The Journal of Business Inquiry 2017, 17, Issue 1 (Special Issue), 55-71
http:www.uvu.edu/woodbury/jbi/articles
ISSN 2155-4072

When Are Crowdsourced Data Truthful, Accurate, and Representative?

By LIAM BRUNT AND ERIK MEIDELL*

We trace crowdsourcing, as a business strategy to gather information, to Britain in the Industrial Revolution, when it was used to create trade directories. We show that the trade directories' occupational snapshot was very highly correlated (≈0.99) with the 1851 census – a valuable objective metric of accuracy. Accuracy of modern crowdsourced data is more difficult to judge, but seems somewhat lower; we make an explicit comparison to Yelp. We rationalize our results by considering: construction of the sampling frame; incentives of the crowd to report correct information; disincentives to report incorrect information (cost of contributing, presence of "gatekeepers"); and sampling strategy.

Keywords: Crowdsourcing, Sampling, Census, Inference

JEL Classification: C81, C83, M55, N01

I. Introduction

Over the last decade, crowdsourcing has become a key strategy for gathering information. Online reviews of products and services present the most obvious example. Consumers can almost costlessly access firsthand information about any product that they want to buy. Typically, there are tens – and frequently hundreds or thousands – of customer reviews for virtually any product offered on Amazon, or the website of any major retailer. In fact, there are so many more reviews than anyone could feasibly read that they have to be aggregated into summary statistics: as well as star ratings, many websites provide average scores for durability, ease-of-use, value for money, and so on. While retailers offer reviews as a convenience for their customers – i.e., it is sideline to their main business – many websites now exist *only* based on crowdsourced information. An obvious example is TripAdvisor. Visitors check the site primarily to see other people's reviews of places that they themselves are considering visiting; TripAdvisor then makes money by selling advertising for associated products and services. The most extreme case is Wikipedia – a platform consisting entirely of crowdsourced content that makes no profit at all: it exists only to crowdsource.

But crowdsourcing is a key element in many other, less obvious, information collection mechanisms. For example, prediction markets essentially provide a platform for countless individuals to bet anonymously on the outcome of an event, such as the U.S. presidential election. The odds of each candidate winning are derived from these bets, and have proved remarkably accurate at forecasting the winner. In the prediction case, the crowd comes to the platform to share

* Liam Brunt, corresponding author, Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway. Email: liam.brunt@nhh.no. Erik Meidell, Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway. Email: erik.meidell@nhh.no. We would like to thank Lucy White and Jeffrey Williamson for their helpful comments, as well as seminar participants at the London School of Economics, the University of Bergen, and the Economic History Association Meetings in Boston. Any remaining errors are our responsibility.

information. In other cases, the platform actively seeks information from the crowd, such as firms using software to search social media for user sentiment (Evans, 2016). Crowdsourcing is frequently used in China to track people down, such as hit-and-run drivers: someone puts out a blurry photo of a car or an individual on SinaWeibo (the Chinese equivalent of Twitter) and within hours the perpetrator is usually unmasked (Simpson, 2014). The U.S. tried a similar approach after the 2013 Boston Marathon bombings, although that effort was less successful and actually wasted police time by generating a number of false leads (Wadhwa, 2013).

Note that there is a key difference between crowdsourcing to unmask criminals and crowdsourcing for book reviews. In the case of a crime, there is unquestionably a *right answer*. However, when 10,000 people rate a book or a tourist destination, they are not placing the item on an objective scale: if a book is not to one's taste then it may not be liked, even if other readers (with different tastes) gave it five stars. Even with similar preferences, it is not clear that everyone's scale maps directly to the scale of others; one person may give four stars while another gives five stars. Economists refer to this phenomenon as the problem of comparing "interpersonal utility". So, the most that can be said in considering these reviews is that "many readers liked it". In contrast, a certain person is going to win the U.S. election, and certain perpetrators carried out the Boston bombing: there is a clear benchmark against which to judge the truthfulness and overall accuracy of information.

If we are to rely on crowdsourcing to gather business information, then we need to be sure that the information is truthful, accurate, and representative. We can make this assessment only if we have an external metric against which we can evaluate it. An important element of our setting is that we have a clear, objective measure against which we can judge the accuracy of our crowdsourced data. First, we examine the occupational structure of England in 1851, as revealed by trade directories and the U.K. Government census; second, we examine the business structure of Norway in 2017, as revealed by Yelp and Norwegian Government establishment data. When we survey the literature in the next section, we will see that it is almost unique to have such a metric. Our empirical analysis will show that crowdsourced trade directories are remarkably accurate, particularly the historical ones. This then raises the question as to why? We argue that the answer lies in the information structure – how the crowd was tapped for information, how many crowd members reported on each individual fact, and how incorrect information was excluded. The next section gives more detail on the important variations in information structure that we see in crowdsourcing.

II. Crowdsourcing: Approaches to Information Gathering and Assessments of Accuracy

Suppose we use crowdsourcing to collect information about a well-defined aspect of the world. How accurate is this crowdsourced information likely to be? What are the incentives for different people – for example, informed versus uninformed – to participate? Even if everyone is informed, will the respondents be randomly drawn from the population? Are lovers or haters more motivated to give feedback on a product or service? It is hardly an exaggeration to say that crowdsourced information is changing the world. In fact, this is at the heart of the current concern about "fake news" (BBC News, 2016): many people now get their news via links that are shared (typically sourced from a Twitter or Facebook crowd) rather than the mainstream media, where facts have historically been more carefully checked. Even governments have adopted crowdsourcing as a *modus operandi*. For example, the U.S. government set out plans for a prediction market for terrorist attacks in order to try to get advance warning (Yeh, 2006), and

NOAA is testing software to take automated bathymetric readings from private vessels navigating U.S. waters (Reed, 2016).

