2020-09-22

# A Team Composition Approach For Social Crowdsourcing Communities

Zaamout, Khobaib

UNIVERSITY OF CALGARY

# A Team Composition Approach For Social Crowdsourcing Communities

by

Khobaib Zaamout

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

SEPTEMBER, 2020

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled “A Team Composition Approach For Social Crowdsourcing Communities” submitted by Khobaib Zaamout in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

# Abstract

This research takes place in an emerging paradigm of social computation that we name *social crowdsourcing communities* (SCCs). These are moderated online communities where members participate in collaborative activities (*i.e.* queries) designed to elicit their opinions concerning some topics, products, or services. This paradigm is distinct in that it combines the powers of crowdsourcing and social networking (SN) to allow for systematic querying of crowds and synthesizing response data (*i.e.* contributions) into coherent reports for decision-makers. SCCs consist of a beneficiary (*i.e.* the operators, the moderators, the analysts, and the organization that benefits from the reports), queries, a working crowd, and a platform where all activities occur.

We show that it is possible to apply methods and techniques from existing fields to alleviate many of their challenges. One of these challenges is improving teamwork outcomes (*i.e.* contribution quality). Currently, SCC members, who are interested in a specific task, self-assemble into teams without considering any factors that may cause them to exhibit lower levels of productivity, participation, and contribution quality. The growth and query frequency restrictions imposed on these platforms by their operators to control operation costs further exacerbate this challenge.

This thesis demonstrates how member behaviour can guide team formations and identifies specific behavioural characteristics related to improved team performance through an exploratory case study. It accomplishes this goal by capturing member behaviour in a model and using it to explore the compositions of existing teams. In doing so, this thesis identifies the specific compositions associated with increased team performance. The outcomes indicate the validity of this approach and provide a strong foundation for further investigation.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to Professor Ken Barker for his guidance, support and wise counsel throughout my Ph.D. program. I have learned a lot from him; for that, I will always be in his debt.

I also want to thank Professor Güenther Ruhe, Professor Anthony Tang, Professor Thomas O'Neill and Professor Carson Leung for their insightful feedback on my research project.

I would also like to give Professor Carey Williamson and Dr. Robert Collier special thanks for their numerous and valuable advice that helped me navigate many challenges during my Ph.D. program.

Additionally, I would like to thank Chaordix[1] for their valuable collaboration and for providing the data and computational resources that made this thesis possible.

Finally, I would like to thank my wife, Fahima, for her patience, care and encouragement throughout my studies, my parents for all their help and support, and my friends Tom Arjannikov, Maryam Majedi and Muhammad Gaafar for their wonderful companionship during my studies.

Thank you all; this work would not be possible without your support.

[1] https://www.chaordix.com/

# Table of Contents



[2]The results in this chapter are published [1]

[3]The results in this chapter are published [2].
[4]The results in this chapter are published [3]
[5]The results in this chapter are published [3]

# List of Tables

# List of Figures and Illustrations

# List of Symbols, Abbreviations and Nomenclature

| Symbol | Definition |
|---|---|
| U of C | University of Calgary |
| SCC | Social Crowdsourcing Community |
| FA | Factor Analysis |
| EFA | Exploratory Factor Analysis |
| CFA | Confirmatory Factor Analysis |
| FB | Factors of Behaviour |
| SN | Social Network |
| WWW | World Wide Web |
| KS | Kolmogorov-Smirnov |

# Chapter 1

# INTRODUCTION

Involving consumers in an organization's decision-making processes is an essential and well-established tradition in modern organizations [7]. This involvement is immensely beneficial to the organization because it helps identify and understand its target market, deliver relevant products/services to its target market, and anticipate future product/service ideas. Traditionally, organizations involve their consumers using approaches, such as focus groups, surveys, and phone interviews. Today's advanced communication technology allows organizations to engage with more consumers from geographically dispersed locations than ever before using approaches, such as Social Crowdsourcing Communities (SCCs).

SCCs are online communities that organizations create, operate, moderate, and analyze to engage with their consumers for purposes, such as eliciting their opinions concerning some topics, products, or services, and generating ideas for new products or services. SCCs allow organizations to pose questions (*i.e.* queries) to members who in turn collaborate in undertaking these queries by responding to them directly or discussing other members' responses in a discussion-thread fashion. Moderators take part in these discussions to steer them towards answering the queries, and analysts process the discussion data (*i.e.* content) to synthesize reports that address the queries (Chapter 3 provides more details about SCCs).

The SCC operators maintain a specific member-to-moderator ratio to sustain adequate levels of engagement. They also control the number of queries to manage the workload of their data analysts. Thus, an unbound growth of their member-base requires an increase in the number of moderators and data analysts. Since both moderators and analysts are hired and trained professionals, this increase adds to operation costs. In other words, the larger the number of members that undertake queries and the higher the query frequency, the higher the moderation and analysis

efforts, and thus, the operation costs. Therefore, operators tend to constrain the growth of these communities, pose arbitrary constraints on the number of members that can participate in a query (*i.e.* restrict team size), and constrain the frequency at which queries are pushed onto their communities to control their operation costs. These restrictions limit the effectiveness and potential of SCCs.

A SCC "team" is a group of members who voluntarily respond to a query and discuss existing responses. These discussions are the primary mechanism through which members can perfect responses and build on them, which leads to team-like behaviour. Section 7.2.1 discussion SCC teams in more details.

Constraining team size without considering team composition leads to self-assembled teams that are likely to exhibit higher levels of conflict and lower levels of cohesion, productivity, and participation [8, 9], which ultimately affects the quality of their contributions[1]. Therefore, this thesis presents a study that quantifies member behaviour into factors and identifies the behavioural composition of effective SCC teams that produce high-quality contributions regardless of team size. We show that teams with particular behavioural compositions are likely to produce high-quality contributions. This approach shifts the emphasis from constraining teams based purely on arbitrary size constraints to constraining teams based on the behavioural factors of its members. In this way, moderators can compose teams of the desired size that are also likely to produce high-quality contributions.

## 1.1 Assumptions

This research is rooted in the assumption that an individual's behaviour is the key to improved teamwork. This assumption is predicated on the theory that individuals' (online) behaviour indicates many aspects of their lives [10, 11, 12, 13], including their personality [14, 15], and that personality is a viable guide for forming high performing teams [16, 8, 17, 18]. Furthermore, since

[1]This thesis defines a team as a group of SCC members who voluntarily participate in a project. Section 7.2 discusses this definition in more details.

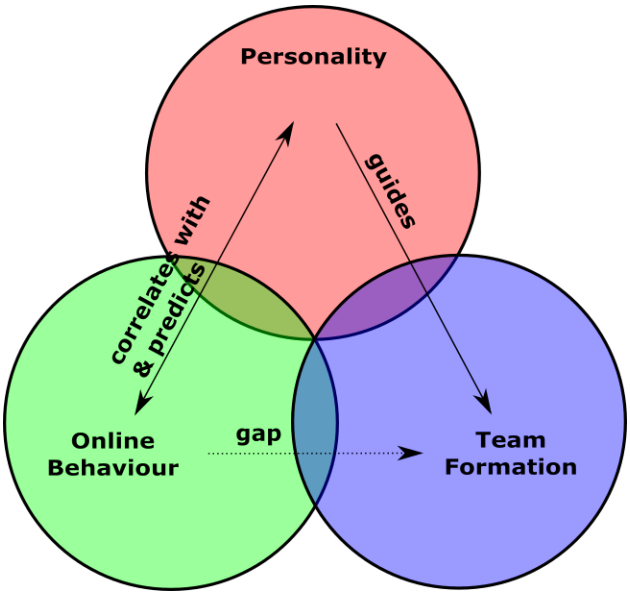



Figure 1.1: Illustration of the research gap this thesis aims to span.

online behaviour correlates with and predicts personality [19, 20], then online behaviour can guide team formation for improved work outcomes (see Figure 1.1).

A few researchers have tackled this research gap, providing further evidence of this assumption. For example, a study conducted on 144 students in 30 self-assembled teams working on a semester-long project concludes that the emergence of highly central members in teams' advice networks, leadership networks, obligation networks, and social networks is critical to their performance. It also finds that a team with a single, highly central member performs better [9]. Another study, conducted on 499 employees in 76 teams, reveals a positive correlation between a team's interaction network density and the quality of resulting projects. [21]. Another contribution [22] argues that socially connected individuals with highly dense local clusters form the majority of entrepreneur teams. They show that composing teams of such individuals restrict creativity, autonomy, and innovation. They further argue that teams should contain members of the same network who are globally prestigious and act as bridges among local clusters. This argument is consistent with previous claims made by others [23, 24].

## 1.2 Related Work and Consideration

Forming teams to improve work quality is a thoroughly-investigated area of research. There are many approaches and various perspectives for tackling this problem. They fall into three broad categories: *competence*, *psychometrics*, and *others* [22]. The competence category consists of approaches that suggest forming teams around the skill set that a project requires. The psychometrics category includes approaches that suggest forming teams around psychometrics, such as personality models, and behavioural tendencies to ensure compatibility. The other category consists of a variety of perspectives that emerged from non-traditional areas of research, such as social networking, and game theory due to the rise of online work environments [25, 9, 21, 22, 26, 27]. Many organizations use approaches from all these categories to ensure their new hires fit their needs.

SCC moderators often screen prospective members before they allow them to join the community or undertake a particular query. This process ensures that a potential member has the required knowledge of the topics, products, or services. Moreover, the queries in this thesis's SCCs inquire about members' opinions and require only basic communication skills. The screening process and the basic skill requirements make the competence approaches unnecessary.

The psychometric approaches, precisely the personality approach, initially seem suitable for SCCs because it addresses improving work quality through member compatibility. However, closer inspection of this approach reveals several challenges in using it as a guide in SCC team formation. First, moderators must administer an extensive survey to each member to obtain their personality traits. Administering such surveys is a costly and time-consuming process, especially when considering the relatively large number of members in SCCs. Secondly, considering that personalities evolve [28], these surveys must be administered multiple times to track these changes, which in the light of the first point further challenges this approach. Thirdly, many individuals have unfavourable views about participating in surveys that uncover private and sensitive information about themselves to others [29, 30]. These views may lead them to report incorrect information about themselves [31], which is a phenomenon known as *self-reporting bias*, or lead them to re-

frain from participating in the survey [29]. Finally, the personality approach to team formation does not apply to short-duration projects [32], which is the primary duration for projects in SCCs.

## 1.3 Scientific Question

Section 1.1 establishes the reasoning upon which this thesis predicates the assumption that an individual's online behaviour can guide the formation of high-performance teams and positions this research in the broader context of existing work.

With this in mind, this thesis's scientific question becomes:

*How do SCC member behaviour guide team formation and what specific behavioural characteristics can, individually or collectively, guide SCC team compositions for improved team performance?*

To answer this question, we carry out a quantitative exploratory case study using historical logs of member interactions to arrive at descriptive team-level behavioural characteristics related to team performance. Specifically, following the methodology adopted by personality-based approaches, this thesis derives a model (namely, *Factors of Behaviour* (FB) model) from member behaviour logs. It then uses this model to identify existing compositions of high performing teams, resulting in an approach that falls under alternative methods for team formation [25, 9, 21, 22, 26].

Since this approach uses logs of member behaviour instead of surveys, it sidesteps the challenges of the personality approach that Section 1.2 identifies. Specifically, due to the relatively low cost and high speed of obtaining and tracking member behaviour in SCCs, it is cheaper and quicker to obtain the FB than personality traits and track their changes over time. Furthermore, capturing and using member behaviours happen under the "Term of Service" agreement that members agree to upon signing up to a SCC. Therefore, this approach sidesteps the situations where some members refrain from participation in surveys. Moreover, member behaviour logs are factual records that are not prone to self-reporting bias. Therefore, this approach is more scalable than the personality approach and more cost and time effective.

However, this approach's dependence on logs makes it exhibit what is commonly known as the *cold-start* problem [33]. In this case, the cold-start problem means that it is not possible to derive FB for new members because they have little or no activity. We eliminate these members to mitigate the effect of this problem on our analysis.

## 1.4 Methodology

This thesis performs five steps to answer the scientific question. First, it defines SCCs and describes their common and unique characteristics compared to existing crowdsourcing and social networking systems. Second, it employs social network analysis methods to analyze the network structure generated from several SCCs' member interactions and associations (*i.e.* interaction and association networks) to identify common and unique patterns in these communities in comparison to existing social networking systems. Third, it employs various methods for an in-depth analysis of the SCC members' interactions to identify member interaction patterns, how these patterns evolve, and the factors that affect them. Fourthly, using the previously identified patterns, this thesis captures the member behaviour into variables then use Factor Analysis (FA) [34] to derive the FB model. It then analyzes this model's factors, names them, and explains them based on the findings from the previous steps. Finally, it uses FB to obtain team-level behavioural characteristics and identifies the relationship between specific team compositions and team performance.

The first step starts by coining the term SCC. It details the components that make up SCCs and the standard process they follow. It also identifies the conceptual commonalities and differences between SCCs and existing crowdsourcing and SN systems and provides examples to support them. The second step analyzes several SCCs and SNs to identify how these conceptual commonalities and differences manifest in these systems' member behaviour. For each SCC and SN, it extracts an interaction network from its members' interactions and uses social network analysis to analyze its structure and capture its properties. It compares these networks, identifies common and distinguishing structural properties, and hypothesizes the member behavioural patterns among

SCC members that underpin these common and distinguishing structural properties. The third step builds upon the previous steps' findings by analyzing the interactions among member pairs in the largest SCC (*i.e. HCC*). This analysis identifies specific member interaction patterns that gave rise to the previously identified structural properties. It does so by analyzing member responsiveness, interaction rate decay over time, and duration of interactions and assessing the effects of certain SCC features on these interactions. It also examines the volatility and evolution of the interaction network over time and arrives at the specific behavioural patterns underpinning its stability.

This thesis presents a process for capturing behaviours in variables and aggregating them into factors to accomplish the fourth step. It determines a set of relevant member behaviours that capture the common and distinguishing SCC interaction patterns that the previous steps identify and counts their occurrences for each member in *HCC*, creating a dataset. By performing FA on this dataset, it derives the FB model factors, names these factors, analyzes them, and discusses their relation to previous findings. The thesis uses the FB model to obtain member FB scores for the fifth step. Then for each response to a query (*i.e.* project) in *HCC*, it calculates its team-level factor scores by aggregating all contributing members' scores using various aggregation methods. It then calculates this project's quality as the sum of endorsements all textual contributions within this project received from none team members. It then shows the relationship between the team-level factor scores and project quality scores using correlation and machine learning and identifies specific team compositions associated with high performing teams using hypothesis testing techniques.

## 1.5 Thesis Contributions and Outline

This thesis's main contribution is in describing a novel strategy that utilizes member behaviour to form high-performance SCC teams and identifying specific team compositions of such teams. It arrives at this contribution by identifying the conceptual and practical commonalities of SCCs and their distinguishing characteristics, deriving a member behavioural model that captures these

aspects, and using this model to analyze existing team formations.

Specifically, it answers the scientific question by demonstrating how member behaviour can guide team formation through capturing relevant member behaviour into a behavioural model (*i.e.* the FB model) using FA and using this model to identify the behavioural makeup of high-performing teams. It captures member behaviour in three dimensions: *Impact*, *Activity*, and *Policing/Rowdiness*, which represent the endorsements members receive, the amount of work they perform, and their involvement in policing or rowdy behaviour, respectively. It also identifies that teams whose collective activity scores are diverse and has a member with a substantially higher impact score than the rest of the team tend to outperform teams that lack this composition.

The following is an outline of this thesis's chapters and their relevance to the research.

- Chapter 2 provides detailed background information necessary for understanding this thesis.

- Chapter 3 presents a comprehensive definition of SCCs and introduces them as a new research frontier. It also details their process and components and distinguishes them from SNs. This chapter provides a conceptual framework that guides the remaining chapters' analyses and supports their outcomes.

- Chapter 4 presents a thorough investigation of the structural makeup of several SCC interaction networks and a comparative structural analysis between SCCs, SNs, and the WWW. This investigation results in describing common structural properties among SCCs that results from the interactions among SCC members and outlines common and distinguishing structural properties between SCCs on one side and SNs and the WWW on the other. It relates these findings to the conceptual differences between SCCs and SNs, informing the subsequent chapters' analyses.

- Chapter 5 presents a detailed analysis of the interactions among member pairs in an SCC and members of a popular SN platform. This analysis identifies specific

interaction patterns that underpin the conceptual differences between SCCs and SNs. It also describes the evolution of these patterns and identifies the factors that affect them. This chapter's findings explain the mechanisms that give rise to SCCs' structural properties and provide insights for the following analysis.

- Chapter 6 builds upon the previous chapters' findings and presents a generic process for quantifying *HCC* member behaviour into factors, and interprets, analyzes and names these factors. These factors are the FB model that forms the foundation of the subsequent analysis.

- Chapter 7 presents an exploratory case study of existing teams' FB compositions and their performance in *HCC*. It defines projects, project quality and teams, and sets up explicit inclusion/exclusion criteria for team members and projects. It derives member FB scores, aggregates each team's member FB score to team-level using several aggregation methods, and uses project quality as a proxy for team performance to create a new dataset. This chapter uses this dataset to demonstrate the correlative relationship between team-level aggregates and project quality and the interactions among these factor aggregates, which results in several hypotheses regarding specific compositions' role in team performance. This chapter demonstrates this role by comparing the performance of teams with a specific composition against others using an appropriate hypothesis testing method. It also demonstrates the utility of the team-level FB aggregates in predicting team performance.

- Chapter 8 concludes this thesis by providing an overview of the key results and their practical implications, discussing the limitations of the methodology and results, and identifying the avenues for future work.

This chapter overviews this thesis's research work. It describes the research environment, the

research problem, and the assumptions of this thesis. It also introduces the scientific question and situates it in the context of existing work. It breaks down the process of answering the scientific questions and details the methodology of each step. It also provides an overview of this thesis's contributions relative to each chapter. The next chapter provides the background information regarding concepts used in this thesis.

# Chapter 2

# BACKGROUND

This chapter provides detailed background information concerning the concepts used in this thesis. It answers the following questions:

- What is a social network?
- What is crowdsourcing?
- What is FA?

## 2.1 Social Network

In 1735, Euler [35] proved the insolvability of the Seven Bridges of Königsberg problem by using graphs (*i.e.* networks). In his problem formulation, representation, and analysis, he laid the foundations of graph theory and the concept of topology. However, it was not until the turn of the twentieth century that the study of graphs and networks was applied to social and other scientific disciplines, which placed more emphasis on using networks for modelling problems and exploring hidden relations and meanings. The first disciplines to pick up the concept of graphs for modelling and simulation were social psychology, sociology, and anthropology [36, 37]. Scientists in these areas coined the term *social networks* and studied interactions of vertices within networks rather than network topology.

A social network (SN) is a graph made up of individuals, represented by vertices or nodes that connect by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, or relationships of belief, knowledge, or prestige [38]. Researchers from the fields of sociology, social psychology, and anthropology sparked the interest in studying SNs [39, 40, 41, 42, 43, 44]. Tönnies (1887), a social psychologist, discussed this as a topic of

interest in his book *Community and Society* [39]. A prominent psychiatrist (Moreno) and psychologist (Jennings) carried out one of the first SN analysis research efforts on a group of prison inmates [40] in 1932 and 1934, respectively. Later, Moreno produced another work on social network analysis by studying the social network of a reform school [41].

Other work in this area carried out at the same time [42, 43, 44] gave a strong push for mathematicians, physicists, computer scientists, and scientists from other disciplines to get involved in formalizing and establishing this field as a multidisciplinary research area [36, 37].

The work of Watts and Strogatz [45], and Barabási and Albert [46] materialized this interest when they studied properties of various networks and deduced important descriptive measures. Work in this area continues to flourish and branches into many fields, paving the way for businesses and enterprises to make use of it.

### 2.1.1 Social networking platforms

Online social networking platforms, or SNs, are websites and applications dedicated to networking between people, either directly by forming friendships or following news, or indirectly by sharing media and blog posts. In recent years, many of these platforms have emerged as influential players on the Web. Facebook, Twitter, LinkedIn, YouTube, Orkut, and Live Journal are among the most popular social networking platforms. Members can effortlessly create an online presence on these platforms and engage in social and individualistic activities, such as connecting with others, viewing, and sharing content and gaming.

## 2.2 Crowdsourcing

The Internet allows people from dispersed geographical locations to interact, share ideas, and influence each other. This widespread information exchange and efforts to harness it produces a new form of collective intelligence [47, 48, 49, 50] and co-creation [51], namely *crowdsourcing*. Crowdsourcing has emerged as a model of collective intelligence and a practical method

of co-creation over the Internet [52]. It is an online collaborative model in which an organization proposes tasks over the Internet to a large group of individuals who separately or collectively undertake these tasks [53, 54]. Crowdsourcing has both advantages [55, 56]. Most notably, the quality of work produced by the crowd is comparable and sometimes better than the quality of work produced by professionals [57, 58].

It is challenging to tie crowdsourcing to a specific science or area of research, or to claim a particular work or researcher is the main reason behind its existence. In the following, we note some significant milestones that helped establish this concept.

In 1785, Marquis de Condorcet, a French philosopher, and mathematician produced an essay [59] that argues favourably for collective intelligence. Although he did not coin the term "crowdsourcing", his Condorcet's jury theorem shows that the collective decision reached by a jury is more likely to be correct than that of an individual. He further argues that the likelihood increases with the increase in the number of members of this jury. In 1979, Hofstadter's book [60] had a profound impact on many sciences, including social and computer sciences. In his book, he shed light on the idea of universal conscious by addressing the concept of the *self* and how it can form [61], which ignited interest in understanding collaborative intelligence as well as the emergence of learning and human-like intelligence from interactions.

Similarly, the concepts of *noosphere* by Vernadsky [62] and *world brain* by Wells [63] described a vision of a universal information collection and collaboration that yields superior intelligence and ultimately shapes individuals to appear as a single superorganism. Other significant contributions have come from the work of Russell [64], Atlee [48, 49], and Lévy [50] helped establish the concept of collective intelligence and defined it as *"intelligence that emerges from the collaboration and competition of many individuals."* In 2004, the term *crowdsourcing* emerged as a model of collective intelligence.

Crowdsourcing is a relatively new concept that has become practical due to advancements in electronic communication [54, 57]. Howe coined the term crowdsourcing in his 2006 Wired maga-

zine article The Rise of Crowdsourcing [65]. He proposes it to describe the process of discovering ways to tap into and utilize the hidden powers of the crowd. Howe defines crowdsourcing 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 [52].

Many researchers have defined crowdsourcing [66, 67, 68]. A study by Estellés-Arolas *et al.* [53], similar in manner to the famous work of Tatarkiewicz who attempted to define the term *art* [69], arrived at a cohesive definition of crowdsourcing. Estellés-Arolas *et al.* [53] gathered 40 definitions of crowdsourcing from 32 popular publications, and undertook an analysis to identify the boundaries of the term and derived a comprehensive description:

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate in bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken [53].

One learns from all the definitions that crowdsourcing is a participant-driven, problem-solving, and creativity-producing model that utilizes current communication infrastructure to obtain solutions from a large and undefined group of people [70] as opposed to professionals [71] in exchange for micro-payments, social recognition, or entertainment value [72, 53].

### 2.2.1 Crowdsourcing components

A crowdsourcing system consists of four components: the *beneficiary*, the *query*, the *crowd*, and the *platform*. Beneficiaries are individuals or organizations that cultivate and moderate the crowd, measure and ensure its health and diversity, issue queries to the crowd and incentivize it, and collect responses and analyze them. The query is the task or collection of tasks that the beneficiary would like responses for. The crowd is a group of online users who undertake queries and give solutions to them. The platform is the medium that brings the all the components together by allowing them to carry out their parts [73, 74, 75], effectively and efficiently [76, 77].

Although these components are comprehensive, they are not unique to crowdsourcing systems and assuming so would wrongly classify some SN systems as crowdsourcing systems. Therefore, we note the following subtle differences. First, unlike the beneficiaries of SN systems, the beneficiaries of crowdsourcing systems control and moderate crowdsourcing projects to ensure they achieve specific goals. These goals can range from answering an overarching question through discussions to responding to a survey or a task through structured artifacts. The crowd's responses potentially fully or partially achieve the beneficiaries' goals. Secondly, individuals in the crowd may interact with each other on a crowdsourcing platform in a similar manner to members of any SN system. However, they must also respond to queries issued by the beneficiaries. Finally, crowdsourcing platforms support collaborative work, can distribute a large number of queries to any number of members and collect their responses, and offer various reputation and rewarding mechanisms based on the performance of members in responding to queries. Therefore, without a driving beneficiary, a working crowd, and an equipped platform, a given SN system is not a crowdsourcing system.

### 2.2.2 Functions of crowdsourcing

Crowdsourcing systems serve various functions [78, 70] in many domains [79, 54, 80]. Many authors categorize these functions in various ways, such as the following: *Labour*, *Funding*, *In-*

*novation*, *Research*, and *Marketing*. It is essential to understand the nature of queries on these platforms to be able to analyze them correctly and that current categorizations of the functions describe the "intention" of a given platform rather than function. Therefore, we revise the current categorization of crowdsourcing platforms as follows.

The nature of queries on crowdsourcing platforms falls on a spectrum. On one extreme are labour queries that are well-defined and accounted for by the beneficiary. These queries consist of small and independent tasks that require structured response artifacts, which are easy to integrate, track, and analyze. Examples of these queries are tagging images and translating tweets. On the other extreme are conversational queries that, unlike labour queries, are difficult to breakdown into small, well-defined tasks and are therefore not easily tracked or analyzed. The response artifacts of these queries are discussion threads that require analysis to synthesize into reports. Platforms that support labour queries offer features that allow the beneficiaries to represent query dependencies, distribute queries to members, track and display progress, and integrate responses. However, they have no features for member socializing or interactions. Platforms that support conversational tasks are usually social and therefore tend to form societies. Interactions among members of these communities produce digital footprints like those of social networking platforms. This thesis analyzes crowdsourcing platforms that form societies and refers to them as Social Crowdsourcing Communities (SCCs) (see Chapter 3).

### 2.2.3 Crowdsourcing process

A crowdsourcing query goes through five distinct stages before it is complete. These stages are: *Planning*, *Cultivation*, *Querying*, *Responding*, and *Reporting*. We describe each stage below:

1. *Planning:* This is the stage when the community moderators and beneficiaries plan all aspects of the query, including articulating its description, objectives, suitable response artifacts, performance measures, and tasks and task dependencies.

2. *Cultivation:* This is when the community moderators recruit, and train members on

handling a specific query, which involves various tasks, including:

- advertising the query to the crowd,
- screening, and accepting members interested in undertaking the query,
- training them in undertaking this type of query and familiarize them with its requirements,
- incentivizing them by clarifying performance measures and by using appropriate rewards mechanisms, such as points, badges, leaderboards, and giveaways, and
- policing the crowd by defining community rules and regulations, and by providing necessary tools for reporting of and handling rowdy behaviour.

3. *Querying:* This is when the moderators pose the query to the crowd, which also involves unloading a series of previously designed tasks to the crowd. Conversational queries require moderators to engage members in conversations to clarify responses, probe for more information, and steer discussions towards the objectives.

4. *Responding:* This is when the moderators, or the platform, collect and integrate responses from members. Depending on the response artifacts, the platform may automatically collect and integrate response, such as collecting the number of responses for each choice in a multiple-choice question, or responses may require work to collect and integrate, such as the case of answering open-ended questions through conversational artifacts. Collecting and integrating responses in the latter case requires the moderators to identify relevant and significant responses that are specific to the query and its objectives.

5. *Reporting:* This is when moderators (or analysts) analyze, summarize, and report

on the appropriate member responses. It also involves analyzing the performance of community in undertaking queries.

## 2.3 Factor Analysis

FA is a term that describes a suite of methods - analytical procedures and algorithms - that psychologists developed [34] to factorize and confirm factorization (also known as factor solution) of a given dataset. FA derives factors (called a factor solution) as linear combinations of input variables and offers statistical methods to assess the reliability of the factor solution. These factors give a summary of the original dataset, which makes understanding and interpreting relationships easier [81]. In other words, FA simplifies complex measured variables by deriving unmeasured "latent" variables (*i.e.* factors) that reflect the measured variables. The output of FA is a variable-factor loadings matrix ($VF$), which has the loading value each variable has on each factor. Variable loading is the amount that a variable contributes to a factor. Since a variable "loads" on all factors, (because factors are linear combinations of variables) these factors can be thought of as groups that variables belong to as defined by the variable loadings matrix. For simplicity, we say that a variable belongs to a factor on which it has the highest loading value.

Subjectivity plays a crucial role in obtaining and interpreting FA results. This subjectivity is necessary because it allows analysts to use their professional judgment and domain knowledge to tune the analysis process and interpret its results within the context of the research question. FA results are credible if the analyst reports all relevant information about the dataset, the analysis process, and the results to allow others to critique and reproduce them.

### 2.3.1 Factor analysis process

There are two FA approaches [82, 83]: Exploratory FA (EFA) and Confirmatory FA (CFA). EFA is an algorithmic procedure that summarizes a dataset by merging interrelated variables into factors. This process is named *factorization*, and it produces the factor solution [84, 82, 81]. CFA is a FA

procedure that verifies whether a given dataset supports a specific variable grouping (*i.e.* factor solution) through various statistical tests [84, 82, 81].

#### 2.3.1.1 Exploratory factor analysis

Exploratory Factor Analysis (EFA) consists of six steps: *gathering a dataset*, *choosing an EFA method*, *preparing the dataset*, *performing EFA*, *obtaining and interpreting a factor solution*, and *obtaining factor scores*.

1. *Gathering a dataset* is the accumulation of records (*i.e.* samples or observations) in sufficient quantities and with adequate diversity, so the analysis is possible, and the results are stable, credible, and generalizable. The dataset is sufficient if the ratio of observations to variables surpasses some threshold [82, 85]. There are many possible threshold values; the most conservative of which is an observation-to-variable ratio of 30 to 1 [81]. The dataset is an adequate sample if it contains diverse subjects, and if the variables measure various aspects of the problem. Lack of diversity in the data (*i.e.* homogeneity of the subjects) prevents factors from emerging [81].

2. *Choosing an EFA method* is the process of analyzing the dataset for specific characteristics that help determine a suitable method. The choice between the various EFA methods is complex [86, 85]. The argument for the preferred approach of EFA by analysts is contentious from various fields, including statistics [86, 85]. While some analysts suggest that all methods produce the same results when presented with a "sufficiently" large dataset [87], others believe that the differences between the various approaches are significant and suggest different approaches to differently distributed datasets [85, 88, 81].

3. *Preparing the dataset* is the process of cleaning and transforming the dataset in the ways needed by the chosen EFA method. This process involves checking for and

removing variables with multicollinearity (collinearity) and singularity [82, 81] by using Squared Multiple Correlation (SMC). Singular variables have a SMC value close to zero, and collinear variables have SMC value close to one. Removing a variable from the dataset affects the SMC scores of the others.

4. *Performing EFA* is the process of setting the chosen EFA method's required options, feeding it with the cleaned and transformed dataset, and obtaining results (*i.e.* factors) and factor loadings. This process is iterative, where the analyst checks the results for any anomalies that may require changing the method's options and rerunning it. Typically, analysts change two options: the number of factors the method will derive and the rotation method.

    - Deciding on the number of factors to derive is the most crucial decision the analyst must make. Deriving too many or too few factors has a direct effect on the correctness and interpretability of these factors. Many methods help analysts arrive at a number, and many heuristics help analysts decide on the quality of a factor solution. Kaiser's criterion [89, 86, 85], scree plot test [90], and parallel analysis [91] are popular methods for deciding the number of factors to derive. Kaiser's criterion suggests keeping all factors with an Eigenvalue greater than one [85], Scree plot test helps analysts decide the number of factors based on visual cues in the plot [90], and parallel analysis suggests keeping all factors that have an Eigenvalue larger than the average Eigenvalue obtained from a series of random data matrices [91]. Several heuristics assess the quality and feasibility of a factor solution [85, 86, 81]. These heuristics concern four areas: weak variable loadings, cross-loading, weak factors, and factor interpretability. A factor has weak variable loadings if the high-

est variable loading is less than 0.3 (some suggested 0.32 [92] for sample sizes larger than or equal to 300). Cross-loading refers to a variable that has a high loading value (>0.3) on two or more factors. Cross-loading may be acceptable if both factors have a high correlation with each other. A weak factor is a factor that has fewer than three variables that load highest on it. Some suggest that a factor that has two variables that load highest on it is acceptable if the loading values are high and the two variables are both highly correlated with each other and have low correlation with others [82]. In other words, if we change the number of factors and rerun the FA and this factor remains the same, then this is a stable and acceptable factor. Factors interpretability refers to the ease of interpreting the meaning of the factors and the degree of which they make sense. Since the aim of EFA is to produce interpretable and straightforward factors, some argue [86] that the emergence of simple structures (*i.e.* factors with few high loadings that are easy to interpret) is a vital sign of the success of the FA.

- Rotation aims to simplify the factors by minimizing the number of variables that factors load on while maximizing the loading value of each factor [85, 81], which improves the factor solution fit to the dataset and therefore improves the interpretability of the factors. There are two types of rotations, orthogonal and oblique. Orthogonal rotation methods produce uncorrelated factors, and oblique rotation methods produce factors that are correlated [85, 81].

5. *Obtaining and interpreting the factor solution* refers to analyzing variable loading values in the context of the research topic to give meaning and justification to the

factors and to name them. A variable that loads highest on a specific factor belongs to that factor. By reflecting on the variables that belong to each factor, one can arrive at the proper naming and meaning behind it, which is ultimately the goal behind EFA and a key to deciding its success.

6. *Obtaining factor scores* refers to the process of deriving factor scores for each SCC member based on a specific factor solution. There are several methods to do this [93, 94]; they fall into one of two categories: refined and non-refined methods [94]. The non-refined methods are computationally simple methods and therefore, are easy to understand and interpret. Refined methods are more complex and provide accurate and standardized factor scores. Two of the common methods are the Bartlett method and the Anderson-Rubin method. The Bartlett method produces unbiased factor scores that correlate only with the corresponding factor, and the Anderson-Rubin method produces uncorrelated and standardized scores [81].

