Chapter 2
Crowdsourcing Information Systems

2.1 Overview

Crowdsourcing information systems can be classified in many different ways. One of the classifications can be the nature of collaboration: explicit or implicit. In explicit collaboration systems (e.g., Wikipedia or Linux), users collaborate explicitly to build information artifacts. On the other hand, implicit collaboration systems let users collaborate implicitly to solve a problem for the system owners. For instance, the ESP game [33] makes users collaboratively label images as a side effect while playing the game.

Second type of classification can be based on the type of the target problem. The target problem can be any problem defined by the system owners, from building temporary or permanent information artifacts to executing tasks. Another dimension can be the degree of manual effort. When building a crowdsourcing system, system owners must decide how much manual effort is required to maintain the system. This can range from relatively little (e.g., combining ratings) to substantial (e.g., combining code), and also depends on how much the system is automated. The system owners must decide how to divide the manual effort between the users and themselves. Some systems ask the users to do relatively little and the owners a great deal. For instance, to detect malicious users, the users may simply click a button to report suspicious behaviors, whereas the owners must carefully examine all relevant evidence to determine if a user is indeed malicious. Some systems do the reverse. For example, most of the manual burden of merging Wikipedia edits falls on the users who are currently editing, not the owner.

The fourth criteria for classification can be the role of human users. Here, we can consider four basic roles for humans in a crowdsourcing system. *Slaves*: humans help solving the problem in a divide-and-conquer fashion, and minimize the resources (e.g., time, effort) required by the owners. Examples are ESP games and finding a missing boat in satellite images using Mechanical Turk-based systems [154]. *Perspective providers*: humans contribute different perspectives that when combined produce a better solution than with a single perspective. Examples are reviewing books and aggregating user bets to make predictions [195]. *Content providers*: humans contribute self-generated content (e.g., videos on YouTube or images on

A. Ranj Bar, M. Maheswaran, *Confidentiality and Integrity in Crowdsourcing Systems,* SpringerBriefs in Applied Sciences and Technology, DOI 10.1007/978-3-319-02717-3_2,

Flickr). *Component providers*: humans function as components in the target artifact, such as a social network, or simply just a community of users (e.g., the owner can sell ads). Humans often play multiple roles within a single crowdsourcing system (for example, slaves, perspective providers, and content providers in Wikipedia) [80].

2.2 Major Crowdsourcing Systems

We introduce the most widely used crowdsourcing information systems by categorizing them into four different groups based on the target problem. For this categorization, we use collective knowledge management, collective creativity, collaborative gaming and collaborative voting as the four groups. Because online social networks are one of the most important class of crowdsourcing information systems, we discuss them separately in the next section.

2.2.1 Collective Knowledge Management

This type of systems allows users to build artifacts often merging user inputs tightly and requiring users to edit and merge one another's inputs. A well-known artifact created by this type of system is textual knowledge bases (KBs). To build such KBs, users contribute data such as sentences, paragraphs, web pages, and edit and merge one another's contributions. The two main examples of crowdsourcing KB systems are Wikipedia and Yahoo! Answers.

Wikipedia is an online encyclopedia that is freely available. The notion of open editing in Wikipedia encourages many people to collaborate in a distributed manner to create and maintain a repository of information artifacts. Wikipedia has more than 17 million registered authors and more than four million articles [2]. It has become a valuable resource and many people cite it as a credible information source. However, the open process that provides popularity to Wikipedia makes it difficult for readers to ascertain the credibility of the content. Similar to other crowdsourcing systems, Wikipedia articles are constantly changed by contributors who can be non-experts or even vandals. On the other hand, Yahoo! Answers is a general question-answering forum to provide automated collection of human reviewed data at Internet-scale.

2.2.2 Collective Creativity

The role of human in creativity cannot be replaced by any advanced technologies. The creative tasks such as drawing and coding can only be done by humans. Here, crowdsourcing is used to tap into online communities of thousands of users to develop original products and concepts in areas such as photography, advertising,

filming, video production, graphic design, and apparel design. As a result, several entrepreneurs have setup crowdsourcing systems (e.g., Sheep Market) to harness the power of the crowds to cheaply complete creative tasks. The Sheep Market is a web-based artwork to implicate thousands of online workers in the creation of a massive database of drawings. It is a collection of 10,000 sheep created by MTurk workers, and each worker was paid US$ 0.02 to draw a sheep facing left [3, 121].

Another example is Threadless which is a platform of collecting graphic T-shirt designs created by the community. Although technology advances rapidly nowadays, computers still have no clue about how to creatively solve a specific problem to develop a new product. In Threadless, different individuals may come up with different design ideas for T-shirts [55] from which the most appropriate ones could be selected. Moreover, Leimeister [130] proposed crowdsourcing software development tasks as ideas for competitions to motivate user participation in crowdsourcing. Well known software such as Apache, Linux, Hadoop are produced and maintained by crowdsourcing systems.

