Outsourcing labour to the cloud

Jonathan R. Corney*,
Carmen Torres-Sánchez and
A. Prasanna Jagadeesan

Department of Design, Manufacture and Engineering Management,
University of Strathclyde,
75 Montrose Street,
Glasgow G1 1XJ, UK
Email: jonathan.corney@strath.ac.uk
Email: carmen.torres@strath.ac.uk
Email: ananda.jagadeesan@strath.ac.uk
*Corresponding author

William C. Regli

Department of Computing Science,
Drexel University,
3201 Arch Street,
Philadelphia, PA 19104, USA
Email: regli@drexel.edu

Abstract: Various forms of open sourcing to the online population are establishing themselves as cheap, effective methods of getting work done. These have revolutionised the traditional methods for innovation and have contributed to the enrichment of the concept of 'open innovation'. To date, the literature concerning this emerging topic has been spread across a diverse number of media, disciplines and academic journals. This paper attempts for the first time to survey the emerging phenomenon of open outsourcing of work to the internet using 'cloud computing'. The paper describes the volunteer origins and recent commercialisation of this business service. It then surveys the current platforms, applications and academic literature. Based on this, a generic classification for crowdsourcing tasks and a number of performance metrics are proposed. After discussing strengths and limitations, the paper concludes with an agenda for academic research in this new area.

Keywords: crowdsourcing; open innovation; cloud computing; micro-outsourcing; creation; evaluation; organisation; nature of the crowd; manufacturing; geometric reasoning; measurement; analysis of results.

Biographical notes: Jonathan R. Corney is a Professor in the Department of Design Manufacture and Engineering Management (DMEM) at Strathclyde University in Glasgow, where he has been involved in the research of crowdsourcing applied to industrial manufacture. In the past, he investigated

various topics in mechanical CAD/CAM (i.e. 3D feature recognition, 3D content based retrieval, subdivision for layer manufacture and automated digital painting).

Carmen Torres-Sánchez is a Research Fellow in the Department of Design Manufacture and Engineering Management (DMEM) at Strathclyde University in Glasgow. Before that, she worked for the pan-European Confederation of Junior Enterprises, based in Brussels, and was involved in information management and passover of know-how in SMEs. Although her main interests are in bio-inspired systems, her contribution to that work with young entrepreneurs and start-ups continues today.

A. Prasanna Jagadeesan is a Research Fellow in the Department of Design Manufacture and Engineering Management (DMEM) at Strathclyde University in Glasgow and a Software Engineer at Shapespace Ltd where he develops a software product called 'Part Browser', which helps to search and browse 3D CAD models by shape. He works in artificial intelligence and robotics, with extensive commercial experience in C/C++, Builder C++ and Visual C++/C#.NET.

William C. Regli is a Professor in the Department of Computer Science in the College of Engineering at Drexel University. He pursues interdisciplinary research spanning a number of computer science subfields (artificial intelligence planning, knowledge-based systems, evolutionary computation, solid modelling and graphics, internet computing, databases and human–computer interaction) as well as several engineering disciplines (mechanical, electrical, civil and software engineering).

1 Introduction

There is growing evidence that the internet can be used to efficiently distribute work to a global work force at almost zero cost. Those labourers do not belong to a group, a corporation or a network and do not necessarily even communicate among themselves. Thanks to internet technology, they can access tasks, execute them, upload the results and receive various forms of payment using any web browser. 'Cloud computing' has made irrelevant both the physical location of those workers and where the resources are hosted. This is a labour market open 24/7, with a diverse workforce available to perform tasks quickly and cheaply. Crowdsourcing has the potential to revolutionise the way jobs requiring human judgement are performed by offering a 'virtual automation' of tasks that might appear simple for a human but extremely complicated for a computer to solve (e.g. 'is there a dog in this picture?'). With this reality in mind, crowdsourcing is emerging as a tool to enable 'open innovation' in firms that look to advance their technology or improve their products using external contributors. It offers flexibility and versatility to facilitate those collaboration approaches that open innovation currently utilised (Pisano and Verganti, 2008).

The term 'crowdsourcing' was defined by Jeff Howe in 2006 as 'the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call' (Howe, 2006). The two main ingredients for an activity to be considered 'crowdsourced'

are an open call and large number of labourers. Mistakenly, the crowdsourcing label is sometimes attributed to activities such as open-source production or collaborative learning by the online communities. These activities are executed by people who come together and self-organise to provide a meaningful participation. In this paper, we make a distinction between bottom-up or 'ad-hoc' ventures that form collaborative communities (where typically each makes an incremental contribution to an on-going task, e.g. programming) and crowdsourcing. For example, it is well known that the videogames market or open-source software (e.g. Linux, Ubuntu, etc.) have benefited from the work of a few enthusiasts that improved coding and add-ins with no expectations of getting rewards. Their satisfaction came from knowing they were contributing to the improvement of a piece of software, in which everyone worked collaboratively and none got paid. A dichotomy rises when the community's creation becomes business and is harnessed by a corporation in order to obtain a benefit. As Howe (2006) indicates, crowdsourcing applies a different strategy than those ruling such online communities. Commercial crowdsourced initiatives are clearly for-profit, top-bottom initiatives (i.e. open calls) where a single company not only owns but also sells the results that the crowd (i.e. the large network) generates. The user, by using a communication channel, becomes productive, a worker, a labourer. That medium can be any vehicle that allows the user to be connected to a network that distributes and harvests the results (e.g. Amazon's MTurk system, Figure 1, where tasks are called HITs). Typically, the medium is the internet and the PC the tool, but there are other initiatives where mobile phones are used (Eagle, 2009) by those without access to the internet or a broadband connection, e.g. African countries (Figure 2).