#### 2.3.1.2 Confirmatory factor analysis

Confirmatory Factor Analysis (CFA) is a form of FA that measures the degree of which the data supports a given factor solution. This analysis is particularly useful when the factor solution is based on theory or analytical research that is not rooted in the data that the CFA will use. Precisely, CFA measures the reliability of the factor solution. Given the factor solution, and the dataset, CFA quantifies the reliability of the factor solution by calculating the degree of which this factor solution "fit" the dataset.

This chapter provides detailed background information on SNs, crowdsourcing, and FA. The next chapter describes SCCs and introduces them as a new research area.

# Chapter 3

# SOCIAL CROWDSOURCING COMMUNITIES

This chapter defines SCCs and introduce them as a viable new area of research. Specifically, it answers three questions to create a conceptual model through which this thesis's results can be interpreted: 1) *What is a SCC?* 2) *What is not a SCC?* 3) *Why study SCCs?*

## 3.1 What is A Social Crowdsourcing Community?

Organizations that seek to elicit their consumers' perspectives on their products or services have historically done so via focus groups, surveys, phone interviews and similar methods [7]. These methods are quite costly and, consequently, are restricted in various ways, including limiting the number of participants, their geography, and the type of response artifacts. The advancements in communication technologies (and specifically the rise of social networking and crowdsourcing) brought forth the opportunity to alleviate these limitations and to innovate new methods for consumer engagement. SCCs are an expression of this opportunity. They allow organizations to engage with their consumers for purposes, such as eliciting their opinions concerning topics, products, or services, and generating ideas for new products or services.

Specifically, a SCC is an online, moderated, and gamified system created and moderated by a specific organization. It combines the powers of crowdsourcing and social networking to allow systematic query of crowds and synthesis of their response data (*i.e.* contributions) into coherent reports for decision-makers within the organization (*i.e.* beneficiaries). As the name suggests, SCCs are social communities that have all the components of crowdsourcing (see Section 2.2.1), follow the same crowdsourcing process (see Section 2.2.3), and focus on some functions of crowdsourcing (see Section 2.2.2) that are relevant to the type of tasks offered in these communities. SCC platforms provide social features to its members that allow them to identify and interact with

each other. The social aspects of these systems necessitate that their "crowds" be well-defined, which diverges from the typical definition of crowdsourcing[1].

SCCs are based on the emerging paradigm of human social computation systems [96] where the members (*i.e.* human crowd) perform tasks (*i.e.* computations) using the features of a social platform (*i.e.* discussions) for the beneficiaries. As a crowdsourcing system, SCCs consist of the same four components mentioned in Section 2.2.1: the beneficiary, the queries, the crowd, and the platform.

The beneficiary, in this case, refers to individuals or organizations who are interested in the crowd's responses. They cultivate, moderate and incentivize the crowd, measure and maintain its health and diversity, issue queries to the crowd, and collect and analyze responses. The beneficiaries center their communities on a specific interest, such as a product, services, or social or political issue, which underpins all aspects of their crowdsourcing process. For instance, members of the SCCs featured in this thesis participate in activities designed to elicit their responses or opinions regarding specific topics, products, or services.

A crowd is a group of online users who are members of a SCC. The beneficiaries recruit these members by employing various strategies, including marketing, partnerships with other organizations, and technological tools that bring "traffic" to their communities. Before the recruited members can undertake queries and respond to them, they typically have to go through a screening process to gather demographic information and assess their knowledge of the community's concerned products or services.

The queries in SCCs are often open-ended questions that moderators plan and conduct, and members carry out through a process. Moderators initiate this *social crowdsourcing process* by posting new queries. These queries contain task-relevant information, such as background knowledge, the question(s) that needs to be answered, the duration of the query, and any activities that need to be performed. The platform immediately notifies existing members, via their preferred no-

[1]In this context, the term *social* describes platforms that enable their members to identify each other through profiles, to associate with one another in some way (friendship, follow, *etc.*), and to interact with one another through messaging, voting, discussions, *etc*, which takes into account the existing definitions of social networks [95, 38].

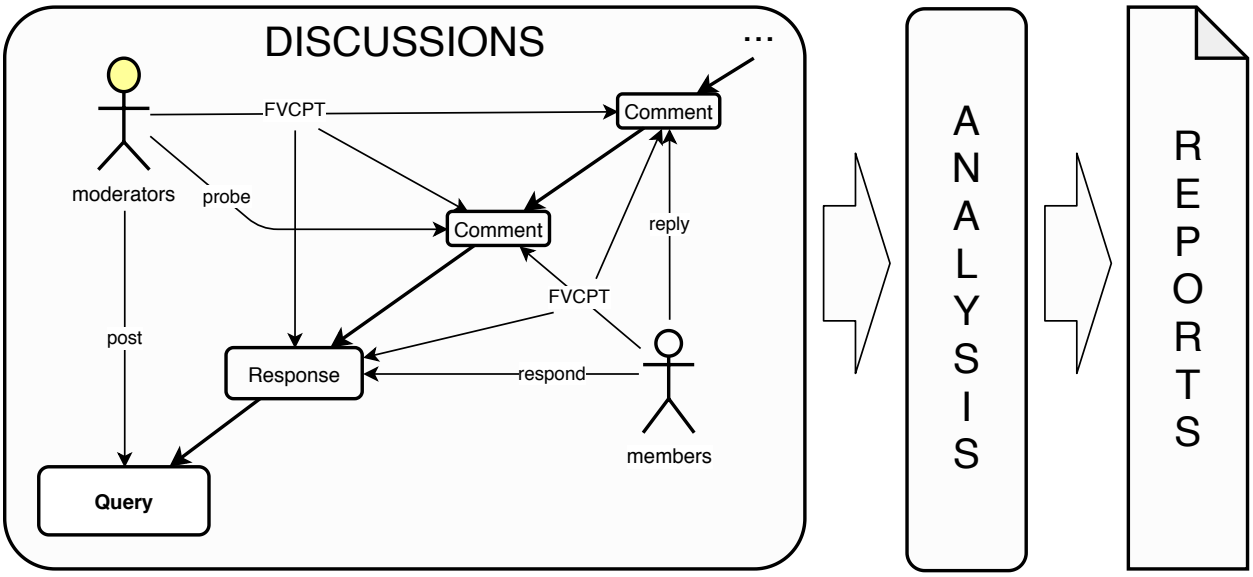



Figure 3.1: The social crowdsourcing process. FVCPT represents flagging, voting, censoring, pinning, and tagging interactions [2].

tification method, of the newly posted query. Interested members voluntarily collaborate either by responding to the queries directly or by building on others' responses through discussions. Moderators steer these discussions by probing and asking clarifying questions, and they reward members for their participation and the quality of their contributions. After the end of the discussions, analysts compile the contributions into reports (see Figure 3.1 for an illustration of this process). This process yields behavioural logs that are captured by the following ERD (see Figure 3.2). Members of the SCCs featured in this thesis participate mainly in these types of activities. The platform, typically a website, is a medium that brings these all the components together by allowing them to carry out their respective parts.

Although members join and participate in SCCs voluntarily, these communities offer their members various rewards and incentives for their participation. For example, the SCCs this thesis uses offer their members "insider" information about the organization they "work for", in addition to many other incentives in return for their contributions. Moreover, all these SCCs employ some gamification techniques to keep them engaged and feeling appreciated, which includes frequent giveaways, points, badges, and recognition through leader-boards and announcements.

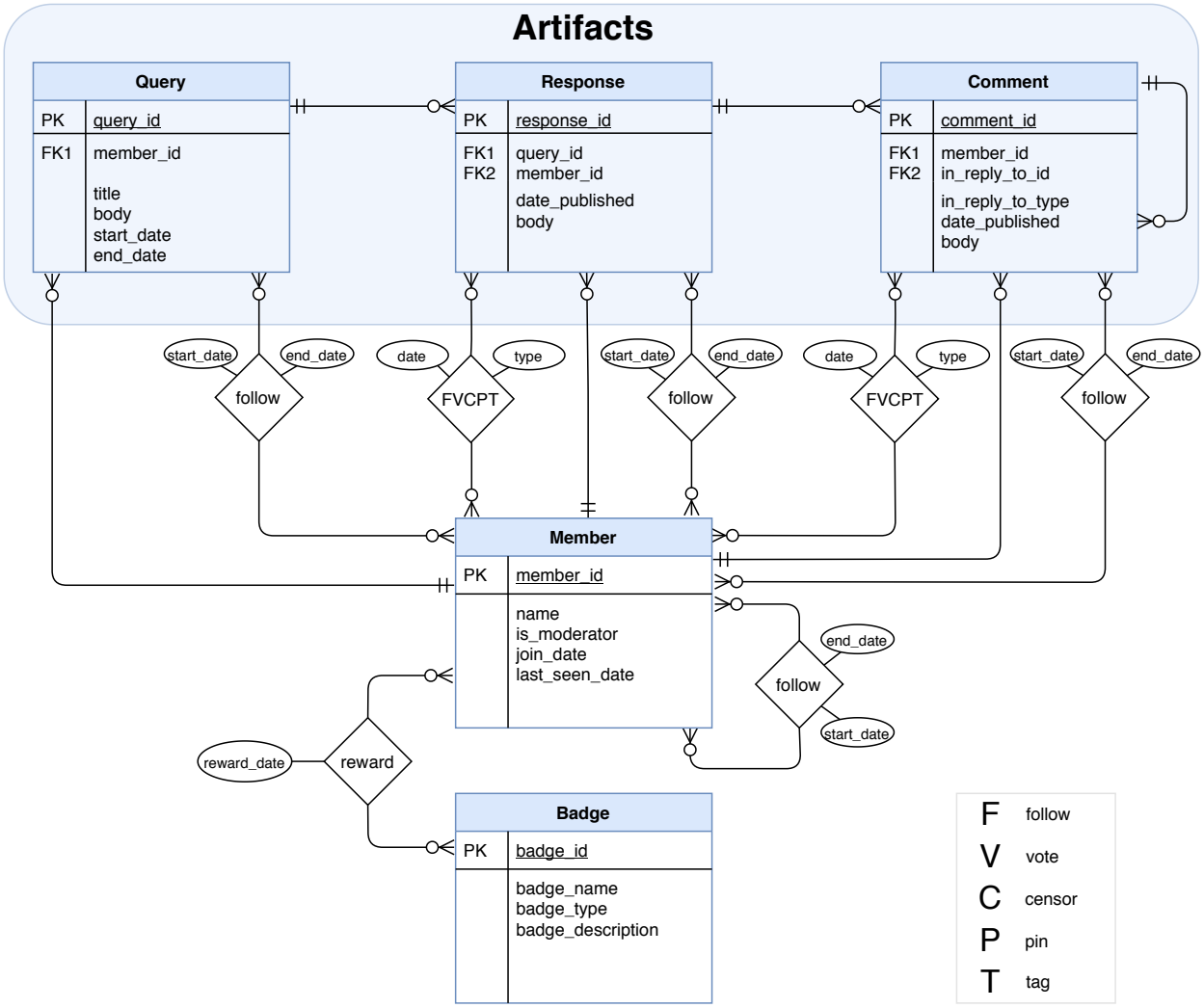



Figure 3.2: A simplified entity-relationship diagram of the data generated from member behaviours in a SCC platform [2].

For a more concrete example, consider the following: A smart-phone manufacturer can run a SCC to engage with their clientele to identify any concerns regarding their current phone models and to identify ideas for future models. The queries pushed onto this SCC's crowd would reflect the manufacturer's goals. The SCC's moderators would work closely with the other SCC members to steer discussions, increase member engagement, and ensure that all queries achieve the desired objectives. These objectives can range from answering an overarching question, such as "what is your favourite thing about this phone", to responding to a survey or a task through structured artifacts. Members are rewarded for their "work" in various ways, including points, badges, and

giveaways.

StackOverflow and Quora are popular and simple examples of SCC platforms. The beneficiary in these platforms is the individual or group of individuals who post queries. The crowd is those undertaking the queries. These platforms offer social features, gamification features, and collaboration features that allow the beneficiary to push queries onto a crowd and gather their responses. Due to the nature of queries on these platforms (*i.e.* questions with clear answers), and their highly refined SCC process, some of these platforms offer features that streamline the synthesis of responses into a single "best" response that represents the synthesis of all the responses. For example, StackOverflow allows its members to refine their responses further based on others' comments and allows other members to vote on the response. It also allows the beneficiary to select the "best" response.

## 3.2 What is Not A Social Crowdsourcing Community?

The number of non-SCC systems is practically infinite; this section does not aim to list them. Instead, this section clarifies what a SCC is by elaborating on what it is not, especially regarding seemingly similar and popular SNs. As stated previously, SCCs and SNs share many similarities due to their social nature. They both allow their members to identify each other through profiles, to associate with one another through friendship or follow features, and to interact with one another directly (through direct messages) or indirectly (through comments, voting, *etc*). However, SCCs differ from SNs in several ways (see Chapter 4), including the following.

As Section 3.1 states, SCCs consist of four components (*i.e.* the beneficiary, the query, the crowd, and the platform). The SCC beneficiaries own the SCC platform and benefit from the outcomes of the SCC process. They recruit and query the crowd and oversee the SCC process to ensure it achieves its goals. The queries are questions that interest the beneficiary. They make up the bulk of these communities' activities. The crowd consists of members who explicitly join these communities to respond to the queries. The platform supports collaborative work by allowing

the beneficiary to distribute queries to members, collecting their responses, and offering various reputation and reward mechanisms based on the performance of members in responding to queries. Therefore, a system with a driving beneficiary, queries, a working crowd, and a collaboration-equipped platform is an SCC system.

SN platforms, such as Facebook, Twitter, and YouTube, are communication and entertainment platforms where members produce content based primarily on their interests. The owners of these platforms do not interfere in the content generation process beyond offering features that facilitate content creation and management. They use the content to obtain insights regarding individuals' interests for marketing and advertising purposes [97]. Furthermore, the intimate nature of SNs requires their platforms to offer sophisticated privacy features that allow members to control their friendship and association circles, and restrict access to their content.

Though SCCs utilize communication and their members may join them for entertainment, they are primarily "work" platforms. They are rooted in the "collective intelligence" concept (see Chapter 2.2), so SCC platforms emphasize openness and maximal exposure of members to ideas and content as a means to arriving at (innovative) responses to queries. Since the beneficiaries of these platforms are directly interested in the query responses, SCC platforms follow a structured process (see Figure 3.1) that leads to the synthesis of these responses.

Thus, social systems that are not designed for "work", and do not follow the SCC process or missing any of the four components lack the appropriate structured process, the driving beneficiary, the working crowd, or the collaboration-equipped platform. Thus, they are not SCC systems.

For example, Facebook, Twitter, and YouTube are not SCCs because they lack a driving beneficiary and consequently have no queries or working crowd, and do not adhere to the SCC process. Nevertheless, organizations and individuals can use SNs to carry out SCCs tasks. For example, consider a situation where a group of individuals "query" members of a SN group regarding some topic by posting their "query" on this SN platform (*e.g.* Facebook Groups and Reddit). The querying-individuals then join the discussions around member responses and synthesize an answer

to the query, which they operationalize in some way. These querying-individuals are the beneficiaries who query and moderate the crowd. The crowd, in this case, is the queried individuals, and the platform is the SN. Therefore, it is possible to use a SN platform to carry out SCC tasks.

Despite it being possible, using SN to carryout SCC tasks is not sustainable for prolonged and continuous use. Such use requires the underlying SN platform to evolve, in many ways, to an SCC. For example, it must offer SCC-specific features that facilitate the SCC process, including moderator tasks, member engagement, and data analysis and reporting. Examples of these features include data capture/retrieval features or streamlined process for synthesizing responses, gamification features (*e.g.* leaderboards, points and badges), and customizable moderator, monitoring, and screening tools.

Consider the StackOverflow and Quora examples mentioned above. Like SN owners, the owners of these platforms do not interfere in the content generation process and use responses for marketing and advertising purposes. However, they designed these platforms as "work" platforms that consist of the four SCC components and follow a structured SCC process. These platforms allow members to become beneficiaries by allowing them to push queries to the crowd. Through their highly refined SCC processes, these platforms allow the beneficiaries to moderate and synthesize responses. The members of these communities explicitly join them to participate in queries. Through the predefined SCC process, they respond to queries or discuss existing responses, which ultimately leads to refining and selecting the "best" answer that represents the synthesis of all responses.

## 3.3 Why Study Social Crowdsourcing Communities?

SCCs are new. Therefore, naturally, they present substantial opportunities for researchers from various fields to advance them. For example, artificial intelligence can help automate or augment many moderator tasks to help scale these communities. This help can be done by devising machine learning algorithms to identify influential members [98, 99], acts of bullying and harass-

ment [100, 101], offensive language [102], spam and spambots [103, 104, 105], and inappropriate images [106]. Similarly, the psychology field can help increase the effectiveness of SCCs and increase member engagement through the application of various psychology models to predict personality [107, 108, 19], performance [109, 10, 110, 11, 18], and identify signs of addiction [111] to understand member behaviour, and motivations. Moreover, this field can also help improve the quality of work outcomes through its various team formation strategies [25, 9, 21, 22]. Furthermore, the application of social network analysis methods and measures can help categorize the structure of SCC networks [112] to place SCC networks in a specific class of networks. This categorization has many implications, including SCC security [113, 114], information diffusion [115], modelling growth [46, 116], determine signs of their deterioration [117], and identify signs of critical transitions [118] to serve as an early warning system for their health. Similarly, many other fields can contribute to the advancement of SCCs.

SCCs in themselves require research attention to help standardize and improve their processes, and identify their unique requirements and address them; especially with the increase in the number of organizations and individuals that are leveraging SCCs for their benefit, financially or otherwise. Naturally, there exists a need to maximize gain and minimize risk in these systems. By identifying standard practices and processes, it becomes possible to analyze, test, and improve these practices and processes to ensure its effectiveness and efficiency in accomplishing its designated goals. One way to do this is by analyzing and comparing the structure of various SCCs to identify common structural properties and then shed light on the underlying processes or behavioural patterns that caused these structures to emerge and helps identify standard processes in SCCs.

Another benefit from such analysis is identifying where SCCs structurally differ or are similar to existing social platforms, which allows for identifying their unique requirements, which in turn allows for altering existing algorithms or innovate new ones to serve their unique needs. One example of such requirements would be producing an algorithm capable of ranking SCC members based on several factors, including their influence on other members, competence in accomplish-

ing tasks, trust in their results, and truthfulness in their contributions, instead of existing ranking algorithms that base their rankings only on network structure [119]. Other examples include producing a method for automatic identification of contributions that address a query and rank them based on their relevance to this query and creating a method for automatic team formation that can recommend specific teams of members based on their competence, experience, and compatibility to work with one another to undertake specific queries. Currently, little work is reported to address these requirements.

SCCs are a new frontier for the application of various methods and techniques from other disciplines. Such applications will further advance SCCs and make them more feasible to organizations.

This chapter defines SCCs and highlights some areas for future research. The next chapter presents a comparative analysis between the structure of SCC networks and several other networks. This analysis leads to identifying structural similarities between SCCs and structural differences and similarities with SNs and the WWW.

# Chapter 4

# STRUCTURE OF SOCIAL CROWDSOURCING COMMUNITIES[1]

This chapter analyzes several SCCs' interaction and association networks structure to find common structural properties among them. These networks are a direct consequence of member behaviours. Thus, they contain insight into how SCC members behave and what common behaviours they demonstrate, which informs subsequent chapters' analyses. We also analyze the network structure of several popular SNs and the WWW and compare them to SCC networks, which allow us to distinguish and find the commonalities between SCCs and these existing networks. Although both SCCs and SNs are social platforms in that they offer social features (see Section 3.1), which allow their members to behave socially, these two types of platforms are different in many ways. We show that these differences reflect on the behaviours of their members. Specifically, they reflect on members' associations and interactions. The results of this chapter show the unique behaviours that SCC members exhibit as well as those that they share with the members of SNs and pages of the WWW. To do this, we produce networks from the members' associations and interactions for several SCCs and from the associations of members and pages from several SN platforms and the WWW. We then analyze the SCC association and interaction network structures and compare them to the network structures of popular SNs and the WWW.

This investigation of the topology of SCCs is similar to several studies that have already investigated network structure of many existing man-made and natural systems, such as the WWW [120, 121], scientific collaboration networks [122, 123], SNs and media sharing networks [124, 4, 125], and neurological and biological networks [126, 127]. Similar to current work [124, 4, 125, 128], we perform an in-depth analysis of the structural properties of well-known and successful SCCs

[1] **The results in this chapter are published [1]**

to confirm or refute recent findings as a first step in the process of categorizing these networks. Unlike past work in this area, we do not focus the analysis on member association networks only. Instead, we analyze both association and interaction networks. The results of this chapter allow us to determine the SCC network type (*i.e.* random, lattice, power-law, *etc.*) and place it in the broader field of SN analysis. It also helps in understanding the structural properties of these networks, which assists future researchers and developers in improving current designs and practices of crowdsourcing, in identifying and implementing new algorithms tailored for the needs of SCCs, and in modelling their networks.

We present a large-scale measurement study and analysis of the topology of five crowdsourcing communities. We have access to anonymized databases of five SCC collectively contain two million vertices and six million edges. This data allows us to understand and categorize SCC networks as well as to compare them with SNs and the WWW network by inspecting structural properties, including the degree of link symmetry, adherence to power-law, correlation of in-degree and out-degree[2], to find common and distinct structural properties of SCC networks.

## 4.1 Data

We use data from five SCCs, five SNs, and a sample of the WWW. We obtained the SCC data from Chaordix[3], a Calgary-based crowdsourcing, and crowd-intelligence company famous for their crowdsourcing platform. We obtained SNs data from Mislove [4], and WWW data from Stanford's WebBase Project[4].

### 4.1.1 Social crowdsourcing communities data

This data consists of anonymized interactions and associations logs in five SCCs[5]. The logs contain a total of 1,918,382 vertices and 5,744,222 edges. 44% of vertices are member vertices, and 10%

[2]In-degrees and out-degrees are the numbers of incoming and outgoing edges to and from a vertex, respectively.
[3]`www.chaordix.com`
[4]`http://dbpubs.stanford.edu:8091/$\sim$testbed/doc2/WebBase/`
[5]The community's name must remain confidential to be consistent with our access agreement.

of the edges are association edges. Table 4.2 shows the number of vertices and edges in both association and interaction networks after the elimination of all singleton member vertices.

Table 4.1: SCC categorization.

| SCC | Access | Life span | Goals |
|---|---|---|---|
| *HCC* | Private | Persistent | Multi |
| *KCC* | Private | Constrained | Single |
| *LCC* | Public | Persistent | Single |
| *MCC* | Public | Persistent | Multi |
| *VCC* | Private | Persistent | Multi |

Table 4.1 shows a brief categorization of these communities in terms of their access type, life span, and goals. SCCs can be private - closed access communities - or public - open access communities. A private community allows only approved individuals to join, browse, and contribute to their community while public communities do not have this restriction. Public communities have fewer restrictions on who can join their community (*i.e.* have lenient screening processes) and browse their content. Although public communities require individuals to register to contribute, they allow and sometimes encourage sharing of contributions outside their platform on other social media.

SCCs can have constrained or persistent life span. Constrained SCCs run for a particular period or until specific outcomes are achieved; then they are shut down. Persistent communities are not restricted to a specific timeframe or outcomes. They are typically multipurpose crowdsourcing platforms with continually evolving goal(s).

SCCs can have a single goal or multiple goals. These goals can be open or close-ended. They also can be tasks or questions. For example, the *LCC* community[6] is an open-ended single goal community where members attempt to accomplish a task by producing new product ideas or designs. Although many succeed in producing such designs, this task stays open. *HCC*, on the other hand, is a multi-goal community where members contribute answers to open and closed-ended questions and tasks. There may be structured responses, such as responses to multiple-choice

[6]We use *HCC*, *KCC*, *LCC*, *MCC*, and *VCC* as pseudo names to protect the anonymity of these SCCs.

questions, or unstructured, such as taking part in discussions. Tasks can be posting reviews, propagating news, and updates on social media.

The five SCCs are identified by *HCC*, *KCC*, *LCC*, *MCC*, and *VCC*. *HCC* is a persistent and private community of brand loyalists of a major mobile phone company; focused on brand advocacy, brand loyalty, and marketing. Moderators allow carefully screened and selected members to join this community to take part in various research and labour activities, and discussions related to mobile devices and accessories in return for exclusive inside information from the company and frequent giveaways of highly desired company branded merchandise.

*KCC* is a constrained private community of health care experts from around the globe who were carefully selected to take part in discussions related to the future of sustainable health care. These highly targeted subject matter experts from a wide range of related industries engaged in an investigation of potential changes to health care systems worldwide and best practices to ensure high-value health care systems.

*LCC* is a large, open, and persistent online community for one of the largest toy manufacturers in the world. The community acts as a crowdsourced idea-curation platform for new product suggestions. Members can create listings of new product ideas (projects) or can vote on existing projects. Operators of this platform further study projects that receive a certain number of votes. If they approve the project for production, members who submitted the project receive a royalty for their creation.

*MCC* is an open persistent online community aimed at discussing the challenges and future of mobility. Visitors can create an account and become members of this platform. The platform presents members with engaging thought starters, articles, or blog posts, and invites these members to take part in discussions related to how society will be affected by the coming changes to mobility, *e.g.* self-driving cars, car sharing, commuting, and impact on the environment.

*VCC* is a persistent private community run by a large university in Denmark. Membership is restricted to university students and staff only. Members provide creative and feasible business ideas

on how to better solve a proposed business challenge or improve existing solutions.. Members submit their business ideas, and others offer critical feedback or suggestions for improvements or vote on their favourite ideas. Winning ideas receive monetary prizes and recognition in the university community.

### 4.1.2 Social networks and the World Wide Web data

SN data consists of 11,360,930 member vertices and 328,496,316 edges obtained by crawling four popular social networking platforms. Flickr is a photo and video sharing platform where members freely share photos and videos for others to view. Members can follow each other. Following is the only association feature of this platform. LiveJournal is a popular platform where members keep a diary, journals, or blogs and share them with others. Members of this platform can follow each other. Orkut is a social networking platform where members can form friendship links with one another, chat, and share content. Friendship in Orkut is a reciprocal relationship. YouTube is prominently a video sharing platform where members upload, share, like, and comment on videos. Members can follow one another.

We obtained the WWW data from Stanford WebBase project. It consists of 18,208,130 page vertices and 64,095,092 edges. Webpages link to each other through URLs. Since there can be several URLs in one page that link to another page, we aggregate these links into one and placing a weight value equals to the total number of URLs on this link.

It is important to note that since SN and the WWW data is a result of crawling the largest Weakly Connected Component (WCC) of the association networks, they are not an accurate reflection of the reality of their corresponding SN or the WWW because they include partial association networks and do not include interactions. Therefore, we compare them only to SCC association networks.

### 4.1.3 Extracting association and interaction networks

We obtained anonymized databases of five SCCs. Since part of this study is comparative, we distinguish between two types of networks: *association networks* and *interaction networks*. We create association networks from following and befriending actions in both SCCs and SNs. We create it based on hyperlinks among webpages in the WWW. We place a directed edge, with a weight of 1, between members who initiate the associations and target members. Association networks reflect relationships among members (or pages) based on foreseeable benefits. Thus, it encodes valuable information concerning association patterns in these communities.

We create interaction networks from member interactions regardless of association status. For example, a member who made a contribution, *e.g.* a post, or comment, may receive an interaction from another member in the form of a comment or a vote on this contribution. In this case, we place a directed edge, with a weight of 1, from the interacting member to the contributing member. Multiple edges can exist between any two members. We aggregate all edges to a single edge with a weight value that equals the number of edges. Thus, a frequently interacting pair of vertices will have an edge weight that reflects their interaction frequency. We do not have records of member interactions in the SN data and certainly not in the WWW data. We extract and analyze SCC interaction networks to identify common structural properties among SCCs.

Given that this analysis focuses on associations and interactions among members (or pages in the case of the WWW network), we eliminate all singleton vertices (*i.e.* vertices with no in- and out-degrees).

## 4.2 Differences between SCCs, and SNs and the WWW

Upon analyzing the networks, we found key differences (and similarities) between the networks of SCCs, SNs, and the WWW. We found that SCCs are smaller and denser than SNs and the WWW. Members of SCCs are less likely to reciprocate association links. Even though the associations in SCCs, SNs, and the WWW follow a power-law distribution, they differ in several ways. SCC mem-

bers tend to associate less, but few members associate significantly more than others do. They also tend to associate with members who have significantly different association degrees. Furthermore, members who associate with many (*i.e.* active members) and those that many others associate with (*i.e.* popular members) are often not the same members.

### 4.2.1 Difference 1: SCCs are smaller and denser than SNs and the WWW

Table 4.2: High-level statistics derived from all SCC Association (ASSO) and Interaction (INTR) networks available to this thesis. Association networks have fewer vertices and edges than interaction networks because of the frequent and casual nature of interactions, in comparison to associations, which connects more vertices to the network.

| | SCC name: | *HCC* | *KCC* | *LCC* | *MCC* | *VCC* |
|---|---|---|---|---|---|---|
| | Launch Date | 11/30/2011 | 3/10/2014 | 10/7/2004 | 12/19/2014 | 10/21/2014 |
| | Cut-off Date | 4/17/2015 | 6/19/2014 | 4/17/2015 | 5/5/2015 | 5/6/2015 |
| | All Vertices | 9,693 | 801 | 830,400 | 2,666 | 2,457 |
| ASSO | # Vertices | 4,657 | 91 | 50,761 | 149 | 160 |
| ASSO | # Edges | 52,122 | 216 | 175,071 | 698 | 714 |
| ASSO | Average Degree | 22.384 | 4.747 | 6.898 | 9.369 | 8.925 |
| ASSO | Density | 2.40E-03 | 2.64E-02 | 6.79E-05 | 3.17E-02 | 2.81E-02 |
| ASSO | Degree of Link Symmetry | 36.0% | 15.7% | 9.8% | 33.8% | 51.5% |
| INTR | # Vertices | 4,868 | 106 | 592,703 | 175 | 493 |
| INTR | # Edges | 152,890 | 347 | 2,622,917 | 516 | 1,721 |
| INTR | Average Degree | 62.814 | 6.547 | 8.851 | 5.897 | 6.982 |
| INTR | Density | 6.45E-03 | 3.12E-02 | 7.47E-06 | 1.69E-02 | 7.10E-03 |
| INTR | Degree of Link Symmetry | 51.4% | 24.8% | 7.1% | 14.3% | 30.6% |

Table 4.3: High-level statistics of all SNs and the WWW Association (ASSO) networks.

| | WWW | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|---|
| Crawl Start Date | 12/03/2003 | 01/09/2007 | 12/09/2006 | 10/03/2006 | 01/15/2007 |
| Crawl End Date | 12/03/2003 | 01/092007 | 12/11/2006 | 11/11/2006 | 01/15/2007 |
| # Vertices | 18,208,130 | 1,846,198 | 5,284,457 | 3,072,448 | 1,157,827 |
| # Edges | 64,095,092 | 22,613,981 | 77,402,652 | 223,534,301 | 4,945,382 |
| Sampling Percentage | - | 26.90% | 95.40% | 11.30% | unknown |
| Average Degree | 7.04 | 24.48 | 29.294 | 145.509 | 8.543 |
| Density | 1.93E-07 | 6.63E-06 | 2.77E-06 | 2.37E-05 | 3.69E-06 |
| Degree of Link Symmetry | 10% | 62.40% | 73.40% | 95.20% | 79.10% |

Tables 4.2 and 4.3 show that SCC association networks have fewer vertices than SCC interaction networks do, except for *MCC*. SCCs have fewer vertices and edges than SNs. The average degrees in SCC association and interaction networks are smaller than the average degrees in SNs. Moreover, the densities of SCC interaction networks are sometimes smaller than association networks and both are many orders of magnitude denser than SNs and the WWW.

SCCs have significantly fewer vertices and edges in comparison to SNs (see Tables 4.2 and 4.3), which means that SCCs are generally smaller than SNs. This is because of the restriction SCC operators impose on the growth of their communities to control the amount of effort and finance required to moderate these communities and analyze responses (see Chapter 1 for more details about this restriction). Moreover, the higher specificity of SCCs in comparison to SNs places more constraints on who can join these communities. As Chapter 3 describes, SCCs are created by organizations that seek to connect with the consumers of their products or services. Therefore, members must have relevant experience and knowledge of this organization and its offerings in order to contribute meaningfully. Another reason is the relatively low appeal and popularity of SCCs, which means that few seek them out. It also means that joining and participating in these communities for prolonged periods is less attractive for people in comparison to SNs. This is because joining a SN that offers entertainment, communication, *etc*. is more appealing than joining a SCC to "work" on tasks that are not directly beneficial to the workers. Thus, SCC operators face the challenge of seeking out, recruiting, and retaining members for their communities.

The average degrees of SCC association networks are, in most cases, smaller than the average degrees SNs and the WWW, which shows that SCC members tend to associate with fewer members than the members and pages of the SNs and the WWW. However, considering the small number of members in SCCs, these average degrees become remarkably high. The densities of SCCs, which are many orders of magnitude larger than SNs and the WWW, further confirms this and implies that SCC members associate with each other with higher intensity than members of SNs and pages of the WWW. Moreover, interactions also exhibit high densities. Therefore, although SCCs have

fewer members than SNs, they exhibit all characteristics of liveliness, which rivals and sometimes exceeds those of the SNs. We attribute this to SCCs' highly active moderators who produce content, engage members, and replace inactive members with new ones to keep the community lively.

### 4.2.2 Difference 2: SCC members are less likely to reciprocate associations than members of SNs

We investigate the link symmetry in SCC association and interaction networks through the degree of link symmetry measure. A high degree of link symmetry in a network (or graph) implies a high degree of reciprocal linking (the presence of a two-way connection between any two vertices), which is an indicator of the strength of socializing (*i.e.* associations and interactions) in a community. Moreover, networks with high degrees of link symmetry have smaller radius and diameter than networks with low degree [4], which is useful for several areas, such as information dissemination and search but problematic for other areas, such as trust and influence determination [4]. High link symmetry is an immediate consequence of reciprocal linking behaviour. Some SNs force reciprocal linking. Orkut and similar platforms enforce it as a means of control to one's circle of friend. Thus, a member must request to befriend another. Upon acceptance of this request by the other member, the SN creates a two-way friendship connection between them. On the other hand, in Flickr, LiveJournal, and YouTube, a member can follow any other member without consent, but the followed member may choose to block specific members from following.

