View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CORE

Wright State University

CORE Scholar

Browse all Theses and Dissertations

Theses and Dissertations

2019

Data-driven and Knowledge-Based Strategies for Realizing Crowd Wisdom on Social Media

Shreyansh Bhatt
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Computer Engineering Commons, and the Computer Sciences Commons

Repository Citation

Bhatt, Shreyansh, "Data-driven and Knowledge-Based Strategies for Realizing Crowd Wisdom on Social Media" (2019). *Browse all Theses and Dissertations*. 2297.
https://corescholar.libraries.wright.edu/etd_all/2297

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

DATA-DRIVEN AND KNOWLEDGE-BASED STRATEGIES FOR REALIZING CROWD WISDOM ON SOCIAL MEDIA

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

By

SHREYANSH BHATT
M.S., Wright State University, USA, 2015.
M.Tech, Dhirubhai Ambani Institute of Information and Communication Technology,
India, 2011.
B.E., Maharaja Sayajirao University of Baroda, India, 2010

2019
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

06-13-2019

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY SHREYANSH BHATT ENTITLED DATA-DRIVEN AND KNOWLEDGE DRIVEN STRATEGIES FOR REALIZING CROWD WISDOM ON SOCIAL MEDIA BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Dr. Amit Sheth
Dissertation Director

Dr. Michael Raymer
Director, Computer Science and
Engineering Ph.D. Program

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Committee on Final Examination:

Dr. Amit Sheth, Ph.D.

Dr. Keke Chen, Ph.D.

Dr. Krishnaprasad Thirunarayan, Ph.D.

Dr. Valerie Shalin, Ph.D.

Dr. Brandon Minnery, Ph.D.

Abstract

Bhatt, Shreyansh. Ph.D. Department of Computer Science and Engineering, Wright State University, 2019. Data-driven and Knowledge-Based Strategies for Realizing Crowd Wisdom on Social Media.

The wisdom of the crowd is a well-known example of collective intelligence wherein an aggregated judgment of a group of individuals is superior to that of an individual. The aggregated judgment is surprisingly accurate for predicting the outcome of a range of tasks from geopolitical forecasting to the stock price prediction. Recent research has shown that participants' previous performance data contributes to the identification of a subset of participants that can collectively predict an accurate outcome. In the absence of such performance data, researchers have explored the role of human-perceived diversity, i.e., whether a human considers a crowd as a diverse crowd, to assemble an intelligent crowd. In fact, diversity among participants and independent decision making are the two most important criteria for a crowd to provide an accurate aggregated judgment. However, perceived diversity based crowd selection does not scale. This dissertation explores whether we can infer the diversity and independence from user generated social network data to inform intelligent crowd selection.

This dissertation first provides a data-driven bottom-up diversity measure and shows that participant diversity can be inferred from social media data and that it can be used to perform diverse crowd selection. It then provides a multi-objective optimization based diverse crowd selection method using this measure. The results show that the diverse crowds significantly outperform both randomly selected and expert crowds. A top-down approach then provides explainable diversity measures to select such a diverse crowd.

The data-driven diversity measures do not utilize the social media profile and link information. Community detection using shared content and link information can both inform diverse crowd selection. However, the existing methods do not consider "contextual" similarity that could play a crucial role in identifying and characterizing contextual communities. This dissertation provides a state-of-the-art contextual similarity measure and a knowledge graph-enhanced community detection approach to select a diverse crowd as well as explain the domain-specific diversity that could affect crowd wisdom. It is shown that such a diverse crowd can accurately predict the outcome of real-world events. These results have implications for numerous domains that utilize aggregated judgments - from consumer reviews to econometrics, to geopolitical forecasting and intelligence analysis.

Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Ph.D. has been an amazing journey filled with a lot of learning experiences. I am grateful to all my friends, peers, family and a diverse set of mentors for their unconditional support.

First and foremost, I want to express my gratitude to my academic father, guru, advisor, mentor Dr. Amit Sheth. I remember several email exchanges, Skype conversations, and 45 minutes of in-person conversation to express my passion to work with and learn from him. If it was my persistence or sheer luck or his kindness or all of the above, he agreed to be my Ph.D. advisor. Dr. Sheth always made sure all his students are well funded throughout the Ph.D. and gave me ample opportunities to become an in-dependant researcher. Technical abilities are only a small fraction of learning I inherited from Dr. Sheth. Dr. Sheth made sure each one of the us are equipped with excellent soft skills - effective communication, team work, importance of asking right questions, identifying right opportunities to name a few. Dr. Sheth has invested immense amount of energy, emotions, sleep, and even hair color to make sure I am equipped with all these things including think big and aim high. I came out as a different person at the end of the process. I found a lifelong mentor in him and seek for his advise even after graduating from Ph.D. I can not be more grateful for his selfless support.

