Crowdsourcing and its Impact on Future Internet Usage

Phuoc Tran-Gia, Tobias Hoßfeld, Matthias Hartmann, Matthias Hirth, University of Würzburg

1 Introduction

Crowdsourcing is a newly emerging service platform and business model in the Internet. In contrast to outsourcing, where a job is performed by a designated worker or company, with crowdsourcing jobs are outsourced to a large, anonymous crowd of workers, the so-called human cloud, in the form of an open call. The rise of crowdsourcing and its seamless integration in current workflows may have a huge impact on the Internet and on society, and will be a guiding paradigm that can form the evolution of work in the next years.

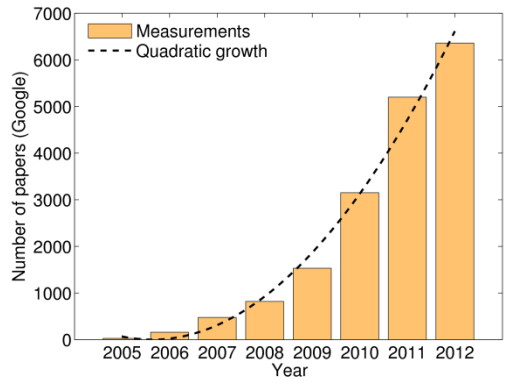
In this article, we discuss applications and use cases of crowdsourcing to demonstrate the impact on Internet usage. A particular focus is on novel measurement approaches of (mobile) Internet traffic and crowdsourcing services which improve current Internet applications or allow creating new services. Then, the impact of crowdsourcing on Internet traffic is considered. Although the current amount of crowdsourcing traffic is still relatively small, there is a significant growth of crowdsourcing in terms of number of users, platforms, and data to be transmitted. To understand the characteristics of the resulting traffic, measurements on the activity of a single crowdsourcing platform are performed, which turn out to be non-constant, time-varying and world-wide distributed. Technical solutions are necessary for the operation of efficient, distributed crowdsourcing platforms taking into account these traffic features. Special attention is drawn to the integration of machine clouds and human crowds. When the cloud meets the crowd, appropriate inter-cloud solutions must be developed for (cost-) efficient automated operation. Finally, we discuss the current research challenges from the platform provider's point of view but also from a scientific point of view

2 Applications and Use Cases of Crowdsourcing

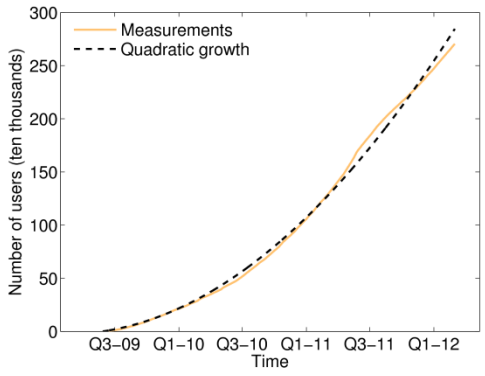
The number of Internet users constantly increases all over the world and, in recent years, this increase was further fostered by the Internet becoming more and more ubiquitously available in using mobile devices and tablet PCs. This growing number of Internet users forms a huge, globally distributed, and manifold workforce, which can be accessed using the Crowdsourcing approach: Small portions of work, so called *tasks*, are distributed to Internet users all over the world willing to complete these tasks. Users completing tasks are in the following referred to as *workers*, and user who generate the tasks and pay the workers are referred to as *employers*. The workers and employers interact with each other using mediator platforms which provide means to distribute the work in a cost and time effective manner. These platforms are called *Crowdsourcing Platforms*. Due to its potential in changing the future development of work organization and its impact on the future Internet, the Crowdsourcing approach gains a lot of attention, both from commercial companies as well as from the academia, as can be seen by the number of papers related to this topic, cf. **Fehler! Verweisquelle konnte nicht gefunden werden.** Figure 1a).