We measure the degree of link symmetry using the concept of *parallel edges*. Two edges are parallel if the source and sink vertices in one edge are the same as the sink and the source of the other edge (see Figure 4.1).

The degree of link symmetry $d(E)$ in a directed graph $G$ is the fraction of parallel edges from all edges. Thus, given the set of all parallel edges, $E'$, where $E' \subseteq E$, the degree of link symmetry is the fraction of the total number of parallel edges from the total number of edges. The following equation derives the degree of link symmetry:

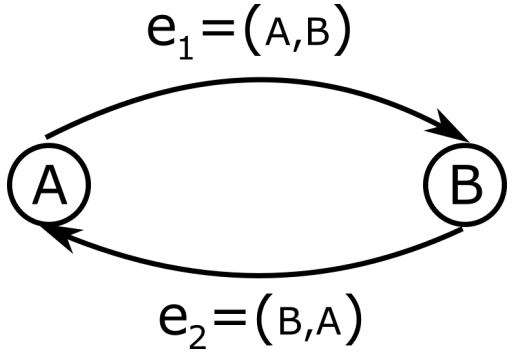



Figure 4.1: Illustration of parallel edges

$$d(E) = \frac{|E`|}{|E|}$$

Tables 4.2 and 4.3 show no consistent patterns emerge from the degree of link symmetry of SCC association and interaction networks, SNs, and the WWW network. However, SNs tend to have a much higher degree of link symmetry than SCCs or the WWW. This higher degree of link symmetry means that association linking in SCCs tends to be less symmetrical than SNs, which means members of SCCs are less likely to reciprocate association links (*i.e.* follows) in comparison to SNs. This result is expected given that SCCs do not enforce reciprocal linking, as is the case in SNs. This low reciprocity means that associations in SCCs are not diluted and are therefore indicative of influence and trust. It also is evident that the lack of "social space" in these SCCs is a contributing factor to the low reciprocal association linking. This insight is supported by the low degree of link symmetry of *LCC*, which is the lowest in comparison to the other SCCs. *LCC* does not offer its members space for socializing, but the others do offer such spaces to some extent. Offering socializing spaces and other community-building techniques increase the likelihood of reciprocal linking.

Members of SCCs are likely to reciprocate interaction links. Table 4.2 show reasonably high levels of reciprocation in interactions, even though some interactions are not possible to reciprocate, such as votes, and post follows.

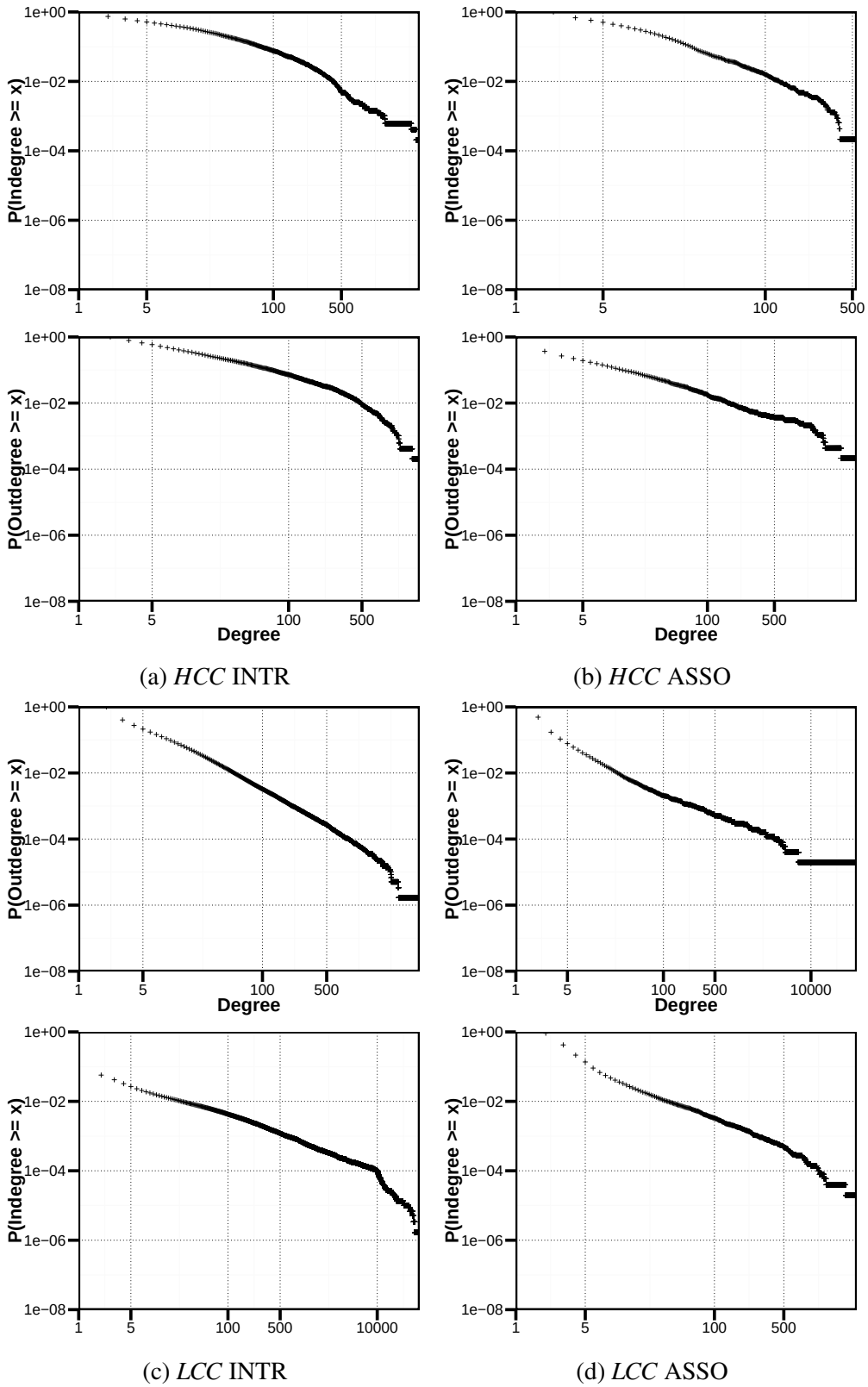
### 4.2.3 Difference 3: members' association patterns in SCCs differ from SNs

From Section 4.2.2, we identify a pattern for associations in SNs based on the high reciprocity. The pattern is that members who associate with others get associated with to the same degree. Since SCC members are less likely to reciprocate associations compared to SNs, this raises a question about the manifestation of this association pattern on the degree distribution of the SCC, SNs, and the WWW association networks. Therefore, we compare the association (and interaction) patterns of all SCCs by comparing their in-degree and out-degree distributions as well as their distributions' inequality scores. First, we identify the most likely distribution that fits our distributions, and use the Kolmogorov-Smirnov (KS) goodness-of-fit test to quantify the degree these distributions fit the observed distribution. We also calculated the Gini coefficient to measure the distributions' inequality scores.

Figure 4.2 shows the log-log plots of in- and out-degree complimentary cumulative distribution function (CCDF) for the association and interaction networks of the two largest SCCs. A straight line on a log-log plot indicates the existence of power-law distribution, while the slope of the line indicates the power-law exponent. The figures show that the association and interaction networks of these two SCCs adhere to a power-law pattern. Thus, a relatively small number of vertices own the majority of edges, both incoming and outgoing, while a majority of vertices possess only a relatively small number of edges, which is similar to SNs and the WWW networks.

We omit the plots of the other SCCs because they are too small for visual assessment of adherence to power-law[7]. Instead, we use the KS goodness-of-fit test and report the results in Table 4.4. The small KS goodness-of-fit test values show that SCC association and interaction networks fit a power-law distribution with a high confidence value ($p$). Few member vertices receive most incoming edges (and outgoing edges), which shows that a small segment of each community is highly popular (and highly active), while a larger segment receives only a little interest. SNs and the WWW networks show adherence to power-law distribution, except for Orkut's in-degree dis-

[7]Visual assessment of adherence to power-law through log-log plots has a known limitation when dealing with small number of points on the plot [129].




(a) *HCC* INTR (b) *HCC* ASSO

(c) *LCC* INTR (d) *LCC* ASSO

Figure 4.2: Log-log plot of in- and out-degree (top then bottom respectively) complementary cumulative distribution function (CCDF).

tribution ($p \leq 0.05$), which is due to the crawling procedure that tends to under-sample vertices with low degrees [4].

SCCs, SNs, and the WWW networks adhere to a power-law distribution. This similarity between SCCs, SNs, and the WWW suggest that all these networks, like many other offline and online networks, have degree distributions that follow a power-law distribution [130, 131, 125, 4]. Networks with a power-law degree distribution tend to form cores and thus are more tolerant to random error but are vulnerable to targeted attacks [113, 114]. They exhibit some useful and well-known patterns in information dissemination, expansion and growth, stability, and tolerance.

Despite the apparent difference in the association patterns in SCCs, SNs, and the WWW (see Section 4.2.2), these networks adhere to a power-law distribution. This similarity raises a question regarding the impact of the association pattern on the distribution. The answer is in the difference between power-law coefficients of the in-degrees and out-degrees of all networks. It is evident from Table 4.4 that the out-degree of SCC and the WWW association networks are more concentrated than the in-degrees (*i.e.* have higher power-law coefficients) showing that few members (or pages) associate a great deal with most other members. This high concentration of out-degrees shows the presence of highly active members in the community who associate with many others. SNs show a weak resemblance to this trend. We investigate this further using the Gini coefficient.

We calculate the Gini coefficient [132] (see Table 4.5) for in- and out-degrees of all networks[8]. The Gini coefficient encodes the degree of inequality in the distribution of edges among vertices. It ranges between zero and one[9] where a value of zero indicates a perfect "equality" of edge distribution to vertices. Thus, each vertex owns the same percentage of edges as the others. A Gini coefficient of one indicates complete inequality where a tiny percentage of vertices own nearly all edges.

Table 4.5 shows that in SCC association networks out-degrees are far more concentrated than

[8]Since power-law coefficients are estimates that depend on several other estimated parameters, such as xmin, alpha, xmax, *etc*., calculated using Maximum Likelihood Estimation (MLE) optimization to minimize KS's D (distance) statistic [129], it is not suited to explain this pattern although it may indicate it. The Gini coefficient is better suited to demonstrate this point.

[9]Although larger values are possible in practice if negative degrees are present, they did not exist in this case.

Table 4.4: Power-law coefficient estimates ($\alpha$) and KS statistics (D) obtained through Maximum Likelihood Estimation procedure, which minimizes D. * indicates $p \leq 0.05$ the WWW results are from [4].

| Network | | $\alpha$.in | D | $\alpha$.out | D |
|---|---|---|---|---|---|
| *HCC* | INTR | 3.0 | 0.063 | 2.1 | 0.063 |
| | ASSO | 2.4 | 0.027 | 2.0 | 0.035 |
| *KCC* | INTR | 2.0 | 0.106 | **3.7** | 0.087 |
| | ASSO | 2.4 | 0.056 | 2.4 | 0.092 |
| *LCC* | INTR | 1.9 | 0.025 | 2.6 | 0.006 |
| | ASSO | 2.1 | 0.016 | 2.1 | 0.015 |
| *MCC* | INTR | 2.4 | 0.052 | 2.2 | 0.029 |
| | ASSO | 2.8 | 0.101 | 2.6 | 0.085 |
| *VCC* | INTR | 2.3 | 0.079 | 1.9 | 0.038 |
| | ASSO | **3.7** | 0.125 | **5.5** | 0.109 |
| WWW | | 2.1 | - | 2.7 | - |
| Flickr | | 1.6 | 0.023 | 1.6 | 0.02 |
| LiveJournal | | 1.4 | 0.038 | 1.4 | 0.059 |
| Orkut | | 3.0* | 0.016 | 2.7 | 0.015 |
| YouTube | | 1.5 | 0.04 | 1.5 | 0.033 |

in-degrees indicating that few members associate a great deal with the majority of other members (an average of 27% difference in inequality). The WWW network shows the same pattern as SCC association networks (20% difference in inequality), which is reasonable since websites tend to have a catalogue page, usually the main page or a sitemap page, that links all other pages. SNs show a similar, but smaller-scale, pattern as SCC association networks (average 5% difference in inequality). With such a small difference, it follows that SNs tend to have a similar distribution of in- and out-degrees (within 5% of similarity) among their members.

These differences and similarities in out- and in-degrees confirm the existence of different association patterns between SCCs, SNs, and the WWW. SCCs have few highly active members in the community who associate with many others. The low reciprocity in Section 4.2.2 means that these members are not equally popular (*i.e.* they are not associated with by others as much as they associate). We investigate this point further in Sections 4.2.4 and 4.2.5.

Table 4.5: Gini coefficients representing the degree of inequality of edge distribution among vertices.

| Network | | Gini.In | Gini.Out |
|---|---|---|---|
| *HCC* | INTR | **0.9238** | 0.9021 |
| | ASSO | 0.7019 | **0.9447** |
| *KCC* | INTR | **0.7879** | 0.6500 |
| | ASSO | 0.4815 | **0.8642** |
| *LCC* | INTR | **0.9964** | 0.7593 |
| | ASSO | 0.6758 | **0.9129** |
| *MCC* | INTR | **0.7796** | 0.6935 |
| | ASSO | 0.5757 | **0.7787** |
| *VCC* | INTR | **0.8853** | 0.7137 |
| | ASSO | 0.5883 | **0.6647** |
| WWW | | 0.7576 | **0.9540** |
| Flickr | | 0.8671 | **0.8920** |
| YouTube | | 0.6898 | **0.8666** |
| LiveJournal | | **0.7203** | 0.7142 |
| Orkut | | 0.5611 | **0.5666** |

In SCC interaction networks, in-degrees are more concentrated than the out-degrees, which indicates that interactions among members in SCCs tend to be more focused on a popular few (an average of 19% difference in inequality). Since interactions take place on members' contents, this shows that fewer members produce popular content, which draws interactions from the remaining members of the community. This finding indicates that popular and active members in these communities are clear and distinguished, which is expected given that moderators engage members in discussions resulting in continuous production of outgoing edges in interaction networks. Moreover, moderators of these communities tend to promote popular content by "pinning" exceptional contributions in the most visible areas, keeping scoreboards and leaderboards, and announcing winners of competitions. Therefore, and given the competitive nature of these platforms, popular members tend to associate with many other members to increase their social capital.

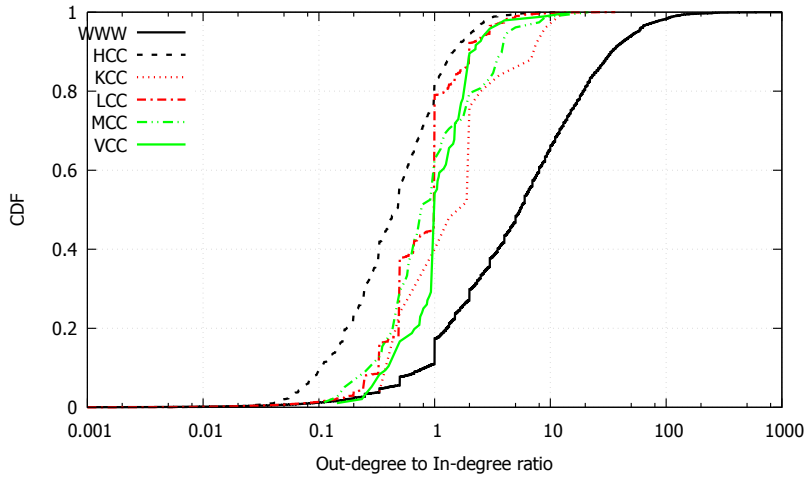
### 4.2.4 Difference 4: unlike SNs, popular members in SCCs are not the same as active members

Section 4.2.3 shows the presence of SCC members who are highly active (*i.e.* associated with many), and that these members are distinct. Section 4.2.2 shows that members do not have a tendency to reciprocate associations. Together, these two sections suggest that many members of SCCs fall in two extreme groups, those who are active, and those who are popular (associated with by many) and that these two groups have fewer overlap than SNs. We verify this finding in each network by, first, plotting the out-degrees to in-degrees ratios of all vertices. This plot visualizes the differences between the out- and in-degrees and the proportion of vertices that fall in each of the two extreme groups. We then plot the degree of overlap between the top x% of popular vertices (*i.e.* highest in-degrees) and active vertices (*i.e.* highest out-degrees) in all networks to identify distinctions in these groups.

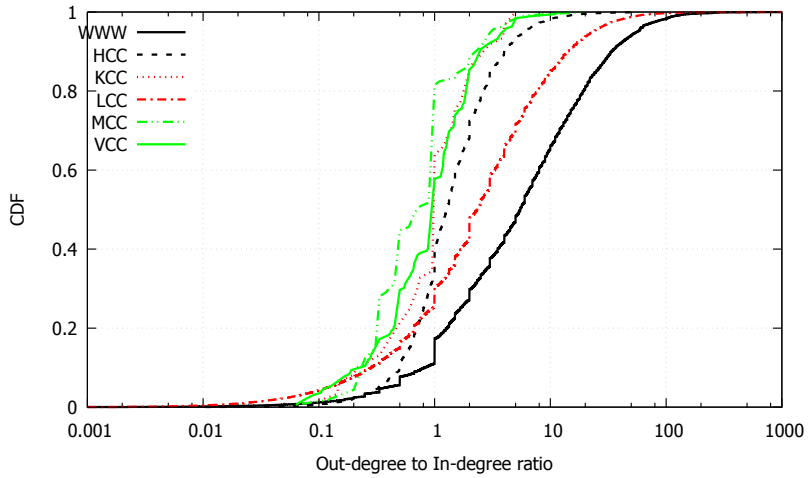
Figure 4.3 shows the out-degree to in-degree ratios in ascending order plotted against the cumulative probability of occurrence for each network. The x-axis is scaled by $log_{10}$ for ease of visualization. The figure shows that SCC association networks demonstrate a similar trend, where most vertices have larger in-degrees than out-degrees. This trend suggests the presence of members who associate with many while others refrain from associating. The WWW network shows an opposite trend where most vertices tend to have larger out-degrees than in-degrees. SN networks show an almost identical trend where vertices have similar out-degrees and in-degrees (an out-degree to in-degree ratio of 1). Similar to the WWW network, most vertices in SCC interaction networks have larger out-degrees than in-degrees. In other words, more vertices interact with others than those receiving interactions.

On average, 18% of vertices of SCC association and interaction networks have equal in-degree and out-degree values in comparison to SNs' 37%. Nearly 25% of vertices in SCC association and interaction networks have in-degree values that are within 20% of their out-degree values. In SNs, on the other hand, over 65% of vertices have in-degree values within 20% of their out-degree values. Nearly 94% of vertices in the WWW have an out-degree to in-degree ratio that is larger or

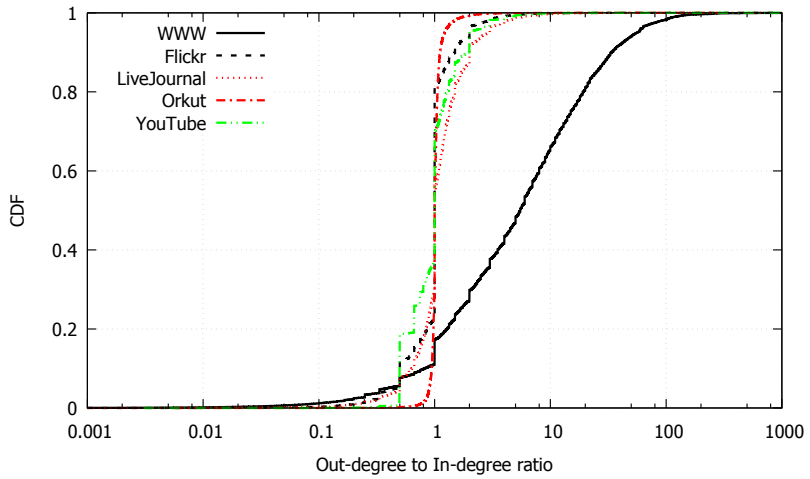
Figure 4.3: Ratio of out-degree to in-degree.




(a) SCC ASSO vs. WWW




(b) SCC INTR vs. WWW




(c) SNs vs. WWW

smaller than one, and almost 89% of these vertices have out-degrees larger than their in-degrees (ratio >1). These facts coupled with the high concentration of in-degrees or out-degrees - see Table 4.5 - mean that in SCCs and the WWW some vertices have an impressively large number of in-degrees or out-degrees, which suggest that they may be highly active members or moderators in SCCs or hubs in the WWW (pages that catalogue other web pages).

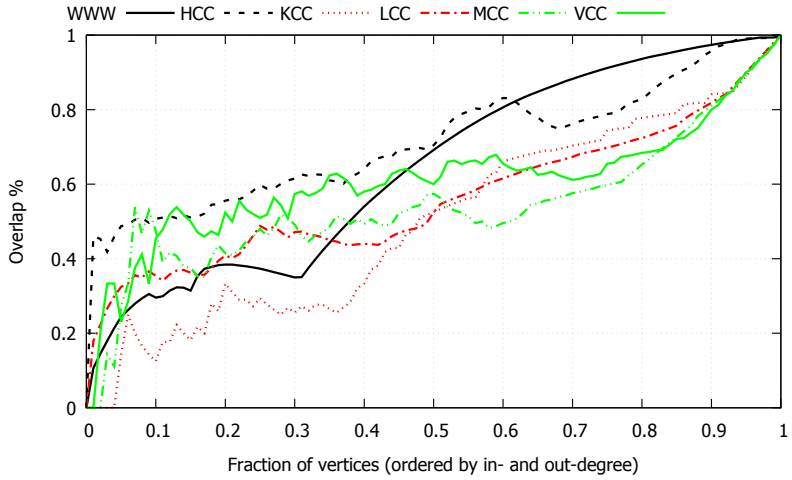
In general, most vertices in SCC association and interaction networks tend to have either larger in-degrees or larger out-degrees. This extreme behaviour further supports the earlier conclusion that active and popular vertices are distinguishable in SCC networks. The difference between in-degrees and out-degrees in SCC networks and the WWW network suggest the distinction of influential vertices in these networks, which is not the case in SNs. We investigate this further by plotting the extent to which the highest in-degree vertices overlap with the highest out-degree vertices.

Figure 4.4 shows the overlap between the top x% vertices in terms of in-degree and those with x% highest out-degree in SCC networks, SNs, and the WWW. Let $V$ be the set of all vertices, and $Vin_{x\%}$ and $Vout_{x\%}$ be the sets of vertices with the highest x% in-degree and out-degree, respectively, where $Vin_{x\%} \subseteq V$ and $Vout_{x\%} \subseteq V$. We calculate the overlap using the following equation:

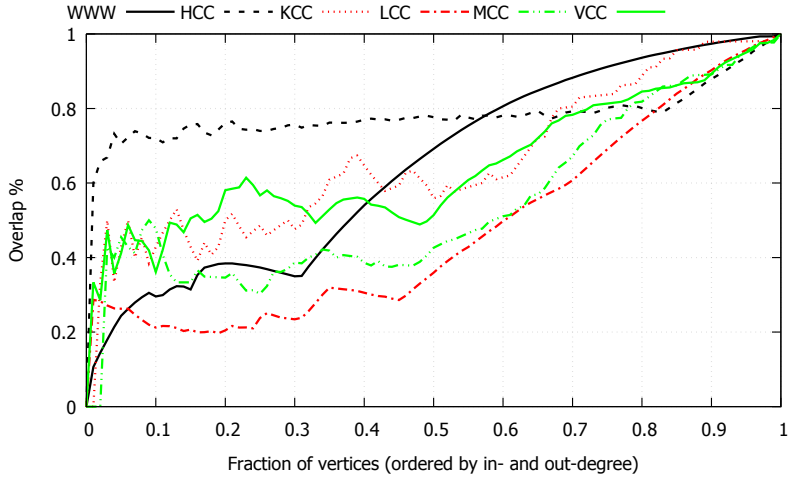
$$f(Vin_{x\%}, Vout_{x\%}) = \frac{Vin_{x\%} \cap Vout_{x\%}}{Vin_{x\%} \cup Vout_{x\%}}$$

In SCC association networks, vertices with high in-degrees do not necessarily have high out-degrees, which is also the case for SCC interaction networks. For association networks, the top 5% in-degree vertices correspond to only 29% of the top 5% out-degree vertices on average. For interaction networks, the top 5% in-degree vertices correspond to 45% of the top 5% out-degree vertices on average. This difference is expected given the reciprocal nature of interactions; recall the higher degree of link symmetry in interaction networks in Table 4.2. In the WWW, the top 5% in-degree vertices correspond to 24% of the top 5% out-degree vertices. In SNs, the top 5%

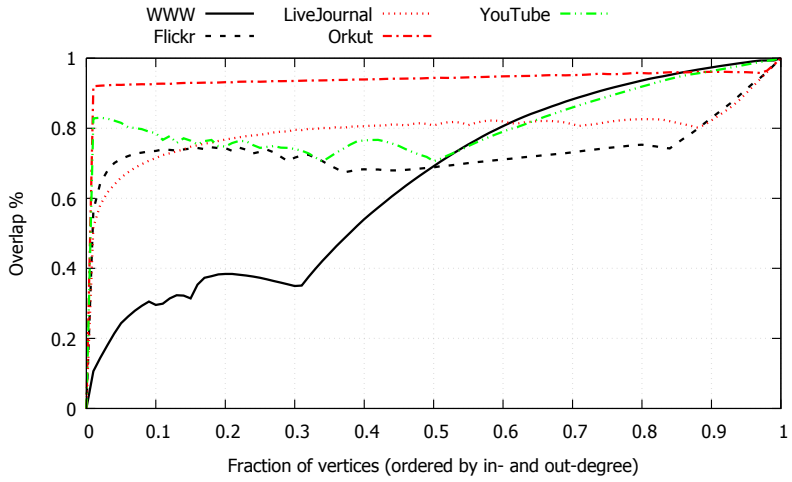
Figure 4.4: Overlap of the top x% vertices in terms of in-degree and the top x% in terms of out-degree.




(a) SCC ASSO vs. WWW




(b) SCC INTR vs. WWW




(c) SNs vs. WWW

in-degree vertices correspond to 78% of the top 5% out-degree vertices on average, which is a significant overlap. Therefore, SCC association and interaction networks are more like the WWW network than SNs. The smaller percentages of overlap between the vertices of the highest in-degrees and highest out-degrees in SCC and the WWW suggest a clear distinction between active and popular vertices.

Most vertices in SCC association and interaction networks tend to have larger in-degrees or out-degrees but not both. This tendency is identical in the WWW. On the other hand, SNs vertices have similar out- and in-degree, which means that in SCC (and the WWW) there is a clear distinction between active and popular members. The low overlap between vertices with the highest in-degree and those with highest out-degree in SCCs and the WWW networks as opposed to the high overlap in SNs further supports this finding.

The implications of in-degrees and out-degrees in a given network differ with the context of the network. For example, in-degrees and out-degrees in SNs imply popularity and activity. These also have applications in influence determination and information dissemination. In the WWW network, in- and out- degrees imply authority, and are used in ranking webpages and search results [133, 134, 38, 99, 135, 98, 136]. Understanding the relationship between in-degrees and out-degrees is important for SCCs for the same reasons as those of SNs and the WWW. Furthermore, understanding the relationship between in-degrees and out-degrees in SCCs is of particular importance because it captures patterns in association and interaction behaviour of members. These patterns provide moderators with insight into association and interaction preferences and tendencies of the members, which in maintaining the moderators can further engage and retain their members. It also helps them identify "normal" association and interaction behaviour, which helps in detecting deteriorations in associations and interactions.

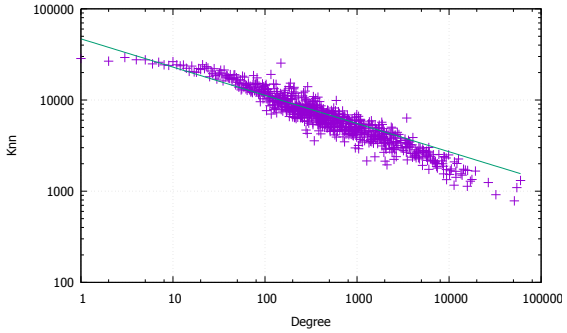
### 4.2.5 Difference 5: unlike SNs, members in SCCs tend to associate with members who have different degrees

One important aspect of analyzing the structure of networks is to understand the connection tendencies of members. Two questions regarding connecting behaviour of vertices arise. The first is regarding the tendency of vertices to connect to others with similar degrees, and the second is regarding the tendency of vertices to connect to others with different degrees. We answer these questions through the measure of degree correlation. Degree correlation is a crucial measure of the tendency of vertices to connect to other vertices with similar or dissimilar degrees. There are several ways to calculate or approximate this measure. For this thesis, we use Joint Degree Distribution (JDD).

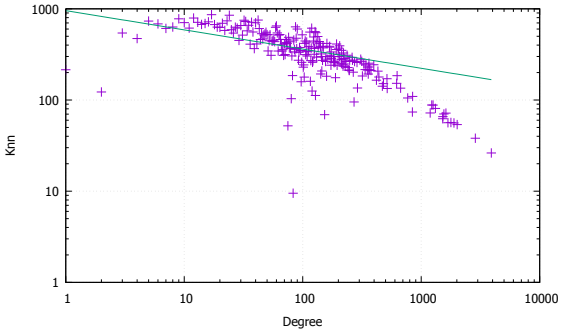
JDD is an analysis of the degree distribution that quantifies the likelihood of vertices to connect to similar-degree vertices. The basic idea is to produce an $n \times n$ matrix $M$, where the rows and columns are the ascending order of degrees. Each cell $m_{ij}$ contains the total number of vertices of degree $i$ who connect to vertices of degree $j$. If similar degree vertices tend to connect, the matrix will have many large values along the diagonal. If vertices of extremely different degrees tend to connect, then the matrix will have large values along the anti-diagonal. In other situations, the matrix will have other distributions. The computational expense of JDD can be quite high, so we use an approximation algorithm, $K$-nearest neighbour (Knn), to produce a more condensed version of the results. Knn obtains the average degrees of $k$ neighbours of a vertex, for all vertices. We plot average degrees of all vertices of a particular degree value in a log-log plot.

Figure 4.5 shows that vertices in SCC networks tend to associate and interact with others with significantly different degree values (*i.e.* higher or lower degree values), which is opposite to the trends in SNs and the WWW [4] (except for YouTube that follows the same trend as SCCs). This finding is in line with our expectations; we attribute it to the competitive nature of members in SCCs where new members associate and interact with well-established and influential members to learn from them, and influential members associate and interact with new members to establish

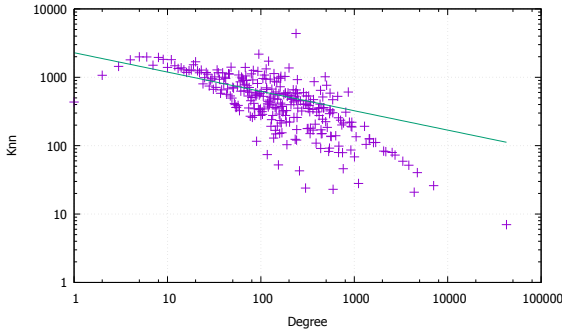
Figure 4.5: Log-log plot of the average degree of all neighbours of vertices of particular degree value. We omit some SNs and SCCs plots because they show similar trends and therefore add no new information.



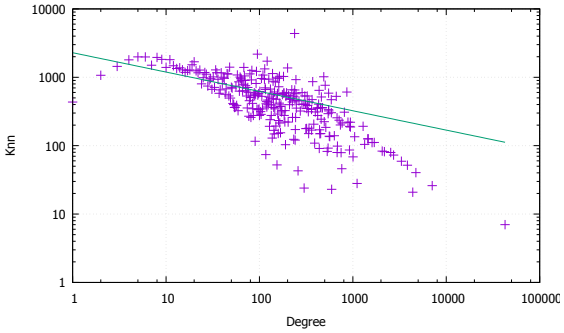

(a) *HCC* INTR



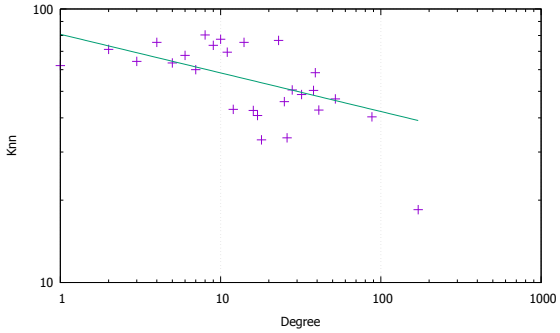

(b) *HCC* ASSO



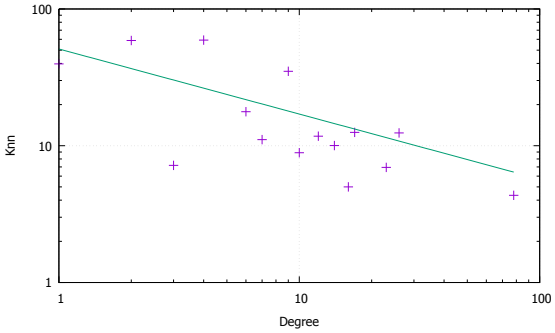

(c) *LCC* INTR



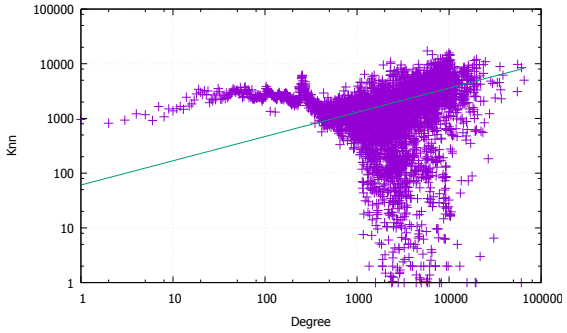

(d) *LCC* ASSO



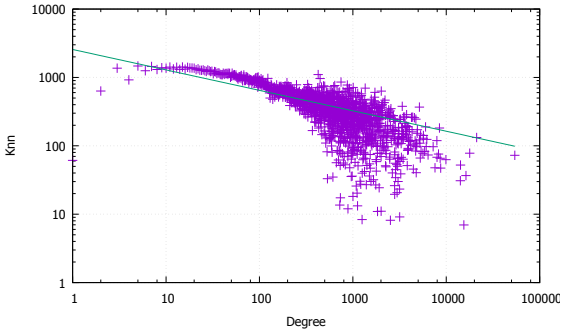

(e) *KCC* INTR




(f) *KCC* ASSO




(g) WWW




(h) YouTube

a social capital or a support base. This way, both members find themselves in an advantageous position, which leads to influential SCC members being less likely to associate and interact with other influential members.