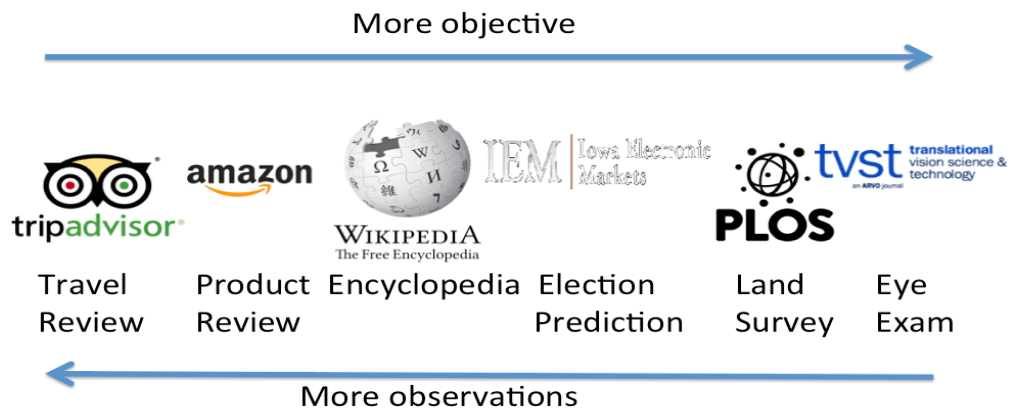
There has been some discussion of the issues that we raise here (Surowiecki, 2004). But remarkably little research has examined the process of crowdsourcing, how it might be structured, and the accuracy of the output. The exception is Wikipedia, where several studies have sampled articles and given them to experts to have their accuracy assessed; they have sometimes been compared to matched articles drawn from established reference works, such as *Encyclopedia Britannica,* in which all the articles are supposedly written by experts (Giles, 2005). The accuracy of Wikipedia is generally considered to be good. However, Wikipedia is not a typical example of crowdsourced information. Note that each article is parsed by many people who collaborate to refine it (i.e., a large crowd is asked to agree on one item). This is analogous to Galton's original discovery – which he himself found surprising – that many people estimating the weight of an ox at an agricultural show together get very close to its true weight (i.e., the average estimate is accurate, even though the individual estimates vary widely; see Galton, 1907). But in many crowdsourcing contexts, *one* member of the crowd is asked to provide *one* piece of information – like a mosaic tile – and this is placed next to others to build a picture of the overall situation. The statistical properties of this approach are clearly very different: there is no averaging effect at work. Many examples of this exact situation are found in the geography literature, where researchers have tried to use the presence of the crowd "in the field" to report local mapping information (Al-Bakri and Fairbairn, 2010) or have used volunteers to categorize land use based on a mosaic of aerial photographs (See *et al.*, 2013; Salk *et al.*, 2016). Accuracy is generally low (only 62 percent of photographs were correctly categorized in the 2013 study, and the correlation of the crowd with a sub-set of expert evaluations was low in the 2016 study). Importantly, accuracy is also inferior to traditional methods (the OpenSourceMaps were significantly less accurate than the Ordnance Survey equivalent in the 2010 study). In our historical case from the British Industrial Revolution, we have a combination of formats: there is a mosaic effect, in that data are collected on businesses located in different towns; but there is also a kind of averaging effect, in that the business list for each town is parsed by multiple members of the crowd (so individual errors – in particular, omissions due to ignorance – may be eradicated as the crowd becomes larger).

Better crowdsourcing results have been reported in medical studies. For example, a volunteer crowd proved no worse than experts at detecting severe eye abnormalities from retinal scans (Mitry *et al.*, 2016), although the crowd performed substantially worse with cases of mild damage (around 60 percent, depending on which measure is used). Importantly, we are not told the accuracy of either group (crowd or expert) compared to the actual clinical condition of the patients (i.e., there is no truly objective measure used in the study). Better results have also been reported with prediction markets, although these conclusions have been challenged. In particular, it has been claimed that prediction markets are superior to traditional polling techniques in forecasting the outcomes of presidential elections (Berg *et al.*, 2008). But this is true only if we compare the forecasts throughout the election campaign; if we compare prediction markets and polls on the eve of the election, then polls are better. Why would you wish to disregard earlier information? Because voting intentions may change through the election campaign. Since we have no objective measure of voters' intentions *before the election date*, we cannot assume that the prediction markets were more accurate than the polls before the election date. It could be the case that the polls were correct – and the prediction markets incorrect – at the time the polls were taken. Notably, neither of them was very accurate in the 2016 U.S. presidential election. This takes us back to the general problem: assessing the accuracy of crowdsourcing requires an objective metric

against which to compare it, and this is typically absent (as in the case of reviews) or prohibitively expensive to obtain (the output of the crowd may need to be somehow sampled by experts or a clinical analysis to gauge its accuracy). We overcome this problem in our study by comparing the crowdsourced data to the objective measure of the 1851 census, kindly prepared for us by the U.K. Government.

Some of these issues are summarized in Figure 1. We often have many observations of something that is not objectively verifiable (such as TripAdvisor telling us that a holiday destination is "five star"). There are also some instances where we have very few observations of something that is objectively verifiable (such as whether a photo displays symptoms of eye disease). But relying on very small numbers of observations – typically one – is not harnessing the power of the crowd: Galton's original insight was that averaging the crowd's estimates greatly increases accuracy, compared to relying on any single individual. Moreover, the settings in which the crowdsourced information is objectively verifiable have generally not effectively tested its accuracy against the available external metric (such as whether or not the patient really has eye disease). The only real test of crowdsourcing has been in the context of Wikipedia, where crowds have been used to parse every piece of information and where the facts can be checked against alternative information sources. The results for Wikipedia have been promising. But Wikipedia has another peculiarity of its information structure which we believe is crucial to its accuracy and which we will discuss in detail in the next section: "gatekeepers" (i.e., article editors) who can reject information that they know to be incorrect.

Figure 1: A Spectrum of Crowdsourcing Types



Finally, a natural assumption might be that effective crowdsourcing requires modern technology, such as the internet or mobile phones, because it reduces the cost of contributing. The historical account that we offer in this paper shows that such an assumption would be false. (Indeed, we will argue later that making it somewhat costly to contribute is a benefit because it discourages the contribution of incorrect information.) In the next section, we trace crowdsourcing back to the creation of trade directories in Britain in the 1790s, when people were working with paper and quill pens. We discuss how the crowd was tapped for information, and why this was likely to result in accurate data; we contrast this with government efforts to collect information. In the succeeding section, we compare the occupational structure represented in trade directories to "hard" information from the 1851 occupational census, and show that the data are highly correlated at the town level. Our final section draws out the key lessons from our story.