I would like to Thank Dr. Valerie Shalin for being source of support throughout of my Ph.D. I met Valerie (Yes. we call her Valerie and Yes. She likes it) in 2013 and if I could describe her in one word it would be “powersource”. Valerie has always motivated in every situation. While all my mentors and peers contributed in achieving excellent publications, it would not have been possible for me to get even a single paper without Valerie. Valerie has been up till the very last minute of the deadline, editing paper, to make sure the paper is in a right shape. However, that is just a small fraction of her contribution towards my Ph.D. Valerie always made me feel confident about myself and that I can achieve the best (of course I can with such a mentor on my side, but even with a minimal support). I would like to thank Dr. Keke Chen for providing excellent technical mentoring. Dr. Chen did not give up on me, even at times I gave up on one of the core work in the dissertation. I specifically remember him calling for a meeting when I did not show up because I kinda of gave on the work. Dr. Chen always pushed me to my limits in coming up with the best technical solution to the problem. I want to express my gratitude to Dr. Chen for such an amazing support. I am also grateful to Dr. Brandon Minnery for showing trust in me and letting me work on his projects. The projects that funded half of my Ph.D. Dr. Minnery's vision provided a unique and concrete application for the work proposed in this dissertation. In fact, it was his vision that drove all the approaches contributed as part of the this dissertation. It was so great to work with Dr. Minnery and Valerie that I waited for Friday afternoons every week to have our weekly project meetings. Dr. Krishnaprasad Thirunanrayan has been another mentor for me who provided an excellent technical guidance. Discussions with Dr. Prasad shaped my work in the best possible direction. His down to earth personality and happy to help attitude makes everyone comfortable to discuss all the technical aspects of the work. Thank you Dr. prasad for having your doors always open for me for valuable discussions.

I have also been blessed to have excellent mentors including advisors. I would like to thank my first mentor, Dr. Hemant Purohit who I met when I started working at Kno.e.sis. Hemant has a unique perspective for looking at a problem. I learnt the importance of attention of details from him. I could not have a pleasure to work with Dr. Pavan Kapanipathi and Dr. Delroy Cameron. However, his suggestions always helped my Ph.D. journey. I would like to express my gratitude to Dr. David Haglin for being an amazing mentor during my tenure at PNNL. I learnt a number of important aspects of communicating ideas from Dr. Haglin and asking right questions. I would like to Thank Dr. Jesse Weaver, Dr. Alessandro Morari, Dr. Daniel Chavarria, and Dr. Vito Castellana for providing an excellent technical mentoring. Finally, I would like to thank Mr. Yogesh Bhatt - my guru -

for his advises that shaped my academic and professional life.

I made amazing friends during my Ph.D. These friends cum mentors cum gossip buddies defined a way to enjoy ardous Ph.D. journey. I am always grateful to to you guys - Hemant, Pavan, Delroy, Sanjaya, Sujan, Sarasi, Pramod, Wenbo, Lu, Manas, Utkarshani, and Vinh. I had a great pleasure to work with all the people at Kno.e.sis., especially Swati. You have been an immense support and a great friend and I can't be more grateful to you.

I would also like to thank all the collaborators - Srikanth, Beth, Abhishek, Pavan, and Rakshith. It has been a great pleasure to work with all of you.

Finally, I would like to Thank my family for their unconditional support throughout this journey. I am thankful to god for blessing me with amazing parents. No words can express my gratitude for their support throughout this journey.

Dedicated to Satvik and Shruti

1 Introduction

Thesis Statement: *Cognitive diversity inferred from social media enables intelligent crowd selection for decision making without requiring analysis of past judgment-outcome data.*

One of the well-known examples of human collective intelligence is the “wisdom of crowds” (WoC) in which an aggregated judgment of a group of individuals is, in some cases, better than an average and even an expert individual. First described in the scientific literature over a century ago [1], WoC more recently vaulted into the public consciousness with the publication of James Surowiecki’s best-selling book of the same name [2]. The book describes the following as essential conditions to realize the wisdom of crowd,

- Domain Knowledge: The individuals are expected to have some degree of domain knowledge.
- Motivation: There should be a reward associated with making a correct judgment.
- Aggregation: There should be a proper mechanism to aggregate a crowd judgment.
- Diversity of Opinions: A crowd must consist of individuals bringing diverse information into the decision making.
- Independence: Individuals should not be influenced by each other.

Existing studies exploring wisdom of crowd, without using historical performance data of the individuals, are human driven and require assembling a crowd in one place, or involve complex monitoring techniques designed by humans. These techniques do not scale. On the other hand, social networks are increasingly used to share opinions regarding some real-world events. The significant contribution of this dissertation is to explore whether social network data can be used to infer diversity and influence so as to assemble a crowd that satisfies all the conditions defined above. This shall enable large scale wisdom of crowd studies in various domains ranging from geo-political forecasting to fantasy sports.

For this purpose, this dissertation first explores whether the social network data has enough signal to be used to infer diversity. Then it examines whether such diversity can inform intelligent crowd selection. An explainable diversity measure is also explored as one of the diversity inference techniques. We studied several attributes from social network data that can be used to infer diversity. Moreover, the different context of attributes can lead to different diverse crowds. For example, a group of individuals may be diverse based on their location attribute associated with their profile. Or location attribute can be considered in the context of housing-price, leading to a different diverse crowd than location attributes considered for example in the context of their interest in sports teams. This dissertation proposes and develops an algorithm that can automatically select important attributes and contexts that are relevant to a given network of individuals in forming a group.

We use the domain of Fantasy Sports to evaluate diverse crowd selection. Fantasy sports such as (and in the case of our study) Fantasy Premier League (FPL) soccer can serve as a useful domain for studying WoC effects precisely because participants (called fantasy team "owners") make frequent, on-the-record judgments about future real-life outcomes (i.e., a real-life player's performance in an upcoming game week)[3]. These judgments are then scored in accordance with the rules of the fantasy league. A detailed description of FPL rules is beyond the scope of this dissertation: here we provide only a brief overview. As in most fantasy sports, an FPL team owner selects a fantasy team composed of realworld players. An FPL team (or squad) consists of 15 players: two goalkeepers, five defenders, five midfielders, and three forwards. Points are accrued based on each player's real-life performance in each game week (e.g., goals scored, total number of minutes played, and number of assists). A fantasy team owner's selections are constrained by various factors such as budget caps, a limit on the maximum number of players that can be selected from any one real life team, and so on. While the selection of an initial squad occurs only once per season, a team owner may repeatedly make changes to his/her team roster throughout the season via player transfers. For each game week, a team owner chooses 11 of the 15 players on his/her roster to serve as the starting lineup. Points are assessed only for these 11 players. From among the starting lineup, an owner nominates one player to serve as the team captain. The captain earns double points for that week, so an owner is strongly motivated to select the captain whom he/she believes is most likely to perform well for that particular game week.