## 4.3 Findings Summary

We conduct an extensive study of the structure of several SCCs. We study various graph-theoretic properties of association and interaction networks, and compare them to the networks of SNs and WWW network. We obtained data for five successful SCCs with nearly 2 million vertices and almost 6 million edges, as well as data for four popular SNs: Flickr, YouTube, Orkut, and LiveJournal with a more than 11 million vertices and over 328 million edges. We also obtained WWW data containing over 18 million vertices and over 64 million edges.

We extracted association networks based on the association actions (*i.e.* follows and befriending actions) of SCC members, and SNs and linking between pages in the WWW. We extracted interaction networks based on interactions between members of SCCs. We analyzed these networks and compared them with each other. We found that although SCCs have fewer members than SNs and SCC members are less likely to reciprocate linking, SCCs show higher association densities than SNs and the WWW. SCCs show a different association pattern than SNs, one where members connect to others with significantly different degrees. We also found that popular members (those with many followers) are less likely to be active members (those who follow many others) and that popular and active members are less likely to associate with each other. The following is a list of this chapter's key findings:

SCCs are smaller and denser than SNs and the WWW:

1. SCCs have significantly fewer members and connections in comparison to SNs, which means that SCCs are smaller than SNs.

2. SCC associations and interactions are denser than the WWW and SNs.

SCC members are more likely to reciprocate interaction than association links:

3. Members of SCCs are less likely to reciprocate association links in comparison to SNs.

4. Members of SCCs are likely to reciprocate interaction.

Similar to SNs and the WWW, SCC networks follow a power-law distribution. However, the findings suggest that SCC members have unique linking behaviour. Specifically, a few members tend to associate with many and are therefore considered highly active. Furthermore, many members interact with a few popular members:

5. SCC association and interaction networks, SNs, and the WWW follow a power-law distribution.

6. SCC association networks' out-degrees are more concentrated than their in-degrees.

7. SCC interaction networks' in-degrees are more concentrated than their out-degrees.

8. SNs have a similar distribution of in- and out-degrees.

9. The WWW network's out-degrees are more concentrated than their in-degrees.

10. Most members in SCC association networks have larger in-degrees than out-degrees.

11. Most members in SCC interaction networks have larger out-degrees than in-degrees.

12. Most members in SNs have similar out-degrees and in-degrees.

13. Most pages in the WWW network have larger out-degrees than in-degrees.

Our findings further confirm the unique linking behaviour of SCCs. Specifically, SCC members tend to link with others who have different degrees rather than those with similar degrees:

14. Members in SCC association networks tend to connect to others who have significantly different degree values.

15. Members in SCC interaction networks tend to connect to others who have significantly different degree values.

16. SN members tend to associate with others with similar degree values.

17. Pages in the WWW network tend to associate with others with similar degree values.

Interestingly, active SCC members are not the same as those who are popular:

18. Members in SCC association networks who have high in-degrees are not likely to have high out-degrees.

19. Members in SCC interaction networks who have high in-degrees are not likely to have high out-degrees.

20. SN members who have high in-degrees are likely to have higher out-degrees than the WWW and SCCs.

21. Pages in the WWW network that have high in-degrees are not likely to have high out-degrees.

As future research, we will:

1. assess the adherence of SCC interaction and association network to the small-world property [45],

2. assess the degree of which members form sub-communities [137] through interactions and associations,

3. assess the degree of which members exhibit assortative mixing [138] in their interactions and associations,

4. assess the adherence of SCC interaction and association network to the scale-free property of power-law networks [139, 140],

5. identify patterns in the evolution of these networks, and

6. produce a topology-generating model that mimics these networks.

This chapter presents a details structural analysis of several SCC interactions and association networks as well as several SNs and the WWW networks. The analysis results in identifying common structural properties among SCC networks. It also results in identifying several structural properties that SCC networks share with SNs and the WWW and several structural properties that are unique only to SCCs. The next chapter presents a detailed analysis of SCC member interactions, their evolution, and factors that affect them. It also presents a comparison of these patterns with interaction patterns on Facebook.

# Chapter 5

# INTERACTIONS IN SOCIAL CROWDSOURCING COMMUNITIES[1]

As Section 3.1 states and Chapter 4 shows, SCCs differ from SNs in several ways, including the purpose and goals of these communities, the reasons underpinning the interest of members to join, and the features offered by their platforms, which reflect on the behaviour of their members.

These differences raise a question regarding the effect they have on the way members interact. Thus, we perform a detailed analysis on pairwise interactions among members of *HCC* in this chapter, which results in identifying interaction patterns among SCC members, the evolution of these interactions, and the key factors that affect them. We compare some of the findings from this analysis with findings from an analysis of Facebook's interactions. This analysis provides insight into how and why SCC members interact in the way they do, which is critical for capturing and modelling behaviour in SCCs.

We adopt and adapt various methods, including the methods used by Viswanath *et al.* [128], to analyze this SCC across the five years starting from the day of its inception. The improvements and adaptation of existing analysis methods to SCCs is an essential contribution of this chapter. Moreover, answering these particular questions is a stepping-stone to solidifying understanding of interaction patterns, factors that affect them, and how they manifest on a larger scale, which demonstrates the importance of interactions in SCCs, and identifies the key features a SCC platform should have to encourage interactions. Answering these questions is vital to establish the expectations around interactions, which has benefits in tracking the health of these communities [141, 142], and in modelling the evolution and growth of interactions within them [143].

The data description, the definitions of relevant concepts, and the high-level analysis of the

[1] **The results in this chapter are published [2].**

interactions appear in Section 5.1. This initial step helps us identify important characteristics of the data that dictate the design of the upcoming analysis and the choice of methods. First, we confirm a key difference between SCC interactions and other commonly studied interactions. Specifically, the associations among the SCC members do not bind their interactions, and the interactions are not indicators of the degree of friendship; instead, they convey information about shared interests among co-workers. Second, we find that there are two types of member-pairs in our SCC, namely *frequent* and *infrequent*. Because of this, we consider the two separately when we look at the interaction patterns in Section 5.2 and the factors that affect interactions in Section 5.3. Specifically, Section 5.2 explores member responsiveness, artifact preference, and interaction rate and duration while Section 5.3 looks at the relationship between SCC members' response time to a new query or artifact and associations and rewards. Third, we examine the overall distribution of interactions over time and find a trend and seasonality in our data that shapes the study of interaction network evolution in Section 5.4. This section shows that the interaction network is characterized by high interaction volatility and remarkable structural stability as it evolves. This stability, despite the high volatility, is underpinned by a high replenishment rate of interacting members and the tendency of members to interact with new members.

Note that this work is limited to investigating only one SCC. However, the outcomes from this chapter may apply to other SCCs because of the structural similarities among SCCs in general (see Chapter 4). We leave answering this question to future work.

## 5.1 DATA OVERVIEW AND HIGH-LEVEL ANALYSIS

This section provides an overview and high-level analysis of our data, which influence the choice of methods in the following sections. After providing the data description, we start the analysis by analyzing the relationship between associations and interactions in Section 5.1.2. This analysis helps us confirm a key difference between the commonly studied interactions and SCC interactions - the associations among SCC members do not bind their interactions. The analysis also reveals

a relationship between member associations and member interactions, where an increase in one corresponds to an increase in the other. Section 5.3 investigates this relationship further. As Section 5.1.3 details, our SCC members interact at varying rates that in their totality adhere to the Pareto principle, where nearly 20% of interacting member-pairs produce nearly 80% of interactions. Thus, we derive two interaction groups using the Pareto principle and investigate each group separately in the following sections. Additionally, Section 5.1.4 provides a simple time-series decomposition, which indicates that there is a recurring yearly pattern (seasonality), and a trend in our data. These findings guide our analysis of the interaction network evolution in Section 5.4.

### 5.1.1 Data Overview

Our data comes from an active private[2] SCC, namely *HCC*, which consists of brand loyalists of a major mobile phone company. The SCC is focused on brand advocacy, brand loyalty, and marketing. This typical SCC has all the key SCC characteristics described in Section 3. It is comprised of 5,436 carefully selected and screened members who participate in activities related to mobile devices and their accessories in return for extrinsic virtual and physical rewards, such as exclusive access to company information and frequent giveaways of highly desired merchandise. This chapter presents an analysis of the anonymized log of member interactions, which consists of 779,673 interactions between 128,255 distinct member pairs (we refer to them as *pairs* for short) over five years starting from January 13, 2012, until January 31, 2017. The end date was chosen arbitrarily.

We identify two types of interactions, *direct* and *indirect*. The direct interactions occur when members send private messages to each other. This chapter focuses on the indirect interactions, which occur when members perform actions on *artifacts* (explained below) that other members and moderators create. These actions include responding to a query, commenting on a response, or another comment, and flagging a comment. The query, the response, and the comment in this

[2]We obtained the SCC data from Chaordix (`http://www.chaordix.com/`). Our data access agreement mandates that the data remains confidential.

example are the artifacts.

In total, our SCC has four types of artifacts with which members can interact in thirteen different ways. The artifacts are hierarchical because they are produced in response to other artifacts at a higher level. At the top of this hierarchy are queries made by SCC moderators. Second-level artifacts are responses (comments) to these queries, and third-level artifacts are comments on responses and other comments. These artifacts are created at different stages of the social crowdsourcing process where responses are created in reply to queries, and comments in reply to responses and other comments. Section 5.2.1 explores the members' propensity to join the social crowdsourcing process through these artifacts.

### 5.1.2 Association and Interaction Out-degrees

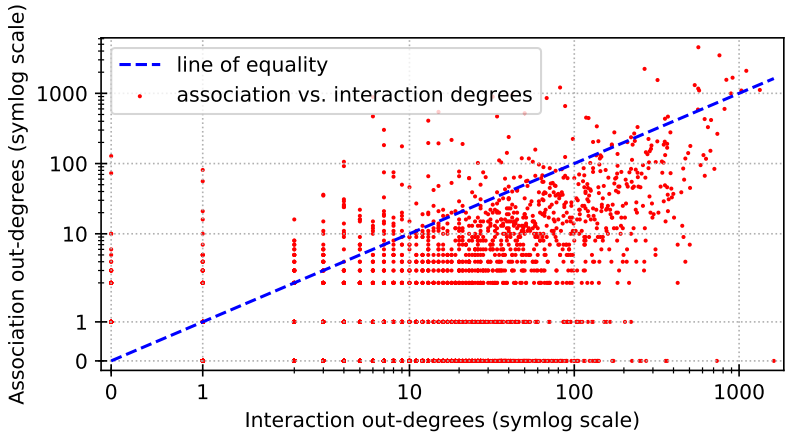



Figure 5.1: The members' association out-degrees and interaction out-degree.

Because of the default privacy settings, some interactions in SNs are only possible among friends [128, 144], such as posting on someone else's Facebook Timeline (commonly known as the Facebook wall) while others, such as discussions in Facebook groups, or over artifacts, do not have this restriction. The collaborative nature of SCCs requires unrestricted interactions. Therefore, the interactions in this SCC can occur between members who are not associated with each other. To capture this observation, we plot each member's number of associations made (*i.e.* the number of members followed by this member) against the number of members with which this member

interacted over the entire period in Figure 5.1. We name these numbers *association out-degree* and *interaction out-degree*, for short.

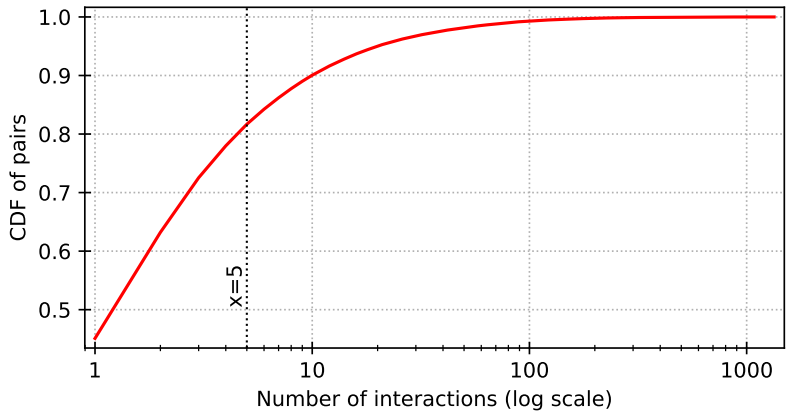
The figure shows that the majority of members in this SCC interact with more members than they associate, which is consistent with other SCCs (see Chapter 4). Specifically, 93.5% of members have higher interaction out-degree than association out-degree. The figure also shows a linear relationship between association out-degrees and interaction out-degrees. Precisely, these two have a 0.56 Spearman rank-order correlation coefficient[3] with a 2-tailed $p$-value of zero. These observations show that associations do not bind interactions and therefore interactions are not indicative of the degree of friendship among interacting members; instead, this kind of interaction conveys the degree of common interest in the content of artifacts among the interacting members. Nevertheless, this correlation coefficient implies the presence of a relationship between associations and interactions, so we investigate this relationship further in Section 5.3.

### 5.1.3 Interaction Distribution over Pairs

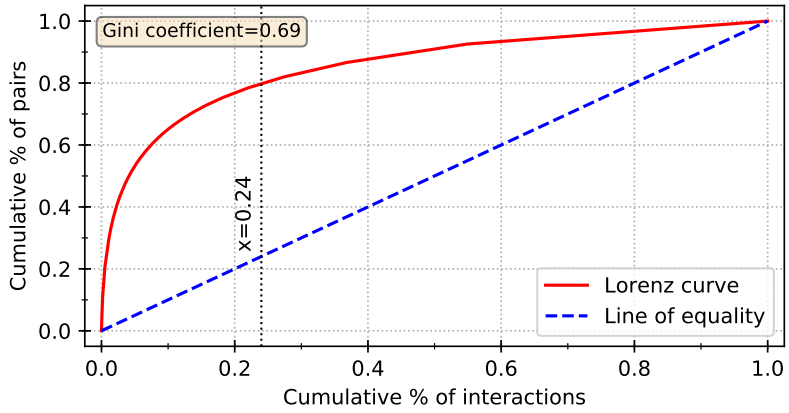
We now consider the distribution of interactions over pairs. Similar to Viswanath *et al.* [128], we plot the CDF of interactions among all pairs. Figure 5.2(a) shows that this distribution is highly skewed with the vast majority of pairs having few interactions. This distribution means that relatively few highly active pairs produce the majority of interactions. To confirm this, we compute the Gini coefficient [132] and plot the Lorenz curve [145] in Figure 5.2(b).

The Gini coefficient encodes the degree of inequality in the interaction distribution over pairs. It ranges between zero and one. A value of zero indicates perfect equality where each pair owns the same percentage of interactions as do the others, and a value of one indicates a complete inequality where a tiny percentage of pairs own nearly all interactions. The Lorenz curve is a graphical representation of the inequality in the distribution of interactions among pairs. The Gini coefficient value is high (0.69), which means that the distribution of interactions over pairs exhibits

[3]We chose this approach because the distributions of the association and the interaction degrees do not have a defined mean and finite variance.




(a) The cumulative distribution function (CDF) of the number of interactions among pairs.




(b) The Lorenz curve of interaction ownership among pairs (axis flipped to ease comparison with Figure 5.2(a)).

inequality. Furthermore, Figure 5.2(b) shows that the lowest 80% of pairs, in terms of interaction, produce 24% of interactions and the top 10% produce over half of all interactions. This distribution is in line with the Pareto principle [146], which states that approximately 80% of the effects (*i.e.* interactions) come from 20% of the causes (*i.e.* pairs).

To accommodate the high degree of skewness and to ensure consistent analysis in the following sections, we split our pairs into two groups, based on their number of interactions. This split also helps identify patterns relating to interaction frequency and identify interaction patterns of the most active pairs as well as the majority of pairs. Since the distribution of interactions over pairs adheres to the Pareto principle, we set the threshold at the $80^{\text{th}}$ percentile, or five interactions, as indicated in Figure 5.2(a). We refer to these two groups as *infrequent* (five or less) and *frequent* (six or more).

### 5.1.4 Interaction Distribution over Time

Similar to Viswanath *et al.* [128], we examine the distribution of all interactions over time. Figure 5.3 shows rapid growth in the number of interactions within the first 16 months culminating in a large spike, which is followed by a gradual decline across the remaining months with additional smaller spikes whose amplitude also consistently declines over time. We conjecture that this pattern reflects the reality of the start of a new and growing community and the adoption of SCC as new technology. However, analysis of other SCCs starting at their inception is required to confirm this conjecture.

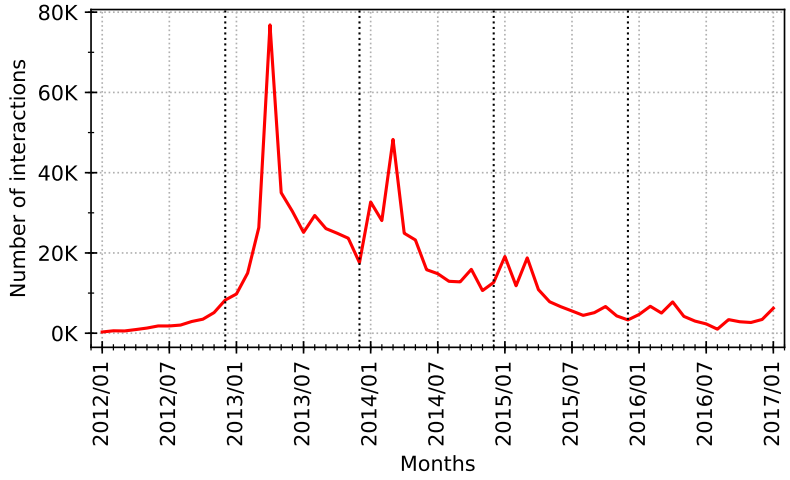



Figure 5.3: The interaction distribution over time.

There are two patterns in the data that are visible in Figure 5.3. There is an initial ramp-up in activity that culminates in a large spike at the 16th month, which is followed by a gradual decline for the remainder of the data. Additionally, there is a recurring (*i.e.* a seasonal) pattern that happens roughly every year, except for the first year where the interaction frequency strictly increases. To confirm these patterns, we perform simple time-series decomposition[4] [147] presented in Figure 5.4. Since the magnitude of the data changes as time progresses, we choose the multiplicative decomposition model. As can be seen from the peaks and the valleys in the data, the seasonal cycle

[4]We use the *seasonal_decompose()* method, which is included with version 0.9.0 of the Python StatsModels package (`https://www.statsmodels.org`).

is roughly 12 months. We confirm this number by computing a periodogram from the de-trended data and use it as the period parameter in the time-series decomposition above.

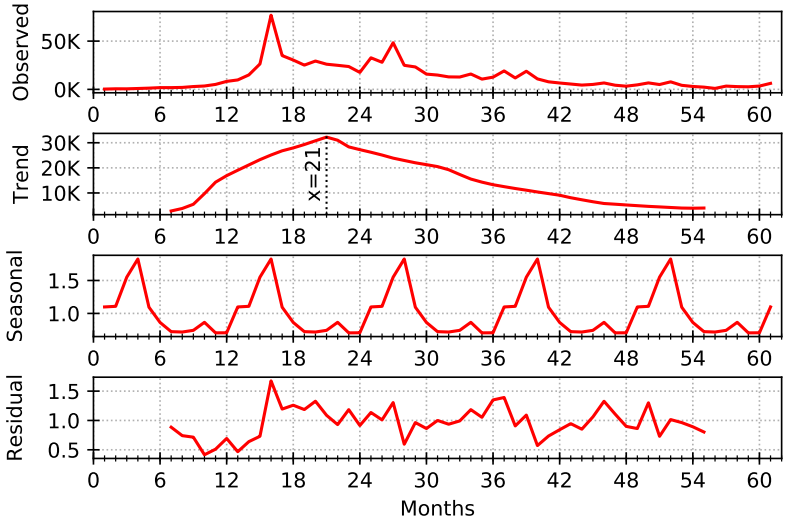



Figure 5.4: The time-series decomposition of the interactions data.

## 5.2 PATTERNS OF INTERACTION

This section builds on the distribution inequality finding from Section 5.1.3 by investigating interaction patterns that the most active interactors exhibit and those of the remaining majority. Specifically, this section identifies patterns in four aspects of interactions: response time, artifact preference, rate, and duration. The analysis of members' response times and preferred artifacts in Section 5.2.1 shows that all pairs are highly responsive and that the frequently interacting pairs were distinctly more responsive than the infrequent pairs. Furthermore, the analysis shows that the pairs of each interaction groups play unique (and crucial) roles in the social crowdsourcing process. Frequently interacting pairs play a primary role in responding to queries, and infrequently interacting pairs play a primary role in initiating discussions around these responses. The analysis of interaction rate and duration in Section 5.2.2 shows that interactions of frequently interacting pairs decay at a slower rate and last longer than interactions of the infrequently interacting pairs, while the latter pairs tend to have bursts of interactions that end quickly.

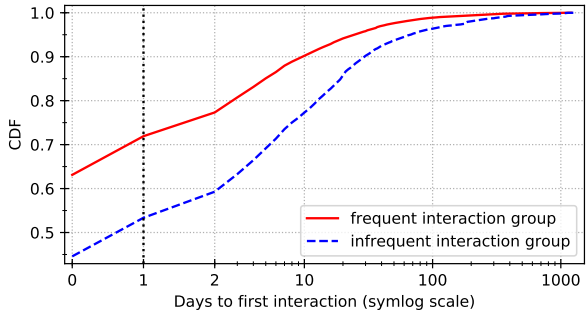
### 5.2.1 Response Time and Artifact Preference

First, we identify patterns in the response time of both interaction groups over different artifact levels: queries (first-level), responses to queries (second-level), and comments (third-level). Specifically, we identify the artifact levels that attract members of each interaction groups to join the social crowdsourcing process, and we examine the time that it takes for members to act in response to interactions (*i.e.* their response times). Since members can interact with an artifact multiple times, we only consider the first interaction of each member with each artifact. We measure each member's response time to an artifact as the number of days between the time that artifact was created and the time that member interacted with it for the first time. We define the response time of members to artifacts that were created before they joined the SCC as the number of days between the dates these members joined the SCC and the dates they made their first interaction with the artifacts.

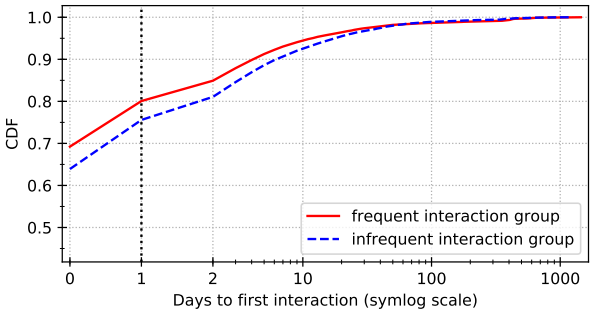
The pairs in the frequent interaction group have distinct preferences in the levels of artifacts over which they interact from those in the infrequent interaction group. Frequently interacting pairs are 1.63 times more likely to interact with queries than the infrequently interacting pairs, while the latter pairs are 1.59 times more likely to interact with responses to queries than the former. In addition, frequently interacting pairs are 1.62 times more likely to interact with comments on responses and comments on comments than pairs of the infrequent interaction group. We obtain these comparative ratios by, first, obtaining the percentage of first-interactions that occur over a specific artifact level from all interactions for each interaction group. We then obtain the ratio for each artifact level by dividing the corresponding percentages.

To compare the response times between frequent and infrequent groups, we plot the cumulative density function (CDF) of the days it takes members to make their first interaction with first-level artifacts (see Figure 5.5(a)), and second- and third-level artifacts (see Figure 5.5(b)). The vast majority of members make their first interaction within the first week. We see in Figure 5.5(a) that 45% the first interactions in the infrequent group occur during the same day on which the

query is created, and 53% occurred within two days. Compared to that, members in the frequently interacting pairs are more likely to contribute on the same day (63%) or by the following day (72%). During the first two days, there is almost a 20% difference in the percentages of interactions that occur with first-level artifacts and about 5% difference in the percentages concerning the second- and third-level artifacts between the two interaction groups.




(a) Interactions with first-level artifacts.




(b) Interactions with second and third-level artifacts.

These results indicate that both frequently and infrequently interacting pairs have short response times and have a distinct tendency to interact with specific levels of artifacts. This tendency implies that each interaction group plays a distinct and crucial role in the social crowdsourcing process. Specifically, those in the frequent group are more likely to produce responses to top-level queries, and they are quicker at doing so. Infrequently interacting pairs are more likely to kick off discussions around these responses, which naturally draws further interactions from the community and motivates additional discussions. These discussions are essential to the SCC process because they allow members to build on each other's ideas.

### 5.2.2 Interaction Rate and Duration

To quantify the degree to which the interaction rate changes over time for the different interaction groups, we plot the percentage of interactions made by each interaction group. To demonstrate the relationship between interaction frequency and interaction rate over time, we split the frequently interacting pairs into subgroups based on their number of interactions. For each pair in each subgroup, we count the number of interactions both members of the pair had with each other every

30.5-day interval[5] since their first interaction. For each subgroup, we aggregate the number of interactions its pairs made each monthly interval, and then derive the monthly percentages. We report the results in Figure 5.6, along with the statistically significant ($p$-value $\leq 0.05$) linear regression coefficients, referred to hereafter as the slope ($\beta$).

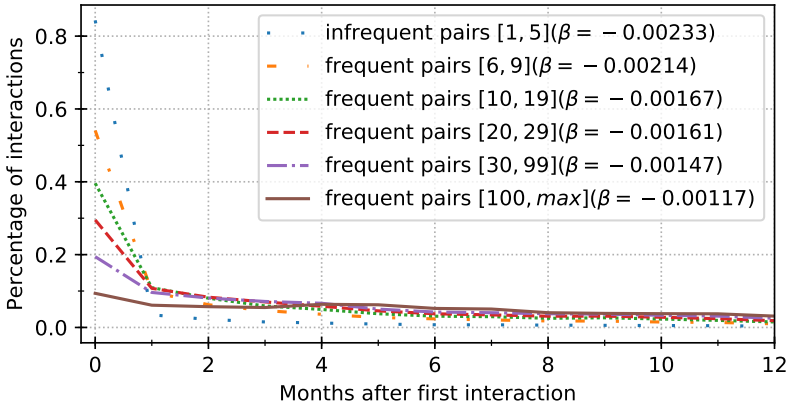



Figure 5.6: The percentage of interactions made each month for each interaction subgroup. The ranges and the $\beta$ in the legend represent the range of interactions and the slope of the linear regression line in the respective subgroup. The x-axis stops at 12 months to enhance the visualization, but the pattern does not change significantly after that, and the downward trend continues until the last day in the data

We make three observations from Figure 5.6. First, there is a general decline in the interaction rate over time, which indicates that pairs interact less as more time passes since their first interaction. The negative $\beta$ values in the figure's legend quantify the interaction rate decay of each group. Second, the interactions of frequently interacting pairs decay at a slower rate than lower interaction frequency, which is why $\beta$ values decrease as the interaction frequency increases. Third, the majority of interactions (˜85%) in the infrequently interacting group occur within the first month, which is not consistently the case in the frequently interacting group. In the latter interaction group, the percentage of interactions that occur within the first month is significantly smaller in the subgroups with higher interaction frequency than those with lower interaction frequency. Overall, all pairs have the highest number of interactions within the first month following their first interaction.

These observations mean that pairs who interact frequently tend to be more consistent in their

[5] 30.5 is the average number of days in one month.

interactions and for longer periods than the infrequently interacting pairs, who tend to have bursts of interactions that decay quickly. To confirm this, we find the interaction duration by plotting the percentage of pairs who no longer interact after a certain number of months have passed (see Figure 5.7). There is a clear relationship between the interaction duration and frequency among the pairs. Specifically, as the interaction frequency increases, so does the interaction duration. Approximately 9% of pairs with more than 100 interactions stop interacting within a year; on the other hand, nearly 97% of infrequently interacting pairs stop interacting within the first year and ˜83% cease to interact after one month.

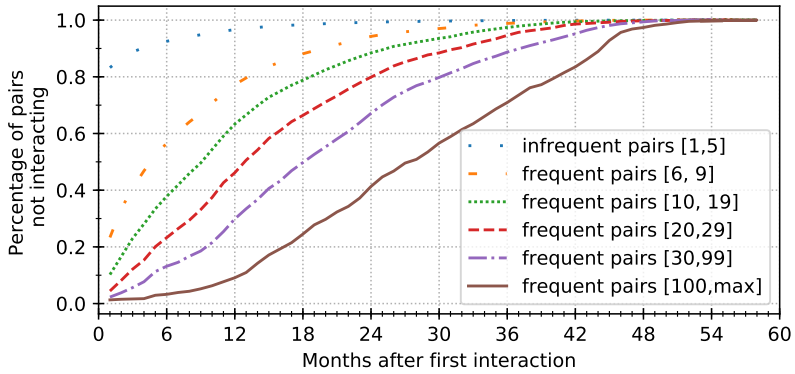



Figure 5.7: The percentage of pairs who no longer interact after a given number of months.

## 5.3 FACTORS THAT AFFECT INTERACTION RESPONSE TIME

Section 5.1.2 identifies a positive correlation between association and interaction, where an increase in one corresponds to an increase in the other. Furthermore, Section 5.2.1 shows that our pairs, in general, are highly responsive, with the pairs in the frequent interaction group being more responsive than the infrequent ones. These two findings raise a question about the effect that associations (and other factors) have on responsiveness. Therefore, this section investigates this relationship by analyzing the effect that *follow* and *rewards* have on response times in each interaction group.

The hypothesis is that *follow* improves member responsiveness by design because following a

specific member or artifact increases the follower's exposure to "interesting" artifacts and notifies the follower immediately of new activity. In addition to increasing member responsiveness, *follow* also increases the number of interactions by increasing the likelihood that members see more artifacts of interest. Moreover, *rewards* increase member responsiveness because they increase engagement in knowledge sharing [148]. "Rewards" in this SCC specifically refer to badges. SCCs employ different gamification techniques, including badges, point systems, and leader boards, to increase member engagement in the SCC. Badges have predefined meanings and requirements. The SCC platform automatically awards these badges to members who fulfill their requirements. For example, badges are awarded to members who log in for a certain number of consecutive days or respond to a certain number of queries. Members can receive a badge multiple times and are notified each time they receive it.

Section 5.3.2 examines the effects of *follow* and *rewards* on the responsiveness in each interaction group using statistical hypothesis testing and Section 5.3.3 further examines this effect through visualization. The results indicate that *follow* and *rewards* are associated with shorter response times for both frequently and infrequently interacting pairs; *follow* more so than *rewards*. This effect is most substantial when both factors are present. We also show that *follow's* effect on response times diminishes relatively quickly in both interaction groups and that the effect of *rewards* remains consistent over time. These findings are clear indicators of the vital role that social and gamification aspects play to encourage interactions.

### 5.3.1 Data

Recall from Section 5.1 that our data contains both direct and indirect interactions. Here, only the indirect interactions are relevant because *follow* is undefined for the direct ones, and there are no *rewards* for them either. Additionally, a member may have multiple interactions with the same artifact; thus, similar to Section 5.2.1, we only consider the first interaction the member has with the artifact. The response time of a member to an artifact becomes the difference between the creation time of this artifact and the time of the member's first interaction with it. However, in this

analysis, we consider the difference in hours instead of days because the analysis and visualization require this finer granularity.

We partition the interactions in each interaction group into four test groups based on *follow* and *rewards*, and label each accordingly. The first letter in the label corresponds to *follow*, and the second corresponds to *rewards*. Each letter can be either "T" or "F", denoting "true" and "false". They indicate that before the interaction, the interacting member was:

- TT - following the source and had previously received rewards,
- TF - following the source and had *not* previously received rewards,
- FT - *not* following the source and had previously received rewards,
- FF - *not* following the source and had *not* previously received rewards.

### 5.3.2 Statistical Significance of the Effect

First, we attempt to identify the appropriate distribution that fits each test group's data by using Kolmogorov-Smirnov (KS) test for goodness of fit to test against 80 known distributions and find that the data is not consistent with any of them. We also perform Levene's test [149] to determine if any of the test groups have identical variance, which they do not. Therefore, we use the non-parametric KS test[6] to answer four questions about the effects of *follow* and *rewards* on member responsiveness:

1. FF_TT[7]: *Do follow and rewards (together) have any effect on response time?*
2. FF_TF: *Does follow alone have any effect on response time?*
3. FF_FT: *Do rewards alone have any effect on response time?*
4. TF_FT: *Which one of the two has more of an effect on response time?*

[6] We use *ks.test()* function offered in R's [150] stats package version 3.4.3.
[7] Labels are described in the following narrative.

To determine if the factors associate with reduced response times, we use a one-tailed KS test for each direction. We perform each set of tests for both interaction groups and report the results in Table 5.1.

Table 5.1: The results of the one-tailed KS hypothesis tests. The "greater *p*-value" column contains the *p*-values of the KS test with the alternative hypothesis that first test group (group 1) is greater than second test group (group 2). The "lesser *p*-value" column is for the opposite alternative hypothesis.

| **test** | **interaction group** | **group 1 size** | **group 2 size** | **greater *p*-value** | **lesser *p*-value** |
|---|---|---|---|---|---|
| FF_TT | Frequent | 53,915 | 191,343 | 0.997148 | **0** |
| | Infrequent | 19,606 | 28,046 | 0.999388 | **9.37E-245** |
| FF_TF | Frequent | 53,915 | 1,027 | 0.998370 | **3.30E-44** |
| | Infrequent | 19,606 | 1,461 | 0.979509 | **4.50E-34** |
| FF_FT | Frequent | 53,915 | 203,232 | 0.890705 | **0** |
| | Infrequent | 19,606 | 113,049 | 0.999739 | **1.09E-113** |
| TF_FT | Frequent | 1,027 | 203,232 | **1.04E-13** | **9.19E-05** |
| | Infrequent | 1,461 | 113,049 | **9.35E-16** | 0.112772 |

To answer question (1), we compare the distribution of the test group with response times that are not affected by either *follow* nor *rewards* (FF) with that of the test group that was affected by both *follow* and *rewards* (TT). The distribution of the TT group is significantly greater than that of the FF group, so we conclude that both *follow* and *rewards* together are associated with shortening of response times in both frequent and infrequent groups. To answer questions (2) and (3), we compare the response time distributions of the FF group with TF and FT and find that the distributions of the FF group for both interaction groups are significantly shorter than that of FT and TF groups. This result means that regardless of the interaction group when the SCC members utilize only one of the two factors, we expect to observe a statistically significant reduction in their response time.

