# Mining Reliable Information from Passively and Actively Crowdsourced Data

Jing Gao[1], Qi Li[1], Bo Zhao[2], Wei Fan[3], and Jiawei Han[4]
[1]SUNY Buffalo, Buffalo, NY USA
[2]LinkedIn, Mountain View, CA USA
[3]Baidu Research Big Data Lab, Sunnyvale, CA USA
[4]University of Illinois at Urbana-Champaign, Urbana, IL USA
{jing,qli22}@buffalo.edu, bozhao@linkedin.com, fanwei03@baidu.com, hanj@illinois.edu

## ABSTRACT

Recent years have witnessed an astonishing growth of crowd-contributed data, which has become a powerful information source that covers almost every aspect of our lives. This big treasure trove of information has fundamentally changed the ways in which we learn about our world. Crowdsourcing has attracted considerable attentions with various approaches developed to utilize these enormous crowdsourced data from different perspectives. From the data collection perspective, crowdsourced data can be divided into two types: "passively" crowdsourced data and "actively" crowdsourced data; from task perspective, crowdsourcing research includes information aggregation, budget allocation, worker incentive mechanism, etc. To answer the need of a systematic introduction of the field and comparison of the techniques, we will present an organized picture on crowdsourcing methods in this tutorial. The covered topics will be interested for both advanced researchers and beginners in this field.

## 1. INTRODUCTION

The crowd-contributed data have become a powerful information source that covers almost every aspect of our lives, including traffic conditions, environmental conditions, health, public events, and many others. Traditionally, such information is provided only by specialized sources such as authoritative agencies, professional media, and expensive sensing devices, which might not update frequently or provide complete coverage. With the proliferation of mobile devices and social media platforms, now any person can publicize his observations about any activities, events or objects anywhere and at any time. The confluence of these enormous crowdsourced data can contribute to an inexpensive, sustainable and large-scale decision system that has never been possible before. To build such a system, many research studies have been deployed from various perspectives of crowdsourcing, such as aggregation, budget allocation, and task allocation. In this tutorial, we will give a full picture of the state-of-the-art crowdsourcing researches. We will focus on two types of crowdsourcing, referred to as "passive" crowdsourcing and "active" crowdsourcing. In passive crowdsourcing, users are sharing what they observe and experience, typically via social media platforms, discussion forums and smartphone apps. These platforms are usually serving general-purpose information sharing, but we can extract relevant information regarding a specific task (e.g., traffic, environment, or drug effect monitoring) from such platforms. On the other hand, actively crowdsourced data usually come from the platforms and apps that are designed to actively solicit users' reports and answers for specific tasks, such as Amazon Mechanical Turk[1] (mTurk) and CrowdFlower[2]. In this tutorial, we will introduce and compare approaches to handle both types of crowdsourced data for the tasks including data aggregation, budget allocation, etc.



*KDD '16 August 13-17, 2016, San Francisco, CA, USA*



ACM ISBN 978-1-4503-4232-2/16/08.
DOI: http://dx.doi.org/10.1145/2939672.2945389

## 2. INTENDED AUDIENCE AND PREREQUISITES

This tutorial is intended for researchers and practitioners in data mining. While the audience with a good background on data mining and algorithms would benefit most from this tutorial, we believe the material to be presented would give general audience and newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more.

## 3. PRESENTERS

**Jing Gao** is currently an assistant professor in the Department of Computer Science, University at Buffalo (UB), State University of New York. She received her Ph.D. from Department of Computer Science, University of Illinois at Urbana-Champaign in 2011, and subsequently joined UB in 2012. She is broadly interested in data and information analysis with a focus on truth discovery, information integration, ensemble methods, mining data streams, transfer learning and anomaly detection. She has published more than 60 papers in referred journals and conferences and her work has received over 1800 citations. She has served as program committee member of many conferences including KDD, ICDM, SDM, ECML/PKDD and CIKM. She is the recipient of NSF CAREER award and IBM faculty award.

[1]https://www.mturk.com/mturk/
[2]http://www.crowdflower.com/

**Qi Li** received the BS degree in Mathematics from Xidian University and the MS degree in Statistics from University of Illinois at Urbana-Champaign, in 2010 and 2012 respectively. She is currently working toward the Ph.D. degree in the Department of Computer Science, University at Buffalo. Her research interest includes truth discovery, data aggregation and crowdsourcing. She has published papers on these topics in SIGMOD, VLDB, KDD, and WSDM.

**Bo Zhao** is a senior engineer at LinkedIn, before which he was a researcher at Microsoft Research Silicon Valley, prior to which he received his Ph.D. from University of Illinois at Urbana-Champaign. His research interests include truth discovery, data integration, knowledge bases, crowdsourcing, and more recently recommender systems.

**Wei Fan** is currently the senior director and deputy head of Baidu Big Data Lab in Sunnyvale, California. He received his PhD in Computer Science from Columbia University in 2001. His main research interests and experiences are in various areas of data mining and database systems, such as, deep learning, stream computing, high performance computing, bioinformatics, social network analysis, novel applications and commercial data mining systems. His co-authored paper received ICDM06/KDD11/KDD12/KDD13/KDD97 Best Paper & Best Paper Runner-up Awards. He led the team that used his Random Decision Tree (www.dice.com) method to win 2008 ICDM Data Mining Cup Championship. He received 2010 IBM Outstanding Technical Achievement Award for his contribution to IBM Infosphere Streams. He is the associate editor of ACM Transaction on Knowledge Discovery and Data Mining (TKDD). During his times as the Associate Director in Huawei NoahâĂŹs Ark Lab in Hong Kong from Aug. 2012 to Dec. 2014, he has led his colleagues to develop Huawei StreamSMART a streaming platform for online and real-time processing, query and mining of very fast streaming data. StreamSMART is 3 to 5 times faster than STORM and 10 times faster than SparkStreaming, and was used in Beijing Telecom, Saudi Arabia STC, Norway Telenor and a few other mobile carriers in Asia. Since joining Baidu Big Data Lab, Wei has been working on medical and healthcare research and applications, such as deep learning-based disease diagnosis based on NLP input as well as medical dialogue robot.

**Jiawei Han**, Abel Bliss Professor, Department of Computer Science, University of Illinois at Urbana-Champaign. His research areas encompass data mining, data warehousing, information network analysis, etc., with over 600 conference and journal publications. He is Fellow of ACM, Fellow of IEEE, the Director of IPAN, supported by Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and the Director of KnowEnG: a Knowledge Engine for Genomics, one of the NIH supported Big Data to Knowledge (BD2K) Centers.

# 4. ACKNOWLEDGMENTS

The work was supported by the National Science Foundation under Grant NSF 1553411 and 1319973. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.