Following [3], we focused on the team owners' captain selection as our judgment of interest. Because captain selection is a categorical judgment, one cannot simply average the judgments of multiple owners. For this reason, we interpreted an owner's choice of a particular captain as a "vote" for that captain (similar to [3]). We then assembled virtual crowds of team owners and computed each crowd's captain choice as the captain receiving the most votes. The benefits of diversity still apply with voting because a diverse crowd's choice of captain reflects a variety of valid, performance related predictors, such as a captain's recent performance trends, recent injuries, opponent's strength and playing style.

1.0.1 Diversity and wisdom of crowds

In order for individual judges' errors to cancel one another out (thereby producing a collective judgment close to the truth), the spread of judgments must fall on both sides of (or bracket) the correct answer. Hence, the likelihood of bracketing is increased when individual judges are diverse. Diverse judges are likely to produce such an accurate judgement as they bring diverse viewpoints, knowledge, and perspectives in the decision making. This in turn leads judges to produce uncorrelated errors. Indeed, diversity is such a critical ingredient for WoC that crowds of diverse agents can, under certain conditions, outperform crowds of experts [4] [5].

Given the essential role of diversity in WoC, surprisingly little work has focused on measuring and manipulating diversity for the purpose of enhancing WoC effects (for an exception, see [6]). Perhaps this is because diversity - which has been formally modeled in domains ranging from machine learning [7] to genomics [8] to biological ecosystems [9] - remains an “analytically neglected” [10] concept with respect to human social systems. This is an important gap because if one were able to measure crowd diversity *a priori*, then one could purposefully select crowd members who possess a broad range of information and analytic approaches. Moreover, a validated method for quantifying diversity would empower researchers to explore additional questions concerning the role of diversity in collective intelligence, such as the relative importance of (and tradeoffs between) diversity and expertise as factors for crowd selection.

This dissertation first explores whether we can infer diversity from openly available social media communications using data-driven methodologies. It then presents a knowledge driven methodology that complements social media data with real-world crowd sourced knowledge to explore a better and moreover, explainable, crowd selection technique. Social influence may trigger individuals to revise their estimates, which can have a substantial impact on the statistical wisdom of crowd effect [11]. Hence, the knowledge driven methodology also considers influence captured by retweet/follower/friends relationships to form a crowd of judges that are diverse as well as independent.

1.1 Data-driven diversity quantification

We discuss data-driven diversity quantification using bottom-up (word embedding based) and top-down (captain-selection strategy based) user representation techniques.

1.1.1 Word embedding for diversity representation

While multiple formal definitions of diversity have been proposed (see [10] for a review), we chose as our starting point a simple measure based on the semantic distance between user content (text). Specifically, we measured the distance between crowd members (i.e., Twitter users) by applying a popular word embedding technique, Word2Vec [12]. Word2Vec represents each user within a high dimensional semantic vector space such that a measure of crowd diversity can be computed based on the distance between users within this space. The farther apart users are, the more diverse they are. A weak but statistically significant result indicated that diverse crowds, formed using this measure, are likely to produce a better judgement than crowds generated at random. Next, we explored crowd selection using this user representation technique.

1.1.2 Word embedding, clustering, and multi-view objective optimization based crowd selection

We propose a diverse crowd selection approach (SmartCrowd) based on social media posts (tweets). Each user is represented by the collection of their FPL tweets; the diversity of users is reflected in the topic and other latent communication patterns between their tweet collections. We adopt word2vec [12] to summarize a user's set of tweets, generating one equal-length summary word vector for each participant and then clustering these vectors to derive user clusters.[1] With the summary word2vec vector for each user, we cluster

[1] Other text summarization methods (e.g. TF-IDF and LDA [13] for topic extraction and summarization) might also work.

their vectors to generate user clusters. Multiple clustering strategies have been tested here, such as cosine distance and Euclidean distance measures, single-view spectral clustering and multi-view clustering. The best strategy is multi-view clustering that synthesizes views based on both cosine distance and Euclidean distance. Finally, to select optimal representatives from the clusters to compose the final crowd, we employ a multi-view objective optimization method using both distance measures as the objectives.

The crowds selected with our SmartCrowd approach beat a random crowd 85% of the time and outperformed individual participants 93% of the time. We also compared our approach to the Goldstein et al. [3] method, which forms crowds based on users' "expertise" derived from the performance history in the past seasons. The crowds generated with SmartCrowd outperformed the expert crowds consisting of the top-10% experts and the top-20% experts, and did only slightly worse than crowds of the top-2% experts.

A bottom-up diversity measure combined with crowd selection can produce crowds that achieved significantly better wisdom score than crowds generated at random and even experts. However, these methods do not explain the kind of diversity that plays a role in crowd selection. Next, we developed a more explainable, top-down, diversity measure.