While commercial companies are mainly focusing on new business cases, Crowdsoucing is interesting from a scientific point of view due to the challenges evolving from numerous possible use cases of this approach. Use cases can be classified by the *type of work* or by their *operational conditions*. The *type of work* includes use cases from sensing to problem solving. **Crowdsensing** considers in particular mobile or participatory sensing, like the project NoiseTube [Mai09] which aims to generate maps of noise pollution by utilizing smartphone measurements of its participants. Crowdsourcing can also be used to acquire users to run specialized

measurement applications, e.g. for distributed network analysis. This can also be seen as Crowdsensing in a broader context. **Crowdsolving** lets users perform tasks in fields such as research and development or data analysis. Existing platforms hosting Crowdsolving task are Innocentive [1] or kaggle[2] which are used by companies like Colgate-Palmolive, Microsoft or NASA to crowdsource the development of mechanisms to inject fluoride powder into a toothpaste tube, algorithms to recognize hand and arm gestures, or algorithms to detect distortions in galaxy images caused by dark matter. **Crowdtesting** utilizes the human cloud for conducting scientific studies, e.g. in the context of user perceived quality [Hoß11], or product and software evaluation. On platforms similar to uTest[3], companies like Google or BBC can easily perform large scale functional, usability or load tests. Other relevant task categories are **data extraction**, like e.g. contact data collection of business customers and content tagging, or **crowdvoting** for gathering opinions, which are offered by multiple service providers like CrowdFlower[4] and CrowdSource[5]. **Crowdwisdom** tasks in turn can be used to gather knowledge e.g. like Wikipedia, and more generally **creative tasks,** which include logo or web page design as offered by 99designs[6], or text production for online market places or blogs.



a) Number of publications related to key word "crowdsourcing" according to Google.



b) Number of registered users at Microworkers platform from platform launch until today.

Figure 1 Interest in crowdsourcing is significantly increasing and reveals a quadratic growth.

An *operational condition* is for example **enterprise crowdsourcing** for usage within an enterprise. In that context, the separation of efforts that engage employees vs. public crowds has to be considered. One example of an enterprise crowdsourcing use case is the usage of internal IT experts to execute a large scale IT Inventory Management exercise in an efficient manner [Vuk11]. **Real-time crowdsourcing** addresses the completion of work in real-time and invokes additional challenges, like assuring the concurrent availability of numerous workers [Ber11]. Further, also technical challenges arise here, because the infrastructure has to cope with massive requests at the same time, i.e. flash crowd effects as known from P2P networks. As a result of ubiquitous connectivity and advances in mobile technologies, **ubiquitous crowdsourcing** emerges with mobile users seamlessly forming interactive networks and participation in a variety of tasks involving gathering, analyzing and sharing data, such as reporting security threats, natural disasters[7], or information for location-based services. Furthermore, general challenges of crowdsourcing are the quality of work, automated collection and processing of information, incentives engaging users. Usually a trade-off between costs and quality has to be found.

[1] https://www.innocentive.com/
[2] http://www.kaggle.com/
[3] http://www.utest.com/
[4] http://crowdflower.com/
[5] http://www.crowdsource.com/
[6] http://99designs.com/
[7] http://www.ushahidi.com/

Crowdsoucing can be utilized to acquire users who run distributed measurement test in the field of network research. Involving end-users in network measurements has already been subject to many research efforts. For example Faggiani et al. [Fag12] developed a measurement framework to perform traceroute or UDP probes from users' smartphones, and Gember et al. [Gem12] designed a prototype to perform active measurements while users interact with their mobile device. These tests can provide valuable information for Internet service providers to optimize their resource allocation and to identify bottlenecks. Also services like OpenSignal[8] or RootMetrics[9] use volunteer smartphone owners, to gather information about the quality and the availability of wireless network services. Also in wired networks, end user measurements can be used to detect network events and their impact on the service of certain services. These measurements can be realized using specialized software or plugins e.g. for BitTorrent [Cho10] or Firefox [Dha12]. However, the presented approaches focus mainly on voluntary participation and there are only little means to control the number of participants or gain measurements from dedicated location, respectively from dedicated devices. Here paid crowdsourcing can be used to provide monetary incentives to recruit tailor fit users in the required locations or with the required devices, consequently leading to better results from user based network measurements.