Note that the *p*-values are approximated due to the presence of ties among the test groups. Thus, directly comparing *p*-values resulting from different tests is not valid. For this reason, to answer question (4), we perform a separate test to determine whether *follow* or *rewards* have

a stronger effect on responsiveness. The results show that for the infrequent interaction group, *follow* had a more significant impact on reducing response time than *rewards*. However, in the frequent group, the KS test shows that the distribution of FT is different from TF in both directions. More precisely, the maximum positive difference between the empirical distributions of TF and FT groups and the maximum negative difference are both statistically significant. We investigate this further in Section 5.3.3.

### 5.3.3 Visualizing the Effect

The statistical hypothesis testing shows that the response times in both interaction groups reduce when one or both factors are present. However, it does not indicate the strength of this effect. Moreover, it raises questions regarding the nature of this effect when comparing the TF and FT groups for the frequent interaction group, where the one-sided KS tests indicate that the two distributions are different in both directions simultaneously. To elucidate this and to gain more insight into the other results presented in Table 5.1, we plot the response time CDF for each of the four test groups of each interaction group and present them in Figure 5.8.

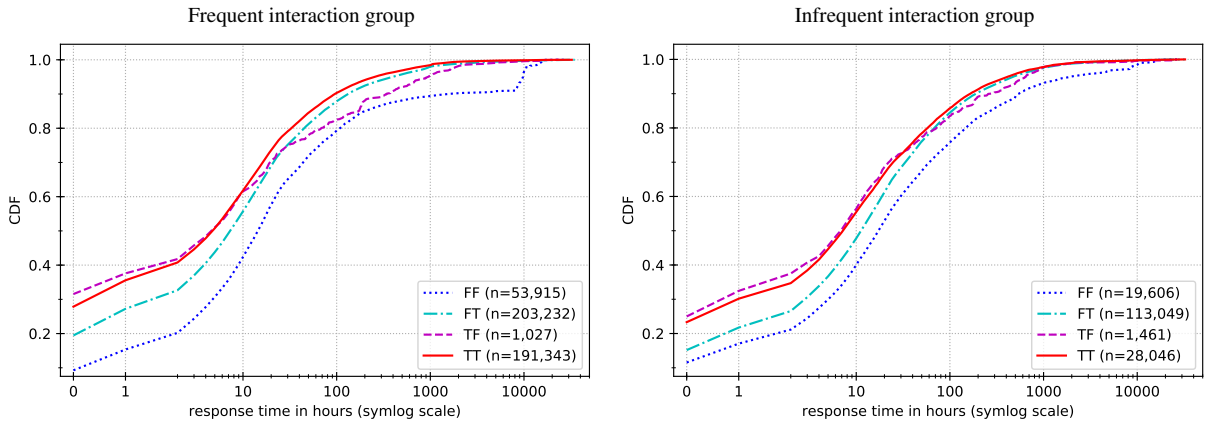



Figure 5.8: The CDF of response times of the frequent (above) and infrequent (below) interaction groups.

The figure shows that the pairs of both interaction groups are highly responsive and the majority of members in both frequent and infrequent interaction groups make their first interaction with an

artifact within 20 hours of its creation time. It also shows that frequently interacting pairs are more responsive than the infrequently interacting pairs, except during the first nine hours in the FF groups where the opposite is true. There is also a clear ranking of the test groups' CDFs in both interaction groups, where the TT groups consistently have the shortest response times, followed by FT then FF, which is evidence of the effect the two factors (separately and collectively) have on response times.

The TF distribution is inconsistent with the others in both interaction groups. The CDFs of both TF groups start with the highest values as compared to the other corresponding test groups indicating that members who *followed* but did not receive *rewards* (TF) are quicker to respond than those who *followed* and received *rewards* (TT). After several hours, the CDF lines for both TF groups cross over TT and later FT also. This pattern is more pronounced in the frequent interaction group than the infrequent one, which is consistent with the TF_FT KS hypothesis test result of the frequent interaction group that indicates that TF is both larger and lesser than FT (see Table 5.1). This cross-over does not happen in the test groups where *rewards* are present with *follow* (TT), nor that it occurs when only *rewards* are present (FT), which means that *rewards'* effect over time differs from that of *follow*. Specifically, the effect of *follow* alone (TF) is stronger in the short term than the effects of the combined *follow* and *rewards* (TT), and the effect of *rewards* alone (FT) for both interaction groups. However, this effect diminishes faster than the others' effect.

Both *follow* and *rewards* associate with shortening of response times in both interaction groups. The frequent interaction group exhibits shorter response times than the infrequent group. Since we separate the frequent from infrequent interaction groups in Figure 5.8, it may be difficult to see the difference between the two interaction groups. To make this clearer, we present the number of hours that have elapsed since artifact creation until 50% and 75% of the responses have occurred within each interaction group in Table 5.2. From the table, we see that the frequent interaction group has consistently shorter response times than the infrequent interaction groups, regardless of *follow* and *rewards*.

Table 5.2: The number of hours passed before 50% and 75% of responses were made.

| interaction group | 50% | | | | 75% | | | |
|---|---|---|---|---|---|---|---|---|
| | FF | TF | FT | TT | FF | TF | FT | TT |
| frequent | 14 | 5 | 7 | 5 | 66 | 30 | 30 | 22 |
| infrequent | 17 | 7 | 11 | 7 | 93 | 40 | 47 | 38 |

## 5.4 INTERACTION NETWORK EVOLUTION

Section 5.2.2 shows that the interaction rate diminishes quickly for most pairs and interactions last a short period, which suggest that interactions are highly unstable. However, Section 5.1.4 shows that the interaction distribution has an overall trend and seasonal pattern, which indicate that the effects of the rapidly declining interaction rate and short duration of interactions do not manifest as instability. Therefore, this section investigates the stability of the interactions over time and identifies the factors that affect it by analyzing the structure of a series of networks constructed from the interactions over time to determine how interaction patterns we previously identified, manifest on the highest level. In other words, this section analyzes the structure of the interaction network as it evolves across the five years. Specifically, Section 5.4.1 assesses the stability of the interaction network structure over time by examining the distributions of several graph-theoretic properties. Section 5.4.2 assesses the persistence of the interaction network edges over time by examining the degree to which edges remain active over time.

These analyses find that the interaction network exhibits lower edge persistence than that of Facebook [128], yet similar to Facebook, it exhibits stable structural properties across time. This stability, despite the low edge persistence, is due primarily to an underlying phenomenon, which leads to the production of new edges (*i.e.* the formation of new pairs that did not exist before) in the interaction network. The underlying reasons for this phenomenon are two: 1) the tendency of members to interact with others whom they never interacted with before, and 2) the high replenishment rate of interacting members. We show that this phenomenon has a quick and direct effect on the interaction network in this SCC.

To do this analysis, we split our interactions into bins based on the month during which they occurred. We represent the interactions in each bin as a simple undirected graph, where nodes correspond to the SCC members and an edge between two nodes represents one or more interactions between a pair of members, so we name it the *interaction network*. In other words, we produce a monthly snapshot of the interaction in the SCC, which yields 60 snapshots for the following analysis.

Interactions in our SCC are directional; one member acts on another member's artifact. Moreover, there could be more than one interaction in the same direction between the same two members. Thus, when representing the interactions as a graph, one would consider a directional graph with weighted edges. However, similar to Wilson *et al.* [151] and Ahn *et al.* [125], when considering the purposes of this section, we find that undirected unweighted graph is sufficient considering the information that is encoded by the edge weight and its direction. The direction of edges encodes information from the member perspective, such as the levels of popularity, activity, influence, and reachability of members, while the weight of the contributing edges encodes the strength of the relationship between interacting members. In this section, we are concerned with overall connectivity, regardless of the member perspective. Thus, we represent our interactions as a simple undirected graph with no weights.

### 5.4.1 Structural Properties over Time

To identify the effect of the interaction patterns that we previously identified on the stability of the interaction network over time, we compute several popular graph-theoretic structural properties of the interaction network for each snapshot and compare them across time (see Figure 5.9). These properties are path length, diameter, density, global transitivity (clustering coefficient), betweenness, and node degrees[8]. Diameter, density, and global transitivity are global properties, so we report them directly. Path length, betweenness, and node degrees are properties concerning one or more nodes, so we aggregate them to obtain global properties for each snapshot. The distribution of

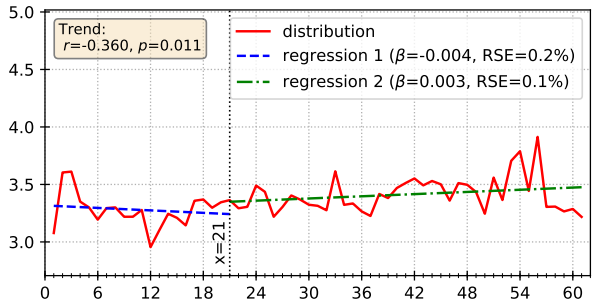
[8]We used python-igraph package version 0.7.1 to calculate these properties.

path lengths follows a normal distribution, so we report the average path length for each snapshot using arithmetic mean. The betweenness and node degrees follow a heavy-tailed distribution for which we cannot measure central tendency using the arithmetic mean, so we use median instead because it is resistant to outliers.

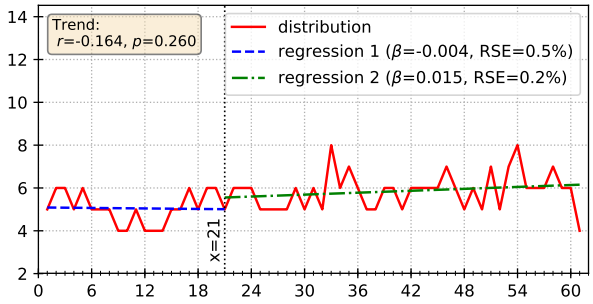
We split each property's distribution at the 21$^{st}$ month where the interaction ramp-up period ends, as the overall interaction distribution trend indicates in Figure 5.4. We perform linear regression on each half and report the slopes ($\beta$) and Relative Standard Errors (RSE) in the corresponding figure. Furthermore, we report on each figure the correlation coefficient ($r$) and the 2-tailed $p$-value ($p$) that result from a Pearson correlation analysis of the corresponding graph-theoretic property with the overall trend reported in Figure 5.4.

Figure 5.9 shows that most structural properties are sensitive to the overall trend. Specifically, we observe two trends in these distributions where some structural properties increase in value for the first 21 months then decrease for the remaining period, while other properties show the opposite trend of decreasing in value for the first 21 months then increasing for the remaining period. The statistically significant $r$-values show that these two trends relate to the overall trend. Recall that the overall trend shows a ramp-up of interactions for the first 21 months. This increase in interactions increases the network connectivity, which manifests in the increase of node degrees, betweenness, density, and transitivity properties.

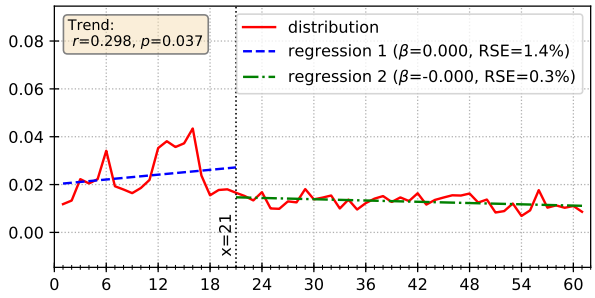
Conversely, this increase in interactions (and consequently in the network connectivity) reduces network diameter and path lengths. The remaining period of the overall trend shows a decrease in interactions, which corresponds to increased path lengths and diameter; and decreased node degrees, betweenness, density, and transitivity. We note that the median betweenness values (see Figures 5.9(e)) fluctuate across the different snapshots more than the other properties. These fluctuations correlate with the seasonality pattern we identified in Figure 5.4, where every year exhibits a peak betweenness value around the 3$^{rd}$ month except for the second year, where the highest value is at the 11$^{th}$ month. A statistically significant $r$-value of 0.406 further confirms that this property



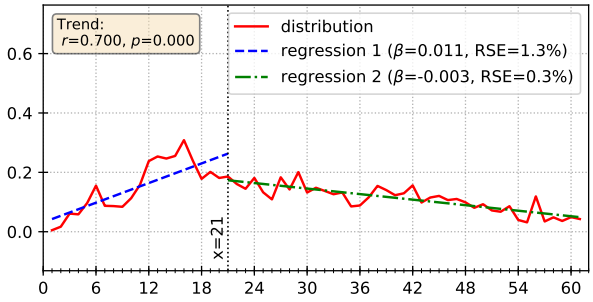

(a) Average path length.



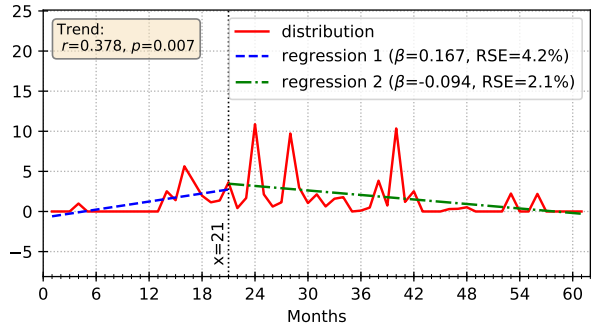

(b) Diameter.



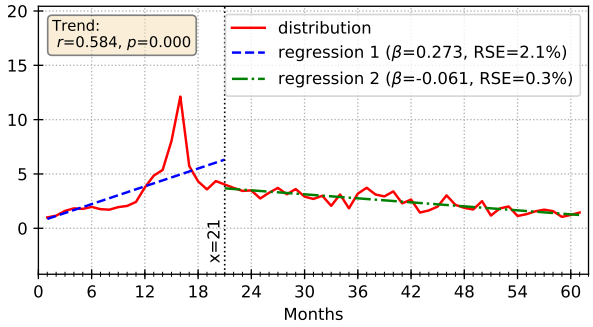

(c) Density.




(d) Transitivity (clustering coefficient).




(e) betweenness.




(f) Median node degrees.

Figure 5.9: The distribution of several graph-theoretic properties of the interaction network over the time. We scale the y-axes using the range $[\mu + 10\sigma, \mu - 4\sigma]$ to minimize bias. We arrived at standard deviation coefficients (*i.e.* 10 and 4) empirically through trial and error. We chose them, such that they are the smallest integer values possible that keep all the distributions from overlapping with their respective legend or text, or extend beyond the boundaries of their respective figure. The $\beta$ value is relative to y-axis units. Therefore, they are not comparable across figures. Hence, we visualize the regression line.

is also sensitive to the seasonality pattern.

The small RSE values, in all the properties, indicate that the interaction network has stable structural properties over time by indicating that the fluctuations around the regression line are small. The RSE values after the ramp-up period are smaller than the values before the ramp-up, which is indicative of the remarkable stability of this period. Further investigation of this observation for each property's distribution reveals that the majority of values after the ramp-up period in each distribution are within one standard deviation of the arithmetic mean. These properties remain stable at various levels of snapshot granularity, including bi-monthly and quarterly.

### 5.4.2 Edge Persistence over Time

To examine the edge persistence of the interaction network over time, we calculate the "resemblance" value for every two consecutive snapshots using Equation 5.1 [128].

$$R_t = \frac{|P_t \cap P_{t+1}|}{|P_t|} \tag{5.1}$$

Resemblance ($R_t$) is a measure between two consecutive snapshots ($P_t$ and $P_{t+1}$) of a network that quantifies the percentage of edges from the first snapshot that remains in the second snapshot. In our case, the resemblance is the percentage of pairs from the first snapshot who continue to interact in the following snapshot. A value of zero means none of the pairs from the first snapshot continued to interact in the second snapshot, and a value of one means all the pairs continue to interact in the second snapshot. We report the resemblance values over time in Figure 5.10, along with their arithmetic mean and standard deviation.

The figure shows a stable pattern where 75% of the resemblance values are within one standard deviation of the mean. The resemblance value grows during the first 16 months, which corresponds with the growth of interactions (see Figure 5.3). This initial growth is followed by a relatively stable three-year period. Furthermore, an average of 29% of pairs continue to interact across consecutive snapshots; hence, an average of 71% of pairs from one snapshot do not interact in the

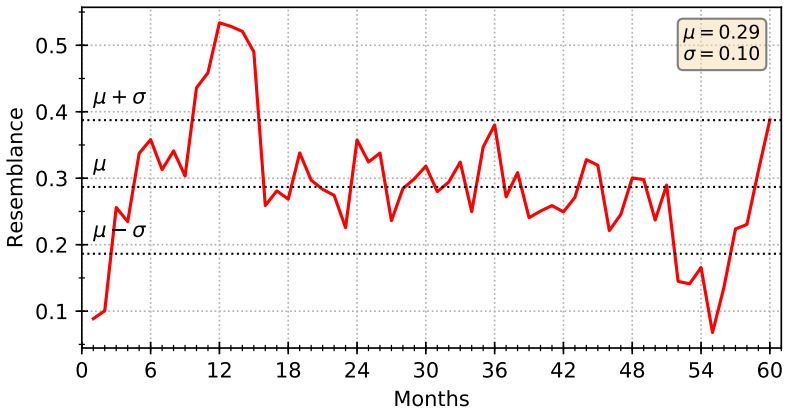



Figure 5.10: The resemblance of the interaction network. We exclude the resemblance value between the 61$^{st}$ and 62$^{nd}$ because the latter is an incomplete month.

following one. Since the resemblance is sensitive to snapshot granularity (*i.e.* snapshot duration), we measure the resemblance at different levels of granularity and find that the arithmetic mean of the resemblance values across weekly and quarterly snapshots are 12% and 39%, respectively. These small changes in the resemblance values show that the interaction network exhibit low edge persistence at different levels of granularity, which also means that the interaction network is highly volatile. This finding is consistent with our observation in Section 5.2.2 in that the majority of pairs have short interaction durations.

The low edge persistence raises a question regarding the presence of a core group of pairs who persistently interact across all five years. We find that our SCC does not have such a core group. However, there is a long period starting in the 9$^{th}$ month, where 0.43% of pairs continuously interact every month for 45 months. Similar to the resemblance, the definition of the core is dependent on snapshot granularity. By examining various levels of granularity, we find a core at the yearly level that consists of 3.6% of pairs, which further confirms the high volatility of the interaction network.

The high volatility of the interaction network and the interaction distribution in Figure 5.3 imply the presence of a phenomenon that underpins this network's stability that we observe in Section 5.4.1. Specifically, as the interaction network loses edges due to pairs who cease to inter-

act, new pairs start interacting causing the interaction network to gain new edges at a rate that is sufficient to stabilize the interaction network. These new pairs form because of 1) new members starting to interact with others, or 2) existing members having the tendency to interacting with others who they never interacted with before.

We find that our SCC has a higher number of new members who start interacting at each snapshot than those who cease interactions. We arrive at this conclusion by plotting, over time, the ratio of the number of members who start interacting in a snapshot over the number of members who cease interactions by the previous snapshot in Figure 5.11. This ratio over time is essentially the rate at which the network replenishes its interacting members. The figure shows that the interaction network exhibits high replenishment rates in the first 14 months where the members who start interacting in one snapshot are many orders of magnitude more than those who ceased interactions in the previous snapshot. For the remaining period, the replenishment rate fluctuates and eventually drops below 100%, which means the network is losing more interacting members than gaining new ones. On average, the network gains 129 interacting members for every 100 interacting members it loses.

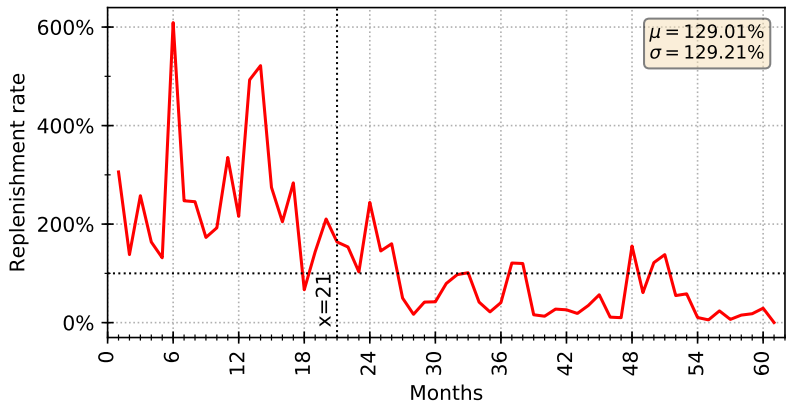



Figure 5.11: The interacting members replenishment rate in the interaction network.

Furthermore, we find that our members have a clear tendency to interact with others with whom they have not interacted before. On average, 47% of pairs at each snapshot are pairs of members who never interacted before. These pairs account for an average of 35% of interactions at each

snapshot. Furthermore, on average, 72% of pairs at each snapshot did not interact in the previous one, and they account for an average of 55% of interactions at each snapshot. This tendency is consistent with our finding in Chapter 4 where SCC members tend to interact with others who have radically different interaction degrees. The high replenishment rate and the members' interaction tendency are the underlying reasons that lead to the formation of new pairs in a manner that stabilizes the interaction network. Interestingly, this interaction tendency further supports our finding in Section 5.1.2 that these interactions convey the degree of common interest in the content of artifacts among "co-workers" rather than friendship.

Closer inspection of Figure 5.11 reveals that the replenish rate distribution over time is similar to that of the number of interactions in Figure 5.3. We assess the similarity between both distributions by decomposing the replenishment rate distribution and compare its trend and seasonality with those of the interaction distribution. We decompose the replenishment rate distribution using the same time-series decomposition technique we use in Section 5.1.4 and plot the results of this decomposition in Figure 5.12. The analysis reveals an annual seasonal pattern that repeats five times, which closely resembles the seasonal pattern of the interaction distribution shown in Section 5.1.4. It is interesting to note that the replenishment rate increase around the beginning of each season could be an indicator of the SCC operators' effort to cultivate the crowd.

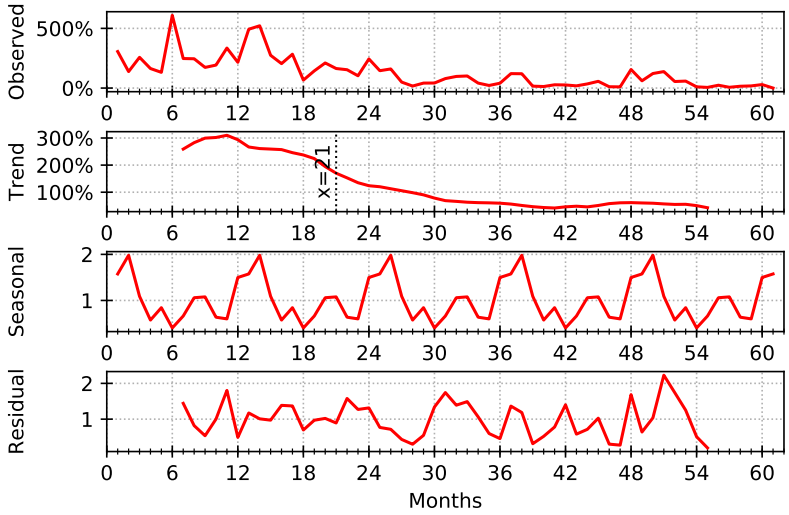
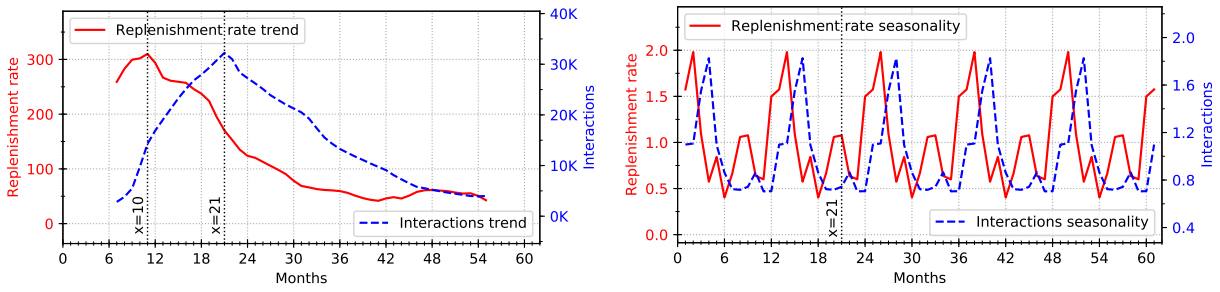



Figure 5.12: The replenishment rate decomposition.

Figure 5.13 shows that the trends and seasonality patterns of both distributions are remarkably similar. The replenishment rate trend appears to lag behind the overall trend. The peak of the overall trend comes 11 months after the peak of the replenishment rate, which implies that the increase in interacting members corresponds with a 10-month increase in interactions. Interestingly, the median node degrees (see Figure 5.9(f)) also spikes up to its highest value between months 10 and 21, which indicates that, on average, members interacted with more members during this period. The high replenishment rate and increased node degrees during this period contribute to increasing the number of interactions in this community to the highest historical value (see Figure 5.3). We also find that both seasonal patterns are remarkably similar, with the replenishment rate seasonal pattern lagging behind the interactions seasonal pattern by two months. This lag means that the new pairs that form within a monthly snapshot spike up the interactions during the following monthly snapshot, which is consistent with the finding from Figure 5.6 where we see that all new pairs make their largest number of interactions within one month from their first interaction.




(a) The trends.

(b) The seasonal patterns.

Figure 5.13: Comparison between replenishment rate trend and seasonality pattern and the (overall) interaction distribution trend and seasonality.

The similarities between the replenishment rate distribution and that of the interactions suggest the presence of a direct relationship where the replenishment rate causes a substantial increase in interactions. This effect is also quick to manifest where an increase in replenishment rate causes an increase in interactions in as little time as one month. Although there are potentially many factors that affect the interactions in this SCC, we showed that the replenishment rate and the interaction tendency of the members are primary factors that affect interactions quickly.

## 5.5 Findings Summary

Various social network analysis methods and those of other related fields analyze and capture the evolution of interactions in a SCC and answer questions relevant to SCCs. We analyzed a log of 779,673 direct and indirect interactions that span five years of a SCC, starting at the time of its inception. The analysis showed that a minority of pairs generate the majority of interactions and that the generation of these interactions follows a yearly pattern. We found that 29% of pairs consistently interact over monthly periods, yet no pairs interact every month throughout the entire log period. Furthermore, we observe a general decay in the interaction frequency among pairs over time. Surprisingly, we found that various global structural properties of the interaction network are stable over the monthly periods. We found that this stability is due to a high replenishment rate in SCCs and the tendency of its members to interaction with others with whom they never interacted before. We also found that pairs who frequently interact tend to "pace" their interactions and tend to interact longer as opposed to infrequently interacting pairs, which tend to have bursts of interactions that end quickly.

Moreover, we found that the frequently interacting pairs are more likely to join the social crowdsourcing process at the beginning by responding to queries while the infrequently interacting pairs are more likely to join later by initiating discussions around these responses. We also found that both *follow*, and automated and virtual *rewards* features have a statistically significant effect on interaction response times. Finally, we found that *follow's* effect is stronger than that of *rewards* but diminishes quicker than the effect of *rewards*. We provide a listing of all findings below:
SCC interactions are not bound by associations, which is a fundamental difference between the interactions analyzed in this chapter and the common interactions analyzed in literature:

1. The majority of members interact with more members than they associate.

2. Interactions are indicative of the degree of common interest among the interacting members.

Our finding identified a skewness in the interaction distribution that suggests that a few interacting-pairs produce the majority of interactions:

3. The top 20% of interaction-producing pairs produced the nearly 80% of interactions, which adheres to the Pareto principle.

SCC members tend to be highly responsive, but those who interact frequently tend to be exceptionally responsive. Despite the difference between them, both seem to play a unique role in the SCC process:

4. All pairs are highly responsive.
5. The frequently interacting pairs are more responsive than the infrequent pairs.
6. Both interaction groups are essential for the social crowdsourcing process, where frequently interacting pairs play a primary role in responding to queries, and infrequently interacting pairs play a primary role in initiating discussions around these responses.

Interactions tend to exhibit two main patterns: a burst of interactions followed by a quick decay, and a smaller burst of interactions that decays at a much slower rate:

7. The interactions of frequently interacting pairs tend to decay at a lower rate and last longer than the interactions of the infrequently interacting pairs.
8. The interactions of infrequently interacting pairs are more likely to have bursts of interactions that end quickly.

Associations and rewards have a clear relationship with member responsiveness. Both are associated with reducing response times. Associations (*i.e.* follow) reduce response times more than rewards, but its effect seems to last a shorter duration than does the effect of rewards:

9. *Follow*, and *rewards* are associated with faster response times for both frequently and infrequently interacting pairs.

10. *Follow* is associated with faster response times for both interaction groups than *rewards*.

11. Response times are fastest when both *follow*, and *rewards* are present.

12. *Follow's* effect on response times diminishes relatively quickly in both interaction groups.

13. The effect of *rewards* remains consistent over time.

Previous findings suggest that SCC interactions may be unstable. However, various graph-theoretic properties of the interaction network remain stable. This stability is due to the replenishment rate of interactions and the tendency of SCC members to interact with new members:

14. Various graph-theoretic properties of the interaction network are stable over time.

15. The interaction network exhibits a high refresh rate where a minority of interacting pairs in one snapshot continues interacting in the subsequent one.

16. The interaction network does not have a core of pairs that persistently interact over the entire period.

17. This SCC has a high replenishment rate where the number of new members who start interacting at each snapshot exceeds those who cease interactions.

18. Members of this SCC have a clear tendency to interact with others with whom they have not interacted before.

19. The high replenishment rate and tendency to interact with new members underpin the stability of the interaction network.

The future research of this chapter involves:

1. repeating this analysis using multiple SCCs to confirm the generality of these findings, and

2. produce a topology-generating model that captures the interaction patterns, their evolution patterns, and factors that affect them, and

3. use this model to assess to studying the effect of random and targeted attacks [113] on SCC interactions as well as identifying critical points of transition in the interactions and the best ways to deal with them.

This chapter present a detailed analysis of the interactions among SCC members. It results in identifying common patterns in these interactions and the key factors that affect these interaction patterns. The following chapter builds on the findings from Chapter 4 and this chapter by presenting an analysis of SCC member behaviour, which culminates in the production of the behavioural model that we use throughout the rest of this thesis.

# Chapter 6

# QUANTIFYING BEHAVIOUR IN A SCC[1]

This chapter develops and tests a behavioural model that quantifies interactions in *HCC*. We present a process for capturing behaviours in variables and aggregating them into factors using FA. This process is sufficiently generic that it applies to future studies, which helps develop an understanding of behaviour in various online environments systemically and reliably. The process involves five steps: determining an appropriate set of behaviours to capture in variables, cleaning these variables, setting up FA, running FA to obtain the model, testing the model, and interpreting the model. In this chapter, this model forms the basis to analyze and reason about the relationship between team performance and composition.

## 6.1 Data

We use an anonymized back-end database of an active and diverse SCC from Chaordix[2], namely *HCC*. Recall that this SCC is a private community of brand loyalists of a major mobile phone company[3]. Members of this community take part in various activities related to mobile devices in return for exclusive information from within the company and frequent giveaways of desired company branded merchandise. This database has member demographic information and their actions. From this database, we produce a dataset for this analysis.

Member demographic data consists of identification number, age, sex, country of citizenship, number of spoken languages, occupation, and level of education for 7,524 members. 89% of members are male, and 10% are female. These members are from 119 countries on six continents (see Table 6.1). The average age in this community is 31 years, with 9 years standard deviation

[1] **The results in this chapter are published [3]**

[2] `www.chaordix.com`

[3] The community's name must remain confidential to be consistent with our access agreement.

and over 73% of members within 1 standard deviation.

Table 6.1: Distribution of members on continents.

| Continent | Percentage |
|---|---|
| Africa | 0.9% |
| Asia | 18.9% |
| Australia (Oceania) | 2.0% |
| Europe | 33.0% |
| North America | 44.4% |
| South America | 0.3% |
| Unspecified | 0.6% |

68.1% of members speak English as a primary language, and the majority can speak more languages. 83% of members specified their occupation. The most substantial portion of members is students, and the smallest portion of members is stay-at-home parents. Table 6.2 shows the percentages for the highest level of education achieved by members.

Table 6.2: Education level of members.

| Level of education | Percentage |
|---|---|
| Unspecified | 1.5% |
| Less than High School | 1.4% |
| High School or equivalent | 16.0% |
| Some college (but no degree) | 26.0% |
| Associate degree | 8.7% |
| Bachelor degree | 32.9% |
| Graduate degree Masters | 12.2% |
| Graduate degree Doctorate | 1.5% |

## 6.2 Method

We produce a dataset that quantifies all relevant actions that members perform on this SCC (see Section 6.2.1). We then use FA to derive FB then analyze these factors, name them, and interpret their meaning (see Figure 6.1).

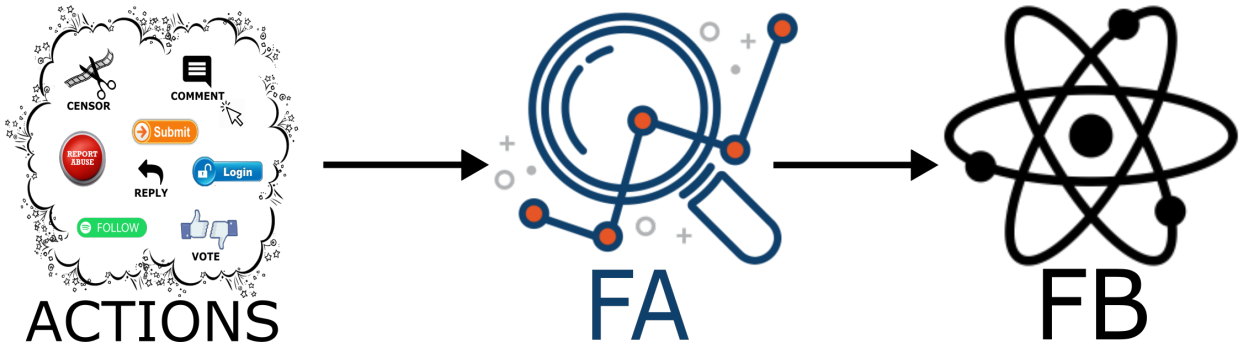



Figure 6.1: The process of obtaining FB.