III. The Creation of Trade Directories

Samuel Lee prepared the first British trade directory in 1677, but the entries covered only 1,953 wholesale merchants living in London (Goss, 1932). It seems to have met with limited success, since the exercise was not repeated until Henry Kent produced a new directory in 1734. Kent followed the same format as Lee but included 693 fewer names – so either London had shrunk or Kent's directory was very incomplete, the latter seeming more plausible. Coverage seems to have improved over the first few editions (up to 2,006 entries in 1740), but Kent's ambitions remained very limited in his subsequent annual revisions. Osborn's London directory first appeared in 1740 and offered a wider range of information, but was seemingly still very incomplete. The bar was finally raised in 1763 with the appearance of Mortimer's *Universal Directory*. He included not only the merchants and bankers of London but also people in other trades and professions: artists, musicians, doctors, lawyers, booksellers, shopkeepers, and so on. By the early nineteenth century, the *Post Office London Directory*, which first appeared in 1800, contained around 11,000 entries; and Johnstone's 1817 directory was up to 27,000.

Importantly, Sketchley produced a directory for Birmingham in 1763 – the first for a town outside London (Norton, 1950). The first two editions of Sketchley's directory have not survived, but the third edition (1767) has a format very similar to Mortimer's *Universal Directory* for London. Directories soon appeared for many other towns around England and up to 50 new directories were produced between 1763 and 1790. These covered ten towns, and some also attempted to cover larger areas, with county directories appearing for Hampshire (1784) and Bedfordshire (1785). William Bailey, in 1784, was the first to attempt a national directory that covered the principal towns throughout the kingdom. Wilke's *Universal British Directory*, which appeared in eight volumes between 1791 and 1798, raised the bar again by including many smaller towns.

In the early nineteenth century, town and county directories became common. In total, Norton's exhaustive survey (1950) counts 878 provincial (i.e., non-London) directories published before 1856. Many of these directories are readily available in electronic format because they are of interest to genealogists; therefore, they constitute one of the most accessible historical sources. Over time, directories became more thorough and complete and were produced to a higher standard. Famous names – such as Pigot's and White's – started to appear in the 1810s; they set out to cover the whole country both systematically and repeatedly. Repetition is a key ingredient in generating a worthwhile data source. First, it may enable us to trace changes over time using a consistent source. Second, it probably generates a more accurate directory. How does repetition increase accuracy? First, the directory producer had an extra incentive to ensure that his directory was accurate because he had a reputation to maintain to generate future sales. Second, he had experience in producing directories and thereby a better idea of how to elicit accurate information (as we discuss further below). Third, the directory producer already had local knowledge when preparing his directory (i.e., the data base generated by the previous edition).

The issue of accuracy is, of course, crucial. First, consider what we mean by accuracy. It is obviously not the case that the entire population was listed in a trade directory. Poor people would not have been listed; nor would many better off people who were not involved in trade (for example, retired people or military officers or noblemen). In fact, it is highly unlikely that even all the traders were recorded. There may be systematic omissions – such as dung collectors, who might not have wanted to advertise their trade – as well as random omissions and errors. In that sense, the directories are incomplete. But this does not make the directories useless. If we want to

track business development over time, or map variations across the country, then we do not necessarily need a complete register of all traders and producers. What we desire is transparency and, preferably, consistency. If we know the likely sources of error – so that we can correct them – or if we know that they remained constant over time, then we may be able to say something worthwhile about changes or variation in business structure.

So how did directory producers compile their data? Several approaches seem to have been adopted (Norton, 1950). Early producers, such as Bailey and Pye, claim to have visited every house in the locality to elicit information from the householder. Pye, in fact, states that he gave up this approach in his later directories because it was too expensive. It may also have been counterproductive because people knocking unexpectedly at the door and asking about the nature of the householder's business might be suspected of being tax collectors – and therefore lied to, or told to go away. In any case, personal interviews could not have been a practical mode of compiling county or national directories because the task was simply too vast for a private entrepreneur. Thus, it became common to use local agents to collect information. For his *Universal British Directory* – which remained the most ambitious directory undertaking for several decades – Wilkes first enlisted local printers and booksellers as his agents. This was a natural step, given that he must have had contacts in the publishing world; in a moment, we discuss the merits of this approach, in terms of information accuracy and completeness. Wilkes then crowdsourced in order to improve the quality of the directory further. A draft of the local directory was left with a prominent resident of the local town and people were invited to inspect and correct it. Of course, as well as being a way of collecting information, this was also a form of advertising: people would be aware that the directory was going to appear, and might even be more likely to buy it because they had had a hand in preparing it.

Wilkes' approach provided several incentives for agents to furnish accurate information. The local printers and booksellers that Wilkes recruited were remunerated in the form of offprints for local sale, so they had a stake in generating an accurate and complete product. Logically, the first thing that a potential purchaser would examine to gauge the quality of a national directory would be his own town: if the local entries were accurate and complete, then he might be willing to believe that the rest of the directory was of similarly high quality; if the local entries were no good, then it would be difficult for the local bookseller to persuade the customer that the other entries were better. Thus, each local bookseller was likely to be able to retail his free offprints of the national directory only if he did a good job of collecting the data in his own town. We can think of Wilkes and the booksellers as "frame makers": Wilkes constructed the sampling frame (in a statistical sense) by choosing which towns to include in the directory; and the booksellers created a basic framework for each town, which could then have layers of information added to it by the townsfolk.

Now consider the actions of the crowd. When the draft was opened for correction by the townsfolk, the traders and professional people had an obvious incentive to ensure that the information about them was accurate and up to date – just as they have an obvious incentive today to check their name in credit registries (such as Experian) to make sure that no erroneous record is driving away customers. Not only might appearing in the directory attract business from out of town, but one could also imagine that there was a certain cachet derived from being in the directory. The same tactic is used by *Who's Who in Academia* – they persistently write to academic staff and ask them to complete a form with biographical details in order that they appear accurately in the next edition (which they can then buy for a special discount, of course). Friends, family, and business associates would also have an incentive to ensure that each business was correctly

recorded, since they might benefit from any additional income. One can think of this tactic as using the crowd to minimize Type I errors – that is, erroneously rejecting (or omitting) a correct piece of information. Then we have the local "prominent person", who acted like a Wikipedia editor. He would have had a fair idea of who was in business in his neighborhood and this would have discouraged people from adducing false information – such as crossing out the names of competing businessmen on the basis that they had "gone to Texas" when really, they had not; or else writing that they themselves were a "cotton manufacturer and banker", when they were only a cotton manufacturer, in order to make themselves look more reliable. It is worth noting that all businesses in this period were sole proprietorships or small partnerships because joint stock companies were outlawed: thus all businesses traded under a personal name and were not anonymous in the way that modern businesses are. One can think of this tactic as using gatekeepers to minimize Type II errors – that is, erroneously accepting an incorrect piece of information.