Figure 1 Schematic of the Amazon's MTurk system for crowdsourcing tasks (see online version for colours)
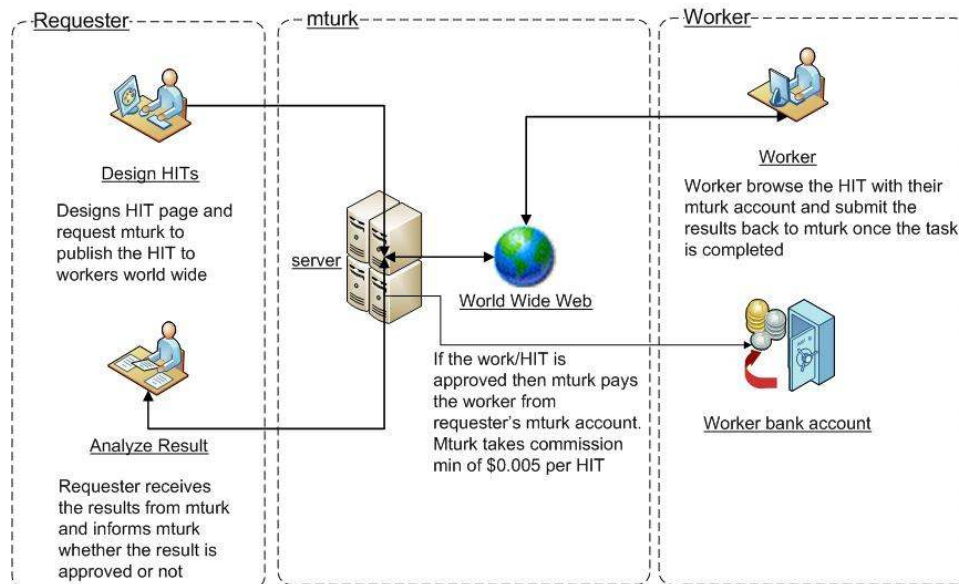
Figure 2 Levels of internet usage in the world (Wegen, 2008) (see online version for colours)
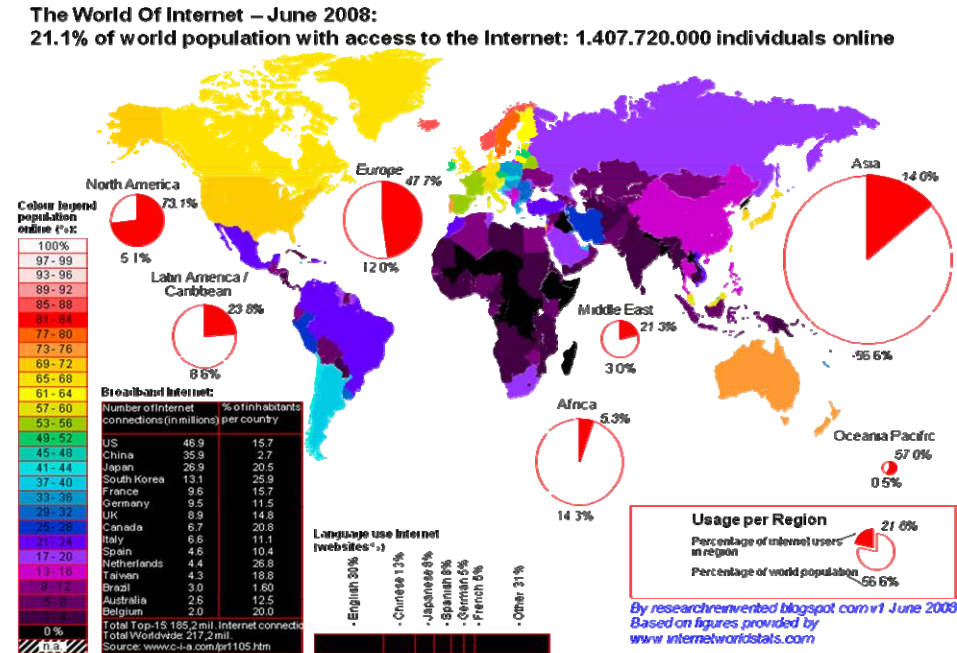
This paper is structured as follows: Section 2 proposes some classifications of crowdsourcing activities and case studies. Exemplars of some categories are then presented in Section 3 before the problems of assessing, measuring and quantifying the performance of crowdsourcing tasks (Section 4). Lastly, the discussion, in Section 5, presents an agenda for academic research in this area, and Section 6 draws some conclusions. A section that amalgamates notes with brief descriptions of the crowdsourcing activities mentioned in this paper has been added at the end.

2 The taxonomy of crowdsourcing

In many established disciplines, experience has allowed the classification of tasks, or problem types, so that the most appropriate methodologies or analysis methods can be easily identified. However, because of its recent origins, a classification of crowdsourcing tasks has not yet been established. Consequently, in this section we propose three possible categorisations based upon: nature of the task, nature of the crowd and nature of the payment. Later in the paper we discuss the value of these distinctions when considering analysis methods and site metrics.

2.1 Nature of the task

Perhaps the most obvious classification for crowdsourced work can be done on the basis of the nature of the tasks. Three main types of tasks used in crowdsourcing are as follows:

1   *Creation tasks*: This is probably the most lucrative of the three categories of a crowdsourcing activity. In it, corporations and companies align their efforts to harvest the crowd's intellect and obtain a solution to their problem. Clothing (e.g. Threadless[1]) and furniture (e.g. Muj[2]) design, advertisement and PR campaigns [e.g. Unilever (Sweney, 2009)] have been designed by anonymous labourers in a new spin of user-centred innovation. Brabham (2008b) claims that the 'desire to make money, develop individual skills and have fun' are the strongest motivators for participation in these sort of activities. Examples are not limited to form and appearance. Design problems have also been presented as games (DeOrio and Bertacco, 2009) and the obtained results used to automate decision-making process in electronic design.

2   *Evaluation tasks*: Market surveys and feedback calls fall under this category. The case study presented in Section 3.3 is an example of this type of task where workers evaluate the quality of each view. Similarly Dolores Labs' work on 'attractiveness' (DoloresLab, 2009a) (where the crowd is asked to assess the attractiveness of people in photographs) is another example of an evaluation task.

3   *Organisation tasks*: This category is probably one of the first applications of the crowdsourcing strategy in 'open innovation'. Image tagging (Google,[3] Galaxy Zoo,[4] etc.), websites rating (e.g. StumbleUpon, etc.), text recognition (e.g. reCAPTCHA[5]), spatial layout ordering (e.g. FoldIt[6] and packing) are examples of organising micro-tasks performed by the crowd. Transcription and translation tasks can also be listed under this heading. Since they tend to be tasks that are only a couple of sentences long, these can be performed in a few minutes and in an automatic fashion.

2.2   Nature of the crowd

It is also possible to categorise crowdsourcing tasks on the basis of the type of workers sought. On this basis, we have discerned three possible classifications (Figure 3).

1   *'Any individual' tasks*: Trivial tasks that anyone can do (e.g. image-tagging, identification of objects, shapes, etc.), or tasks seeking simple assessments based on human judgement (e.g. rating, relevance assessment, feedback on a service, etc.).

2   *'Most people' tasks*: These are tasks which most people can do (e.g. quantify a property such as functionality, readability, etc.). The final outcome is drawn from the aggregation of many individual responses. Typically the items to be quantified have ill defined properties that require judgement. For example, the crowd might be asked to assess the attractiveness of a new car profile on a scale of 1 to 10. The sought result is the combination of a few hundred responses. Tasks such as translation or proof-reading might fall into this class when a number of translations are compared and the results compiled. The best response will be the aggregation of all responses to the task.

3   *'Expert' tasks*, for which people with a unique ability, a specialisation or a specific skill are sought. Typically this assignment is a difficult problem, and due to its nature, it does not lend itself to aggregating a number of responses. In general, these are difficult problems such as protein folding (e.g. FoldIt[6]) or geometric packing problems (e.g. case study 3, Section 3.4) where some individuals exhibit an exceptional ability and computational approaches are NP-complete (i.e. have infinite search space that cannot be fully explored by even the most powerful computers).