1.1.3 Top-down diversity quantification

We explore whether top-down diversity quantification can help assemble an *intelligent* crowd. We define diversity in terms of the solution strategies employed by a participant for generating a prediction. Specifically, we hypothesize that diverse solution strategies lead to a more robust aggregated crowd prediction, where solution strategies are inferred using participants' social media posts. We provide real-world evidence that such diversity can help achieve an accurate prediction. We first characterize each participant according to whether his/her tweets refer to a particular strategy by classifying individual tweets. Using a binomial test-based participant categorization, we then identify a set of participants employing *similar* solution strategies. Finally, we form a diverse *virtual* crowd by selecting participants from each category.

We found that a diverse crowd determined by strategy is likely to perform better in the FPL captain prediction task than 90% of the individual participants. We also compared a diverse crowd with a randomly selected crowd of comparable size and found that a diverse crowd is 63% likely to outperform a randomly selected crowd. Crowds based on diverse strategies also perform favorably relative to standard word2vec methods for clustering users.

To explain the diversity in captain selection strategies and its effect on captain selection, we used a domain specific knowledge graph extracted from DBpedia[14]. The extracted knowledge graph is a concept hierarchy where a parent concept subsumes child concepts. To explain diversity, we mapped the keyword features used in classification to the knowledge graph and investigated the parent concepts that maximally subsume these keywords. We found that features identifying both - Popular choice and Differential choice tweets mapped to two parent soccer players who happened to be the top performers in terms of scoring FPL points. This supports the claim that diverse strategies ensure that a selected captain is effective from both perspectives.

1.2 Knowledge-driven diversity quantification

These data-driven methods use only tweet content for clustering, missing an important attribute for the wisdom of crowd–influence. Social influence can affect crowd wisdom. For social network, we have the potential influence information available in form of retweet or follower/followee relationship. Hence, we developed a method that clusters users based on their shared content as well as “link” between these users. Such a group detection is solved as community detection in node attributed networks where nodes are users, node attribute can be their tweet content, and link can be reweets.

1.2.1 Knowledge graph enhanced community detection and crowd selection

“Does interest in sports or music form conversational communities among participants?” Recent approaches model such problems as community detection and characterization. These approaches report both state-of-the-art community detection accuracy and effective community characterization with node attributes driving community detection[15][16]. These approaches increase edge weights between nodes belonging to the same community if these nodes share similar node attribute values. While such techniques detect whether communities form around the *particular* sports teams or music bands referenced explicitly, they fall short on identifying whether communities are formed from participants’ *general* interest in sports or music. Such problems require meaning-oriented community characterization with an assessment of accuracy that combines network nodes, edges, and node attributes. Instead of relying on apparent attribute relations, i.e., exact matching for nominal attributes and Euclidean distance for numeric attributes, we seek contextual relations between attribute values. The resulting meaningful community detection is also crucial for applications such as network visualization [17] and online-marketing[18].

Consider the friendship network of participants shown in Figure 1.1 with the available node attributes expressed as the city in which a participant lives. The existing approach to community detection on such a network considers“Austin”, “Dallas”, and “Houston” as different attribute values [15][16], missing the important subsuming relationship (i.e., they are in the same state). Considering such relationships can improve community characterization. Moreover, detecting such relationships provides a basis for updating edge weights.

We explore the use of domain-specific knowledge graphs to find such contextually meaningful attribute relationships. Domain-specific *hierarchical* knowledge graphs (HKGs) provide particularly relevant real-world clustering information. The domain-specific HKG in Figure 1.2 indicates that all states of United States are subsumed by “States in United

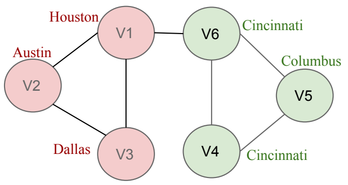


Figure 1.1: Friendship network with nodes representing user, edges representing friendship, and node attribute as the home-city.

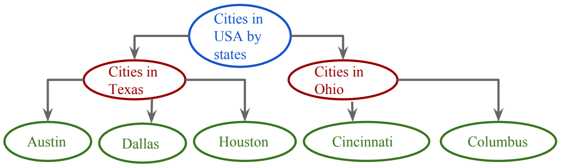


Figure 1.2: Hierarchical knowledge graph for USA geo-location.

States". The decomposition starting from each concept of such an HKG provides a *context*. E.g., all the concepts subsumed by "Cities in Ohio" along with "Cities in Ohio" provides a *context* "Ohio". Such knowledge graphs can be generated automatically with demonstrated benefit to applications such as personalization [19]. HKGs provide complementary real-world information regarding communities or clusters that may not be explicit in the network but are nevertheless useful in finding and characterizing communities. However, incorporating domain-specific HKGs in community detection raises three key challenges. 1. There is no clear measure for computing node similarity using an HKG characterization. For example, at the city level, "Austin", "Dallas", and "Houston" are different, although they are similar in the context of "Cities in Texas". Additionally, we need to determine the optimal context characterizing the community structures. E.g., in Figure 1.1, "Cities in Texas" characterizes Community 1(V1, V2, and V3). 2. Optimal Context reflects multiple factors. Moving up the hierarchy towards the root, we obtain a more general context subsuming lower level attribute values. However, the generalization disguises the differences between attribute values, potentially losing details that distinguish node groups. 3. Optimizing context generalization should coordinate with the discovery of topological structure, which should reflect computed community contexts.

We develop an algorithm that iteratively optimizes two tasks: (i) Optimal community label assignment while keeping the community context unchanged, (ii) Optimal community context assignment while keeping the community labels constant. For the first task, we propose a contextual similarity measure for defining node pair similarities to capture community contexts. We employ a widely used community label assignment algorithm, the Louvain community detection algorithm [20], which finds community labels for nodes using modularity maximization.