### 6.2.1 Produce dataset

We produce a dataset that consists of members and base variables. Base variables hold simple counts of actions for each member. We do a FA on this dataset to derive FB.

Why use derived factors instead of using this dataset directly? Behaviour is the set of all measurable actions that members can perform on the SCC through features offered by the SCC platform, such as following other members, creating posts, commenting, voting, *etc*. Thus, this definition implies that this dataset quantifies behaviour. However, it consists of elaborate and platform-dependent variables that are not conducive to producing generalized conclusions. FA yields factors (*i.e.* dimensions) that are latent (*i.e.* unmeasured), and generalized (*i.e.* applicable to more than a single SCC platform).

Why are base variables adequate? Why not use derived variables? For two reasons: First, derived variables, such as rates, ratios, and inter-arrival times are more specific and complex than base variables, which make them harder to understand and make interpreting their relationships with other variables difficult. They also pose a challenge in FA because they tend to cross-load[4], and to be redundant[5][83]. Second, the number of derived variables is enormous, so we only consider base variables. We also ensure that the base variables collectively form a fulsome representation of SCCs' behaviours and patterns that we previously identified in Section 4.3 and Section 5.5.

[4]Variables that have a loading value $\geq$0.3 on multiple factors.

[5]Variables derived from multiple base variables tend to have high correlations with the base variables, which is problematic for FA.

Specifically, we did the following:

1. due to SCC members' tendency to reciprocate interactions, the disparity between in-degrees and out-degrees in their interactions (see findings # 4, 7, 11, 15, and 19 in Section 4.3) we capture interactions made and interactions received as separate variables;

2. due to the distinct preferences of some SCC members to interact with specific artifacts-levels (see finding # 6 in Section 5.5) we capture responses to queries (*i.e.* second-level artifacts), and discussions on responses (*i.e.* third-level artifacts) separately; and

3. due to the effect that follow has on response times (see findings # 9-13 in Section 5.5) we capture artifact following. However, we exclude member following or rewards because, as explained in the next paragraph, they are out of the analysis scope.

We focus on artifacts and actions concerning one type of textual contributions, namely, submissions. Submissions are the highest level of written content that members produce as a direct response to a moderator's newly initiated discussion topic or question. Members can remove their submissions and can tag them with descriptive tags. These tags help members and data analysts to find and organize submissions. Moderators can remove and tag any member's submissions. Any member can vote, follow, and comment on submissions and any other member can reply to any of the comments. A reply to a comment creates a *comment thread*. Members can also flag submissions as "inappropriate", which notifies moderators to review and potentially censor the flagged submission. Moderators can censor (*i.e.* delete) submissions and can flag submissions as "significant contributions". They also can pin - and unpin - submissions to display them in a highly visible area. Pins indicate that the submission is significant.

We capture actions that a member performs, such as creating submissions and voting on others' comments, and actions that a member receives on written contributions, such as comments received

from others on their submissions or comments and votes received on submissions or comments. From these actions, we calculate simple quantitative variables that have clear meaning and mapping to the SCC actions and avoid any complex variables that address several aspects of behaviour. All base variables have a minimum value of zero and an unbound maximum value.

Let $A$ denote actions that *HCC* offers (see Table 6.3 for details). Let $M$ be the set of members, $M = \{m_0, m_1, \ \ldots, m_n\}$. A dataset, $MA$, is a matrix that holds values for each member action combination. Thus, $MA = M \times A$ where each cell, $ma_{ij}$, contains the total number of times a member $m_i$ performed an action $a_j$ since joining the SCC. Since members join the SCC at various times, they have been on the SCC for different periods and they login and participate at different rates, we normalize the dataset by dividing each member's variables by the number of days this member logged in to the SCC, which results in the average number of actions this member performed per day. In other words, we normalize the data by dividing each $ma_{ij}$ by the number of days member $m_i$ logged into the SCC, which produces an average $a_j$ per days logged in, which neutralizes the effect of time on member action counts.

We calculate these variables for all members who logged in three or more days, and exclude the rest. We choose this threshold value empirically with the intuition that members with less than 3 day-logins have too few data points to establish their behavioural tendencies, which is a consequence of the inherent cold-start problem of our approach that Section 1.3 explains. We choose the smallest value possible that satisfies the aforementioned intuition without introducing the selection bias threat to validity (*i.e.* retrains as many members as possible).

### 6.2.2 Derive factors of behaviour

To derive the FB from the dataset $MA$, we put the $MA$ dataset through FA to yield a new dataset $VF$[6]. $VF = A \times F$, where $A$ is the set of actions and $F = \{f_0, f_1, \ldots\}$ is the set of factors we arrive at through FA. Thus, $vf_{ij}$ is the loading value variable $v_i$ has on factor $f_j$.

Recall from Section 2.3.1 that there are two types of FA: Exploratory FA (EFA) and Confirma-

[6] $VF$ stands for variable-factor loading matrix, which contains variable loadings on factors.

Table 6.3: Variables and their descriptions. None of the variables has univariate or multivariate normality (tested using KS goodness of fit test [5] and Henze-Zirkler's Multivariate Normality Test [6]. * indicates that a variable's distribution does not fit power-law distribution ($p$ <0.05).

| Variable name | Description |
|---|---|
| *comment made by me on my submissions* | replies made to comments received on submissions |
| *comment made by me on others submissions** | comments made on others' submissions |
| *comment others made on my submissions* | comments received on submission |
| *submission** | submissions made |
| *submission censors made* | censors a moderator made |
| *submission censors received* | censors a member received |
| *submission flags made* | flags made on others' inappropriate submissions |
| *submission flags received* | flags received on inappropriate submissions |
| *submission follows made** | follows made on others' submissions |
| *submission follows received* | follows a member's submissions received |
| *submission pins made* | pins a moderator made on others' submissions |
| *submission pins received* | pins a member received on submissions |
| *submission significant sets made* | times a moderator set others' submissions to significant |
| *submission significant sets received* | times a member's submissions were set to significant |
| *submission tags by me on me** | tags made on submissions |
| *submission tags made* | tags a moderator made on others' submissions |
| *submission tags received* | tags received on submissions |
| *submission votes made* | number of votes made on others' submissions |
| *submission votes received** | votes a member received on submissions |

tory FA (CFA). EFA produces a factor solution from the data while CFA tests the "fit" of a factor solution to the data. To the best of our knowledge, no rigorous theoretical treatment of SCCs appears in the literature. Thus, there is no factor solution based on preconceived understandings. Therefore, we perform EFA only on the data to produce a factor solution. We tune the EFA process based on our understanding of the data, and analysis procedure and document every step and decision throughout this chapter.

EFA consists of six steps: gathering a dataset, choosing an EFA method, preparing the dataset, performing EFA, obtaining a factor solution and interpreting its meaning, and obtaining member factor scores.

We gathered a dataset that contains sufficient and adequate samples. Section 6.2.1 shows that dataset has 5,593 observations and 19 variables, which exceeds the most conservative 30:1 ratio and Section 6.1 shows the data contains diverse subjects, which makes the findings applicable to a

diverse population. We choose to use the Maximum Likelihood EFA method due to its popularity in the specific *R* package we used to perform FA.

We clean the dataset by removing singular and collinear variables. We identify these variables by calculating their SMC scores. Singular variables have SMC values "close" to zero, and multicollinear variables have SMC values "close" to one.

To avoid subjectivity in choosing a "close enough" threshold value, we remove a variable if FA is unable to run on the data. We choose to remove the variable that is closest to one or zero. Moreover, since removing a variable from the dataset affects the SMC scores of the others, we perform this process iteratively. That is, if FA fails to run, we remove the variable closest to one or zero, then attempt to rerun FA. We repeat the process until FA can successfully run (see Algorithm 1).

The iterative process stops when FA can run to completion on the dataset. If FA does not run, we run SMC on the dataset. We sort the variables by their SMC scores in ascending order. Variables that suffer from collinearity or singularity will be at either the bottom or top of the list. We start by removing a collinear variable. Collinearity affects groups of variables. Thus, if the dataset has this problem, then two or more variables have it. Multicollinear variables have similar SMC scores. We inspect these variables together and remove the variable that is least meaningful to the analysis. If there are no collinear variables, we remove the variable with the least SMC score (see Algorithm 1).

We ran the data cleaning algorithm (see Algorithm 1), which resulted in the removal of *submission pins made*, *submission significant sets made*, *submission significant sets received*, and *submission tags made* variables because of low SMC scores. The dataset contained no multicollinearity within the variables.

We use oblique rotation methods because they produce simpler factor solutions [152] that reflect "real-life" and are therefore easier to interpret. Human behaviour is rarely partitioned into "neatly packaged units that function independently of one another" and therefore using oblique ro-

**Algorithm 1:** Algorithm for cleaning a dataset for FA.

**Data:** $MA$ a matrix that contains member actions frequency (*i.e.* variable).
$FA(MA)$ FA function. Return a factor solution.
$SMC(MA)$ SMC function. Return a 2D list that contains variable-SMC scores.
$MCL(a)$ a function that takes in a variable-SMC scores list and checks the existence of a group of variables with similar SMC scores that are also close to 1. If such a group of variables exists, it returns the name of the least meaningful variable that ought to be removed (this process is done visually). Otherwise, it returns null.
$remove(MA, target)$ a function that removes the variable $target$ from the dataset $MA$ and returns a new dataset $MA'$.
$pop(a)$ a function that removes the top element in a 2D array and return the element and the updated 2D array.
**Result:** A cleaned $MA'$ that contains no problematic variables for FA.

$MA' \leftarrow MA$;
**while** $FA(MA') \neq null$ **do**
    $st \leftarrow sort(SMC(MA'))$;
    $target\_var \leftarrow MCL(st)$ ;
    **if** $target\_var \neq null$ **then**
        $MA' \leftarrow remove(MA', target\_var)$;
    **else**
        $target\_var, st \leftarrow pop(st)$;
        $MA' \leftarrow remove(MA', target\_var)$;
    **end**
**end**

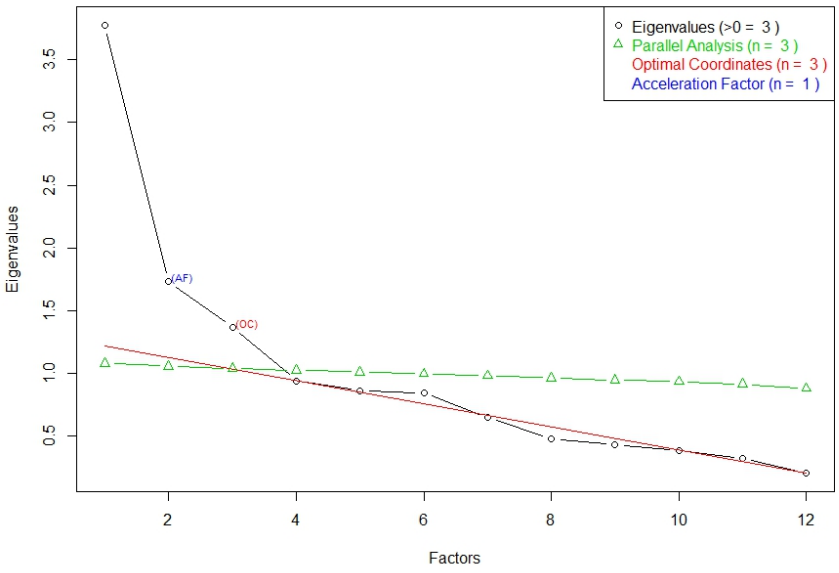
tation would produce a more accurate, interpretable, and reproducible outcome [85, 81, 86, 153].

We perform the FA process (see Section 2.3.1) using *oblimin factor rotation* method. The ideal number of factors is three because it results in the most interpretable factor solution and adheres to the heuristics (see Section 2.3). Furthermore, all three methods for estimating the number of factors converge to the same number (see Figure 6.2). After running this iterative process, we removed three variables: *submission flags made*, *submission follows made*, and *submission tags received* because they consistently loaded at less than 0.3 on any factor.

## 6.3 Results and Discussions

We ran the Maximum Likelihood Estimation (MLE) FA method on the dataset *MA* to obtain the factor solution. The next sections discuss this factor solution.

Figure 6.2: Scree plot test, parallel analysis, and Kaiser's criterion methods all converge to the same number of factors.




#### 6.3.0.1 Variable loadings

The FA on the initial dataset yields three factors. We name these factors: *Impact*, *Activity*, and *Policing/Rowdiness*. Table 6.4 shows the three factors, along with their variable loadings. *Impact* is a factor that represents the attention a member's submissions receive through comments, votes, follows, and pins from other SCC members. Thus, variables concerning received reactions load highest on the *Impact* factor. This factor is a positive factor because its variables are positive in value and sentiment. In other words, the smallest value a member can have in any of these variables is a zero, and these variables have no negative sentiments (*i.e.* no downvotes, no negative number of comments, *etc.*). *Activity* is a factor that represents the amount of work a member performs in the SCC. Thus, the variables that load highest on this factor are ones concerning the number of submissions, tags, comments, and votes members make on others' submissions. Similar to *Impact*, *Activity* is a positive factor.

Some of the variables that load the highest on the *Activity* factor capture two unique aspects of the SCC process. The two highest-loading variables: *comment made by me on others submissions*

Table 6.4: Factors and loadings. The highest loading values for each variable is in bold and cross-loadings are underlined.

| Variables | Impact | Activity | Policing |
|---|---|---|---|
| *comment others made on my submissions* | **0.86** | 0.14 | 0.02 |
| *submission follows received* | **0.81** | -0.19 | -0.02 |
| *submission votes received* | **0.50** | <u>0.37</u> | 0.03 |
| *submission pins received* | **0.31** | -0.16 | 0.031 |
| *comment made by me on others submissions* | -0.02 | **0.76** | -0.01 |
| *submission votes made* | -0.073 | **0.67** | -0.02 |
| *submission* | 0.20 | **0.50** | 0.08 |
| *comment made by me on my submissions* | <u>0.40</u> | **0.48** | -0.01 |
| *submission tags by me on me* | 0.01 | **0.47** | 0.043 |
| *submission censors received* | -0.01 | -0.02 | **1.00** |
| *submission censors made* | 0.04 | 0.02 | **0.52** |
| *submission flags received* | -0.01 | 0.05 | **0.49** |

and *submission votes made* quantify actions performed on others' content, which increases their *Impact* scores. The *comment made by me on my submissions* variable quantifies the degree of which a member generates content in response to others' reactions. Thus, members whose *Activity* scores predominately comprise their score on this particular variable will have a correlated *Impact* scores.

The latter finding is further supported by the cross-loading of *Submission votes received* and *comment made by me on my submissions* on both *Impact* and *Activity*. These variables' inner correlations explain this cross-loading. Specifically, *submission votes received* has a correlation coefficient of 0.56 with *comment made by me on my submissions*. Furthermore, *comment made by me on my submissions* has correlation coefficients of 0.66 with *comment others made on my submissions*. These correlation values are in tune with the reciprocal nature of interactions in SCCs we identify in Section 4.2.2). That is to say, others' comments on a member's submissions prompt this member, to some degree, to respond to these comments, forming a discussion. This discussion is a product of the interest the other members have in this member's submission, which manifests as votes and comments.

This observation raises an important question about the interaction of factors. In practice, a

member must have some degree of activity to generate submissions and comments through which this member may receive endorsements that increase his/her *Impact* score. This practicality suggests that both *Impact* and *Activity* may be correlated, so we investigate this further in the next section (see Section 6.3.0.2).

*Policing/Rowdiness* is a factor that quantifies members' involvement in the process of producing and censoring inappropriate submissions. High *Policing/Rowdiness* scores imply high degree of involvement in this process, which occurs when members produce many inappropriate submissions that are later flagged and potentially censored, when moderators censor many submissions, or both. Two of the three variables that load highest on this factor concern rowdiness: *submission flags received* and *submission censors received*. These two variables represent the number of inappropriate submissions a member made, which other members flagged and moderators censored. These variables are positive in value but negative in sentiment, so a member with high values in these variables tends to produce content that often gets flagged or censored. The third variable, *submission censors made*, is also positive in value and negative in sentiment because it represents the number of inappropriate submissions a moderator censors.

We inspect this factor further by performing Pearson correlation analysis on its highest loading variables. We find that *submission flags received* has a moderate correlation (a correlation of 0.50) with *submission censor received*. This correlation indicates that those who receive many flags on their submissions tend to have their submissions censored by moderators. The reasons that this correlation is moderate, as opposed to high, are three. First, the moderators will not censor all flagged submissions because, after inspecting the submissions, they may find that some of them are appropriate. Second, many members can flag the same submission, but this submission may be censored once or not at all. Third, moderators may censor some submissions before members flag them.

We also find that *submission censors made* has no correlation with *submission flags made* or *submission flags received*. This lack of correlation means that censoring submissions is not related

to producing or flagging inappropriate submissions. This is expected given that moderators are trained professionals who rarely produce inappropriate submissions, and they can directly censor submissions without having to flag them first. Interestingly, *submission censors made* has a moderate correlation (a correlation of 0.52) with *submission censors received*. This correlation means that moderators who censor submissions also have their own submissions censored, which indicates that moderators censor each other's submissions to a considerable degree. This correlation explains why *submission censors made* is part of this factor and reveals a process where moderators "police" other moderators. Since moderators rarely produce inappropriate submissions, then the lack of correlation between *submission censors made* and *submission flags received* despite the 0.52 correlation between *submission censors made* and *submission censors received* indicates that moderators censor other moderators' submissions considerably but these submissions are not consistently flagged. This suggests that this censoring occurs for reasons that are known to mostly by moderators, such as violating moderator-specific roles of which non-moderator members are not privy.

#### 6.3.0.2 Factor correlations

We perform correlation analysis using member factor scores to identify interrelated factors. Specifically, we derive member factor scores using the process described in Section 7.1.1. We then perform a Pearson correlation analysis to obtain the correlation coefficients. The correlation analysis results show that *Impact* and *Activity* are not correlated (a correlation coefficient of -0.125[7]), which means that a member's level of impact is not relevant to the amount of content (s)he goes through, reacts to or produces, and vice versa. This lack of correlation despite the previously identified practical relationship between one's *Impact* and *Activity* scores (*i.e.* for a member to have impactful content, this member must have produced content first, which reflects in this member's *Activity* score), points at a predominant behaviour that dilutes this relationship. Specifically, members with small activity scores produce much of the impactful content, and highly active members

[7] All factor correlations significance tests are two-tailed tests with $p$-value = 0.01.

produce few impactful content. This behaviour is consistent with our previous finding regarding the distinguishability of popular and active members in terms of interactions (see Section 4.2.4).

The results also show that *Activity* correlates with *Policing/Rowdiness* (a correlation coefficient of 0.471), which is expected given that moderators are highly involved in policing communities and tend to be active members. Furthermore, given the high amount of content they go through and react to, and the high volume of content they produce, active members are likely to encounter (and report) inappropriate content. They are also more likely to produce content that others find inappropriate. Factors *Impact* and *Policing/Rowdiness* are not correlated (a correlation coefficient of -0.0756), which means that *Impact* scores are irrelevant to policing activities or rowdy behaviour.

## 6.4 Findings Summary

We created a detailed process for quantifying individual behaviour in *HCC*. This process involves capturing the actions of members about textual contributions using measurable variables, performing FA on these to produce FB, and testing, interpreting and naming these FB. We found that member behaviour has three dimensions/factors: *Impact*, *Activity*, and *Policing/Rowdiness*. We found that *Activity* correlates with *Policing/Rowdiness*, which means that active members are more likely to see and report bad behaviour. It also means that members with high *Activity* scores tend to produce a large amount of content so are more likely to produce content that others find inappropriate. *Impact* and *Activity* are not correlated, which means that a member's level of *Impact* is not relevant to the amount of content accessed, reacted to, or produced. *Impact* and *Policing/Rowdiness* are not correlated, which means that *Impact* scores are irrelevant to policing activities or rowdy behaviour. The following is a summary of the findings:

1. We identified three factors with remarkable consistency: *Impact*, *Activity*, and *Policing/Rowdiness*.

2. *Impact*: represents the incoming endorsements (incoming interactions) on a mem-

ber's contributions.

3. *Activity*: represents the amount of work a member performs, including producing submissions, and reading and endorsing others' work (outgoing interactions).

4. *Policing/Rowdiness*: represents a specific type of outgoing and incoming interactions. Specifically, it represents the degree of which a member is involved in producing inappropriate submissions and censoring inappropriate submissions.

5. *Impact* and *Activity* are not correlated.

6. *Impact* and *Policing/Rowdiness* are not correlated.

7. *Activity* correlates with *Policing*.

The natural next steps of this research involves addressing its generalizability. To this end, we identify the following concrete steps:

1. apply the process of quantifying behaviour to several SCCs to obtain a general FB that is common across SCCs;

2. include more base variables in the quantification process to produce a holistic model that involves all aspects of behaviour in SCC; and

3. assess the utility of the new FB in predicting various aspects of individual behaviour in teams, such as propensity to lead, and participate in discussions.

This chapter presents a process for quantifying member behaviour in a model using FA. It shows a detailed analysis and discussion of this model's factors and that this model could produce member-level scores that are indicative of various aspects of member behaviour. The next chapter will demonstrate the utility of these factors in guiding team formation.

# Chapter 7

# A TEAM COMPOSITION IN A SCC[1]

This chapter examines the FB model at member and team levels. It shows that member-level factor scores correlate with other member actions that were not considered when deriving the FB model. It also shows that team-level aggregated factor scores related to team performance (*i.e.* project quality). Furthermore, this chapter shows that specific mixtures of these aggregated factor scores in teams are associated with increased team performance and demonstrates the team-level factor aggregates' utility in predicting team performance.

## 7.1 How Do Member Factor Scores Relate To Member Behaviour?

To answer this question, we calculate members' factor scores and then perform correlation analysis between these factor scores and base variables concerning other actions in *HCC* that were not included in the FA. These are actions concerning albums, blog posts and comments artifacts, and point and badge rewards that are directly relevant to these artifacts (see Table!7.1 for details). Albums are collections of images uploaded by community members, and blog posts are textual artifacts that members produce without being prompted by a query. Comments are textual artifacts that members produce in response to albums, blog posts, or other comments. The points and badges included in this analysis are automatic rewards received by members for conversation-related activities.

### 7.1.1 Dataset

We derive a member's factor scores using a specific mathematical method that converts this member's actions to corresponding factor scores using the FB (see Figure 7.1).

[1] **The results in this chapter are published [3]**

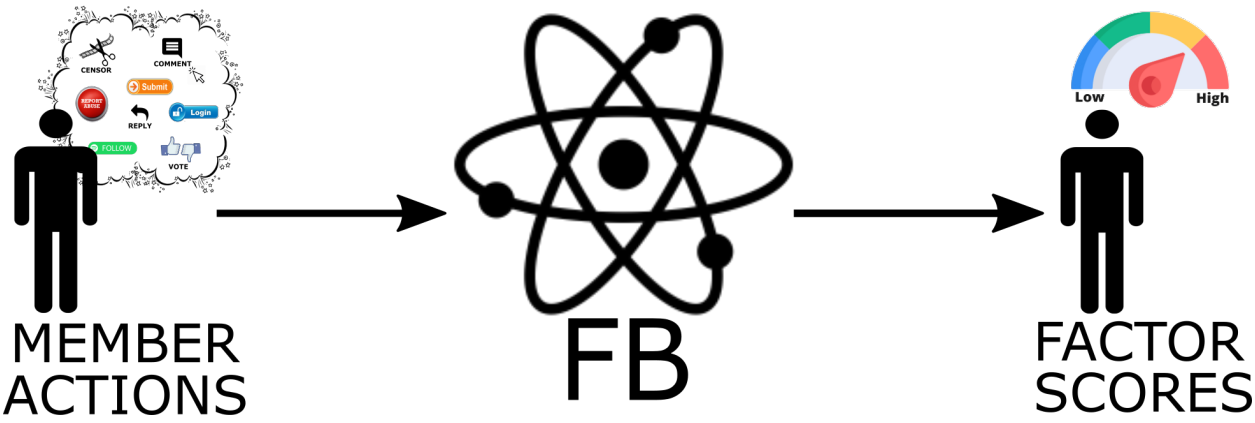



Figure 7.1: The process for producing factor scores.

There are many methods to calculate member factor scores. We tested the two common methods (*i.e.* Anderson-Rubin and Bartlett methods) and found that they had no significant effect on the analysis results. Thus, we choose to use the Anderson-Rubin method [94] to calculate factor scores. This method requires $VF$ and $MA$ datasets to produce these scores. We create a Members Factors Scores ($MFS$) dataset $MFS = Anderson(MA, VF)$, where $MA$ is the member actions dataset and $VF$ is the variable-factor loadings dataset. $MFS$ is a matrix that contains the score each member has on each of the factors. Thus, $mfs_{ij}$ is the factor score that member $m_i$ has on factor $f_j$.

### 7.1.2 Analysis and results

Table 7.1 shows the correlation coefficients between the three factors and other base variables. Overall, we found statistically significant high correlations between the factors and other base variables. *Impact* factor correlates positively with all variables concerning the quality of contributions, such as receiving comments, votes, follows, pins, and points and badges concerning conversations. *Activity* correlates with all variables concerning the quantity of contributions, such as the total and the unique number of badges, the number of blog post submissions, comments made on others and in response to others' comments, following others, voting, tagging, and points concerning community activities. Interestingly, *Policing/Rowdiness* factor did not correlate significantly with any

Table 7.1: Factors' correlations with other base variables using Pearson's Correlation. ** correlation is significant at the 0.01 level and * at the 0.05 level (two-tailed). We kept and bolded correlation records with absolute values $\geq$0.3.

| Variable | Impact | Activity | Policing | Description |
|---|---|---|---|---|
| *album flags made* | **.398**** | -.157** | 0.011 | flags made on others' albums |
| *album pins received* | **.519**** | -.211** | 0.018 | pins received from others |
| *badge conversation* | **.430**** | **.569**** | .044* | conversation badges received |
| *badge conversation unique* | .244** | **.326**** | 0.028 | unique conversation badges received |
| *blog post submissions* | .136** | **.341**** | 0.015 | blog posts made by a member. |
| *blog post follows received* | **.477**** | -.175** | 0.014 | follows a member's blog posts received |
| *blog post pins received* | **.409**** | -.160** | 0.015 | pins a member's blog posts received |
| *blog post tags by me on me* | **.470**** | -.196** | 0.014 | tags made on own blog posts |
| *blog post votes received* | **.518**** | -.208** | 0.017 | votes a member's blog posts received |
| *comment by me on me* | **.669**** | **.456**** | .070** | replies made to comments received |
| *comment by me on others* | .238** | **.760**** | .035* | comments made on others content |
| *comment by others on me* | **.718**** | -.098** | .053** | comments received on my content |
| *comment follows made* | **.438**** | **.523**** | .037* | follows made on others' comments |
| *comment follows received* | .236** | **.537**** | .035* | follows received on comments |
| *comment votes made* | .290** | **.496**** | .053** | votes made on others' comments |
| *comment votes received* | **.501**** | **.463**** | .049** | votes received on own comments |
| *points* | **.347**** | .202** | .073** | points earned |
| *point community* | .234** | **.313**** | .084** | points for community related activities |
| *point conversation* | **.568**** | .224** | .047** | points for conversation related activities |
| *tag made* | .224** | **.415**** | .067** | tags a member made |
| *tag received* | .235** | **.444**** | .075** | tags a member received |

of the variables. These correlation results indicate that the FB, although derived from a small set of variables concerning one type of textual contributions, are indicative of member actions in this SCC.

## 7.2 How Do Aggregated Factor Scores Relate to Project Quality?

This section investigates the relationship between team-level aggregated factor scores and team performance. It shows that specific team-level aggregates correlate with team performance. Furthermore, it shows that teams that are composed of specific compositions tend to outperform others. To this end, the following sections define teams in the SCC context, create a dataset consisting of member factor scores that have been aggregated to team-level and the quality score of every

project a team worked on, and perform several univariate and multivariate analysis to demonstrate the relationship between the team-level factor scores and team performance.

### 7.2.1 What is an SCC Project, Team, and Performance?

This thesis uses the term "team" as a shorthand for collaboration among a subset of SCC members in a virtual environment. A traditional team is a group of interdependent and skilled individuals who share the responsibility and commitment of accomplishing a task [154, 155]. Virtual teams are traditional teams that utilize electronic information and communication technologies [156] as the primary communication medium. The team members are recruited/admitted to a team based on the need for their skills, among other qualities.

SCC teams are virtual groups of interdependent and skilled members who voluntarily utilize the SCC platform to work on tasks. They consist of moderators and members. While traditional teams are explicitly formed based on the projects' needs, SCC teams form implicitly by their members' contributions on projects they find interesting. In order words, unlike the traditional teams, SCC members strictly self-assemble into teams by voluntarily contributing to any ongoing projects they find interesting, which means that SCC teams can grow in size quickly during the project period.

Furthermore, since SCC members are volunteers, they are not obligated to participate in projects and are not responsible for accomplishing them. The moderators are responsible and committed to seeing projects to completion, which underpins their coordination and engagement efforts.

We consider discussion threads to be projects. A discussion thread consists of a submission and comments. A submission is a direct response to a query, and comments are replies to the submission or other comments under this submission. We use the term *textual contributions* to refer to the submission and comments collectively. A team consists of members who contribute - textually - to this project (*i.e.* discussion thread). We determine that a member is part of a team if this member makes two or more textual contributions to the project, and define a team as three or more members who participate in a discussion thread.

We choose these threshold values somewhat arbitrarily based on the following considerations.

First, we base the team size threshold value on the fact that aggregating member factor scores to team-level through arithmetic mean, median, and the standard deviation is substantially less meaningful in a population of a size less than 3. Second, we base the other threshold value on the nature of queries in our SCC. Specifically, since this thesis is concerned with open-ended queries requiring discussions (see Section 3.1), we consider a member to be part of a team when this member demonstrates a minimum involvement in this team's discussions by making at least two textual contributions. We do not consider the content of these contributions because of the large number of projects and comments, which makes the manual investigation of their substance impractical.

For each remaining project, we measure its quality as the sum of endorsements it receives from members who are not part of its team. We choose *follow* as the representative form of these endorsements because it conveys a prolonged interest in the project, which better represents project quality than momentary actions, such as a vote. This measure, however, does not represent quality in all its facets. Instead, it serves as an indicator of the interest a specific project generates, which suffices for this work. We leave a more fulsome investigation of the quality measure for future work. We measure team performance by the quality of its outcomes, so we use this project quality measure as a proxy for team performance.

### 7.2.2 Dataset

We create a dataset, $D$, which has 1,235 projects, as follows. First, we retrieve a list of all projects (*i.e.* discussion threads) in $HCC$. For each project, we retrieve the list of members who made two or more textual contributions to it. These members are the team for this project. We include this project if its team size is larger than or equal to three. We end up with 1,235 self-assembled teams of sizes that vary between 3 and 107 members with a long-tail-like distribution (see Figure 7.2). We then calculate the quality score of the project by summing up the number of followers each textual contribution in this project attracted from non-team members. Thus, this dataset contains records of projects, team members, and project quality scores. We join this dataset with the $MFS$

dataset to produce a new dataset that has records of projects, team members, factor scores, and the projects' quality scores. We aggregate this dataset to team-level to create a matrix $D$, which contains records of projects, aggregate factor scores for each team through summations, averaging, standard deviation, *etc.*, and the project quality scores (see Figure 7.3).

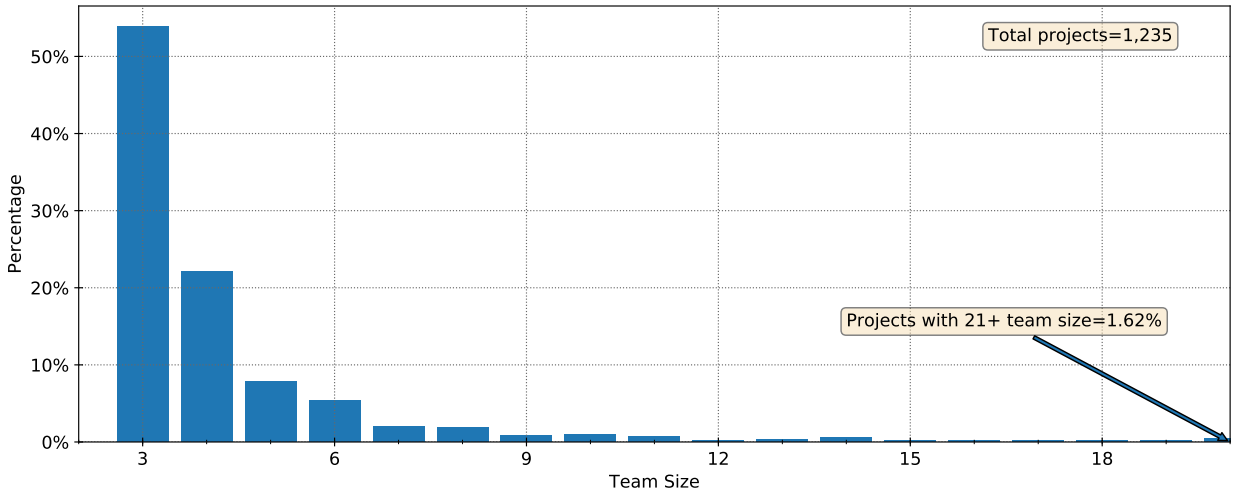



Figure 7.2: The distribution of team sizes in the dataset $D$.