We can contrast Wilkes' data collection approach with that of the government. When the first British census was undertaken in 1801, the Overseers of the Poor were employed as enumerators. They obviously had the advantage of local knowledge (albeit disproportionately of the poorest households); and they had the disadvantage of unpopularity. Moreover, people were always concerned that the government was collecting information for tax purposes. Therefore, England did not take an agricultural census until 1866 – whereas it started in France in 1840, for example – and even then, it contained data only on inputs, such as land and animals; data on outputs began to be collected only in 1885. So, it seems plausible that some people, at least, avoided the census enumerators and gave them the least amount of information possible. In fact, the earliest population censuses were restricted almost entirely to questions on the number and sex of household members. The censuses additionally report numbers of people "chiefly engaged in agriculture", "manufactures", and "otherwise". But these data are essentially worthless. For one thing, the data were recorded at the level of the household, not the individual, which immediately raises the question of what the household head reported when there were multiple people working in different sectors. Since many household heads had multiple occupations themselves – such as agricultural worker and carter – it is not even obvious how they reported their own chief occupation, let alone those of their wives and children. The occupational data are better in 1841, but really become usable only in 1851 (as we discuss below). Using the Overseers of the Poor as enumerators also had the disadvantage that the agents were not well trained – hence there seems to have been some confusion about exactly who was to be recorded and how (Higgs, 2005). Moreover, they did not have particularly strong incentives to be thorough because no one was willing or able to check their fieldwork. Finally, the enumerators had to do all their work on one night of the year, so they were in a big rush compared to the crowd and you might imagine that their returns would be incomplete (for example, if no one answered the door) or inaccurate (if someone was vague about their occupation). In the later censuses, such as 1851, there was an effort made to recruit and train specialized census takers, as used in modern U.K. or the U.S. censuses.

So, by comparison to Wilkes' approach, the census actually uses fewer people (only the enumerators, not a broad body of citizenry); with a lower level of knowledge (being experts only on the poor, not on their own and their neighbors' businesses); and worse incentives for accuracy (having nothing to gain personally from increased rigor); contributing information on more units of observation (every household, not just every business); in a shorter amount of time (just one night, rather than over a period of time). It seems plausible that the census could even be less accurate than the trade directory under such conditions. In fact, this problem is still a hot topic in the U.S. (Sullivan, 2009). The U.S. Census Bureau would like to use sampling in certain areas to

estimate the population because they believe that it is more accurate to survey some areas very intensively and then reflate the survey data than to ask their enumerators to try to make an actual count of everyone (including the homeless and illegal immigrants and others who actively avoid authority figures). The Republican Party opposes this move precisely because it would lead to higher estimates of the number of poor people, which would affect the costs of government relief programs and so on.

IV. Comparing Trade Directories to the Census

If we can establish the representativeness and the accuracy of trade directories, then we can establish the effectiveness of crowdsourcing. One line of attack is to examine how closely the occupational structure recorded in trade directories maps to the occupational structure reported in the census. Note that we are testing a joint hypothesis here: that the trade directories are both representative and accurate. Absence of a mapping may be due to *either* unrepresentativeness *or* inaccuracy (or both); if we find no correlation then we cannot be sure which part(s) of the hypothesis is (are) rejected. But if we reject the alternative hypothesis (i.e., that there is no correlation), then we can be sure that the two requirements (representativeness and accuracy) are both met.

Comparing trade directories to the census is difficult for several reasons. First, trade directories report the number of businesses operating in each occupation in each town, whereas the census reports the number of workers employed. We therefore need to divide the total number of people in each occupation by the average number of employees per business (in that occupation) to infer the number of businesses in each occupation. This generates a sort of national trade directory for Great Britain (albeit a trade directory with the street addresses and names of the businesses removed). Census data are broken down by county and by major town, which enables us to match the data to many town-level trade directories.

Second, the quality of the occupational data collected in the census was very poor up to 1841, so if there were a low correlation with the trade directories then we would not be able to tell whether this was due to the low quality of the directories or the low quality of the census. By contrast, the Registrar General devoted an enormous amount of effort to systematizing the collection of occupational data in 1851, and it really represents a high point in the collection of occupational data (i.e., the data became coarser in subsequent censuses). A huge amount of groundwork had been laid, in terms of preparing and categorizing a list of 1,089 occupations that covered all the major employments of the nation (British Government, *Census of Great Britain, 1851: Population Tables II*, vol. 1, lxix-ci). We therefore take 1851 as our benchmark date for comparison to the trade directories. This has the additional advantage that the 1851 census contains a table of employees per business (British Government, *Census of Great Britain, 1851: Population Tables II*, vol. 1, cclxxvi-cclxxix), broken down by occupation, which we need to convert the numbers of workers reported in the census into the number of businesses.

Of course, the procedure turns out to be more complicated than this. First, the 1851 table of employees per business enumerates only those businessmen ("Masters") who have more than zero employees ("Journeymen and Apprentices"). So, we must infer how many businessmen there were who had zero employees. In principle, this is straightforward because, for each occupation, the table reports the number of employers having a number of workers. If we were to multiply all the employers in an occupation by the number of workers that each of them employed, then we should get the total number of people working in that occupation *except those businessmen who employed*

zero. We could then compare this number to the total number of people recorded in the census as having that occupation. Any difference should (in theory) be composed of businessmen who had zero employees. The first problem with this exercise is that the number of employees is given only within certain bounds (1, 2, 3,… 10-19, 20-29,… 50- 74,… 75-100,… 350 and over). We address this problem by assuming that – on average – each firm was located mid-way between its particular set of bounds. For example, we assume that firms in the 10-19 category employed 15 workers; this is the most plausible assumption and – in expectation – will minimize the magnitude of any error.

The second problem is that most occupations have a very large discrepancy between the two estimates of total workers (i.e., the estimated number of workers employed is much lower than that enumerated in the census). This implies that many occupations had an implausibly high frequency of businessmen who employed zero workers. For example, in order to reconcile the two estimates of the number of people working as bakers, it would have to be the case that 75 percent of bakers employed no workers. It is possible that 75 percent of bakers employed no help, but it is not the most plausible suggestion. The census therefore seems to be internally inconsistent. An explanation for such inconsistency is offered on p. cclxxvi of the 1851 census itself. Many employers neglected to complete the part of the form asking about the number of their employees. This would lead us to incorrectly assume that all the missing bakers (who were not recorded as employees) were sole proprietors with no employees. This would lead us to overestimate the total number of bakery *businesses* in Great Britain. For example, if a baker employed three people but neglected to note this in his census return, then those three people would end up be counted as three one-man bakery businesses in our calculations. This could make it impossible for us to match the census with trade directories accurately.