Figure 3 Nature of the crowd (see online version for colours)
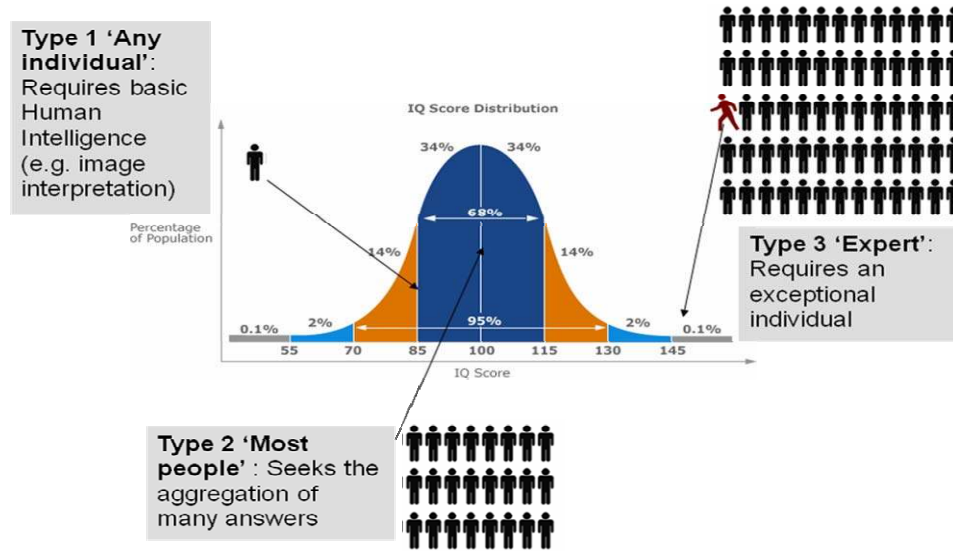
2.3 Nature of the payment

1 *Voluntary contribution*: e.g. clickworkers,[7] GalaxyZoo. The reward is only the satisfaction of having helped a social or humanitarian task (Powell, 2008) or some small amount of fame (e.g. the search for Steve Fosset's plane[8]).

2 *Rewarded contribution at a flat rate*: e.g. evaluation of Wikipedia articles (Kittur et al., 2008). All workers are paid a fixed amount for acceptable work.

3 *Rewarded contribution with a bonus or prize*: (e.g. Cisco Systems[9]). Based on performance, or simply the winners of a competition.

In the rewarded categories, payment can vary from cash to 'in kind' (tokens, etc.). One of the most best known commercial platforms for crowdsourcing is probably Amazon's MTurk (MechanicalTurk, 2009) who offer cash or gift vouchers, but there are others broadly used whose major form of payment is cash [e.g. 'Taskcn' in China (Yang et al., 2008), the freelance portals 'Donanza.com' in Israel and 'Humangrid.de' in Germany, to name just a few examples]. In contrast, 'txteagle' (Eagle, 2009) pays in airtime.

The payment made for a task, no matter how small, has a profound effect on the nature of these activities because the IP for the creation of a product, a classification or a market survey belongs to the company (unlike open-source production). However, even in the voluntary contribution scenario, companies can still benefit of the work done by the users. For users, crowdsourcing is an opportunity to have a democratic consumer participation, to take part in the process production of something they will later consume. For companies, crowdsourcing is an opportunity to reduce their risk in product innovation or the management of a new process. Several companies, for example CambrianHouse (2009) and DoloresLab (2009c), now offer integrated solutions for 'crowdharvesting' of results offered by the labourers to the company that initiated the open call.

Flat rate payments are often very small (i.e. 1 cent to a few dollars per task). Conversely, a bonus or prize can be very large. For example, Cisco saved millions using the winner solution of their 'iPrize' competition among those who attempted to propose an innovative business plan for the company. Although the prize was only $250,000, some authors (e.g. Brabham, 2008a) defend the fact that the company still has to take the burden and risk of manufacture of the product, distribute it and sell it.

To summarise this section, the classifications and examples have been tabulated (Table 1).

Table 1   Classification of some examples of crowdsourcing activities

	Any individual	*Most people (qualification might be required)*	*Experts*
Creation	Crowdspirit[10] (could be rewarded) Associated content[11] (rewarded)	iStockphoto (low reward) Crowdspring[12] (rewarded)	Threadless[1] (rewarded) Cisco Systems[9] (rewarded) InnoCentive[18] (rewarded) Strip-packing (rewarded)
Evaluation	e-Rewards Market Research (rewarded) Canonical views (rewarded)	'txteagle' (rewarded) UserTesting[13] (rewarded)	Nosago[14] (rewarded)
Organisation	GalaxyZoo[4] (voluntary) Clickworkers[7] (voluntary) ESP-Google Image labeller[3] (voluntary) reCAPTCHA[5] (voluntary)	FunSAT (DeOrio and Bertacco, 2009) (voluntary) CamClickr[15] (voluntary) Shape similarity (rewarded)	FoldIt[6] (voluntary) Ushahidi and Wikicrimes[16] (voluntary)

3   Case studies of the open outsourcing of an industrial manufacturing task

The following case studies are examples of rewarded crowdsourced work where evaluation and organisation tasks have been used to support industrial applications. In these cases, a content-based classification had to be performed by the workers. The first case study is the 'Shape Similarity' application. The classification of 3D shapes might appear an easy task to humans, but it is an almost impossible task to perform by a machine. The difficulty in defining, and therefore cataloguing complex shapes or textures has caused engineering companies (Jagadeesan et al., 2009a; Jagadeesan et al., 2009b) to turn to crowdsourcing to be able to classify their stock. In this way, they find a more efficient way to search in their stores for a specific piece or part required in the assembly instead of re-drawing and manufacturing again from scratch, as done traditionally because the search of the part in the store proved to be more tedious and time-consuming that its remanufacture.

The second case study presented here is the 'Canonical view'. In this classification task the crowd was asked to choose which orientation of a series of 3D objects was 'the most representative view'. This application was motivated by CAD/CAM applications in which easy navigation and intuitive appearance are important aspects (e.g. thumbnail

component databases and online store catalogues). The third case study is 'strip packing'. Workers were invited to pack shapes in a limited space in the most compact fashion possible. In the 3D scenario this application is of great interest to online retail and shipping businesses whose postage costs have to be minimised without compromising the integrity of the goods in the parcels. To give an example of 2D applications, the nesting of components on sheets will save material during their manufacture (e.g. metal punching or fabric cutting).