2.2.3 Collaborative Gaming

The concept of *games with a purpose* was pioneered by Luis Von Ahn and his colleagues [32]. The games produce useful metadata as a by-product. By taking advantage of people's desire to be entertained, problems can be solved efficiently through online games. The online ESP Game [33] was the first human computation system, and it was subsequently adopted as the Google Image Labeler. Its objective is to collect labels for images on the Web. In addition to image annotation, the Peekaboom system [34] can help determine the location of objects in images and provide complete outlines of the objects in an image. The concept of the ESP Game has been applied to other problems. For instance, the TagATune system [128], MajorMiner [138] and TheListen Game [206] provide annotation for sounds and music which can improve audio searches.

2.2.4 Collaborative Voting

In this type of crowdsourcing systems, a user is required to select an answer from a number of choices. The answer that the majority of the users select is considered to be correct. Voting can be used as a tool to evaluate the correctness of an answer from the crowd. An example of popular crowdsourcing websites with collaborative voting is Amazon Mechanical Turk (or MTurk) [4]. A large number of applications or experiments are conducted in Amazon's MTurk site. It can support a large number of voting tasks.

2.3 Social Networks

Social networks are social structures that consist of social entities (e.g., individuals, groups, organizations) that are connected to one another by social relationships [215]. Social relationships can be quite broad; examples include friendships, behavioral interactions, biological relationships, or affiliations.

Social network analysis focuses on studying the different patterns of relationships among social entities along with their implications on the behavior and decisions of the social entities [211]. Based on the same concepts used in graph theory, a social network is represented by a graph consisting of a set of nodes and edges. The nodes in the graph represent the social entities, while the edges represent the social ties that link those entities. The resulting graph structures are often complex where the social entities are considered interdependent rather than independent units. This means that in social network analysis the discrete unit of analysis is the combination of social entities and the relationships among them.

Social network analysis has been widely used in recent decades in such diverse areas as sociology, anthropology, biology, economics, and information science [152, 46, 103, 205, 213, 149]. For example, in the area of epidemiology, social network analysis has been used to study the effect of different patterns of social contacts on the spread of human diseases and viruses [152], and also to study the relationship between social and community ties and mortality among people [46]. In the field of sociology, social network analysis played an important role in understanding how information spreads on social networks [103] and how individuals are connected in the physical world [205, 213]. In economics, the influence of social structures on the outcomes of the labor market are analyzed using the tools of social network analysis [149].

The last few years have witnessed the emergence of the second generation of the world wide web—"Web x2.0." By facilitating online collaboration and information sharing for people, Web 2.0 has enabled the development and evolution of online based social networks (communities). This has resulted in an increased use of social network analysis to study the underlying structures of these communities and address the problems and challenges that arise within these online systems. Due to their importance, we introduce online social networks and discuss the different properties associated with them in the next section.

2.4 Online Social Networks

Online social networks are online communities of individuals who share common interests or activities. There are many web portals on the Internet that offer the facilities to create manage online communities with different modalities to socialize among the members. Boyd [54] defines today's social networking websites as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share

a connection, and (3) view and traverse their list of connections and those made by others within the system."

While the definition for social networking websites provided in [54] presents those sites as being mainly profile-based, there exists many social networking sites that offer other types of services. The research report in [5] presents a much broader categorization of the social networking services that exist today. It identifies eight main types, among which are the popular profile-based social networks like myspace.com and facebook.com. Content-based social networks are also among the most popular sites, where the main form of interactions and relationships between users are established through the creation of user content. Examples of such sites include flickr.com, a photo-sharing site, youtube.com, a site for sharing user created videos, and delicious.com, a social bookmarking and tagging site. In addition, other social networks provide micro-blogging services, where the users post status messages allowing other people on their social network to track their status; an example of such a service is twitter.com. What makes these social networking sites interesting is that they eliminate the physical limitations of the traditional social networks, allowing their participants to extend and build their personal social networks by meeting new individuals from across the globe. As a result, we are witnessing the rise of new and different relationship structures that are not related to the offline world.

Some suggest that online social networking can be traced back to 1997 with the launch of the first blogging site [6] and social networking website sixdegrees.com [54]. Since then, the number of social networking sites has increased dramatically, attracting many users and generating high web traffic. Based on the information provided by Alexa (alexa.com—a database of information about sites that includes various statistics), many of the existing social networking websites are ranked in the top 500 websites in terms of traffic generated on the web [7]. Reports from Nielsen Online [8], a company that provides measurement and analysis of online audiences, indicates that nearly half of the biggest social networking sites are also among the fastest growing, with still room for potential future growth.

http://www.springer.com/978-3-319-02716-6

Confidentiality and Integrity in Crowdsourcing Systems

Ranj Bar, A.; Maheswaran, M.

2014, V, 77 p. 8 illus., 4 illus. in color., Softcover
ISBN: 978-3-319-02716-6