We could therefore make one of two extreme assumptions. Either all the missing people in an occupation were one-man businesses; or all the businesses in that particular occupation employed people in the same size distribution that we observe in the table (i.e., for those firms that completed the form). This would be correct if some employers randomly neglected to complete that part of the census return. Logically, the truth will lie somewhere between these two extreme assumptions (i.e., there were actually some Masters who had zero employees and there some who neglected to fill in the form). We made all the calculations that follow using both alternative, extreme assumptions and found that it made no significant difference to our results. How can this be? It is because we are concerned only with the *distribution* of businesses across occupations. If the employers in all trades were equally likely to ignore the part of the form dealing with the number of employees (for example, suppose that 50 percent of all employers failed to complete it), then this will have very little effect on the estimated *distribution* of businesses.

If we make either of these assumptions, then does the census generate an estimate of the business structure that is consistent with the trade directories? Does it suggest that the crowdsourced trade directories are accurate and representative? The census does not report occupational data for every English town, but we can look at a sample of individual towns to shed light on the issue. We downloaded the Chadwyck-Healey pdf version of the 1851 census and matched every town that was reported there against all the available trade directories produced in the years around 1851. This gave us a sample encompassing Whitehaven (Cumberland), Gateshead (Durham), Boston and Lincoln (Lincolnshire), Newark-on-Trent (Nottinghamshire), Kingston-upon-Hull (East Yorkshire), and Leeds (West Yorkshire). We made the calculations described above (based on each of the alternative assumptions) and then compared the total number of businesses estimated from the census to the total number of businesses recorded in the trade

directories.[1] The number of businesses recorded in the trade directories was much smaller, showing conclusively that the directories do not offer an exhaustive list of businesses in operation.

However, we are really interested in the *distribution* of businesses across occupations. Were the distributions of businesses across occupations the same in the census and the trade directories? Yes, absolutely. How can we summarize their similarity in some type of descriptive statistic? Calculate the percentage of total businesses constituted by each occupation in both the census and the trade directory. That is, work out what percentage of businesses were bakeries, tailors, taverns, and so on. Now regress the trade directory distribution on the census distribution. What should you expect to find if the trade directory is a random sample of businesses in a particular town? Then a one percent larger share accruing to a particular occupation in the census will be reflected by a one percent larger share accruing to that occupation in the trade directory (i.e., the coefficient on the census data will be unity). So if bakeries and taverns comprised five percent and ten percent respectively of the population of businesses in a town, according to the census, then they should similarly comprise five percent and ten percent respectively of the businesses recorded in the trade directory.

Of course, to the extent that there is measurement error in the estimated occupational structure, the estimated coefficient in the regression will be biased downwards for standard statistical reasons. Hence, we expect to observe estimated coefficients that are less than unity but hopefully not statistically significantly different from it. If the overall distributions are quite similar, then the fit of the regression (the r-squared) will also be high. Note that some of the trade directories that we matched against the 1851 census were compiled several years after the census; we chose them simply because they were the closest years available. Such temporal mismatch would be expected to induce more measurement error and bias the results towards rejecting the hypothesis that the trade directories and the census exhibit the same occupational distribution. Note further that this need not generally be a problem with using trade directories. We are constrained here to find trade directories as close as possible to 1851 because we are undertaking a direct test against the census. If we were given a free choice of year, and were simply trying to assemble a set of trade directories that gave a good coverage, then there would be less temporal mismatch.

We undertook the regression exercise for our sample of towns and found that the distributions of the census and trade directories were very similar for each town, and the coefficient on the census was not significantly different from unity. We report these regressions in Table 1.

[1] A small number of occupational terms used in the census were not used in the trade directory. For example, no business is listed as a "Fustian manufacturer"; since fustian was a type of fine cotton cloth, those businesses were presumably listed as "Cotton manufacturer". The same is true of "Thread manufacturer" and "Calico and cotton printer". We, therefore, aggregated workers in those industries (as reported in the 1851 census) with cotton manufacturers and calculated one multiplier for all branches of the cotton industry that we applied to each of its components (cotton, fustian, thread, and printing). For "Weaver (material not stated)" we took the multiplier to be the average of cotton, flax, and woolen manufacturers. For "Skinner" we took the multiplier to be the average of other occupations in the sub-class (which were all very similar); and the same for "Fuller".

Table 1: Regressing Trade Directory Occupational Shares on Those of the Census, c. 1851

	Coefficient	*95% confidence interval*	r^2	*N*
Greater Birmingham	0.86	0.75 – 0.97	0.71	97
Boston	0.95	0.79 – 1.10	0.70	64
Gateshead	0.91	0.75 – 1.08	0.66	61
Kingston Upon Hull	0.85	0.70 – 1.00	0.65	70
Leeds	0.92	0.82 – 1.03	0.79	82
Lincoln	1.01	0.86 – 1.15	0.73	72
Newark	1.00	0.83 – 1.16	0.71	60
Whitehaven	0.93	0.75 – 1.12	0.57	76
Pooled Sample	0.99	0.90 – 1.09	0.78	119

Notes: We exclude all occupations for which there are zero workers and all occupations for which there is no multiplier available from the table of employees per business. We aggregated "Builders" with "Mason (pavior)" and "Bricklayer"; we excluded "Merchants" because the multiplier in the 1851 table of employees per business is based on only three observations in the entire country; and we excluded the top five and bottom five occupations (in terms of their distance from the occupational share reported in the census) in each town. Our rationale for the last step was that there were a small number of very large outliers that were drastically and randomly skewing the results, and most of these outliers were obviously problematic. For example, "Coal miners" seem to be massively underreported in the trade directories, compared to the census. But this is easily understood when we see that the table of employees per business reports an average of 49 miners per coal mine, which must surely be a drastic underestimate. In general, it was more or less the same 10 occupations that were problematic in each of the towns (notably, "Straw hat and bonnet maker", "Woollen cloth manufacture", "Flax, linen manufacture", "Coal merchant, dealer", "Shopkeeper (branch undefined)" and "Hosier, haberdasher"). The number of observations differs for each regression simply because some towns have more occupations than others.