For the second task, we find a concept generalization scheme that balances two criteria: 1. *Informativeness*, which is essentially the specificity of a concept in a hierarchical knowledge graph. The lower the concept is in the hierarchy, the more specific information the generalization preserves and 2. *Purity* which is the difference between the number of nodes subsumed by a concept of a given community and neighboring communities.

Our framework has three unique features: (i) It can accept any predefined domain-specific hierarchies for any attributes (numeric or nominal), together with a topological network structure (i.e., nodes and edges). (ii) The algorithm does not assume *a priori* that a domain must correlate with the communities we want to discover. Instead, it will quantify the relationship between a certain domain and communities. If one exists, the algorithm will progressively find it. (iii) It allows us to analyze competing contexts on the same attributes. For example, the location attributes may have multiple different context hierarchies: one based on the geographical concepts, another on housing markets, and the third on household income levels.

As the resulting algorithm can assign more appropriate edge weights than using only attribute values, the algorithm can facilitate the discovery of an accurate community structure. We evaluated community detection accuracy on four real-world networks and five baseline community detection algorithms. The proposed algorithm improves community detection accuracy by nearly 20%. We also evaluated the accuracy of community structure characterization and found that the proposed approach was able to discover correct underlying community "types" for all four datasets while two baseline methods [15][16] failed to characterize communities for at least two datasets. We also demonstrate that contextual community detection and characterization effectively mediates the representation of the original data for two practical problems: Harassment in online social networks and diversity in crowd sampling.

1.3 Organization of the Dissertation

The rest of the dissertation is organized as follows: Chapter 2 differentiates the current research from existing work and positions it from the wisdom of crowds perspective. Chapter **??** describes the dataset and data-collection processes. Chapter 4 provides the details on word embedding-based user characterization. Chapter 5 details the crowd selection approach. Chapter 6 describes the top-down diversity characterization and diversity explanation approach. Chapter 9 introduces the knowledge graph enhanced crowd selection algorithm and provides details on crowd selection and diversity explanation. Chapter 7 identifies possible research directions with Chapter 8 concluding the dissertation.

2 Background and Related Work

In this chapter, we review the existng related literature that covers data-driven wisdom of crowd and community detection.

2.1 Wisdom of crowds

Wisdom of crowd is observed in many real-world applications ranging from guessing a weight of an ox to the web-page ranking using Pagerank[2]. Several research studies explore the domains in which such an effect exists [21][22][23]. Some of the recent studies also explore the conditions in which such an effect may not exist for certain domains [24]. These studies identify the importance of studying the conditions that determine the wisdom of crowds effect.

2.2 Diversity and wisdom of crowds

With the apparent benefits from studying the wisdom of crowds, several studies have explored the effect of diversity and crowd size on wisdom of crowd. This dissertation can be understood in the context of several related research veins. For instance, one implication of our findings is that, by recruiting maximally diverse group members, one might minimize the group size needed to form an accurate judgment without forfeiting the benefits of a larger crowd. In this sense, our work can be situated within a larger body of research that seeks to develop methods for identifying smaller, wiser sub-crowds within larger crowds - most notably recent work by Goldstein et al. [3], whose methods for analyzing fantasy sports data provide the groundwork for the present study (see also [25]). Minimizing crowd size is of practical importance, because many tasks that have been shown to benefit from WoC methods, such as geopolitical forecasting [26], are time and labor-intensive, otherwise depending on rare expertise. Thus it is often desirable to reduce the number of group members tasked with producing a judgment.

A large body of work deals with finding a virtual small and smart crowd from a large set of participants. The traditional wisdom of crowd research has explored the correlation between the diversity and accuracy of the collective judgment in crowd selection [22]. These experiments solicit participants to indicate diversity explicitly. Teng et al. asked participants to define their similarity to other members of groups [27]. They found that more diverse teams were more creative than less diverse teams. Thus *explicitly* indicated participant diversity plays a vital role in generating a smart crowd. In contrast, we *infer* diversity from online social media data to build a smart crowd and compare it with other crowd selection strategies.

A rich organizational psychology literature examines the impact of diversity on group performance. A key research concept in this field is social category diversity, i.e., diversity defined by surface characteristics such as race or gender [28]. This form of diversity – which is more closely aligned with the popular understanding of the term – is not the focus of our study; however, we expect that such “superficial" forms of diversity do in fact correlate with deeper differences in mental models derived from a lifetime of divergent experiences. While we expect that social category diversity would be less likely to contribute to WoC effects in narrowly defined tasks having less of a cultural dimension (such as fantasy sports), the literature suggests that such differences can have important impacts in small group settings in a variety of domains.

This work is also related to a broader research thread within computational social science aimed at modeling self-organizing structures, such as communities and cliques, using social media data [29]. We also developed methods to analyze community structure in this research. Our approach is consistent (in a methodological sense) with diversity modeling techniques used in other disciplines, like ecological science, where diversity measures are often based on how many individuals from various discrete categories (e.g. species) are represented within a system [30] (with a “species" being analogous in the social sense to a well-defined clique). In general, the role of social influence in modulating heterogeneity among group members’ beliefs is a well-explored topic that has previously been investigated in the context of WoC [11]. Twitter, in conjunction with fantasy sports, is a ripe medium/domain for further examining these effects.

