system use behavior and performance," *Information & Management,* vol. 45, pp. 349-358, 2008.

[24] D. C. Brabham, "Crowdsourcing as a model for problem solving", The International Journal of Research into New Media Technologies, Vol. 14, No. 1, 75-90 (2008). "Crowdsourcing as a model for problem solving," *The International Journal of Research into New Media Technologies,* vol. 14, pp. 75-90 2008.

[25] J. Giles, "Internet encyclopedias go head to head," *Nature,* vol. 438, pp. 900-901, 2005.
[26] R. Rosenzweig, "Can History be Opensource? Wikipedia and the Future of the Past " *The Journal of American History,* vol. 93, pp. 117-146, 2006.
[27] A. Kittur and R. E. Kraut, "Harnessing the wisdom of crowds in Wikipedia: Quality through coordination," in *Proc. 2008 Internat. ACM Conf. on Supporting Group Work* San Diego, CA, New York, 2008, pp. 37-46.
[28] F. B. Viégas, M. Wattenberg, and M. M. McKeon, *The hidden order of Wikipedia*: Berlin: Springer, 2007.
[29] Authors, "Governing Open Content Creation," 2009.