Besides enhancing current Internet applications, Crowdsourcing also fosters the development of new Internet-based applications and services. Examples are the virtual assembly lines by Cloudfactory[10], which combine human and automatic data processors, or Soylent [Ber10], a plugin for Microsoft Word that enables it's users to directly delegate tasks like proofreading to the Crowd. Similar to the emerging Machine-to-Machine communication, this Machine-to-Crowd, respectively Human-to-Crowd communication also shows different traffic patterns which have to be explored in the future.

3 Implications of Crowdsourcing on Internet Traffic

Already today, Crowdsourcing platforms accumulate hundreds of thousands of users and process an enormous amount of tasks. Amazon Mechanical Turk (MTurk)[11] reported 500.000 registered users in 2011, Microworkers[12] about 380.000 users end of 2012. In December 2012, the Crowdsourcing Platform Crowdflower stated to have completed about 770.000.000 individual micro tasks since 2007.

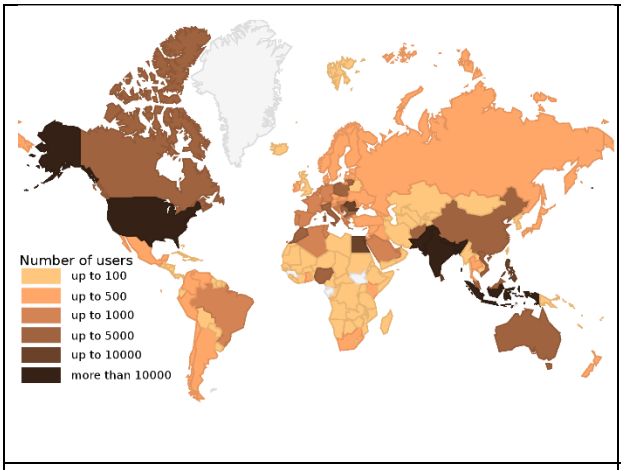
The pure amount of their users and the current growth rate of Crowdsoucing platforms and services make them emerging traffic hotpots in the near future. One example is Google's reCAPTCHA[13] service, which helps to digitalize books by using Internet users to transcribe images. Even if the individual picture send to the users is just about 3KB, a daily traffic of over 90 GB is generated, as there are more than 30 million pictures processed per day. Other crowdsourcing platforms like Microworkers currently generate only relatively small amounts of traffic, however due to the continuous growth of these platforms as for shown in Figure 1b), this will highly increase in the future.

Crowdsourcing will not only generate a large amount of additional traffic, but also leads to traffic distributions that are more difficult to handle. When new tasks are published, the human cloud workers are notified and often immediately want to begin their work. This results in flash-crowd traffic patterns when thousands of workers connect to the platform at the same time to retrieve their task [Hir11].

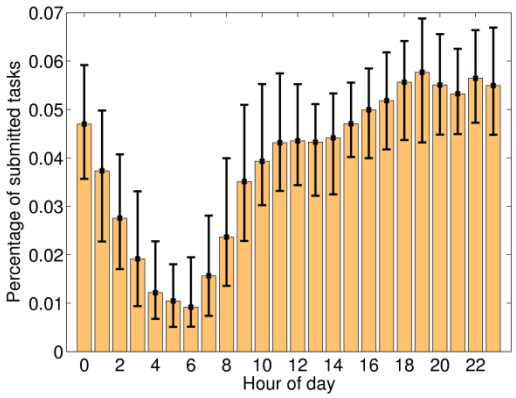
Crowdsourcing platform users are often distributed all over the world [Ros10], as shown in Figure 2a) depicting the distribution of the Microworkers-users around the globe. The majority of the employers origin from western countries, while most workers are from Asia [Hir11]. Even if crowdsourcing platform build an