Thus, $D$ is a matrix, $D = P \times F'$, where $P$ is a set of projects and corresponding quality scores, and $F'$ is the set of aggregated factors scores using total, arithmetic mean (commonly referred to as average), median, standard deviation, minimum and maximum functions [110, 157]. Thus,

$$F' = \{sum(Impact\_score), avg(Impact\_score), std(Impact\_score), \ldots, etc.\}.$$

and each cell $d_{ij} \in D$ holds the $j^{th}$ aggregation of a factor score for $i^{th}$ project.

### 7.2.3 Correlations of aggregated factor scores and project quality

We show the relationship between aggregated factor scores and project quality by performing Pearson's Product-Moment Correlation analysis using SPSS Statistics on $D$. The results of this analysis are in Table 7.2. The correlation coefficients are statistically significant but are not strong. We expect that these correlations are not strong because our teams are self-assembled teams, which have been shown to exhibit low levels of productivity, and participation [8, 9]. These correlations, however, are sufficient for demonstrating the presence of a relationship between some of the FB

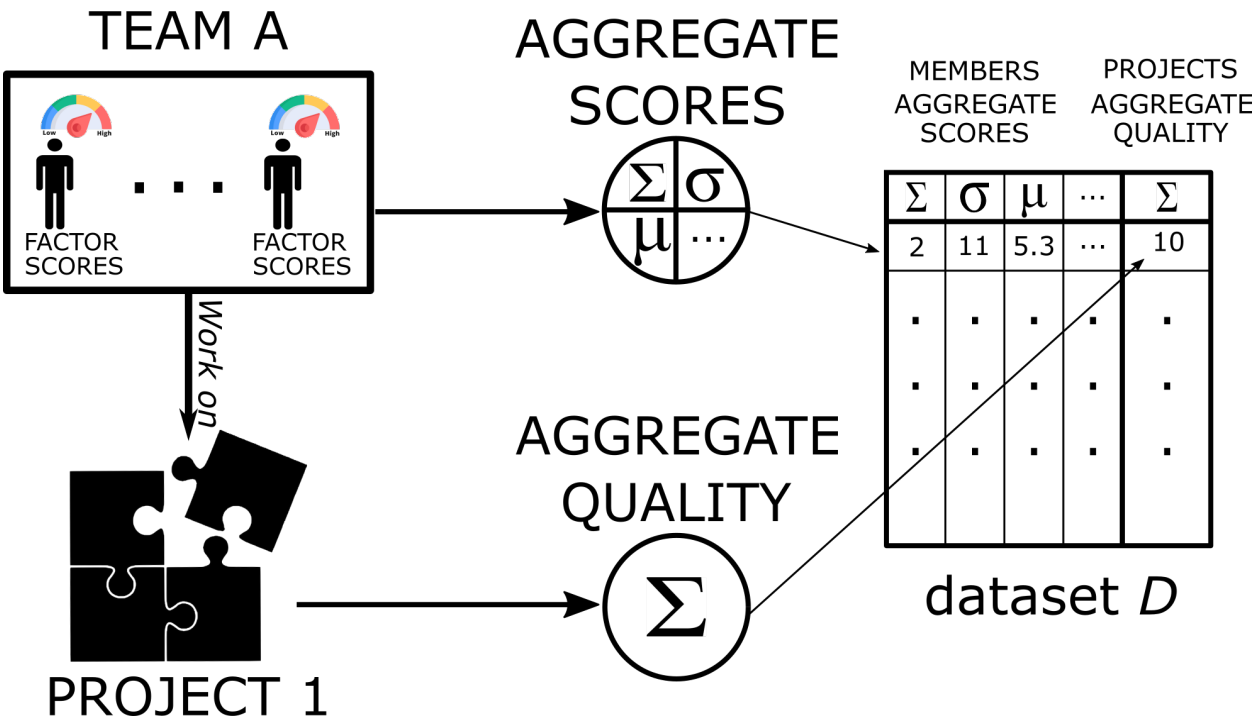



Figure 7.3: The process for producing the dataset *D*.

aggregates with project quality.

We arrive at four findings. First, IMPACT_TTL is the sum of the impact factor scores of all team members. This aggregated factor score (factor aggregate) moderately and positively correlates with project quality (see Table 7.2). This correlation shows that teams that accumulate a high impact total are more likely to produce high quality projects. Second, IMPACT_MAX is the greatest impact factor score in each team. This factor aggregate moderately and positively correlates with project quality (see Table 7.2), which shows that the highest impact score in a team is indicative of the team's propensity to produce high quality projects. Third, ACTIVITY_MIN is the smallest activity factor score in each team. This factor aggregate moderately and negatively correlates with project quality (see Table 7.2). A team having a large smallest activity factor score is a team with the least active member being quite active. The negative correlation shows that highly active teams (as demonstrated by the high level of activity of even the least active member),

Table 7.2: Factor aggregates correlation coefficients with project quality. ** correlation is significant at the 0.01 level and * at the 0.05 level (two-tailed). We bold the correlation coefficients that have an absolute value $\geq$0.3.

| Factor aggregates | Correlation with project quality |
|---|---|
| IMPACT_TTL | **.307**** |
| IMPACT_AVG | .104** |
| IMPACT_MDN | -.086** |
| IMPACT_MAX | **.319**** |
| IMPACT_MIN | -.220** |
| IMPACT_STD | .247** |
| ACTIVITY_TTL | .144** |
| ACTIVITY_AVG | -.196** |
| ACTIVITY_MDN | -.161** |
| ACTIVITY_MAX | 0.009 |
| ACTIVITY_MIN | **-.363**** |
| ACTIVITY_STD | .088** |
| POLICING_TTL | .143** |
| POLICING_AVG | -0.053 |
| POLICING_MDN | -.116** |
| POLICING_MAX | .137** |
| POLICING_MIN | -.058* |
| POLICING_STD | 0.049 |

produce low-quality projects. In other words, a team made exclusively of highly active members is counterproductive. Although such members may read or produce much content; the quality of their content is likely to be low and does not attract endorsement. The lack of correlation between *Impact* and *Activity* factors reported in Section 6.3.0.2 supports this conclusion further. Fourth, *Policing/Rowdiness* factor has no high or moderate correlations with project quality. Its highest correlations are POLICING_TTL and POLICING_MAX, which may show that the higher the total and maximum rowdiness and policing activities of members, the more follows the project receives. In other words, higher policing or rowdiness score in teams generates more attention in the form of following actions. However, given the nature of this factor (represents both policing and rowdy behaviours) and because of the low correlation coefficients, this should be interpreted with caution. We investigate these findings further in the following sections.

### 7.2.4 Interactions of aggregated factor scores

The correlation analysis between team-level factor aggregate scores and project quality shows that a team's IMPACT_TTL, IMPACT_MAX, and ACTIVITY_MIN aggregate scores are important indicators of performance. Thus, this section investigates the presence of team compositions that are likely to exhibit these scores by analyzing these aggregates' interactions in existing teams. To this end, we perform Pearson's correlation on the team-level aggregates and report the results in Table 7.3.

We make several findings in this section. First, teams with high total impact scores tend to contain a single member with a substantially higher impact score than the rest of the team and tend to contain diverse activity scores. Second, teams with low ACTIVITY_MIN scores tend to have low average activity levels and high activity diversity. These findings further elucidate the relationship between a team's IMPACT_TTL, IMPACT_MAX, and ACTIVITY_MIN aggregate scores and project quality.

From Table 7.3, we see that as a team's total impact scores increase, its average impact score (but not median) also increases. Unlike the median, the average is sensitive to outliers. This increase in the average suggests that the total impact score is mostly due to the presence of a highly impactful member. The high correlation between the total and maximum impact scores, between the total and the standard deviation of impact scores, and between the maximum and the standard deviation of impact, supports this conclusion.

Furthermore, the high correlations of the standard deviation of the impact score and the total and the maximum impact scores indicate that teams with high impact scores also tend to have members with varying impact scores. Whether the diversity of impact in a team is necessary for improved project quality or if it is just a consequence of having a member with proportionally high impact score resulting in increased standard deviation, needs further analysis. We know, however, that the diversity of impact (*i.e.* IMPACT_STD) in teams correlates weakly and positively with project quality (see Table 7.2). Since IMPACT_TTL and IMPACT_MAX are indicative of project

Table 7.3: Factor aggregates' intercorrelation coefficients. All correlations are significant at the 0.01 level. We bold correlations outside the [-0.3,0.3] range.

| | **IMPACT_TTL** | **IMPACT_AVG** | **IMPACT_MDN** | **IMPACT_MAX** | **IMPACT_MIN** | **IMPACT_STD** |
|---|---|---|---|---|---|---|
| **IMPACT_TTL** | **1** | **0.67** | 0.216 | **0.931** | -0.091 | **0.809** |
| **IMPACT_AVG** | **0.67** | **1** | **0.541** | **0.751** | 0.216 | **0.867** |
| **IMPACT_MDN** | 0.216 | **0.541** | **1** | 0.131 | **0.478** | 0.161 |
| **IMPACT_MAX** | **0.931** | **0.751** | 0.131 | **1** | -0.154 | **0.937** |
| **IMPACT_MIN** | -0.091 | 0.216 | **0.478** | -0.154 | **1** | -0.163 |
| **IMPACT_STD** | **0.809** | **0.867** | 0.161 | **0.937** | -0.163 | **1** |
| | | | | | | |
| **ACTIVITY_TTL** | **0.401** | 0.08 | 0.095 | 0.228 | -0.19 | 0.103 |
| **ACTIVITY_AVG** | -0.052 | 0.148 | **0.309** | -0.085 | 0.155 | -0.036 |
| **ACTIVITY_MDN** | 0.032 | 0.24 | **0.328** | 0.021 | 0.138 | 0.08 |
| **ACTIVITY_MAX** | 0.28 | **0.341** | 0.226 | 0.251 | 0.003 | 0.271 |
| **ACTIVITY_MIN** | **-0.771** | **-0.470** | 0.101 | **-0.832** | 0.228 | **-0.738** |
| **ACTIVITY_STD** | **0.471** | **0.607** | 0.194 | **0.53** | -0.007 | **0.599** |

| | **ACTIVITY_TTL** | **ACTIVITY_AVG** | **ACTIVITY_MDN** | **ACTIVITY_MAX** | **ACTIVITY_MIN** | **ACTIVITY_STD** |
|---|---|---|---|---|---|---|
| **IMPACT_TTL** | **0.401** | -0.052 | 0.032 | 0.280 | **-0.771** | **0.471** |
| **IMPACT_AVG** | 0.080 | 0.148 | 0.240 | **0.341** | **-0.470** | **0.607** |
| **IMPACT_MDN** | 0.095 | **0.309** | **0.328** | 0.226 | 0.101 | 0.194 |
| **IMPACT_MAX** | 0.228 | -0.085 | 0.021 | 0.251 | **-0.832** | **0.53** |
| **IMPACT_MIN** | -0.190 | 0.155 | 0.138 | 0.003 | 0.228 | -0.007 |
| **IMPACT_STD** | 0.103 | -0.036 | 0.08 | 0.271 | **-0.738** | **0.599** |
| | | | | | | |
| **ACTIVITY_TTL** | **1** | **0.534** | **0.427** | **0.679** | 0.019 | **0.4** |
| **ACTIVITY_AVG** | **0.534** | **1** | **0.859** | **0.798** | **0.506** | **0.501** |
| **ACTIVITY_MDN** | **0.427** | **0.859** | **1** | **0.536** | **0.354** | **0.322** |
| **ACTIVITY_MAX** | **0.679** | **0.798** | **0.536** | **1** | 0.090 | **0.847** |
| **ACTIVITY_MIN** | 0.019 | **0.506** | **0.354** | 0.09 | **1** | **-0.324** |
| **ACTIVITY_STD** | **0.4** | **0.501** | **0.322** | **0.847** | **-0.324** | **1** |

| | **POLICING_TTL** | **POLICING_AVG** | **POLICING_MDN** | **POLICING_MAX** | **POLICING_MIN** | **POLICING_STD** |
|---|---|---|---|---|---|---|
| **IMPACT_TTL** | 0.27 | 0.058 | 0.035 | 0.158 | 0.006 | 0.071 |
| **IMPACT_AVG** | 0.167 | 0.18 | 0.22 | 0.101 | 0.168 | 0.092 |
| **IMPACT_MDN** | 0.063 | 0.133 | 0.29 | -0.015 | 0.292 | -0.017 |
| **IMPACT_MAX** | 0.259 | 0.075 | 0.028 | 0.17 | -0.009 | 0.097 |
| **IMPACT_MIN** | -0.072 | 0.057 | 0.229 | -0.117 | 0.233 | -0.086 |
| **IMPACT_STD** | 0.2 | 0.116 | 0.067 | 0.148 | 0.015 | 0.116 |
| | | | | | | |
| **ACTIVITY_TTL** | 0.268 | 0.092 | 0.110 | 0.170 | 0.088 | 0.075 |
| **ACTIVITY_AVG** | 0.134 | 0.249 | **0.369** | 0.069 | **0.326** | 0.077 |
| **ACTIVITY_MDN** | 0.150 | 0.265 | **0.378** | 0.083 | **0.358** | 0.088 |
| **ACTIVITY_MAX** | 0.209 | 0.202 | 0.264 | 0.138 | 0.166 | 0.109 |
| **ACTIVITY_MIN** | -0.168 | 0.048 | 0.142 | -0.124 | 0.191 | -0.059 |
| **ACTIVITY_STD** | 0.200 | 0.186 | 0.213 | 0.142 | 0.078 | 0.125 |

| | **POLICING_TTL** | **POLICING_AVG** | **POLICING_MDN** | **POLICING_MAX** | **POLICING_MIN** | **POLICING_STD** |
|---|---|---|---|---|---|---|
| **POLICING_TTL** | **1** | **0.844** | **0.465** | **0.93** | 0.231 | **0.867** |
| **POLICING_AVG** | **0.844** | **1** | **0.644** | **0.857** | **0.372** | **0.898** |
| **POLICING_MDN** | **0.465** | **0.644** | **1** | **0.303** | **0.475** | **0.304** |
| **POLICING_MAX** | **0.93** | **0.857** | **0.303** | **1** | 0.122 | **0.971** |
| **POLICING_MIN** | 0.231 | **0.372** | **0.475** | 0.122 | **1** | 0.079 |
| **POLICING_STD** | **0.867** | **0.898** | **0.304** | **0.971** | 0.079 | **1** |

quality, the diversity of impact is possibly a contributor to this quality, as well, so we investigate this finding further in Section 7.2.5.

Table 7.3 shows that most activity aggregates are intercorrelated, except for the minimum activity scores, and the maximum and total activity scores. The intercorrelations among factor aggregates of the same factor is expected because they are derived from the same population and represent the same factor. The lack of correlation between the minimum and the maximum, which appears consistently among all other factors, mean that there is no clear relationship between members' activity scores and their tendency to self-assemble into a team. Teams in our SCC self-assemble based purely on individual interests in a particular project. The lack of correlations between the minimum activity scores and the total activity scores is somewhat counterintuitive. We expect changes in the minimum activity score in our teams to reflect on the total, especially since the minimum activity scores correlate positively with the average and the median activity scores. However, in this case, the minimum activity score across teams does not change sufficiently to correlate with the total.

We also see from Table 7.3 that as the total activity score increases in teams, so does the average, the median and the maximum scores, which indicates that our existing team compositions are less prone to outliers in terms of activity, as is the case with the impact scores. The average and the median activity scores' positive correlations with all other activity aggregates support this finding.

The minimum activity score across teams positively correlates with the average and median activity scores, which indicates that as team compositions increase in minimum activity, so does the teams' activity scores' central tendency.

Interestingly, the minimum activity score negatively correlates with the standard deviation. Thus, as the teams' minimum activity scores increase, the diversity of the activity scores among the same teams reduces. Thus, as teams' minimum activity scores reduce, they become more likely to have low average activity scores and high standard deviation activity scores. Since the teams'

minimum activity scores negatively correlate with project quality, the reduction of the minimum activity score in a team is likely to improve project quality. Moreover, since teams with low minimum activity tend to have reduced average activity and increased standard deviation of activity, it follows that reduced average activity and increased standard deviation may also be related to project quality. Interestingly, our teams' minimum activity scores also negatively correlate with the total and the maximum impact scores; the latter two correlating positively with project quality. Moreover, the standard deviation of activity scores across all teams positively correlates with the total and the maximum impact scores. Both these findings indicate that teams with high impact scores tend to have diverse levels of activity, which further indicates that a team's increased diversity of activity scores is related to team performance.

The degree of this relation is not strong enough to manifest in the correlation between ACTIVITY_AVG and ACTIVITY_STD with project quality directly. Table 7.2 shows that ACTIVITY_AVG (and ACTIVITY_MDN) weakly and negatively correlates with project quality, and ACTIVITY_STD has nearly no correlation. These small correlation coefficients could result from having too few teams with these characteristics for this pattern to manifest. These findings, however, are interesting, so we investigate them further in Section 7.2.5.

The central tendencies of policing and activity correlate positively. Moreover, the central tendency of activity positively correlates with POLICING_MIN score in teams. These correlations imply that as teams' activity increases, so does their policing and rowdiness activities. The correlation between members' *Activity* and *Policing/Rowdiness* factors explains this finding. We show that as member activity scores increase, members' tendency to report or produce rowdy content increases. Similar to *Activity*, *Policing/Rowdiness* aggregates correlate with each other, which is expected.

### 7.2.5 Team compositions and project quality

Our findings so far indicate that three team-level factor aggregates correlate with project quality and that several other factor aggregates are likely to occur with the first three in individual teams.

We hypothesized that teams with specific scores on these factor aggregates are likely to produce high-quality projects. Therefore, this section examines the effect of the presence (or absence) of specific factor aggregate scores, individually and collectively, on teams' performance.

We find that teams with high IMPACT_TTL, IMPACT_MAX, and IMPACT_STD scores tend to produce higher quality projects than those with low scores. We also find that teams with high IMPACT_MDN, IMPACT_MIN, ACTIVITY_AVG, ACTIVITY_MDN, and ACTIVITY_MIN tend to produce lower-quality projects than those with high scores, so lowering these factor aggregate scores in teams would likely improve their performance. Furthermore, we find that teams composed of a specific mixture of these factor aggregates tend to produce high-quality projects.

To this end, we perform two sets of analysis, using hypothesis testing. The first analysis compares the performance of teams with specific factor aggregate scores with those that lack these scores. This analysis focuses only on individual factor aggregate. The second analysis compares the performance of teams that are composed of specific mixtures of factor aggregate scores against others that do not have this mixture.

Since *Policing/Rowdiness* is uncorrelated with project quality (see Table 7.2), we exclude it from the subsequent analysis

#### 7.2.5.1 The effect of individual factor aggregates on project quality

We bin each factor aggregate in the dataset $D$ independently into high and low bins using the quantile-binning technique. The high bin contains the factor aggregate scores above the $50^{th}$ percentile for a particular aggregate while the low contains the rest. Thus, both bins contain roughly the same number of scores. We suffice with only two bins considering the small sample size relative to the number of factor aggregates. Similar to Section 5.3.2, we use the non-parametric KS test[2] to carry out this analysis.

To determine if the high or low scores of a specific factor aggregate associate with increased project quality, we divide the teams into two groups for this specific aggregate: teams with a high

[2]We use *ks.test()* function offered in R's [150] stats package version 3.4.3.

score and a low score. We then perform a one-tailed KS test in each direction (denoted right and left). The small $p$-value in the right direction test means teams with low scores on a particular aggregate tend to produce projects with lower quality than those with high scores. The small $p$-value in the left direction test means the opposite. We do this for all factor aggregates and report the results in Table 7.4.

Table 7.4: The results of one-tailed KS hypothesis tests in each direction. The "right $p$-value" column contains the $p$-values of the KS test with the alternative hypothesis that the CDF of the first test set (teams with low score) is above the second one (teams with high score). The "left $p$-value" column is for the opposite alternative hypothesis.

| **factor aggregate** | **# teams with low score** | **# teams with high score** | **right $p$-value** | **left $p$-value** |
|---|---|---|---|---|
| IMPACT_TTL | 624 | 611 | **0.02816** | 0.62585 |
| IMPACT_AVG | 625 | 610 | 0.67358 | 0.23775 |
| IMPACT_MDN | 627 | 608 | 1 | **0.00488** |
| IMPACT_MAX | 647 | 588 | **0.00004** | 0.99413 |
| IMPACT_MIN | 789 | 446 | 1 | **0** |
| IMPACT_STD | 671 | 564 | **0.00143** | 0.99457 |
| | | | | |
| ACTIVITY_TTL | 623 | 612 | 0.65366 | 0.84382 |
| ACTIVITY_AVG | 637 | 598 | 1 | **0.00028** |
| ACTIVITY_MDN | 623 | 612 | 1 | **0.02308** |
| ACTIVITY_MAX | 627 | 608 | 0.30909 | 0.9551396 |
| ACTIVITY_MIN | 637 | 598 | 1 | **0.00005** |
| ACTIVITY_STD | 629 | 606 | **0.158** | 0.98679 |

The results indicate that teams with high IMPACT_TTL and IMPACT_MAX are associated with improved project quality. Interestingly, IMPACT_STD scores are also associated with improved project quality, which is in line with its high correlation coefficient with IMPACT_TTL and IMPACT_MAX (see Table 7.3). Furthermore, high ACTIVITY_STD scores in teams are also associated, though with a moderate statistical significance, with improved project quality.

The results also indicate that teams with high ACTIVITY_MIN scores tend to produce lower-quality projects than do the teams with a high score on this factor aggregate, which is in line with this aggregate's correlation with project quality (see Table 7.2). This finding is also true of IMPACT_MDN, IMPACT_MIN, ACTIVITY_AVG, and ACTIVITY_MDN factor aggregates.

#### 7.2.5.2 The effect of factor aggregate mixtures on project quality

This section builds upon the previous section's findings by testing the tendency of specific team compositions to produce high-quality projects. Section 7.2.5.1 identifies several factor aggregates that, when present in a team at certain levels, associate with improved or lowered project quality. Specifically, teams with high IMPACT_TTL, IMPACT_MAX, IMPACT_STD, and ACTIVITY_STD scores tend to produce higher projects quality than those with low scores on these aggregates. Moreover, teams with high ACTIVITY_MIN, IMPACT_MDN, IMPACT_MIN, ACTIVITY_AVG, and ACTIVITY_MDN tend to produce projects with lower quality than teams with high scores on these aggregates. Therefore, this section investigates several compositions of existing teams based on the findings so far. These compositions are listed in Table 7.5.

The first composition is based on the direct correlation that IMPACT_TTL, IMPACT_MAX, and ACTIVITY_MIN have with the quality we identify in Table 7.2. We show that these aggregates (independently) correlate with the project quality, so we test if their presence together in teams would increase the likelihood of the team producing high-quality projects. The second composition includes the factor aggregates that we identify in Table 7.3, excluding the ones from the first composition (*i.e.* IMPACT_STD, ACTIVITY_AVG, and ACTIVITY_STD). We show that these aggregates positively correlate with the aggregates in the first composition and test the effect of their presence in teams separately to get a clear indication of their utility.

The third, fourth, and fifth compositions are based on the findings we identify in Table 7.4. We show that teams with high IMPACT_TTL and IMPACT_MAX and MPACT_STD and ACTIVITY_STD factor aggregate scores are associated with improved project quality. We also show that teams with high ACTIVITY_MIN, IMPACT_MDN, IMPACT_MIN, ACTIVITY_AVG, and ACTIVITY_MDN factor aggregate scores are associated with reduced project quality. Therefore, we test both subsets independently in compositions 3 and 4. We also test the negation of the fourth composition in composition 5. The last composition includes all factor aggregates that relate to an increase in the project quality.

For each composition, we separate the teams that adhere to this composition from all others, producing two groups. We then perform a one-tailed KS test comparing the project quality of both groups in each direction. The small $p$-value in the right direction test means teams that adhere to the specific composition have higher quality projects, and the small $p$-value in the left direction test means the opposite. We do this for all compositions and report the results in Table 7.6.

Table 7.5: Team compositions.

| team composition name | IMPACT _TTL | IMPACT _MDN | IMPACT _MAX | IMPACT _MIN | IMPACT _STD | ACTIVITY _AVG | ACTIVITY _MDN | ACTIVITY _MIN | ACTIVITY _STD |
|---|---|---|---|---|---|---|---|---|---|
| composition 1 | high | - | high | - | - | - | - | low | - |
| composition 2 | - | - | - | - | high | low | - | - | high |
| composition 3 | high | - | high | - | high | - | - | - | high |
| composition 4 | - | low | - | low | - | low | low | low | - |
| composition 5 | - | high | - | high | - | high | high | high | - |
| composition 6 | high | low | high | low | high | low | low | low | high |

Table 7.6: The results of one-tailed KS hypothesis tests in each direction. The "right $p$-value" column contains the $p$-values of the KS test with the alternative hypothesis that the CDF of the first test set (teams with no specific composition) is above the second one (teams with specific composition). The "left $p$-value" column is for the opposite alternative hypothesis.

| team composition | # teams with no specific composition | # teams with specific composition | right $p$-value | left $p$-value |
|---|---|---|---|---|
| **composition 1** | 973 | 262 | **0.0** | 0.99826 |
| **composition 2** | 1144 | 91 | **0.0** | 0.99948 |
| **composition 3** | 868 | 367 | **0.00979** | 0.99726 |
| **composition 4** | 1035 | 200 | **0.0** | 1.0 |
| **composition 5** | 1114 | 121 | 1.0 | **0.00043** |
| **composition 6** | 1195 | 40 | **0.00001** | 0.99978 |

The results are consistent with our expectations; the teams with specific compositions (*i.e.* compositions 1, 2, 3, 4, and 6) tend to outperform other teams, except for the fifth composition teams, which, as we expect, yield lower quality projects than other teams. The results suggest that the elevation of the impact score within a team (through the presence of a team member with a relatively higher impact score than the rest of the team members, which results in increased impact score diversity), increased activity score diversity, and reduced overall levels of activity within a team are vital for improving team performance. We discuss the implications on SCC teams in the

next section (*i.e.* Section 7.3.

These findings show that an association between specific team compositions and improved (and deteriorated) team performance exists in our data. Given that SCCs share common characteristics (see Chapter 4) and that this analysis captures these characteristics through the inclusion of representative base variables, the results are potentially generalizable to other SCCs. Thus, the repetition of this analysis across various SCCs is the logical next step to establishing the generalizability of these results. Another necessary step is to investigate the presence (and directionality) of a causal relationship between these team compositions and team performance through focused and adequately controlled case studies with rigorous theoretical reasoning of the results. Therefore, we include these two steps as a viable and necessary future direction of this research. We discuss the threats to this research's validity in Section 8.3.

### 7.2.6 Utility of aggregated factor scores in predicting team performance

The analyses in this chapter thus far investigate the relationship between team-level factor aggregates and project quality using univariate analysis methods, indicating the presence of a relationship between specific aggregates and project quality. Naturally, such a relationship can be operationalized for decision-making. Thus, this section builds upon findings from the previous section by producing an artifact (*i.e.* a machine learning model) that is capable of predicting the potential performance of a team based on its composition, effectively demonstrating the utility of these factor aggregates in predicting team performance. Since this approach considers the aggregates collectively (*i.e.* multivariate analysis), it contributes to our understanding of the interactions between these aggregates. We use the project quality measure as a proxy for team performance. We find that a simple decision tree can predict team performance with remarkably low error.

We train a simple machine learning algorithm (*i.e.* decision trees [158]) on a transformed version of the $D$ dataset and report its performance below. Decision trees are simple models that have a relatively small number of hyperparameters to tune and are robust to many data anomalies with which other learners strugle [159], such as the distributions of the underlying data, the scale

of the various features, and their type. Thus, they require little data processing. However, they are sensitive to the class imbalance problem, where the distribution of classes in the underlying data is skewed towards some class(es), so we address this problem below.

To this end, we transform the dataset $D$ by binning the project quality variable into two bins: high and low so that instances with corresponding project quality values that are larger than the average project quality value are labelled as high-quality projects, and low otherwise. The project quality column then serves as the class variable. Since we bin the instances of this dataset into two classes based on the average value of the project quality variable, the number of instances in each class becomes similar but not equal. Thus, we apply the synthetic minority over-sampling technique (SMOTE) [160] to mitigate the effect of class priors on the learning process. Moreover, we exclude *Policing/Rowdiness*-related factor aggregates from this transformed dataset for the same reason mentioned above. We use this transformed dataset to train the decision tree algorithm using scikit-learn [161][3] and report its performance below.

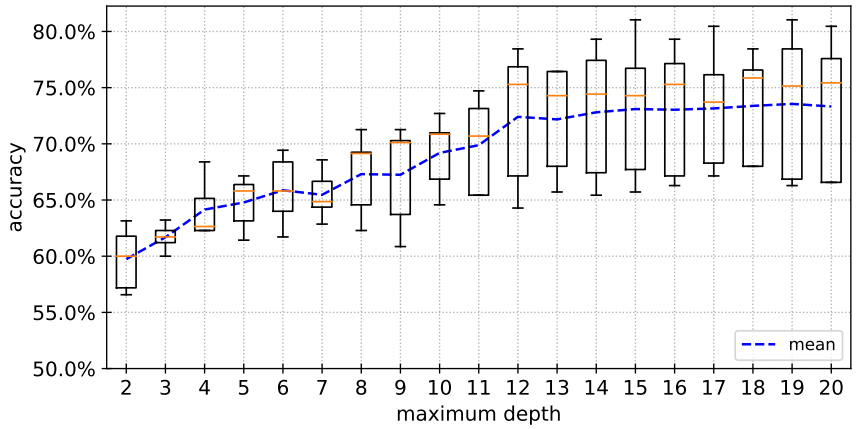
We configure the decision tree to split using the best variable determined by the Gini impurity criterion, and we do not introduce any class weights. The maximum depth hyperparameter caps the depth of the decision tree to some arbitrary value to avoid overfitting. Since it is not possible to estimate its value from the data directly, we perform grid-search to determine its "best" value. For each value of the maximum depth hyperparameter, we perform 5-fold cross-validation and plot the five models' accuracy as a boxplot (see Figure 7.4).

Each box in Figure 7.4 represents the five models' accuracy scores that we trained during the 5-fold cross-validation process using the corresponding maximum depth value and stratified sampling. The boxplot shows the maximum (upper whisker) and minimum (lower whisker) non-outlier accuracy scores along with the median (the orange line dividing the box) and outliers (circles). The dashed line represents the mean accuracy score of all models trained during the 5-fold cross-validation process using the corresponding maximum depth hyperparameter value.

The figure shows that as the trained models' complexity increases (through the increased maxi-

[3]We use scikit-learn version 0.20.3.

Figure 7.4: The 5-fold cross-validation models' accuracy scores as model complex increases.




mum depth hyperparameter), the variance of their accuracies' increases as well, which is consistent with over-fitting behaviour. The models trained using a maximum depth value of seven demonstrate a reasonable trade-off between accuracy and variance. They achieve a reasonable average accuracy value of 66% and a low standard deviation of 1.97%, demonstrating the utility of the FB model in predicting team performance.

## 7.3 Results and Discussions

Existing work shows empirical evidence and theoretical reasoning supporting the claim that teams with diverse personality traits tend to outperform other teams in specific tasks. Precisely, these findings show that teams composed of homogeneous members in terms of personality traits tend to outperform [162, 163, 16, 164] and report a more positive teamwork experience than teams with heterogeneous personality traits [8] when undertaking complex and creative tasks. Furthermore, the findings suggest that the elevation of some personality traits in a team (*e.g.* sum, average, *etc.* of members' conscientiousness scores) and the variability of others (*i.e.* standard deviation, variance, *etc.* of members' extroversion scores) within the same team have a statistically significant positive correlation with team performance. The authors of these studies support their findings by interpreting the implications of these findings on team performance.

For example, conscientiousness is the personality trait that indicates a person's tendency to do what is "right" and perform duties well [165, 166]. Thus, the elevation of conscientiousness in a team offers valuable temperamental qualities to this team, such as organization, planning and task focus that are critical for its success [18, 17]. These qualities are especially valuable for teams undertaking complex and creative tasks. Another example is the extroversion trait, which quantifies individuals' social tendencies and assertiveness [165]. Highly extroverted individuals thrive on social interactions. Thus, extroverts are natural facilitators of inter-team and intra-team communications. Therefore, they are a valuable resource to a team when working on tasks that require such facilitation. However, highly extroverted individuals are talkative, assertive and authoritative, so a team composed primarily of extroverts lacks task-focus and exhibits power struggles [18]. This reasoning explains why the elevation of extroversion in teams does not affect team performance, and its variability (*i.e.* standard deviation, variance, *etc.*) within a team positively relate to team performance and teamwork satisfaction [157].

Similarly, we interpret the implications of our findings on team performance by reasoning about their meaning in SCC teamwork. Members with high-levels of *Impact* scores produce content (*i.e.* artifacts, such as responses to queries or comments) that draw substantial attention from other members. Thus, such members are vital for producing interesting artifacts upon which discussions take place. Since discussions are an essential part of the SCC process, highly impactful individuals are invaluable facilitators of this process. Highly active members view, interact with, and produce large quantities of content. Thus, they naturally entice discussions around artifacts, which, as stated previously, is an essential part of the SCC process. Since SCC interactions are reciprocal (see Section 4.2.2), these members play a crucial role in engaging others in discussions. However, given their tendency to produce large quantities of content, which is mostly impact-less[4] these members may become counterproductive for a team. Such large quantities of impact-less content may reduce a team's task-focus and lead to the loss of ideas within the large corpus of produced

[4]The lack of correlation between *Impact* and *Activity* scores (see Table 6.3.0.2) suggests that the levels of activity are unrelated to the levels of impact for the majority of members.

content.

Given the importance of discussions in the SCC process and the essential roles these temperamental qualities play in facilitating them, it follows that members with such qualities (at appropriate levels) offer invaluable temperamental resources to SCC teams, and naturally form collaborations.

The above reasoning is in line with this chapter's results. We see that the elevation of *Impact* in teams through members with substantially larger impact scores than the rest of their team members (as evident by the positive effect of having increased IMPACT_MAX score, increased IMPACT_STD score, and lowered IMPACT_MDN score), have a positive association with project quality. Yang and Tang [9] identified a similar finding. Their study of team interactions from a social network analysis perspective shows that teams' maximum centrality scores positively correlate with team performance. Since impactful members tend to generate content that draws out interactions from others, it follows that impactful members are at a high propensity to occupy central positions in discussions. Furthermore, we also see that the reduction of overall levels of *Activity* and its increased diversity in teams are indicators of high performing teams. Since this research defines teams as members who made two or more textual contributions to a project, these two findings suggest that despite the clear need for teams to be active, there is a point after which the level of activity becomes counterproductive. This observation requires further analysis.