These results suggest that there is a strong mapping between the business structure revealed by the 1851 census and that reported in contemporary trade directories. This implies that crowdsourcing, when combined with gatekeeping, is an effective way to elicit accurate and representative information – even in a setting with the most rudimentary information technology. We believe that these results offer a satisfactory "proof of principle" of the utility of crowdsourcing. However, the devil may well be in the details and in the final section, where we wrap up, we highlight some key elements.

V. Comparing the Yelp Directory to Government Establishment Data

Yelp is the modern equivalent of the old, paper trade directories. It is obviously a business directory, but we will see shortly that the similarities to the old directories run much deeper than that. We have made an in-depth study of Yelp in Norway and the following discussion is accurate for that market; but the details of Yelp directory construction almost certainly vary across markets

– in response to local laws and data sources – and so we would not want to claim that our characterization is necessarily accurate for all countries. Why choose Norway? In addition to the fact that we are particularly familiar with that market, the Norwegian government is unusually open with microeconomic data pertaining to publicly identifiable units of observation (such as the tax returns of both private individuals and businesses, which are all public information). Amongst the vast ocean of data that the Norwegian government collects – and posts online – is a complete register of all Norwegian businesses. This is crucial for us because it provides an objective metric against which we can judge the representativeness of businesses listed in Yelp. This exercise would not be possible in the U.K. or the U.S., for example, where such a centralized database does not exist or is inaccessible.[2]

Norway is also a nice setting because it is comparable to our historical example in several other dimensions. First, the economies are of similar size – there having been 18 million people in England in 1851 and 5 million people in Norway in 2017. Second, the typical scale of enterprise is very small: 82 percent of Norwegian establishments had fewer than five employees in 2017; in England in 1851, upwards of 44 percent of establishments had fewer than five employees.[3] This is important because you might imagine that large and small firms would have different propensities to list themselves in trade directories. Third, the way that Yelp is compiled is similar to the English historical trade directories. Yelp posts pages for hundreds of thousands of businesses but not all those listings are active. When you find the Yelp page for a business that you know exists, it is often merely a stub and there is no information given except the name and address. It is up to the business owner to claim the listing and then activate it – in the same way that a businessman in 1851 could edit his entry in the draft trade directory in order to add his address and line(s) of business. Fourth, and very importantly, Yelp staff act as gatekeepers: they manually correct information that they believe to be wrong and they can block changes to prevent the infiltration of incorrect information (rather like Wikipedia page editors). In fact, business users sometimes complain that the gatekeepers are too strict in preventing alterations (Kevin, 2012). Fifth, in the case of Yelp, activating the page additionally allows users to post reviews of the business. Yelp then uses artificial intelligence to infer lines of business from customer reviews, thereby using crowdsourced data to adjust for the possibility that owners' classifications may be absent, incomplete, or inaccurate (Tung, 2015). Yelp classifies enterprises into approximately 1,000 different business lines, whereas the 1851 census used a list of 1,089 occupations.

Yelp's business reviews have been a controversial topic (Clark, 2013), particularly the problem of fake reviews. There may be fake positive reviews (primarily business owners posting reviews of themselves, either directly or via employees and relatives); or fake negative reviews (either from people with a personal vendetta against the owner, or from people trying to extort "compensation" – which may or may not be merited – in the form of goods or services). Posting

---

[2] Although there are accessible, official databases of businesses – for example, the one maintained by Companies House in the U.K. – they do not list all enterprises. For example, Companies House tracks only limited companies (whereas most U.K. enterprises would take the form of sole proprietorships or partnerships). This creates obvious sample selection problems, since enterprises of different sizes and sectors tend to choose different business forms.

[3] The 1851 census gives the number of people employed by "Masters" in around 100 different lines of business. This enables us to calculate the percentage of firms in each businesses line having 2, 3, or 4 employees; we then weight these percentages by the frequency of these lines of business (as reported in the contemporary trade directories) to get our overall estimate of 44 percent. However, note that the 1851 census does not tell us how many Masters had 0 employee (i.e., the establishment had only 1 worker in total – the Master himself), so this 44 percent is a lower-bound figure on the total percentage of business having fewer than 5 workers. It is likely that a high percentage of Masters employed no helpers, so a sensible guess for the total figure could well be around 64 percent.

fake positive reviews is known as "astroturfing". Yelp uses algorithms to try to detect and exclude such reviews, although it is nonetheless estimated that around 20 percent of Yelp reviews are fake (Luca and Zervas, 2016). Reviews tagged by Yelp's filtering algorithm are "parked" and not automatically displayed; Yelp users can choose to view them if they wish, but they are still not used when Yelp calculates its star ratings for each business. Importantly, note that the prevalence of fake reviews need not imply that business ratings are biased, even if the fake reviews were to be included in the calculation of star ratings. Fake reviews tend to be either very positive or very negative, thus making the tails of the review distribution fatter, but the mean could remain unchanged.

Of course, the problem of fake reviews – or news – is by no means limited to Yelp. It is known in the political or cultural arena as "opinion spamming": a highly-motivated minority bombards public bulletin boards with messages favoring a particular candidate or viewpoint – typically concealing their true identity by using multiple aliases – in order to try to lead public opinion in a certain direction (Jindal and Liu, 2008). In auctions, it is known as "shilling": bidders in the pay of sellers enter fake bids to force up the price of an object being offered for sale (Grether *et al*., 2015). It would be perfect if we could find an objective metric of business quality to which we could compare Yelp's star ratings to see just how accurate these crowdsourced review data are. Unfortunately, no one has yet managed to find such a quality metric. What we can do, however, is compare the distribution of businesses on Yelp to the actual distribution of businesses – as revealed by Norwegian government records – to infer whether Yelp at least accurately reflects the pattern of economic activity.

We downloaded the entire database of Norwegian establishments ("virksomheter"), which has a total population of 565,054 (Statistics Norway, 2017). An establishment is defined as "a local kind of activity unit, which mainly conducts activities within a specific industry group". They are classified into 100 different industry groups, from 00 ("Unknown") to 99 ("International organizations and bodies"). We also downloaded the entire Yelp database of Norwegian businesses having an activated page, which is effectively a sample containing 128,011 observations in total. We classified the Yelp data on the same basis as the government establishment data. The interesting question is whether the Yelp sample provides an accurate representation of the Norwegian population. So we proceeded as before, first calculating the percentage of total establishments operating in each of 100 industry groups. We then regressed the percentage reported in Yelp on the percentage reported to the government. If the Yelp sample were truly random, then the coefficient should be unity (a 1 percentage point increase in business frequency in Yelp should map to a 1 percentage point increase in business frequency in the government data) and the intercept should be zero. The basic fit is good: an intercept of zero, a coefficient of 0.86 (±0.17, so not significantly different from unity) and an r-squared of 52 percent.