3.1 Methodology for geometric reasoning tasks in MTurk

The crowdsourcing platform used in these series of experiments was Amazon's MTurk (MechanicalTurk, 2009). As shown in Figure 1, the 'requesters' designed and placed a task. In MTurk, tasks given to the workers are called 'HITs'. Requesters could establish a threshold of qualifications to the prospective workers, and they may get tested before engaging in a task. Requesters could also accept or reject the results sent by the workers, which had an impact on the worker's reputation in Amazon's portfolio. As in most of the crowdsourced activities, these workers were well spread around the world. Payments for completing tasks could be redeemed on 'Amazon.com' via gift certificates or alternatively cashed and transferred to a worker's bank account. Details on the MTurk interface design, how an API is used to place the HITs and a thorough description of the participants' characteristics are not on the scope of this paper and these, along with further details of experimental results, can be found elsewhere (Jagadeesan et al., 2009a; Jagadeesan et al., 2009b).

3.2 Case study 1. Organisation task with aggregated answer: crowdsourcing of shape similarity

To investigate the crowd's ability to make subjective judgements about the relative similarity of shapes, workers were asked to sort over 400 thumbnails of 3D shapes into family groups. This open end task was rewarded with a payment of $4 and workers took up to 37 mins 18 sec to complete the task. The results were aggregated using the method described by Jagadeesan et al. (2009b) and displayed using a dendogram (Figure 7). A small example of the possible group of pieces can be seen in Figure 4.

Figure 4 Similar pieces grouped in families by the workers (Jagadeesan et al., 2009b) (see online version for colours)
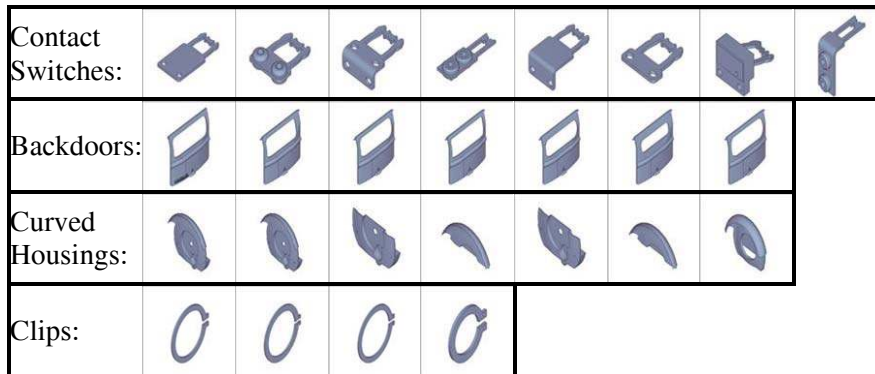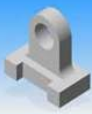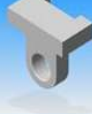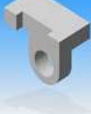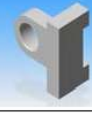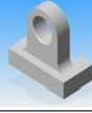
3.3 Case study 2. Evaluation task with a poll on individual results: crowdsourcing of canonical viewpoints

This experiment was designed to investigate the crowd's ability to understand 2D images of 3D CAD models. This was a relatively constrained task involving selection from a limited number of alternatives (Figure 5). The users were paid $0.15 per HIT, and all the results were returned within 1 h 21 min of the task being set. Subsequent rounds were completed in less time (set 2: 23 min 45 sec; set 3: 41 min 22 sec; set 4: 42 min 41 sec), demonstrating the eagerness of the users who, most probably, waited for the different assignments to be published and perform them again.

Figure 5 Results for the 'best' view for this component (in brackets the frequency of selection) (see online version for colours)

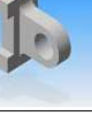| Option 1 (16)	Option 2 (6)	Option 3 (2)	Option 4 (3)	Option 5 (5)	Option 6 (0)
Option 7 (5)	Option 8 (7)	Option 9 (0)	Option 10 (3)	Option 11 (1)	Option 12 (6)

3.4 Case study 3. Creation task requiring expert workers: crowdsourcing of strip packing

In this case, the aim was to investigate the crowd's ability to dynamically optimise a problem with many complex interactions by creating or designing the most compact layout in an area (Figure 6). The results from each individual were compared with the 'packing density' resulting from the established computational algorithms designed for this task. The strip packing HIT is different from the two previous ones in that it was not looking for an averaged or consensus solution, but rather it was seeking the best amongst many attempts. Like many design tasks, results cannot be averaged or aggregated. In this case, users were paid $3 (with a bonus of $0.5 per 0.5% of improvement versus a 'best result' shown during the attempt) and spent between 1 h 24 min and 26 min 20 sec on the task. Plotting the results, packing density and time, allowed a comparison among participants.

Figure 6 Result of a 'Strip Packing' task. This user scored the best result 89.41% packing in 1 h 6 min 58 s) (see online version for colours)

3.5 The crowd in these case studies

Some ethnographic information was also drawn from these case studies and confirms what other authors have reported about the nature of people working on crowdsourcing applications (Howe, 2006; Brabham, 2008a; Kittur et al., 2008):

- *Gender equality*: The workers appeared to be an almost equal number of males and females.

- *Age range*: The age profile was very broad (ranging from teenagers to retired people).

- *Qualifications and skills*: There was no common education or professional background.

- *Location*: Examination of the workers IP addresses suggested that they were literally a global workforce, HITs were returned from India, USA and Europe (depending on the time of day the HIT was made available).

- The workers carrying out the HITs were keen to specialise in tasks they found easy or more rewarded.

- The workforce has a very large capacity for work and responds quickly. Regardless of when in the day work was posted, all the work was done within 2 h.

4 Measuring crowdsourcing output

Although there is a myriad of texts and articles on the potential applications of crowdsourcing for the benefit of both for-profit and non-for profit organisations, these seem to focus on the exploitation of the open calls (i.e. to get the best workers working on their proposed tasks, harvest the best quality of results and in the shortest time) rather than the evaluation of the results. This section attempts to set some recommendations for the evaluation and assessment of results obtained from the crowd. Automation of this analysis is of great interest in those tasks where a large number of results are expected in a short time frame.

4.1 Analysis of crowdsourcing results

In order to set performance metrics for the output obtained from crowdsourced work, the type of the task has to be taken into consideration first. There are several types of questions that can be used to offer work to the crowd. Based on Bloom's definition

(Krathwohl and Bloom, 1956), the cognitive level that needs to be engaged in order to respond to the different questions or tasks can be described as (from least challenging to most): comprehension, application, analysis, judgement and creation. Depending on the level, the questions (or tasks) have to be built accordingly. The responses from the crowd will normally increase in sophistication and complexity as the level increases. Dealing with results originating from tasks that require lower cognitive engagement will be simpler than analysis of responses where an ingredient of judgement, or creation, is implied.