Other research explores the correlations between content diversity and crowd wisdom. For example, Hong et al. showed that opinion diversity derived from participant-generated content is positively correlated with crowd performance[31]. However, they did not explore a crowd selection strategy, and used cosine similarity between traditional word vectored representations to compute participant diversity. This word vectored representation neglects contextual similarity [12] especially for short social media texts. Robert Jr. et al. explored diverse crowd formation for the generation of quality Wikipedia articles[32]. They computed crowd diversity using Wikipedia authors’ stated topics of interest and showed that such diversity could help to form smart crowds. However, they do not explore crowd selection based on such diversity. Moreover, they used explicit participant topic characterization data in their diversity measure instead of inferring diversity from raw social media posts. Several predictive analysis problems, such as the one discussed in this dissertation, do not provide explicit participant indications of topics. For example, we do not have participants’ FPL specific topic affinity listed on the FPL website. Instead, we use social media to infer diversity. Moreover, Ren et al. reported that communication variables also play a key role in defining diverse/smart crowd along with the topic of interests[33]. Word2vec word vector generation captures such latent communication patterns along with topic-specific words. As a result, SmartCrowd had such inputs to sample “wiser” crowds.

The results of this dissertation are also interesting for their application to the Fantasy Sports. Several research studies explore team selection for maximizing reward in a season-long fantasy tournament[34][35][36]. Some studies also explore the maximum number of wins a player will have in sports [37][38]. More recent studies explore team selection for daily fantasy sports[39][40]. These approaches work on sports player data collected by a user. The model selecting a team or predicting a successful player considers specific features. The use of sophisticated features can benefit performance with these models. However, the collection of such broad data can be challenging, e.g., each injury report of a player, player dynamics, player leadership skills, gambling specific knowledge related to Fantasy Sports. Moreover, we consider a different problem from the Fantasy Sports perspective, i.e., a captain selection within a team. Unlike existing approaches, our approach exploits crowd wisdom as a substitute for sports player-specific information.

2.3 Knowledge driven diversity quantification using community detection

Another important contribution of this dissertation is a novel community detection and characterization algorithm. Community detection in node attributed graph has a rich history of work due to its applicability in graph visualization, understanding graph data, link prediction, and graph summarization. Table 2.1 provides a quick summary of the community detection and characterization approaches and position the current work.

Bothorel et al. provides a good summary of community detection methods that incorporate graph attributes [41]. Among the recent approaches, Wang *et al.* works for non-text real-valued node attributed graphs unlike several others [41]. In an approach proposed by Qin *et al.*, link and node attributes are combined at different rates during community detection for improved community detection accuracy. Contrary to the proposed approach, this work does not focus on characterizing community structures. CPCD [42] and UNCut[43] used in the comparison also focus on identifying communities than characterizing these

Table 2.1: Comparison of methods for community detection in node-attributed graphs. NAG: Node attributed graph clustering, SC: Community structure characterization, Non-text: Allows non-textual attributes. Topics: Community detection based on topics related to node attributes.

Method Class	NAG	SC	Non-text	Topics
Clustering [41]	✓	✖	✓	✖
LDA [44][45][46]	✓	✓	✖	✖
SI [15]	✓	✓	✓	✖
JCDC [16]	✓	✓	✓	✖
KDComm	✓	✓	✓	✓

communities.

Several generative models also detect communities and provide information regarding the labels that nodes in a community have in common[44][45][46]. Among recent approaches, He *et al.* finds communities by jointly optimizing over node attributes and links using a generative model [47], similar to Wang *et al.*[48]. These approaches characterize a community structure by revealing latent topics within the textual node attributes of a community. They do not work for non-textual node attributes nor do they find communities along given set of topics. The latent community description is less informative compared to the community descriptions identified by proposed approach.

Community detection in node attributed graphs from Zhang *et al.* [16] and Newman *et al.* [15] inspires our own method. Such methods find communities based on edges and then refine these communities, i.e., by changing edge weights, based on node attribute values. However, Zhang *et al.* and Newman *et al.* do not make use of attribute semantics as we suggest here. Hence, these approaches can not identify communities for different domains as required by the application discussed in section **??**.

Our belief in external knowledge-enhancing community detection in a network is rooted in past work that demonstrated the prominent role of semantics in social network analysis. For example, El *et al.* combines social data with data semantics to create a semantic social network [49]. Pool *et al.* argues that a knowledge graph-based description should inform community structures based on user interests and beliefs [50]. A survey on a semantic social network by Ereto *et al.* summarizes the use of semantics in social network analysis[51]. Palma *et al.* focuses on predicting drug targeted Interaction using semantic similarity and edge partitioning [52]. These approaches integrate the social network links with existing ontologies for generic social network analysis. However, community detection on such combined graphs can be biased with one graph (social graph or ontology) being larger than the other. Wang et al. reported that real-world knowledge represented in knowledge graphs could improve document clustering [53]. Nevertheless, they did not focus on community detection with links connecting nodes and attributes identifying nodes.

3 Data collection

Table 3.1: Description of Twitter Dataset.

Tweet Type	Total, Median*	Description
All Tweets	2M, 2529	Tweets crawled from users' timelines
Soccer Tweets	1M, 591	Tweets with soccer keywords
FPL Tweets	90K, 13	Tweets with keywords (OfficialFPL, FPL, Fantasy Premier League)

*Median tweets per user

Figure 3.1 outlines our tweet and FPL captain pick collection procedure. We used the Twitter streaming API to collect tweets containing FPL related keywords (Table 3.1) over the time period of August-November 2016 corresponding to the first four months of the 2016-17 English Premier League season.