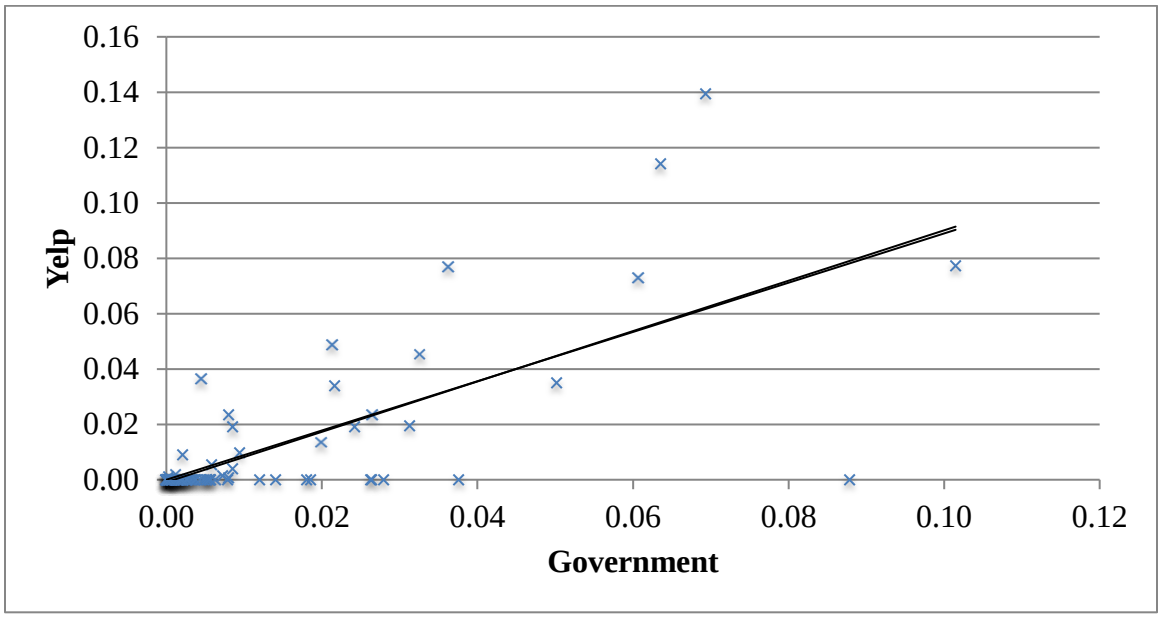
Figure 2 reveals that there is one very big outlier on the lower right of the graph: the category "Crop and animal production, hunting, and related service activities" constitutes 9 percent of Norwegian enterprises but 0 percent of Yelp businesses. The majority of these enterprises would be family farmers: Norwegian agriculture is characterized by smallholders cultivating a few acres and keeping small numbers of animals. We would not generally expect farmers to be listed in Yelp, so it seems reasonable to exclude that category (and there are no farmers in our 1851 data, so it makes a cleaner comparison). Doing so raises the estimated coefficient to 1.04 (±0.16) and the r-squared to 66 percent. Explaining 66 percent of the variation is respectable, though still inferior to our English results for 1851. This is a little surprising because the English estimation involves an extra step: we multiply the number of firms in each business line by the number of workers per

firm in that business line to get an estimate of the occupational structure. We then compare this to the distribution of occupations in the labor census, rather than comparing the trade directories directly to the business census, which we are doing with the Norwegian data. You might expect the extra step to add noise and reduce the r-squared, but it does not seem to do so (or else the Yelp business data are just noisier than the English business data).

Even though Yelp constitutes only a 23 percent sample of Norwegian enterprises (=128,011/565,054) it seems to offer a surprisingly accurate reflection of the distribution of enterprises across business categories (except agriculture). This is consistent with Yelp's own analysis. Their data research team compared the accuracy of Yelp listings to those of competitor sites (such as Google and TripAdvisor) using a hand-collected sample of 1,000 businesses from the U.S. and U.K. (Jason, 2013). Hand-collecting data is obviously time-consuming and expensive: it offers the advantage of very high accuracy but the disadvantage of very small numbers. But if you are trying to judge accuracy against an absolute standard (for example, whether the address, phone number, and website are truly correct) then it is the best strategy. The Yelp team found that their data accuracy was comparable to Google but superior to TripAdvisor and others. We would suggest that Yelp's gatekeeping activity was a crucial component of this success, avoiding the introduction of false information.

In the future, it might be possible to extract firm-specific data (such as the financials, which are publicly available) and take our analysis further by linking them to the Yelp star ratings. It would be interesting, and important, to see whether the crowdsourced star ratings are as accurate as the business categorization, when compared to an objective metric. However, this lies beyond the scope of the current paper.

Figure 2: The Distribution of Norwegian Firms Across Sectors: Government vs. Yelp Data



VI. Discussion and Conclusion

To the best of our knowledge, trade directories represent the first systematic attempt to search for specific information by tapping knowledge embodied in the crowd. It was distinctively different from a census – which was, of course, undertaken in Judaea at least 2,000 years ago – because participation was not compulsory and the information sought did not necessarily pertain to the individual who was reporting it. Our historical scenario shares many key characteristics with modern crowdsourcing – such as the fact that it was a commercial undertaking (and hence participation was voluntary), that accuracy was important, and that the entrepreneur was building up a mosaic of data.

Analysis of both historical and modern data suggests that there is a very tight mapping from crowdsourced (sampled) data to government (population) data – that is, from trade directories to the census. But this may not be a general result for crowdsourced data. The compilers of trade directories have structured their search in clever ways to elicit a broad contribution of accurate information. Businessmen have an incentive to include truthful information about themselves, and gatekeepers have been on hand to discourage the contribution of false information. Contributors have been working within a framework previously formulated by the directory creators (Yelp in the modern setting, Wilkes and the local printers and publishers in the historical setting). The overall structure of the information elicitation scheme is similar to Wikipedia – accepting contributions from the largest possible crowd and then having gatekeepers weed out bad information. Importantly, each piece of information is parsed by multiple members of the crowd, so individual errors are likely to be eliminated (more like Galton, less like researchers who rely on only one member of the crowd to categorize data). The modern and historical directories both seem to accurately reflect the structure of economic activity. However, the reliability of Yelp's more advanced functions – particularly its review and rating system – remains an open issue.