4.1.1 Productivity analysis

The lack of literature on evaluation methods seems to be due to the difficulty in analysing content especially when higher cognitive levels such as creativity or originality are involved. However, some authors (e.g. Huberman et al., 2008) have attempted to quantify productivity levels of users uploading their work to YouTube, a video sharing website on which users can upload videos and visitors can add comments and feedback (www.youtube.com). This study (Huberman et al., 2008) measured productivity of each contributor by the number of videos uploaded during different periods of activity and the number of visits that those items received. These figures rendered dynamic information on each contributor's behaviour towards different levels of 'attention' (i.e. how many people viewed their videos). Following a good and a bad period of attention, an averaged productivity of the users was quantified. It was concluded that contributors tend to become more productive when they receive more views. A lack of attention leads to a decrease in the number of uploaded videos, and consequently a drop in productivity.

4.1.2 Numerical analysis

Crowdsourcing tasks that produce numerical results (like the 'packing task', where parameters such as 'length' and 'efficiency') or tasks where a number of items need to be recorded (e.g. image tagging, counting objects) can be analysed by statistical methods for quantifying frequency or a spread of numbers (e.g. average, standard deviation, etc.). Incoherent answers (e.g. a large error in terms of standard deviations) may indicate that an improvement to the user interface or a threshold of worker's qualifications needs to be introduced (i.e. it is recommended a shift from Type 2, what 'most people' can do, to Type 3, 'expert', Figure 3).

4.1.3 Discrete values analysis

This method is appropriate for crowdsourcing tasks that require a 'yes/no', 'true/false' answer or that can be selected from a list. These results will produce a poll of two or more choices such as the 'canonical viewpoint' task described in Section 3.3. 'A simple way to assess the validity of a response is through majority voting' (Eagle, 2009). This work (Eagle, 2009) reports a method to find out how many users were necessary in order to get to the right answer. The method is based on the maximum likelihood estimate (Dawid and Skene, 1979). Other authors (e.g. Kittur et al., 2008; Alonso and Mizzaro, 2009) have correlated answers provided by workers versus those results offered by a pool of experts. This was possible to automate and analyse because both experts and crowd were given the same information (according to a set of factors) to perform the ratings. Flawed responses could be minimised by rewording of the task instructions or by reducing the number of options in the multi-choice question.

4.1.4 Multi-variable, non-numerical analysis

This category includes tasks that require language processing skills, such as translation, summarising, extraction of keywords, etc. In this case, several workers are asked to do the same task but the answers might differ while still being total or partially correct. In order to assess the robustness of the method and validity of the responses given by the crowd, all workers' answers have to be compared. This might be difficult and time consuming since there are no numerical values to compare directly. In an example of a 'translation' task [like those crowdsourced by texteagle (Eagle, 2009)], the requesters were interested in the variable 'how many errors are there in the translated text received from the worker'. Once this variable was quantified, the validity of the bank of results could be assessed and benchmarked against other banks. Barrow's method (Barrow, 1998) can be used in these cases and a full explanation can be found in the section Notes at the end of this paper.[17] The procedure starts by comparing the differences between the answers of the different workers. In principle, the number of 'unknown errors' reduces to zero when the number of combined responses become very large.

$$u_E = \frac{E_A \cdot E_B}{E_{AB}} \qquad (1)$$

Where u_E is the number of unfound errors, E_A the errors found only by worker A, E_B the errors found only by worker B, and E_{AB} the errors found by both A and B.

Equation (1) can be applied when there is a large pool of proof-readers (i.e. workers). When the crowd of workers becomes very large, the probability of them finding the same errors is large too (C >> 0). Therefore, the number of unfound errors will be very small the larger the pool of workers get. If this is the case, the robustness of the method can be proven. However, if every worker finds lots of mistakes, but none of them found the same mistakes, then it can be concluded that they are not very good at the task and there are likely to be lots of other mistakes that none of them found. Therefore, improvements need to be made to the experiment.

In tasks requiring a multi-variable, non-numerical answer, this method can be used to compare a number of discrete answers. If the results do not converge, the two approaches that can be followed in other to minimise erroneous, or even malicious, responses are to improve the clarity of the instructions and/or apply a threshold for audience qualification.

4.1.5 Analysis and judgement – crowdsourced judgement

In these tasks, where the user is asked to give feedback, evaluate or rate, there is not a right or wrong answer, and it is the amalgamation of all the responses given that generates the 'correct' answer. Often with 'Judgement-type' results, the challenge is to 'display' the results rather than 'perform' an analysis. Case study 1 (in Section 3.2) shows an example of an experiment whose result was the creation of a 'Similarity Matrix' (Figure 7) to show the most frequent shape matching offered by the workers. DoloresLab (2009b) runs a colour tagging activity where workers are asked to label colours (i.e. 'what would you call this colour?'). The responses have been depicted using a colour cloud (Figure 8).

Figure 7   Shape similarity matrix constructed by aggregation of responses and plotted on a dendogram, briefed version (full version in Jagadeesan et al., 2009a) (see online version for colours)
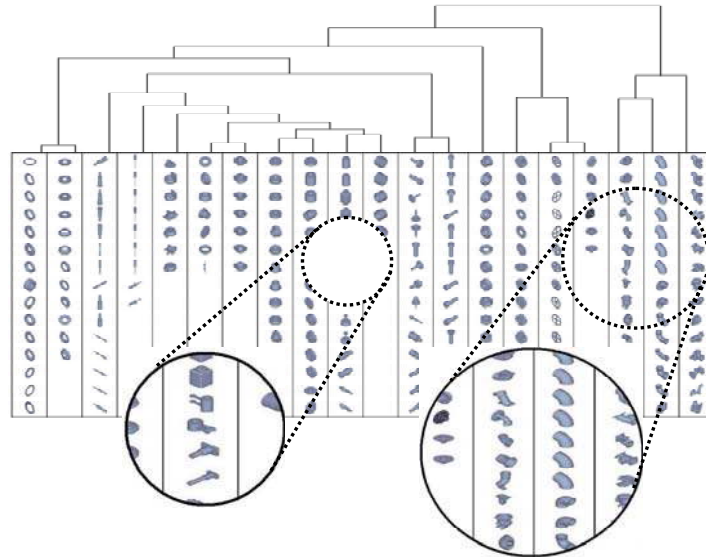
Figure 8   DoloresLab experiment on colour classification (DoloresLab, 2009b) (see online version for colours)
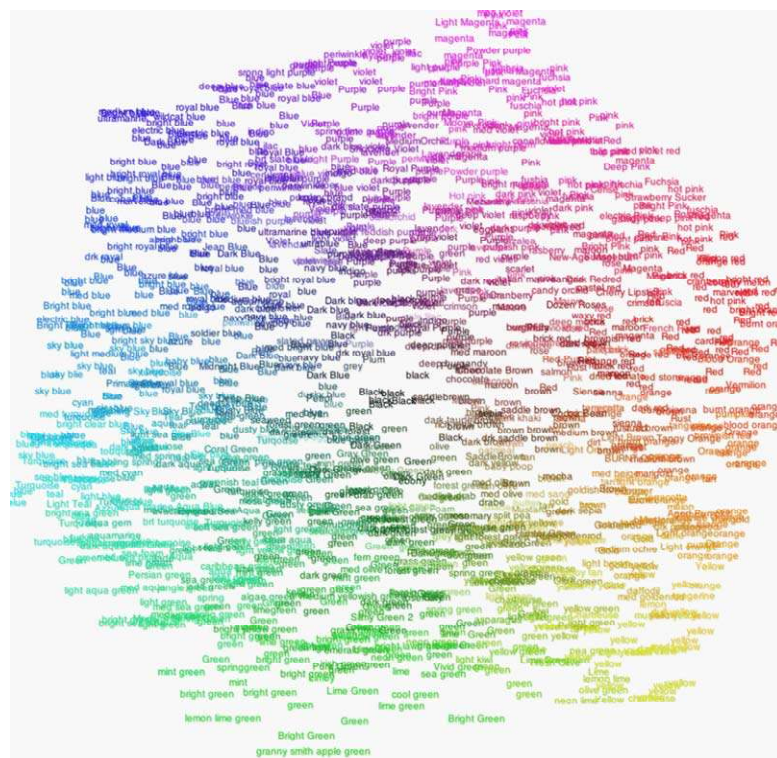
4.1.6 Crowdsourced creation analysis