## 7.4 Findings Summary

This chapter shows significant correlations between FB model and the other activities not considered included in deriving the FB model, which are in line with our interpretation of the FB model. These correlations imply that the FB model is indicative of a boarder range of behaviours than those included in the derivation of the FB model. We then investigate the relationship between the various team-level factor aggregates with team performance, which lead us to identifying various team compositions that are associated with team performance. We then demonstrate the utility

of the FB model in predicting team performance. We do this by deriving factor scores for each member and then, using an abstract notion of teams that is based on the social interactions, we calculate team scores by the aggregating the individual factor scores. The relationship between the team-level factor scores and the corresponding team's project quality is then used as a proxy for team performance.

Overall, the results suggest that a team with a high elevation of *Impact*, high variability of *Activity*, and low overall *Activity* is more likely to produce high-quality projects.

# Chapter 8

# CONCLUSION, IMPLICATIONS, LIMITATIONS, AND FUTURE WORK

This chapter concludes the thesis by overviewing the key findings we identify in the process of answering the scientific question and discussing their practical implications, limitations, and future work.

## 8.1 Conclusion

This thesis identifies specific behavioural characteristics that individually and collectively relate to team performance in SCCs via retrospective quantitative analysis of member behaviour. It identifies that teams of particular behavioural compositions tend to outperform others in producing high-quality contributions. It does so by deriving a model (*i.e.* the FB model) from the historical logs of member behaviour and using this model to identify and test several team compositions' effects on team performance. It concludes that teams consisting of a relatively more impactful member than the rest of the members and the rest being adequately and diversely active tend to outperform teams that lack this consistency.

In the course of answering the scientific question, this thesis coins the SCC term and introduces SCC systems as a new area for research. It does so by identifying the practical commonalities among these systems and their distinguishing characteristics compared to existing crowdsourcing communities and social systems. It shows that SCCs consist of unique components and follow a unique process that reflects in their members' behaviour and consequently manifests as unique interaction and association patterns.

Taking these patterns into account, this thesis produces a behavioural model (*i.e.* FB model)

that quantifies member behaviour into three dimensions: *Impact*, *Activity*, and *Policing/Rowdiness*. It uses this model to identify specific existing team compositions that have a statistically significant association with team performance.

In more details, this thesis identifies association and interaction patterns unique to SCCs and others in common with SNs by applying social network analysis methods to SCCs' association and interaction networks. It also identifies members' interaction patterns, their evolution, and the factors that affect them. Based on the previous findings, this thesis derives the FB model from member interactions in *HCC*. It describes the relationship between teams' aggregated FB scores and their contribution quality, culminating in identifying specific compositions associated with contribution quality. It observes that teams' total and maximum impact scores, and the standard deviation of their activity scores positively correlate with their contribution quality. In contrast, their minimum activity scores negatively correlate with contribution quality. The thesis tests several team compositions and shows that compositions containing high impact scores, as a result of a member having a relatively higher impact score than the rest of the team, and reduced yet diverse levels of activity tend to produce higher contribution quality than others.

## 8.2 Practical implications of the results

This thesis's outcomes can be operationalized to improve SCCs' performance and to monitor their health. Both operationalizations are critical in alleviating the operation cost problem that Chapter 1 describes.

Chapter 7 finds that *Impact* and *Activity*, when present in a team in appropriate quantities, associate with increased contribution quality. This finding provides moderators with a guideline, whereby ensuring that teams have appropriate levels of impact and activity scores (by introducing new members or removing existing ones from teams to alleviate any discrepancies), they can improve their communities' overall performance. Since interactions are the main medium upon which the SCC process is predicated, and the *Impact* and the *Activity* factors are derived, it follows

that moderators must encourage interactions to cultivate these characteristics in their communities. This section identifies the following techniques for encouraging interactions.

First, since associations entice interactions (see Chapter 5), moderators should encourage members to associate with members or projects that they find interesting. To this end, moderators can employ a personalized recommendation engine capable of recommending members and projects to other members. Second, moderators can directly employ rewards, which entice interactions (see Chapter 5). Third, Chapter 4 shows that SCC members tend to reciprocate interactions. Therefore, moderators can directly improve the levels of interactions within their communities by increasing the degree to which they interact with other members.

Fourth, Chapter 5 identifies two factors that underpin the stability of interactions despite their high-volatility: the interaction replenishment rate and the SCC members' tendency to interact with new members. The first factor manifests in interactions between existing members who never interacted before, and the latter factor manifests in interactions among existing members with new members. Moderators can play a crucial role in ensuring that these factors remain present in their SCC. They can invite new members into discussions and invite members to discuss the contributions of others with which they never interacted before. This role can also be carried out systematically via a recommendation engine.

Chapter 4's finding regarding the adherence of both interaction and association networks to the power-law distribution and Chapter 5's finding regarding the manifestation of the Pareto principle in the production of interactions have important implications on community health and security, among others. These findings suggest that SCC interactions and associations follow a naturally occurring pattern in many other natural and human-made systems, the deviation from which is a sign of abnormality that likely indicates deterioration.

Furthermore, the presence of the power-law property in SCCs indicates their resilience to the random loss of members through attrition. However, it also indicates the presence of a few highly connected members, the loss of which is detrimental to the connectively of the network. There-

fore, we recommend that moderators monitor their SCC's health through this property to identify possible signs of deterioration and employ a recruitment and retention strategy that takes the above into account.

## 8.3 Limitations

This work applies several precautions to ensure the validity and generalizability of the results. First, it avoids introducing a selection bias by carefully defining explicit inclusion and exclusion criteria/thresholds for members, teams, and projects. Second, its design avoids introducing a systemic bias. Specifically, the members implicitly self-assemble into teams, and the SCC platform gathers their data without the authors' interference. Furthermore, since this study is retrospective, the members (including moderators) carried out their respective parts while unaware of the study and team formations, making this study blind with consistently applied protocols.

Third, it selects an SCC with demographically and geographically diverse members and a relatively large number of teams compared to similar work. Fourth, it reports all relevant information about the dataset, the analysis process, and the results to allow others to reproduce the results. Fifth, it excludes from the FB model any base variables specific to the organization (*i.e.* Chaordix) or *HCC* platform, such as gamification features, and retains base variables that capture the common characteristics among SCCs, such as interactions and associations. Since Chapter 4 concludes that SCCs share common characteristics and that the FB model captures these characteristics through the inclusion of representative base variables, the results are potentially generalizable to other SCCs.

Sixth, it carefully reprocesses the data, when necessary, to mitigate the effect of group imbalance on the results. Specifically, Section 7.2.6 employs machine learning to demonstrate the utility of the FB model in predicting whether a specific team composition yields a high or low performing team. Since the machine learning algorithm's training process is sensitive to imbalance, this work employs the SMOTE resampling technique to ensure an even distribution of high and low

performing teams in the training and testing data. Furthermore, it employs stratified sampling in the 5-fold cross-validation process to ensure that each training and testing subset consists of the same distribution of high and low performing teams.

The following paragraphs discuss the limitations of this work, including threats to validity. These limitations should be taken into account when interpreting the results.

First, this thesis's results apply only to the specific team performance measure it defines and uses. This measure quantifies a team's performance as the number of followers this team's outcomes generates, which is subjective to members' perspectives and specific to one facet of team performance. Furthermore, this measure can be sensitive to platform features and moderator actions. For example, a platform employing a recommendation engine (or similar moderator actions) can dictate a team's performance by how well it recommends this team's outcomes. An ideal measure for SCCs objectively quantifies a project's contribution towards answering a query, or its utility in leading others to arrive at such answers independent from platform features or moderator actions. Unfortunately, no such measure exists, nor we have access to data that allows us to create one. Thus, we suffice with this measure and leave a more fulsome investigation of this limitation for future work.

Internal Threat - Confounding: a situation where changes in an outcome variable can be thought to have resulted from some third variable related to the treatment you administered Second, despite analyzing a relatively large number of teams (see Section 7.2.2) in comparison to similar research, this number is not sufficiently large to allow for adequate control over other variables that potentially contribute to the "effect" between team compositions and performance. Variables, such as the other factor aggregates that are no part of a composition, team size, project type and complexity, and rewards may have a confounding effect on project quality, which is a threat to internal validity. Nevertheless, this work minimizes the effect of confounding, to some degree, in several ways. First, it considers the effect of a mixture of factor aggregate on project quality (see Section 7.2.5.2). Second, it employs multivariate analysis (*i.e.* decision trees) to accounts for all factor-aggregates

simultaneously when assessing the possibility of predicting team performance based on the factor aggregates. Finally, it ensures that the findings are consistent with the theoretical understanding of SCCs.

Third, Chapter 5 shows that *HCC* has a high turn-over rate, interaction volatility, and interaction replenishment rate. Therefore, this work is likely to be susceptible to the attrition threat to internal validity. Nevertheless, though members may cease interacting, the analysis continues to consider them part of the team because their contributions remain on the platform where others can continue to see and discuss them. Therefore, members do not leave the study in a way that skews the population. Furthermore, the replenishment rate suggests that lost members are replaced with new ones who are likely to join ongoing projects. Therefore, it is unlikely to have a situation where the population becomes primarily composed of members who "remained in the study". Thus, attrition's effect is minimal.

Fourth, members who frequent the SCC platform naturally become more skilled in producing high-quality contributions than those who do not. This is because members' skills improve over time, as they get involved in more queries, responses, and discussions. Consequently, it is feasible that a team's performance may improve as the duration of their project increases. This observation raises an important question regarding the effect of maturation on team performance. Can a team's aggregated "platform age" (*i.e.* number of days logged in) and project duration account for the observed effect that team compositions have on project quality? Section 6.2.1 shows that the effect of "platform age" on the FB model is minimized by averaging each member's base variables over the number of days this member logged in to the SCC, which mitigates this aspect of maturation. However, this work does not consider the effect of project duration, so it remains a potential threat to this work's internal validity.

Fifth, this thesis focuses on describing and demonstrating the relationship between team compositions and performance without any experimental manipulation. It identifies specific team compositions from the data and compares the performance of teams that adhere to these compositions

and those that do not. Therefore, the findings are purely descriptive. They do not imply that specific compositions caused such improvements. Instead, they suggest the presence of a relationship between these compositions and improved team performance. It is possible to strive towards prescriptive team compositions by extending this work to include testing the compositions using controlled experiments on live communities, so we describe this extension in more detail in Section 8.4.

Sixth, despite the steps this work takes toward improving result generalizability, the lack of result replication remains a threat to its external validity. The use of one SCC in Chapters 5 to 7 raises the possibility that these results include phenomena specific to this SCC, such as SCC process, moderator intervention techniques, and nature of work. In this case, the results would not generalize well to other SCCs that do not share the same underlying phenomena.

Considering the above limitations and precautions, this study's implications on forming high performing SCC teams remain interesting and worthy of further investigation.

## 8.4 Future Work

This thesis presents an exploratory case study that utilizes a quantitative approach to the exploration of team performance in SCCs. By adopting simplifying assumptions regarding the definitions of a team, a project, and a project's quality, this thesis arrives at a systemic method that identifies common behavioural characteristics in high performing teams and explains these characteristics' role in team performance. It sufficiently demonstrates the implication of the scientific question in SCC team performance and proves that further investigation, through formal case studies, is both viable and required. Thus, it serves as a stepping-stone for such investigation.

This section identifies two possible avenues of investigation regarding the generalizability and the causality. Since this thesis predicates its findings on a single SCC, the replication of this analysis across various SCCs is the logical next step to establishing the findings' generalizability. Furthermore, since this thesis demonstrates the relationship between team compositions and

performance, investigating the presence (and directionality) of a causal relationship between them, through focused and adequately controlled case studies, is another possible avenue. Therefore, this section identifies these two avenues as a viable and necessary future direction of this research.

As stated above, the replication of this thesis's findings on various SCCs is essential to establishing their generalizability. These replication studies must occur using SCCs from different organizations to ensure that their results do not capture any organization-specific phenomenon. These studies must also account for any confounding variables, such as other team-level factor aggregates that are not part of a specific composition, maturation, team size, member rewards, and project type and complexity (see Section 8.3 for more details). Since these variables are intractably many, it is practically impossible to account for all them in a single study; therefore, it is essential to limit them to variables with theoretical and practical backing of existing work and confirm their sufficiency using statistical approaches, such as mediation analysis [167].

As stated above, investigating the causal relationship between team compositions and performance is another viable future work. This investigation can be carried out through controlled and focused experiments on live SCCs that generate the necessary data for testing the presence of a causal relationship between these team compositions and performance. These experiments must utilize experiment manipulation as the primary process and must account for the confounding variables that Section 8.3 identifies.

Team performance is a multifacet and subjective phenomenon that is difficult to quantify in a single objective measure [168]. Many existing measures capture specific aspects of team performance, relating to the quality of team outcomes, inter-team dynamics, and the team's utility in accomplishing organizational goals. Thus, the careful consideration and reasoning of a selected measure are required to establish its validity in the SCC context for any future research.

Section 8.3 describes an "ideal" SCC team performance measure predicated on team outcomes. It defines it as the measure that objectively captures the utility of a team's output in answering a query or sparking discussions where others arrive at such answers. Thus, we identify the creation

and testing of this measure as a viable and independent future work.

# Bibliography


[1] K. Zaamout and K. Barker, "Structure of crowdsourcing community networks," *IEEE Transactions on Computational Social Systems*, vol. 5, pp. 144–155, March 2018.

[2] K. Zaamout, T. Arjannikov, and K. Barker, "A social crowdsourcing community case study: Interaction patterns, evolution, and factors that affect them," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 659–671, 2020.

[3] K. Zaamout and K. Barker, "Towards quantifying behaviour in social crowdsourcing communities," in *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018.*, p. 203, 2018.

[4] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42, ACM, 2007.

[5] A. R. Henderson, "Testing experimental data for univariate normality," *Clinica Chimica Acta*, vol. 366, no. 1, pp. 112–129, 2006.

[6] S. Korkmaz, D. Goksuluk, and G. Zararsiz, "Mvn: an r package for assessing multivariate normality," *The R Journal*, vol. 6, no. 2, pp. 151–162, 2014.

[7] D. L. Morgan, *The focus group guidebook*, vol. 1. Sage publications, 1997.

[8] R. H. Rutherfoord, "Using personality inventories to help form teams for software engineering class projects," *ACM SIGCSE Bulletin*, vol. 33, no. 3, pp. 73–76, 2001.

[9] H.-L. Yang and J.-H. Tang, "Team structure and team performance in is development: a social network perspective," *Information & Management*, vol. 41, no. 3, pp. 335–349, 2004.

[10] J. F. Salgado, "The five factor model of personality and job performance in the european community.," *Journal of Applied psychology*, vol. 82, no. 1, p. 30, 1997.

[11] B.-C. Lim and R. E. Ployhart, "Transformational leadership: relations to the five-factor model and team performance in typical and maximum contexts.," *Journal of Applied Psychology*, vol. 89, no. 4, p. 610, 2004.

[12] M. Awad, L. Khan, and B. Thuraisingham, "Predicting www surfing using multiple evidence combination," *The VLDB JournalThe International Journal on Very Large Data Bases*, vol. 17, no. 3, pp. 401–417, 2008.

[13] T. Hansen, J. M. Jensen, and H. S. Solgaard, "Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior," *International Journal of Information Management*, vol. 24, no. 6, pp. 539–550, 2004.

[14] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Personality: critical concepts in psychology*, vol. 60, p. 295, 1998.

[15] R. R. McCrae, "The place of the ffm in personality psychology," *Psychological Inquiry*, vol. 21, no. 1, pp. 57–64, 2010.

[16] A. R. Peslak, "The impact of personality on information technology team projects," in *Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future*, pp. 273–279, ACM, 2006.

[17] T. A. O'Neill and N. J. Allen, "Personality and the prediction of team performance," *European Journal of Personality*, vol. 25, no. 1, pp. 31–42, 2011.

[18] M. A. Peeters, H. F. Tuijl, C. G. Rutte, and I. M. Reymen, "Personality and team performance: A meta-analysis," *European Journal of Personality*, vol. 20, no. 5, pp. 377–396, 2006.

[19] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," *Machine learning*,

vol. 95, no. 3, pp. 357–380, 2014.

[20] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Personal and Ubiquitous Computing*, vol. 17, no. 3, pp. 433–450, 2013.

[21] M. Janhonen and J.-E. Johanson, "Role of knowledge conversion and social networks in team performance," *International Journal of Information Management*, vol. 31, no. 3, pp. 217–225, 2011.

[22] H. E. Aldrich and P. H. Kim, "Small worlds, infinite possibilities? how social networks affect entrepreneurial team formation and search," *Strategic Entrepreneurship Journal*, vol. 1, no. 1-2, pp. 147–165, 2007.

[23] R. S. Burt, "The social structure of competition," *Explorations in economic sociology*, vol. 65, p. 103, 1993.

[24] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

[25] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral, "Team assembly mechanisms determine collaboration network structure and team performance," *Science*, vol. 308, no. 5722, pp. 697–702, 2005.

[26] S. Fuger, R. Schimpf, J. Füller, and K. Hutter, "Network structure and user roles of a crowdsourcing community–the context of social innovations for a development project," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[27] S. Fuger, R. Schimpf, J. Füller, and K. Hutter, "User roles and team structures in a crowdsourcing community for international development–a social network perspective," *Information Technology for Development*, vol. 23, no. 3, pp. 438–462, 2017.

[28] J. Wortman, R. E. Lucas, and M. B. Donnellan, "Stability and change in the big five personality domains: Evidence from a longitudinal study of australians.," *Psychology and aging*, vol. 27, no. 4, p. 867, 2012.

[29] R. Tourangeau and T. Yan, "Sensitive questions in surveys.," *Psychological bulletin*, vol. 133, no. 5, p. 859, 2007.

[30] I. A. Junglas, N. A. Johnson, and C. Spitzmüller, "Personality traits and concern for privacy: an empirical study in the context of location-based services," *European Journal of Information Systems*, vol. 17, no. 4, pp. 387–402, 2008.

[31] P. S. Brenner and J. DeLamater, "Lies, damned lies, and survey self-reports? identity as a cause of measurement bias," *Social psychology quarterly*, vol. 79, no. 4, pp. 333–354, 2016.

[32] S. L. Kichuk and W. H. Wiesner, "The big five personality factors and team performance: implications for selecting successful product design teams," *Journal of Engineering and Technology Management*, vol. 14, no. 3, pp. 195–221, 1997.

[33] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260, ACM, 2002.

[34] D. F. Vincent, "The orgin and development of factor analysis," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 2, no. 2, pp. 107–117, 1953.

[35] R. Shields, "Cultural topology: The seven bridges of königsburg, 1736," *Theory, Culture & Society*, vol. 29, no. 4-5, pp. 43–57, 2012.

[36] L. Freeman, "The development of social network analysis," *A Study in the Sociology of Science*, 2004.

[37] L. C. Freeman, "The development of social network analysis–with an emphasis on recent events," *The SAGE handbook of social network analysis*, pp. 26–54, 2011.

[38] S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[39] F. Tonnies, *Community and society*. Courier Dover Publications, 2011.

[40] J. L. Moreno, E. S. Whitin, and H. H. Jennings, *Application of the group method to classification*. National committee on prisons and prison labor, 1932.

[41] J. L. Moreno, *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co, 1934.

[42] F. J. Roethlisberger and W. J. Dickson, *Management and the worker: an account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago, by FJ Roethlisberger and William J. Dickson, with the assistance and collaboration of Harold A. Wright*. Harvard Univ. Press, 1964.

[43] W. L. Warner and P. S. Lunt, *The social life of a modern community.* Yale University Press, 1941.

[44] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep south*. University of Chicago Press Chicago, 1941.

[45] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[46] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[47] E. Bonabeau, "Decisions 2.0: The power of collective intelligence," *MIT Sloan management review*, vol. 50, no. 2, pp. 45–52, 2009.

[48] T. Atlee and K. Mercer, *The First Little Book on Co-Intelligence*. The Co-Intelligence Institute, 1996.

[49] T. Atlee and G. Por, "Collective intelligence as a field of multi-disciplinary study and practice." community-intelligence. https://www.academia.edu/1981875/Collective_Intelligence_as_a_Field_of_Multi-disciplinary_Study_and_Practice. Accessed: 31-07-2020.

[50] P. Lévy, *Collective i/ntelligence*. Plenum/Harper Collins, 1997.

[51] S. M. Lee, D. L. Olson, and S. Trimi, "Co-innovation: convergenomics, collaboration, and co-creation for organizational values," *Management Decision*, vol. 50, no. 5, pp. 817–831, 2012.

[52] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York, NY, USA: Crown Publishing Group, 1 ed., 2008.

[53] E. Estellés-Arolas and F. González-Ladrón-de Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information science*, vol. 38, no. 2, pp. 189–200, 2012.

[54] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Communications of the ACM*, vol. 54, no. 4, pp. 86–96, 2011.

[55] T. Aitamurto, A. Leiponen, and R. Tee, "The promise of idea crowdsourcing–benefits, contexts, limitations," *Nokia Ideasproject White Paper*, 2011.

[56] A. Kulkarni, "The complexity of crowdsourcing: Theoretical problems in human computation," in *CHI Workshop on Crowdsourcing and Human Computation*, 2011.

[57] D. C. Brabham, "Crowdsourcing as a model for problem solving an introduction and cases," *Convergence: the international journal of research into new media technologies*, vol. 14, no. 1, pp. 75–90, 2008.

[58] M. K. Poetz and M. Schreier, "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?," *Journal of Product Innovation Management*, vol. 29, no. 2, pp. 245–256, 2012.

[59] M. d. Condorcet, "Essay on the application of analysis to the probability of majority decisions," *Paris: Imprimerie Royale*, 1785.

[60] E. Gödel, "Bach: An eternal golden braid," *Douglas Hofstadter*, 1979.

[61] G. Worthey, "Excerpts from hofstadters books," 2006.

[62] W.-E. Reif, "Biogeochemistry-biosphere-noosphere. the growth of the theoretical system of vladimir ivanovich vernadsky," 2001.

[63] H. G. Wells, *World brain*. Best Classic Books, 2013.

[64] P. Russell and B. Noetics, *The global brain*. Penny Price Productions, 1985.

[65] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[66] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528, ACM, 2003.

[67] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 341–349, ACM, 2002.

[68] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[69] W. Tatarkiewicz, *A history of six ideas*. M. Nijhoff, 1980.

[70] C. Parvanta, Y. Roth, and H. Keller, "Crowdsourcing 101 a few basics to make you the leader of the pack," *Health promotion practice*, vol. 14, no. 2, pp. 163–167, 2013.

[71] O. Alonso and M. Lease, "Crowdsourcing 101: putting the wsdm of crowds to work for you.," in *WSDM*, pp. 1–2, 2011.

[72] G. Kazai, "In search of quality in crowdsourcing for search engine evaluation," in *Advances in information retrieval*, pp. 165–176, Springer Berlin Heidelberg, 2011.

[73] Y. Zhao and Q. Zhu, "Exploring the motivation of participants in crowdsourcing contest," 2012.

[74] D. C. Brabham, "Moving the crowd at threadless: Motivations for participation in a crowdsourcing application," *Information, Communication & Society*, vol. 13, no. 8, pp. 1122–1145, 2010.

[75] E. Schenk and C. Guittard, "Crowdsourcing: What can be outsourced to the crowd, and why," in *Workshop on Open Source Innovation, Strasbourg, France*, 2009.

[76] M. Vuković, "Crowdsourcing for enterprises," in *Services-I, 2009 World Conference on*, pp. 686–692, IEEE, 2009.

[77] D. C. Brabham, "Crowdsourcing the public participation process for planning projects," *Planning Theory*, vol. 8, no. 3, pp. 242–262, 2009.

[78] F. Kleemann, G. G. Voß, and K. Rieder, "Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing," *Science, technology & innovation studies*, vol. 4, no. 1, pp. PP–5, 2008.

[79] Y. Pan and E. Blevis, "A survey of crowdsourcing as a means of collaboration and the implications of crowdsourcing for interaction design," in *Collaboration Technologies and Systems (CTS), 2011 International Conference on*, pp. 397–403, IEEE, 2011.

[80] A. N. Kittur, "Corralling crowd power: technical perspective," *Communications of the ACM*, vol. 58, no. 8, pp. 84–84, 2015.

[81] A. G. Yong and S. Pearce, "A beginners guide to factor analysis: Focusing on exploratory factor analysis," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 2, pp. 79–94, 2013.

[82] R. D. Froman, "Elements to consider in planning the use of factor analysis," *Southern Online Journal of Nursing Research*, vol. 2, no. 5, 2001.

[83] T. R. Hinkin, J. B. Tracey, and C. A. Enz, "Scale construction: Developing reliable and valid measurement instruments," *Journal of Hospitality & Tourism Research*, vol. 21, no. 1, pp. 100–120, 1997.

[84] D. D. Suhr, *Exploratory or confirmatory factor analysis?* SAS Institute Cary, 2006.

[85] A. B. Costello and J. W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical assessment, research & evaluation*, vol. 10, no. 7, pp. 1–9, 2005.

[86] P. Kline, "Factor analysis and personality theory," *European Journal of Personality*, vol. 1, no. 1, pp. 21–36, 1987.

[87] H. H. Harman, *Modern factor analysis*. University of Chicago Press, 1976.

[88] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.," *Psychological methods*, vol. 4, no. 3, p. 272, 1999.

[89] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and psychological measurement*, vol. 20, no. 1, pp. 141–151, 1960.

[90] R. Cattell, *The scientific use of factor analysis in behavioral and life sciences*. Springer Science & Business Media, 2012.

[91] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965.

[92] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*. Allyn & Bacon/Pearson Education, 2007.

[93] M. Odum, "Factor scores, structure and communality coefficients: A primer.," *Online Submission*, 2011.

[94] C. DiStefano, M. Zhu, and D. Mindrila, "Understanding and using factor scores: Considerations for the applied researcher," *Practical Assessment, Research & Evaluation*, vol. 14, no. 20, pp. 1–11, 2009.

[95] N. B. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of computer-mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[96] F.-Y. Wang, Y. Yuan, X. Wang, and R. Qin, "Societies 5.0: A new paradigm for computational social systems research," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 2–8, 2018.

[97] Facebook, Inc., "Facebook, Inc. Form 10-K." `https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm`. Accessed: 31-07-2020.

[98] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[99] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, pp. 604–632, Sept. 1999.

[100] B. S. Nandhini and J. Sheeba, "Online social network bullying detection using intelligence techniques," *Procedia Computer Science*, vol. 45, pp. 485–492, 2015.

[101] M. Dadvar and F. de Jong, "Cyberbullying detection: A step toward a safer internet yard," in *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, (New York, NY, USA), pp. 121–126, ACM, 2012.

[102] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, (Washington, DC, USA), pp. 71–80, IEEE Computer Society, 2012.

[103] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, (New York, NY, USA), pp. 620–627, ACM, 2009.

[104] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, "Detecting spam blogs: A machine learning approach," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pp. 1351–1356, AAAI Press, 2006.

[105] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Proceedings of the 24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy*, DBSec'10, (Berlin, Heidelberg), pp. 335–342, Springer-Verlag, 2010.

[106] V. Buzuloiu, M. Cuic, R. Beuran, H. Grecu, A. Drimbarean, P. Corcoran, and E. Steinberg, "Automated detection of pornographic images," June 15 2004. US Patent 6,751,348.

[107] Y. Kalish and G. Robins, "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure," *Social Networks*, vol. 28, no. 1, pp. 56–84, 2006.

[108] J. Krause, R. James, and D. Croft, "Personality in the context of social networks," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1560, pp. 4099–4106, 2010.

[109] A. E. Poropat, "A meta-analysis of the five-factor model of personality and academic performance.," *Psychological bulletin*, vol. 135, no. 2, p. 322, 2009.

[110] M. K. Mount, M. R. Barrick, and G. L. Stewart, "Five-factor model of personality and performance in jobs involving interpersonal interactions," *Human performance*, vol. 11, no. 2-3, pp. 145–165, 1998.

[111] D. J. Kuss and M. D. Griffiths, "Online social networking and addictiona review of the psychological literature," *International journal of environmental research and public health*, vol. 8, no. 9, pp. 3528–3552, 2011.

[112] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the national academy of sciences*, vol. 97, no. 21, pp. 11149–11152, 2000.

[113] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[114] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics and endemic states in complex networks," *Physical Review E*, vol. 63, no. 6, p. 066117, 2001.

[115] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, (New York, NY, USA), pp. 33–41, ACM, 2012.

[116] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical review E*, vol. 64, no. 2, p. 025102, 2001.

[117] M. Abufouda, "Community aliveness: Discovering interaction decay patterns in online social communities," *arXiv preprint arXiv:1707.04477*, 2017.

[118] G. Jäger, C. Hofer, M. Kapeller, and M. Füllsack, "Hidden early-warning signals in scale-free networks," *PloS one*, vol. 12, no. 12, p. e0189853, 2017.

[119] N. Duhan, A. Sharma, and K. K. Bhatia, "Page ranking algorithms: a survey," in *2009 IEEE International Advance Computing Conference*, pp. 1530–1537, IEEE, 2009.

[120] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," in *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, pp. 309–320, North-Holland Publishing Co., 2000.

[121] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[122] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[123] H. Hou, H. Kretschmer, and Z. Liu, "The structure of scientific collaboration networks in scientometrics," *Scientometrics*, vol. 75, no. 2, pp. 189–202, 2008.

[124] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link mining: models, algorithms, and applications*, pp. 337–357, Springer, 2010.

[125] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th international conference on World Wide Web*, pp. 835–844, ACM, 2007.

[126] V. Braitenberg and A. Schüz, *Anatomy of the cortex: Statistics and geometry.* Springer-Verlag Publishing, 1991.

[127] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[128] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, (New York, NY, USA), pp. 37–42, ACM, 2009.

[129] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[130] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the national academy of sciences*, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[131] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3, pp. 590–614, 2002.

[132] C. Gini, "Concentration and dependency ratios," *Rivista di Politica Economica*, vol. 87, pp. 769–792, 1997.

[133] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[134] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[135] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[136] M. E. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[137] P. W. Holland and S. Leinhardt, "Transitivity in structural models of small groups," *Comparative group studies*, vol. 2, no. 2, pp. 107–124, 1971.

[138] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[139] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[140] K. Choromański, M. Matuszak, and J. Mikisz, "Scale-free graph with preferential attachment and evolving internal vertex structure," *Journal of Statistical Physics*, vol. 151, no. 6, pp. 1175–1183, 2013.

[141] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel, I. A. van de Leemput, S. A. Levin, E. H. van Nes, M. Pascual, and J. Vandermeer, "Anticipating critical transitions," *Science*, vol. 338, no. 6105, pp. 344–348, 2012.

[142] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, p. 53, 2009.

[143] J. Zhou, M. Heckman, B. Reynolds, A. Carlson, and M. Bishop, "Modeling network intrusion detection alerts for correlation," *ACM Transactions on Information and System Security (TISSEC)*, vol. 10, no. 1, p. 4, 2007.

[144] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 211–220, ACM, 2009.

[145] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The Review of Economics and Statistics*, vol. 54, no. 3, pp. 306–316, 1972.

[146] R. Dunford, Q. Su, and E. Tamang, "The pareto principle," *The Plymouth Student Scientist*, vol. 7, no. 1, pp. 140–148, 2014.

[147] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[148] C.-C. Liu, T.-P. Liang, B. Rajagopalan, V. Sambamurthy, and J. C.-H. Wu, "Knowledge sharing as social exchange: evidence from a meta-analysis," *Pacific Asia Journal of the Association for Information Systems*, vol. 3, no. 4, 2012.

[149] H. Levene, "Contributions to probability and statistics," *Essays in honor of Harold Hotelling*, pp. 278–292, 1960.

[150] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[151] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European Conference on Computer Systems*, EuroSys '09, (New York, NY, USA), pp. 205–218, ACM, 2009.

[152] J. Guilford, "Personality mcgraw hill and company," *New York*, 1959.

[153] R. B. Cattell, *Personality and mood by questionnaire.* Jossey-Bass, 1973.

[154] E. Sundstrom, K. P. De Meuse, and D. Futrell, "Work teams: Applications and effectiveness.," *American psychologist*, vol. 45, no. 2, p. 120, 1990.

[155] J. R. Katzenbach and D. K. Smith, *The wisdom of teams: Creating the high-performance organization*. Harvard Business Review Press, 2015.

[156] N. Ale Ebrahim, S. Ahmed, and Z. Taha, "Virtual r&d teams in small and medium enterprises: A literature review," *Scientific Research and Essays*, vol. 4, no. 13, pp. 1575–1590, 2009.

[157] M. R. Barrick, G. L. Stewart, M. J. Neubert, and M. K. Mount, "Relating member ability and personality to work-team processes and team effectiveness.," *Journal of applied psychology*, vol. 83, no. 3, p. 377, 1998.

[158] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[159] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theory and applications*, vol. 69. World scientific, 2008.

[160] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[161] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[162] M. Omar, S.-L. Syed-Abdullah, and N. M. Hussin, "Analyzing personality types to predict team performance," in *Proceedings of the 2010 International Conference on Science and Social Research (CSSR)*, pp. 624–628, 2010.

[163] M. Omar and S.-L. Syed-Abdullah, "Identifying effective software engineering (se) team personality types composition using rough set approach," in *Information Technology (IT-Sim), 2010 International Symposium in*, vol. 3, pp. 1499–1503, IEEE, 2010.

[164] F. P. Morgeson, M. H. Reider, and M. A. Campion, "Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge," *Personnel psychology*, vol. 58, no. 3, pp. 583–611, 2005.

[165] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[166] D. M. Buss, "Evolutionary personality psychology," *Annual review of psychology*, vol. 42, no. 1, pp. 459–491, 1991.

[167] D. P. MacKinnon, *Introduction to statistical mediation analysis*. Routledge, 2008.

[168] M. T. Brannick, E. Salas, and C. W. Prince, *Team performance assessment and measurement: Theory, methods, and applications*. Psychology Press, 1997.