From these tweets, we extracted the names of the associated Twitter users. To obtain captain pick data, we matched these Twitter users using their names on Twitter with names on the official FPL website [1], on which registered users post their team lineup, including weekly captain picks. Although we could not be 100% certain of a match, we reasoned that an individual tweeting repeatedly about FPL is likely to be the same person as an FPL website user having the same first and last name. To further reduce uncertainty in our matching process, we eliminated from our analyses any non-unique names and their associated data, i.e., names appearing more than once on either Twitter or the FPL website. We then scraped the FPL website to obtain captain picks for these matched users for each

[1] fantasy.premierleague.com

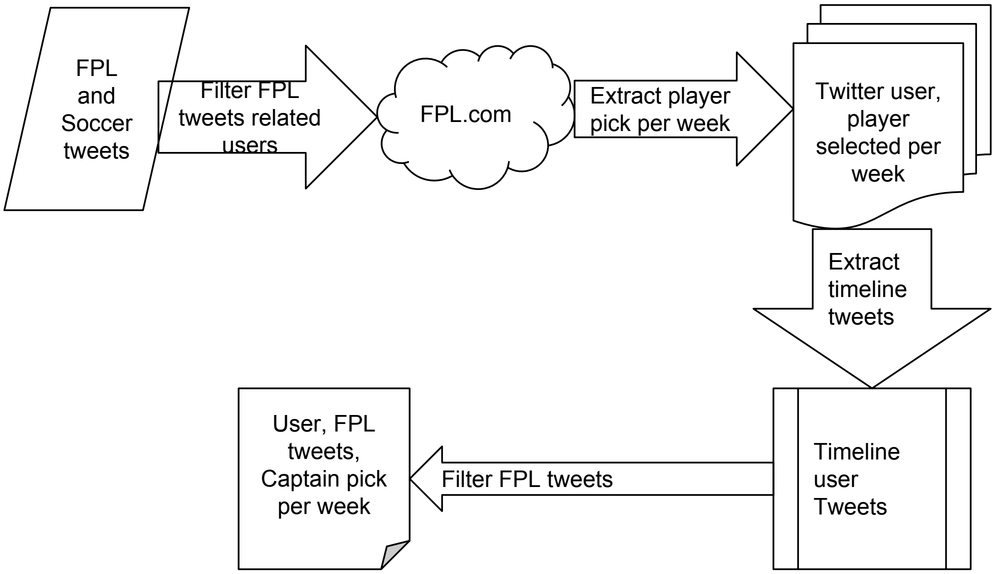

FPL and Soccer tweets
Filter FPL tweets related users
FPL.com
Extract player pick per week
Twitter user, player selected per week
Extract timeline tweets
Timeline user Tweets
Filter FPL tweets
User, FPL tweets, Captain pick per week

Figure 3.1: Data Collection.

of the 25 game weeks that had occurred from the time of our analysis, along with each captain's score for each game week. Because Twitter's streaming API captures only a few tweets for each user, we crawled users' publicly available Twitter timelines to collect additional tweets.

Table 3.1 summarizes the dataset. Column 2 shows the total number of tweets and the median number of tweets per user. Our analysis identified 912 users who tweeted at least five times about FPL. Scraping the timelines of these users resulted in a total of 2M tweets, of which about 1M contained at least one soccer related keyword and about 90K contained at least one FPL related keyword. These 90K tweets form the basis for our group diversity measure.

4 Word embedding for diversity representation

In this chapter, we describe the detailed approach for our bottom-up, data driven user characterization to assemble diverse crowds. We used a word embedding approach to characterize a user who is essentially represented using user generated text in the form of an n-dimensional vector.

A word vectored text representation improves and simplifies Natural Language Processing (NLP) applications such as search, language translation, and information extraction [12][54]. Here, we intend to capture the topical and conversational diversity among these participants. A word vector captures a context of a word, where a context is identified by the surrounding words. Thus word vectors can be used to capture the latent topic as well as the communication pattern of a user's tweet. Specifically, given preceding words, such word vectors predict a probability distribution over the "next" word. Of the available methods, skip-grams represent a word as a vector (known as word2vec) and provide state-of-the-art performance for word similarity tasks [12]. These word-based vectors explicitly encode linguistic regularities and patterns as linear translations. For example, the result of a vector calculation vec("Madrid") - vec("Spain") + vec("France") is closer to vec("Paris") than to any other word vector [55][54]. Hence, word2vec has been used to represent the similarity of social media posts, especially tweets, by averaging tweet word vectors [56].

The next section describes an approach to generate word embeddings and characterize a Twitter user using such word embedding.

4.1 Word embedding generation

Figure 4.1 shows the architecture for our method.

Word2Vec quantifies the semantic similarity between words and has been used in various natural language processing tasks such as sentence completion [57], POS tagging [58], and Twitter hashtag prediction [59]. Word2Vec applies to social data as well, and has been shown to work well in representing short text sentences, social media messages (tweets), and in identifying similar Twitter users [60] [61].

We trained our Word2Vec model with 2M tweets collected as described in Section 2 (Table 1). As a pre-processing step, we converted each word of a tweet to its lowercase, and removed stop words and URLs. We used a Skip-gram model with a negative sampling for training. The negative sampling rate 10 works well with medium sized datasets [12]. Because most of our tweets were short (on average eight words), we used a context window of three so that the training process considered three words to the left and three to the right of the word in question. We ignored words that appeared less than three times by setting min_word_frequency = 3.

We computed user diversity based on FPL related tweets, i.e. tweets with at least one FPL keyword (Table 3.1). After we applied similar pre-processing for the FPL related tweets, we used the trained Word2Vec model to transform each word of a tweet to a 300-dimensional vector. To produce a single vector representation for each user, we aggregated each word vector from each tweet and averaged the vectors, because the average of word vectors has been shown to represent a short sentence effectively, that is, a tweet [60], as well as a set of tweets [61]. Equation 1 formalizes this process.