This is probably the most difficult set of responses to evaluate because the output laid by the users will usually be highly subjective and difficult to quantify. Example for this type of tasks are highly inventive and artistic (such as logos, clothing, furniture design) or highly specialised, for example ideas competitions, business studies (e.g. CiscoSystems, as already mentioned, and InnoCentive[18]) etc. The evaluation process is normally carried out by a group of experts (Leimeister et al., 2009), a jury commonly recruited from within the company who proposed the open call. Other approaches include the use of 'double crowdsourcing', this is, companies return to crowdsourcing to make use of voting and rating systems in order to assess which crowdsourced results are best (particularly in design, e.g. clothing and furniture design).

5   Discussion

Apart from the difficulties in harnessing capabilities from a crowd where an employer–employee relationship does not exist, one of the major challenges that crowdsourcing poses is the assessment of the validity of labourers' response to a given task. Different types of tasks (as per nature of the task, the crowd or the payment) have been described in this paper and each of them, when combined (e.g. a voluntary task of creation for experts, or a rewarded task of organisation for anyone), present their own challenges when evaluating the responses. For micro-tasks, where a 'yes/no' answer is required or a line of text translated, it is simple to aggregate users' input, no matter how large the pool of data might be, and perform traditional statistical operations in order to obtain the best solution. However, those tasks where a more sophisticated answer is sought and a higher cognitive level engaged will be more difficult to assess and compare [e.g. the assessment of Wikipedia articles (Kittur et al., 2008)], the review of a book via Amazon, a movie via IMDb or hotel service in a holiday resort via TripAdvisor, etc.). Furthermore, interferences in the data will occur if malicious users or invalid entries are added to the collection of results. Then, the task initially shifted to the crowd to alleviate the company workload or obtain a better solution, will jeopardise its final objective. Choosing the best result and discarding those entries that are not useful to the process becomes more difficult that the initial task *per se*. Qualitative judgements are more difficult to analyse than quantitative ones. Additionally, the validity of the crowdsourced answer gets compromised when the number of responses has to be sacrificed for the sake of the quality in the answers. Although experience indicates that crowds are better at evaluating and organising rather than creating, remarkable examples of crowdsourced creations with significant payments are in the rise.

While a myriad of examples can be found in the tasks currently outsourced to the crowd, there is a fundamental limitation in this strategy and this has to be taken into account for 'open innovation' activities. Crowdsourcing does not work where there exist an emotional element for a product, e.g. in music, movies and in retail. Human beings are not happy when being told what piece of music to listen to, which meal to choose in a restaurant or which movie to see. Moreover, a recent study on online recommendations (e.g. ratings, reviews, etc.) has explored the practical drawback that crowdsourcing has when it comes to ranking. Kostakos (2009) has studied the rating systems (i.e. stars,

survey, feedback) of several websites and has concluded that a small group of very active members can sway total ratings. In other words, there are significant biases in users' voting behaviour, despite the large size of the online community.

5.1   Academic research agenda

An academic research agenda needs to be established in order to produce a series of recommendations for improving response of the crowd and assess the results more efficiently (also applicable to re-designing tasks that were not performed correctly by the crowd in a first instance) in order to support 'open innovation' activities in a productive and meaningful way.

In following sections we suggest the most important aspects to consider before a crowdsourcing strategy is deployed.

5.1.1   Creating user interfaces that support people solving problems

It is common knowledge among crowdsourcing users that the better the questions (or assignments) are formulated to the crowd, the more valid the results are. Moreover, the design of the interface needs to fulfil basic requirements and be aligned to the aims of the crowdsourced activity. For example, if data is going to be plotted in a poll-type of representation, the questions will have to be designed in a way that is coherent with the question, e.g. a multi-choice question instead of a 'free text' box. The importance of visualisation has been highlighted by various authors (Beynon, 2007) and it is crucial for those crowdsourced tasks where representations and illustrations are the tools that requesters and users have to communicate (e.g. the case studies in Section 3).

It is common that the design of the interface requires several attempts. Once that option has been exhausted and the results to the call are still not satisfactory, a redesign of the call could involve aiming at an audience with more/better qualification, i.e. a shift from Type 1, to Type 2 or 3 in Figure 3, and/or introduce a payment for completed valid response or, if already rewarded, re-organise payment system (e.g. a system of incentives for best answer, etc).

5.1.2   Statistics

Prior to the start of the activity, it is important to think about the format in which the results are going to be harvested, plotted, and consequently which mathematical tools are needed, as listed in Section 4. Statistical methods to analyse crowdsourcing activities are emerging at a slow rate because each type of questions poses an intrinsic challenge when it comes to compiling the results: the greater the cognitive level involved, the more difficult a large bank of results will be to handle. When it comes to evaluate cognitive activities, e.g. text translation (Callison-Burch, 2009), manual evaluations still have to be used to demonstrate the feasibility of the method. For quantifications, especially when obtaining a final result by aggregation of many responses (e.g. surveying and pooling), the requesters have to set a threshold of confidence beyond which the final result will be acceptable (Kittur et al., 2008; Alonso and Mizzaro, 2009). In other words, a minimum number of responses from users is needed to obtain the sought result. There is not a

standard formula for the calculation of that number, although statistical methods can be used in some simple applications, e.g. maximum likelihood estimation (Dawid and Skene, 1979), but it will strongly depend on different design variables such as the cognitive level engaged for the activity, the nature of the crowd at which the activity is aimed, and time constraints on the requesters to obtain the final result.

5.1.3 Business models

There are a number of issues that businesses have to face when deciding to get involved in crowdsourcing activities to facilitate their 'open innovation' and start designing those. Most importantly is the question of how to build this activity into a work flow (rather than simply having a one-off experience), how to promote it and reach the right audience (i.e. the nature of the crowd, experts, most people with a specific qualification, anyone, etc.) versus recruiting your own dedicated group of workers, who will be in charge of validating the responses (e.g. a computer or a group of juries), and how these tasks will be channelled to the workers.