It may seem surprising that crowdsourcing was feasible before the internet age. The cost of contributing was higher because you had to go to the location in person to adjust the record with a pen. Of course, one aspect of our historical setting is that the information collected was local (people were offering information about themselves and their neighbors), which kept the contribution cost low (no one had to travel a great distance to contribute). But, in fact, the non-zero cost of contributing may well have been an advantage: it is plausible that people are less likely to volunteer false or inaccurate information when it is costly to do so. You might write a fake Yelp review from the comfort of your sofa, but you are less likely to bother if you must walk to the other end of town to do it, and then have to hand it to someone who may notice that it is fake. Trade directories demonstrate that crowdsourcing can be an effective way of collecting a vast amount of accurate information. But the design of the information elicitation scheme is likely to prove crucial and there can be no general presumption that crowdsourced data are accurate, truthful, or representative. Given the vast quantity of crowdsourced data becoming available, we need to think very carefully about what – if anything – we can reliably infer from it.

References

Al-Bakri, Maythm, and David Fairbairn. 2010. "Assessing the Accuracy of 'Crowdsourced' Data and its Integration with Official Spatial Data Sets." *Accuracy 2010 Symposium*. July 20-23.

BBC News. 2016. "The Saga of 'Pizzagate': The Fake Story that Shows How Conspiracy Theories Spread." BBC News, December 2, 2016. http://www.bbc.com/news/blogs-trending-38156985 (accessed December 2, 2016).

Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz. 2008. "Results From a Dozen Years of Election Futures Market Research." In *Handbook of Experimental Economic Results*, ed. Charles R. Plott and Vernon L. Smith, Volume 1, Chapter 80, 742-51. New York: North Holland.

Clark, Patrick. 2013. "Yelp's Newest Weapon Against Fake Reviews: Lawsuits." *Bloomberg BusinessWeek*, September 9.

Evans, Patrick. 2016. "Can Social Media be Used to Predict Election Results?" BBC News, November 10. www.bbc.com/news/election-us-2016-37942842 (accessed November 10, 2016).

Galton, Francis. 1907. "Vox Populi." *Nature*, 75: 450-1.

Giles, Jim. 2005. "Internet Encyclopedias Go Head to Head: Jimmy Wales' Wikipedia Comes Close to Britannica in Terms of the Accuracy of its Science Entries." *Nature*, 438 (7070): 900-1.

Goss, Charles William Frederick. 1932. *The London Directories, 1677-1855: A Bibliography with Notes on their Origin and Development.* London: D. Archer.

Grether, David, David Porter, and Matthew Shum. 2015. "Cyber-Shilling in Automobile Auctions: Evidence From a Field Experiment." *American Economic Journal: Microeconomics* 7(3): 85–103.

Higgs, Edward. 2005. *Making Sense of the Census Revisited.* London: Institute for Historical Research.

Jason. 2013. "Data Quality: How Yelp Stacks up to the Competition." https://engineeringblog.yelp.com/2013/11/data-quality-how-yelp-stacks-up-to-the-competition.html (accessed April 16, 2017).

Jindal, Nitin, and Bing Liu. 2008. "Opinion Spam and Analysis." *Proceedings of the ACM International Conference on Web Search and Data Mining*: 219-30.

Kevin B. 2012. "Incorrect Business Information, Locked in Place by Yelp." www.yelp.com/topic/roseville-incorrect-business-information-locked-in-place-by-yelp (accessed April 16, 2017).

Luca, Michael, and Georgios Zervas. 2016. "Fake it till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science*, 62(12): 3412-27.

Mitry Danny, Kris Zutis, Baljean Dhillon, Tunde Peto, Shabina Hayat, Kay-Tee Khaw, James. E. Morgan, Wendy. Moncur, EmanueleTrucco, and Paul J. Foster. 2016. "The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images." *Translational Vision Science and Technology*, 5(5): 6.

Norton, Jane E. 1950. *Guide to the National and Provincial Directories of England and Wales, Excluding London, Published Before 1856*. London: Royal Historical Society.

Reed, Lt. Adam. 2016. "Beta Test of Crowdsourced Bathymetry Holds Promise for Improving U.S. Nautical Charts," NOAA Office of Coast Survey, June 14. https://noaacoastsurvey.wordpress.com/category/crowdsourced-bathymetry (accessed November 8, 2016).

Salk, Carl F., Tobias Sturn, Linda See, Steffen Fritz, and Christoph Perger. 2016. "Assessing Quality of Volunteer Crowdsourcing Contributions: Lessons from the Cropland Capture Game." *International Journal of Digital Earth*, 9(4): 410-26.

See, Linda, Alexis Comber, Carl Salk, Steffen Fritz, Marijn van der Velde, Christoph Perger, Christian Schill, Ian McCallum, Florian Kraxner, and Michael Obersteiner. 2013. "Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts." *PLOS One*, 8(7): 1-11.

Simpson, Jack. 2014. "Chinese Social Media Users Track Down the Pet Owner Who Cruelly Chained Dog to Moving Car." *The Independent*, September 23.

Statistics Norway. 2017. "Establishments." https://www.ssb.no/en/virksomheter-foretak-og-regnskap/statistikker/bedrifter/aar/2017-01-20 (accessed April 12, 2017).

Sullivan, Amy. 2009. "Why the 2010 Census Stirs up Partisan Politics." *Time*, February 15. http://content.time.com/time/nation/article/0,8599,1879667,00.html (accessed December 9, 2016).

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Tung, N. 2015. "Automatically Categorizing Yelp Businesses." https://engineeringblog.yelp.com/amp/2015/09/automatically-categorizing-yelp-businesses.html (accessed April 16, 2017).

Yeh, Puong Fei. 2006. "Using Prediction Markets to Enhance US Intelligence Capabilities." *Studies in Intelligence*, 50(4): 1-12.

Wadhwa, Tarun. 2013. "Lessons from Crowdsourcing the Boston Bombing Investigation." *Forbes.com*, April 22 (accessed November 5, 2016).