[8] http://opensignal.com/
[9] http://www.rootmetrics.com/
[10] http://cloudfactory.com/
[11] http://mturk.com
[12] http://microworkers.com
[13] http://www.google.com/recaptcha

international 24-7 marketplace for work, the workers still stick to local time work shifts. This can be observed in Figure 2b), which depicts the activity in terms of how many tasks are submitted at each hour of the day, normalized by the total number of tasks. The bars depict the median worker activity for countries with more than 100 task submissions in total; the whiskers mark the 25%, respectively 75% quantile.



a) Number of Microworkers.com users per country



b) Average activity of crowdsourcing workers during one day

Figure 2 Crowdsourcing users follow daily diurnal work shifts and der globally distributed.

Due to the different time zones of the workers, the workload on the server is almost constant throughout the whole day. Even though a constant utilization of the hardware resources is generally preferable, this approach imposes unnecessary load on the global Internet, because the data is always send to the currently active worker region. This can be overcome exploiting the regular diurnal effects for designing an allocation scheme of resources in a world-wide infrastructure. This enables the platform provider to migrate the data closer to the currently active workers, reducing their access delay and simultaneously reducing the network load. Furthermore, a distributed infrastructure is also more scalable to account for future growth of the platforms.

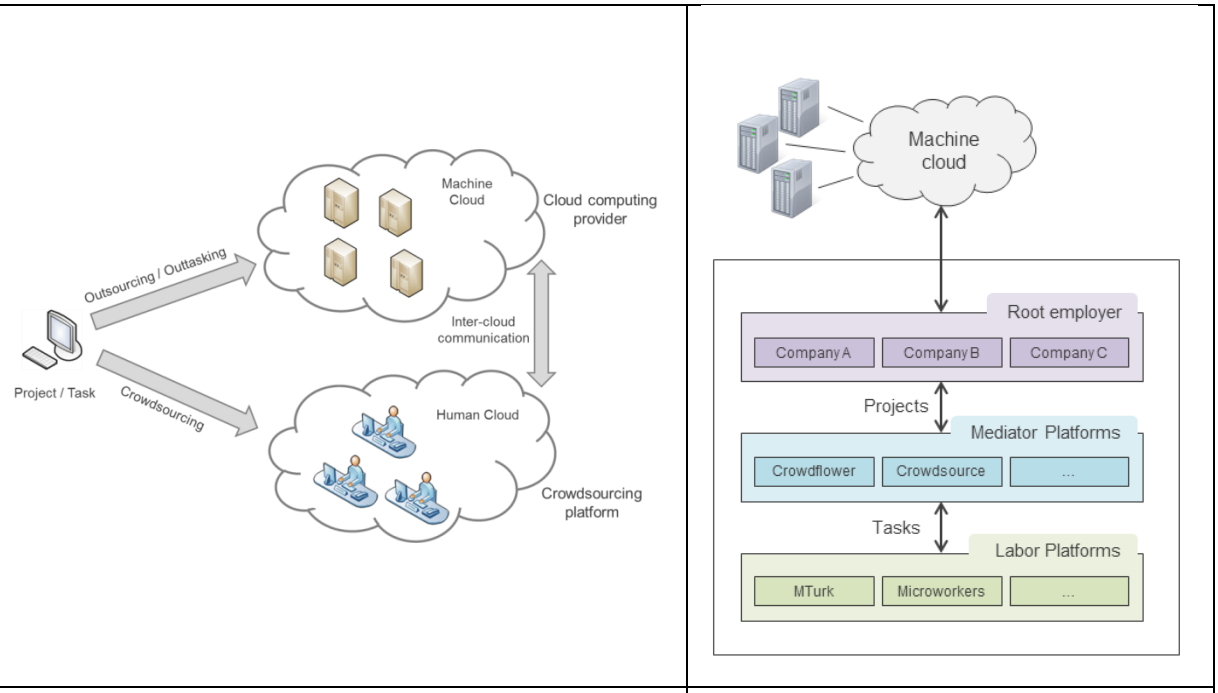
To develop, analyze, and compare new mechanisms that facilitate crowdsourcing, and to dimension future Internet infrastructures that must be able to handle the resulting traffic, it is necessary to develop new models which can realistically represent the crowdsourcing user behavior and resulting traffic patterns.