5.1.4 Social implication

As mentioned in Section 2.3, crowdsourcing activities for 'open innovation' have the potential of having a great impact on business dynamics and sustainability. There are obvious advantages of outsourcing work to the crowd, as discussed in this paper, but there are also important aspects that need to be taken into consideration before a crowdsourcing strategy is adopted. As consumers, we ought to address the issues dealing with moral responsibility and citizenship (Schrader, 2007). For example, social issues of fees and payment (i.e. workers get very low pay in comparison to professionals doing the same job), age of labourers, waived taxes, and whether this activity is fostering underground economic transactions. Legal regulation of these aspects is not only necessary but also compromising and difficult due to the nature of the activities 'in the cloud'.

6　Conclusion

Crowdsourcing is becoming an important tool for leveraging 'open innovation' processes in those firms willing to advance their technology, increase originality and likeability for their products by outsourcing certain tasks to a crowd. Several studies have demonstrated that there is a large, responsive work force available 24/7 capable of carrying out tasks of a range of complexity successfully. Furthermore, our experience, briefly exposed in the case studies in Section 3, suggests that this resource has a very large capacity for work and responds quickly. Beyond the specifics of the results presented, the authors believe that crowdsourcing provides a credible methodology in which a 'human algorithm' (rather than a purely computational one) could be implemented in practical applications with great opportunities for successful in 'open innovation'.

References

Alonso, O. and Mizzaro, S. (2009) 'Relevance criteria for e-commerce: a crowdsourcing-based experimental analysis', *Proceedings of the 32nd International ACM SIGIR'09 Conference on Research and Development in Information Retrieval*, 19–23 July 2009, Boston, MA, USA, pp.760–761.

Barrow, J.D. (1998) *Impossibility: The Limits of Science and the Science of Limits*, Oxford University Press, Oxford, New York.

Beynon, M. (2007, June) 'Visualisation using empirical modelling principles and tools', *AHRC ICT Methods Network Expert Workshop "From Abstract Data Mapping to 3D Photorealism: Understanding Emerging Intersections in Visualisation Practices and Techniques"*, Birmingham UK.

Brabham, D.C. (2008a) 'Crowdsourcing as a model for problem solving: an introduction and cases', *Convergence*, Vol. 14, No. 1, pp.75–90.

Brabham, D.C. (2008b) 'Moving the crowd at iStockphoto: the composition of the crowd and motivations for participation in a crowdsourcing application', *First Monday*, Vol. 13, No. 6, pp.1–10.

Callison-Burch, C. (2009) 'Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk', *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing by Association for Computational Linguistics*, 6–7 August 2009, Singapore, pp.286–295.

CambrianHouse (2009) Available online at: www.cambrianhouse.com/ (accessed on 8 September 2009).

Dawid, A. and Skene, A.M. (1979) 'Maximum likelihood estimation of observer error-rates using the EM algorithm', *Applied Statistics*, Vol. 28, pp.20–28.

Deorio, A. and Bertacco, V. (2009) 'Human computing for EDA', *ACM-DAC*, 26–31 July 2009, San Francisco, CA.

DoloresLab (2009a) *Dolores Lab: age and gender stereotypes*. Available online at: http://blog.doloreslabs.com/topics/faces/ (accessed on 20 August 2009).

DoloresLab (2009b) *Dolores Lab: color flowers, networks, photos, and even 3D*. Available online at: http://blog.doloreslabs.com/topics/colors/ (accessed on 20 August 2009).

DoloresLab (2009c) *Dolores Labs*. Available online at: http://doloreslabs.com/ (accessed on 20 August 2009).

Eagle, N. (2009) 'txteagle: mobile crowdsourcing', *Proceedings of the 3rd International Conference on Internationalization, Design and Global Development: HCI International 2009*, 19–24 July 2009, San Diego, CA, Aykin, N. (Ed.) *Lecture Notes In Computer Science*, Vol. 5623, pp.447–456.

Howe, J. (2006) 'The rise of crowdsourcing', *Wired Magazine*, No. 14. Available online at: http://www.wired.com/wired/archive/14.06/crowds_pr.html (accessed on 17 September 2009).

Huberman, B.A., Romero, D.M. and Wu, F. (2008) 'Crowdsourcing: attention and productivity', *SSRN eLibrary*, 12 September 2008. Available online at: http://ssrn.com/paper=1266996 (accessed on 8 September 2009).

Jagadeesan, P., Wenzel, J., Corney, J.R., Yan, X.T., Sherlock, A., Torres-Sanchez, C. and Regli, W. (2009a) 'Geometric reasoning via internet crowdsourcing', *SIAM/ACM Joint Conference on Geometric & Physical Modeling*, San Francisco, CA.

Jagadeesan, P., Wenzel, J., Corney, J.R., Yan, X.T., Sherlock, A., Torres-Sanchez, C. and Regli, W. (2009b) 'Validation of Purdue engineering shape benchmark clusters by crowdsourcing', *International Conference on Product Lifecycle Management*, Bath, UK.

Kittur, A., Chi, E.H. and Suh, B. (2008) 'Crowdsourcing user studies with Mechanical Turk', *Proceeding of the 26th annual SIGCHI Conference on Human Factors in Computing Systems, ACM*, Florence, Italy, pp.453–456.

Kostakos, V. (2009, September) 'Is the crowd's wisdom biased? A quantitative analysis of three online communities', *International Symposium on Social Intelligence and Networking (SIN09)*, Vancouver, Canada.

Krathwohl, D. and Bloom, B. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*, Longmans, Green.

Leimeister, J.M., Huber, M., Bretschneider, U. and Krcmar, H. (2009) 'Leveraging crowdsourcing: activation-supporting components for IT-based ideas competition', *Journal of Management Information Systems*, Vol. 26, No. 1, pp.197–224.

MechanicalTurk (2009) *Amazon's Mechanical Turk*. Available online at: http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk (accessed on 30 August 2009).

Pisano, G.P. and Verganti, R. (2008) 'Which kind of collaboration is right for you?', *Harvard Business Review*, Vol. 86, pp.78–86.

Powell, D. (2008) 'Amateur hour', *Arts & Sciences Online Magazine of the John Hopkins University*, Vol. 5, No. 2. Available online at: http://krieger.jhu.edu/magazine/sp08/f1.html (accessed on 17 September 2009).

Schrader, U. (2007) 'The moral responsibility of consumers as citizens', *International Journal of Innovation and Sustainable Development*, Vol. 2, No. 1, pp.79–96.

Sweney, M. (2009) 'Unilever goes crowdsourcing to spice up Peperami's TV ads', *The Guardian*, 25th August, UK.