4 When Human Crowds Meet Machine Clouds

The utilization of machine clouds can help to solve the technical challenges like scalability and the global distribution of data. But Crowdsourcing can also be seen as an adaption of the cloud paradigm to human workforces. Similar to machine clouds, Crowdsourcing platforms offer an interface to access a huge easy-to-scale pool of work units which are abstracted to the user of the service. The interconnection of human and machine clouds in turn fosters the development or completely new services. It enables the automation of task which requires both, high computational effort and human judgments or interactions. One example of such a task is the previously mentioned reCAPTCHA service. Here, book digitalization is realized by a combined machine-human base approach. Texts are automatically scanned and afterwards process by an OCR software. If the software is not able to determine phrases in the text, these words are submitted to a web application, which uses humans to transcribe them. This combination allows a cost effective and high quality digitalization, as the majority of the work can be accomplished by a cheap and fast machine based component and additional quality is assured by human contribution, but which in turn takes more time.

In the Crowdsoucing environment exist similar abstraction layers like IaaS, PaaS, and SaaS in the machine cloud environment, shown in Figure 3b). Here we can differentiate between Labor Platforms which accumulate user and provide the actual workforce. These platforms do usually not provide specialized interfaces or workflows

for certain task or categories of tasks. This is usually done by Mediator Platforms. Mediator platforms are specialized platforms which focus on a specific kind of tasks, e.g. image labeling or text transcription. They offer means to decompose projects in their fields into ”crowdsource-able” microtasks and forward these microtasks to the actual labor platforms and their workers. The work itself is generated by root employers, which decompose their project to automatable tasks that are routed to machine clouds and tasks that require human interaction. These are submitted to the corresponding mediator for further decomposition.



a) Human and machine clouds are connected by APIs and exchange traffic across the Internet.

b) Single project split into several tasks conducted in different clouds.

Figure 3 Interconnection of Machine and Human Clouds.

A vision for the future is a ubiquitous Intercloud system that includes both machine clouds and human clouds as shown in Figure 3a). Human clouds can be for example the users of a social network or the workers of a crowdsourcing task. Tasks that can (nowadays) only be completed by humans are only special in the task itself, but most other challenges that come with cloud computing also apply to the human cloud. Currently, different initiatives and standardization organizations like the IEEE, OASIS or DMTF are on their way to standardize an Intercloud architecture: Each task, application, or service that is submitted to the Intercloud must include a detailed and machine-readable subtask-description [Ber09] including technical aspects like requirements, dependencies, or tasks to perform, as well as economical and management aspects like available funds, target QoE, or maximum time until completion. With this information, a special directory and mediation service in the Intercloud environment will take care of finding the matching clouds that are required to compose the new cloud-based application. It also initializes communication between the different clouds and the applications and handles authentication management. Another important task that will be handled by a central directory is the trust and reputation management between cloud services. This applies especially for different labor platforms which provide workforces with diverse skills. A neutral referee must keep track of the reliability and trustworthiness of the different clouds by collecting statistics and including them in the matchmaking for new service requests. Finally, the application must be deployed in the negotiated machine and human clouds.