Wegen, E.V. (2008, June) *The world of internet*. Available online at: http://researchreinvented.blogspot.com (accessed on 30 May 2009).

Yang, J., Adamic, L.A. and Ackerman, M.S. (2008) 'Crowdsourcing and knowledge sharing: strategic user behavior on taskcn', *Proceedings of the 9th ACM conference on Electronic Commerce*, Chicago, Il, USA, pp.246–255.

Notes

1 Threadless: A community-centred online clothing store run by skinnyCorp of Chicago, Illinois, since 2000. Members of the Threadless community submit t-shirt designs online; the designs are then put to a public vote. A small percentage of submitted designs are selected for printing and sold through their online store. Creators of the winning designs receive a prize of cash and store credit (www.threadless.com).

2 Muji: Japanese furniture retailer. This company has a pool of approximately 500,000 people who pre-evaluate designs and short-list the ones that are then passed onto professional designers, who produce the final specifications for the commercial product (www.muji.net).

3 Google Image Labeller: Created by L. von Ahn in 2004 as the ESP Game and later acquired by Google. In this game, players tag images by using words that they think associated to the image. These images get then tabulated and used them to improve 'image search' in Google's search engine (http://images.google.com/imagelabeler).

4 GalaxyZoo: Created by Hopkins researchers Schawinsk, Land and Lindtott in 2006 on an attempt to trying to sort millions of newly discovered galaxies into categories. The galaxies, photographed as part of the Sloan Digital Sky Survey, were too many to be classified by just a few astronomers, and computers were not able to do the task. They created Galaxy Zoo, an experimental website designed to train the public to help them go through the photographs and classify the galaxies. By the end of 2008, users had contributed to the project by looking at 40 million images and classified more than a million galaxies (each galaxy is verified by at least 30 people) (https://www.galaxyzoo.org/).

5   reCaptcha: Created by von Ahn at the Carnegy Mellon University who revisited Blum's earlier work on Captchas (tasks that require the visual skills that humans have but not computers), used to digitise millions of words per day from the Internet Archive and the New York Times archive. Texts from old books, which cannot be deciphered by OCD software, are provided to end users who interpret the images and transcribe the words. This task is used in conjunction with spam filters or to ensure that the end user is indeed human, and not a software application trying to access a site.

6   FoldIt: Protein folding game that requests from the user suggestions for folding strategies in different protein chains (http://fold.it/portal/).

7   Clickworkers: A NASA experiment with volunteers who analysed and identified craters on Mars images obtained via satellites. The experiment started in 2000 with the first images and simple identification tasks, and in 2007 it was still running but at a higher sophistication of analytical activity.

8   Steve Fosset's plane search: After the disappearance of the millionaire adventurer, in September 2007a HIT was placed on Amazon's MTurk with thousand of high-resolution pictures of the area where Fosset was supposed to have crashed his plane. Volunteers could scan through the images and determine whether they could see a plane (or parts) on the pictures.

9   Cisco Systems: This corporation held a Global I-Prize Innovation contest in 2008 and requested teams to use collaborative technologies to create an innovative business plan. More than 2500 people from all over the world entered the competition and the winning team, who created a business plan demonstrating how IP technology can be used to increase energy efficiency, won a prize of US$250,000.

10 Crowdspirit: Launched in 2007, this French portal harnesses any innovative idea from their community. Mainly oriented towards electronic gadgets, they receive suggestions for ideas that their R&D team fine-tunes at later stages (www.crowdspirit.com).

11 Associated content: Founded by L. Beatty in 2005 in Denver, Colorado, this platform aims to become a multimedia-content library. Any individual can contribute with original material in any multimedia format (e.g. video, audio, text, images, etc.) on any topic, as long as the content is original and never published before. The company distributes this content to their partners via their website (www.associatedcontent.com).

12 Crowdspring: Based in Chicago, this venture commenced in 2008 when R. Kimbarovsky and M. Samson set up a network of designers from around the world. Their creative activity is mainly on corporate image (i.e. logos, websites, branding) and merchandising (www.crowdspring.com).

13 UserTesting: This crowdsourcing service, created by D. Garr and D. Benater, offers surveys on websites. The users are pre-screened via qualifying questions, and users with very good communication skills are sought, since the feedback is on their opinions and thoughts while they surf the websites surveyed (www.usertesting.com)

14 Nosago: This service provides recommendations on projects run by companies, government or other organisations. Funded by D. Almour in 2009, seeks higher-than-average professionals who have an expertise in project management and, therefore, can take part in this crowdsourced work (www.nosago.com).

15 Camclickr: Probably encouraged by the success of GalaxyZoo, the Cornell Lab of Ornithology created in 2008 this game where users can classify nest images in different albums. During the tasks, they are briefed on different aspects of animal behaviour and the complexity of the classification increases as the game progresses.

16 Ushahidi and Wikicrimes: These two platforms (Ushahidi in Africa and Wikicrimes in Europe) aim to map crime reports. People who are victim of a criminal offence can enter the information, which gets stored in a database and plots the results on a map. Although the task does not require 'experts' to be done, in the literal sense of the word, we have placed these under that heading since it is only a small proportion of the population who will contribute to this database (see Figure 3).

17   In his book, Barrow (1998) describes the method as follows: *"Strange as it may sound, we can make quantitative statements about the number of things we don't know. So we can answer questions such as: 'How many more bugs are there in this program?', 'How many, unnoticed, errors in the design of our new machine, are there?', 'How many discoveries are there still to be made?', etc."* To understand how this is possible, consider the problem extracted from Barrow (1998) to find how many typographical errors remain unfound while proof-reading an article. Suppose two editors, Jack and Jill, independently read a long newspaper article supplied by one of the journalists. Jack finds 'A' errors, while Jill finds 'B' errors. They compare copies and discover that they found the same error on 'C' occasions. Between them, the total number of errors they found is: $A + B - C$. Let us suppose that the total number of errors in the article is 'E'. This means the number of errors still to be found is: Unknown errors $= E - (A + B - C)$. If the probability that Jack spots an error is 'p', and the probability that Jill spots an error is 'q', then it is expected that: $A = p \times E$ and $B = q \times E$, therefore, $C = p \times q \times E$. Probabilities are multiplied because Jack and Jill search independently, so $AB = p \times q \times E * E$, and, thus $AB = CE$. Finally, the number of unfound errors equals to: Unknown errors $= E - A - B + C = AB/C - A - B + C$. (Unknown errors) $*C = ABC - AC - BC + C^2$, or Unknown errors $= [(A - C)(B - C)]/C$. That is,

$$\text{Number of unfound errors} = \frac{(\text{number found only by Jack}) \cdot (\text{number found only by Jill})}{(\text{number found by both Jack and Jill})}$$

18   InnoCentive: This venture started in 2002. It benefits from the crowd by obtaining suggestions to solve R&D challenges posed by biomedical and pharmaceutical companies (www.innocentive.com).