All these steps pose numerous challenges for future research in cloud computing. In particular, existing open standards and protocols for cloud communications are required to facilitate the composition of services and inter-cloud communications [CMKS10]

An example for a system that implements important features for a future cloud architecture is the recently presented NetStitcher [Lao11], a system for stitching together unutilized bandwidth across different datacenters, using it to carry inter-datacenter bulk traffic for backup, replication, or data migration applications. The variety of often proprietary overlay applications makes it difficult to combine services from different cloud providers to a new service [Pap2011, Be09]. The SmartenIT project[14] addresses composed services from different clouds with new traffic management mechanisms to be influenced by economic criteria and standard requirements, e.g., bandwidth or QoE.

5 Challenges in Crowdsourcing

In order to use crowdsourcing to full capacity, mechanisms are required which allow to improve the current platforms. This includes on the one hand issues especially arising from the Crowdsourcing approach itself and, furthermore, technical and network related issues as mentioned above. In the following, various requirements to the platforms should to be addressed.

- **Recommendation systems**: Task should be recommended to workers according to the skills and interests of the workers. This improves the overall quality of the work. Such recommendation systems may be built by means of advanced machine learning approaches or folksonomy approaches. On the other hand, workers should be recommended for certain tasks because of their skills, reliability, etc. while still respecting the principles of crowdsourcing.
- **Anonymous user profiles** and **specialized crowds**: As a consequence, anonymous user profiles need to be developed which are only accessible by the platform provider. This way, the platform provider, acts as neutral mediator between worker and employer which has the interest to operate the platform successfully for all stakeholders involved. The creating of specialized crowds out of user profiles may speed up the completion time of campaigns while at the same time the quality improves. In this direction, the question arises how to derive technical mechanisms in such a platform to automatically create and evaluate the profiles depending on the existing tasks in the platform.
- **Automated task design**: To support the mechanisms mentioned above, an automated task design may be beneficial which allows tagging campaigns, leveraging the interaction with machine clouds, and automatically finding appropriate human processing units.
- **Incentive design**: The key factor of success of crowdsourcing systems is the user participation. Therefore, incentive schemes need to be developed to encourage user participation and high quality work. This includes the relation between rewards, completion time of campaigns, and quality of work. In this context, the evaluation of incentives, both monetary and non-monetary, is also an interesting question.
- **Quality assurance / Reliability**: The work accomplished by the human workers should be evaluated automatically to ensure a high quality. But because the work itself cannot be automated, an automatic result evaluation imposes problems, too. Thus, new quality control schemes have to be deployed. Especially in the context of mobile sensors, wrong sensing values may be obtained which have to be automatically detected. While it is easy to repeat the same task several times, this invokes some extra costs. Theoretical models for finding optimal but practical guidelines need to be derived.
- **Scalability:** Today, Crowdsourcing platforms accumulate hundreds of thousands of users. With a future growth of these platforms similar scalability issues arise like in current online social networks. However, in contrast to social networks, crowdsourcing platforms impose different requirement on the underlying systems as described before.
- **Global distribution of data**: Alike social networks, Crowdsourcing platform users are distributed all over the worlds, resulting in a need for global availability of the platforms data.

Besides the challenges for the crowdsourcing platform itself, many scientific research questions remain to be solved. As the field of crowdsourcing is novel, a common terminology, classification and taxonomy of crowdsourcing systems, as well as evaluation frameworks are required. In particular, the experiences from industry (as employers or platform operators) are an important input. To this end, measurements and databases, e.g. for automated detection of “bad” work or “unreliable” workers, are interesting in order to test novel methods in that area. We conclude this paper by sketching various open research questions :

[14] http://www.smartenit.eu

- What is an appropriate taxonomy for classifying and analyzing crowdsourcing systems? How can crowdsourcing tasks be grouped based on their task complexity and along key challenges in order to successfully harvest expertise of large human networks?
- Which use cases and applications will exploit the potential of crowdsourcing?
- How does the research community approach improve crowdsourcing mechanisms e.g. for quality and cost control or reliability of users and devices? Which requirements and challenges occur for particular operational conditions, like ubiquitous crowdsourcing due to the user mobility in time and space?
- How can incentive schemes be designed for coordinated problem solving of the crowd among individual humans with their own goals and interests? How to realize gamification of work for improved user engagement? How to identify expertise of users? How to implement such incentive schemes technically?
- How can the experiment and task design be standardized? Which kinds of APIs or templates are promising and useful in practice?
- What are the objectives to be fulfilled and the necessary capabilities of platforms towards the provision of Future Internet services built on top of crowdsourcing facilities?
- How can crowdsourcing systems be evaluated? Which common research methodologies are applicable? Which theories and models from a number various fields are applicable, including artificial intelligence, multi-agent systems, game theory, operations research, or human-computer interaction? How to include human-centric measures such as costs, availability, dependability and usability, including device-specific properties in evaluation frameworks?
- How does the research agenda for crowdsourcing look like in the next years?

The understanding of crowdsourcing and its potential is currently at a very early stage. Many open research questions in different areas must yet be answered to reveal the true effects of and new solutions for a crowdsourcing-assisted social Internet.

6 References

[Ber09] D. Bernstein, E. Ludvigson, K. Sankar, S. Diamond, M. Morrow, Blueprint for the Intercloud - Protocols and Formats for Cloud Computing Interoperability, ICIW '09, pp. 328-336, 2009

[Ber10] M. Bernstein, R. Miller G. Little, M. Ackermann, B. Hartmann, D. Karger, K. Panovich: “Soylent: A Word Processor with a Crowd Inside”. Symposium on User Interface Software and Technology, New York, USA, 2010.

[Ber11] M. Bernstein, J. Brandt, R. Miller, D. Karger: “*Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces*”. Symposium on User Interface Software and Technology, St Andrews, UK, 2011.

[Cho10] D. Choffnes, F. Bustamante, Z. Ge: “Crowdsourcing Service-Level Network Event Monitoring”. SIGCOMM, New Delhi, India, 2010.

[Dha12] M. Dhawan, J. Samuel, R. Teixeira, C. Kreibich, M. Allman, N. Weaver, V. Paxson: “Fathom: A Browser-Based Network Measurement Platform”. Internet Measurement Conference, Boston, USA, 2012.

[Fag12] A. Faggiani, E. Gregori, L. Lenzini, S. Mainardi, A. Vecchio: “*On the Feasibility of Measuring the Internet Through Smartphone-Based Crowdsourcing”*. Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, Paderborn, Germany, 2012.

[Gem12] A. Gember, A. Akella, J. Pang, A. Varshavsky, R. Caceres: “Obtaining In-Context Measurements of Cellular Network Performance”. Internet Measurement Conference, Boston, USA, 2012.

[Hir11] M. Hirth, T. Hoßfeld, P. Tran-Gia: “*Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com*”. Workshop on Future Internet and Next Generation Networks, Seoul, Korea, 2011.

[Hoß11] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia: ”*Quantification of YouTube QoE via Crowdsourcing*”. Workshop on Multimedia QoE - Modeling, Evaluation, and Directions, Dana Point, USA, 2011.

[Lao11] N. Laoutaris, M. Sirivianos, X. Yang, P. Rodriguez: "Inter-Datacenter Bulk Transfers with NetStitcher". ACM SIGCOMM 2011, Toronto, ON, Canada, 2011

[Mai09] N. Maisonneuve, M. Stevens, M. Niessen, L. Steels: “*NoiseTube: Measuring and mapping noise pollution with mobile phones*”. ICSC Symposium, Thessaloniki, Greece, May, 2009.

[Ros10] J. Ross, L. Irani, M. Silberman, A. Zaldivar, B. Tomlinson. “*Who are the Crowdworkers? Shifting Demographics in Mechanical Turk*”. Conference on Human Factors in Computing Systems, Atlanta, USA, 2010.

[Vuk11] M. Vukovic, J. Laredo, S. Rajagopal: “*Challenges and Experiences in Deploying Enterprise Crowdsourcing Service*”. Conference on Web Engineering, Vienna, Austria, 2010.