Let $t = \{w_1, w_2, w_3, \ldots, w_p\}$, where t is a tweet and w_i is a word vector representation of a tweet word. Let Twitter user $U = \{t_1, t_2, \ldots, t_l\}$ where each t_i is a tweet for a user U,

l is total tweets of a user, p is total words per tweet. Then U can be rewritten as a collection of words,

$$U = \{w_1, w_2, \ldots, w_p, w_{p+1}, w_{p+2}, \ldots, w_n\}.$$

For each user we then define a vector, UV, which locates the user within a 300-dimensional semantic vector space, such that users who are close together in space are deemed similar. We define UV as,

$$UV = \frac{\sum_1^n w_i}{n} \tag{4.1}$$

This vector representation is then used to compute a diversity value for any group of two or more users. We use average pairwise cosine distance as our diversity measure of a group. For a group $G = \{UV_1, UV_2, \ldots, UV_n\}$, the diversity value of G is computed by Equation 2,

$$DG = \frac{\sum_{i,j\epsilon n} COS(UV_i, UV_j)}{\binom{n}{2}} \tag{4.2}$$

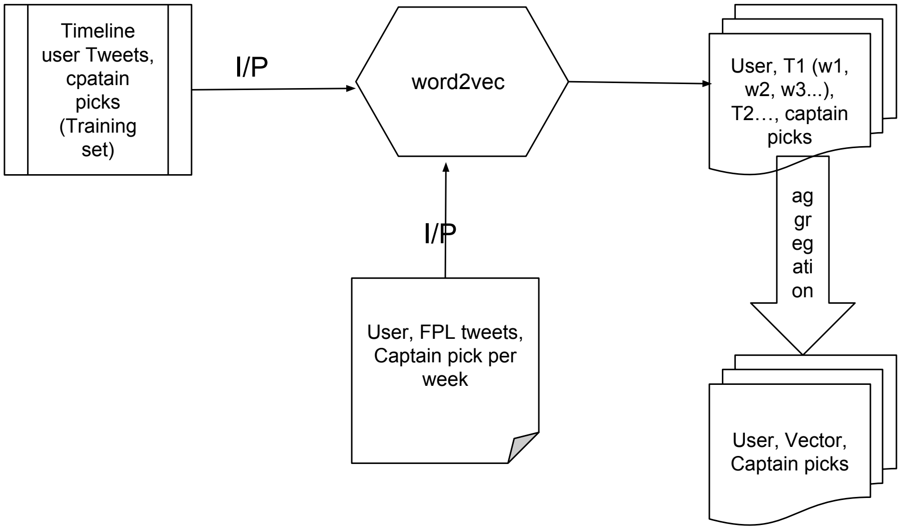

Timeline user Tweets, cpatain picks (Training set)
I/P
word2vec
User, T1 (w1, w2, w3...), T2…, captain picks
I/P
ag gr eg ati on
User, FPL tweets, Captain pick per week
User, Vector, Captain picks

Figure 4.1: Method for Computing Diversity.

4.2 Results

To avoid confusion, we hereafter refer to crowds as groups. Virtual groups of various sizes were composed as follows. Because the total number of unique groups was prohibitively large for the group sizes examined here, we chose a sampling process for constructing groups. For each group of size n, we first generated 5000 unique groups selected randomly from 912 total users. We then computed the diversity of each of these 5000 groups as described earlier. To ensure that the 5000 randomly generated groups were representative of the total universe of potential groups in terms of their diversity values, we repeated this process 100 times for each group size n. We then computed (for each of the 100 runs) the difference in average diversity values between the top 10% of the most diverse (500) and bottom 10% of the least diverse groups (500). We then ranked all 100 runs based on this difference and selected the run with the median difference.

For each group, we measured a group's “wisdom" – i.e., the score obtained by the group's “elected" captain in a given game week (hereafter referred to as the group's “wisdom score"). For each group $G = \{U_1, U_2, \ldots, U_n\}$, we generated C where $C = \{c_1, c_2, \ldots, c_n\}$ and c_i = {captain picked by u_i}. Equation 3 formulates the group wisdom score as,

$$GS = \frac{\sum_1^{25} Mod(C_i)}{25} \tag{4.3}$$

Here, $Mod(C_i)$ represents the score of the individual captain receiving the most votes from the group in the i^{th} game week. In cases where there was a non-unique mode – i.e., a tie – we used a tie-breaker strategy that selected a mode randomly from the set of non-unique modes. A group's wisdom score was then computed as the average of its scores over all 25 game weeks considered in our analysis.

In addition to a group's diversity and wisdom score, we measured its diversity in *judgements* (captain picks). We defined a group's pick diversity (PickDiversity) as the total number of unique picks divided by the total number of picks. For each $G = \{U_1, U_2, \ldots, U_n\}$ and their corresponding captain picks $C = \{c_1, c_2, \ldots, c_n\}$,

$$PickDiversity = \frac{Unique(C)}{n} \tag{4.4}$$

Here, Unique(C) is the number of unique captains picked by a Group G.

We found that most of the groups of size 10 or larger had at least two group members who agreed on a captain pick. This was also reflected by average PickDiversity <0.5 for groups size $\geq$ 10. In contrast, smaller groups (sizes 3 - 6) often ended up picking unique captains (PickDiversity $\geq$ 0.9), i.e., in most cases, we did not observe any captain pick agreement. To avoid this, and to ensure we could compute a meaningful wisdom score for our small groups, we employed a modified approach for generating small groups. For each group size n (3-6), we generated 5000 unique groups in which *all* group members picked the *same* captain. We reasoned that this approach would not interfere with our ability to

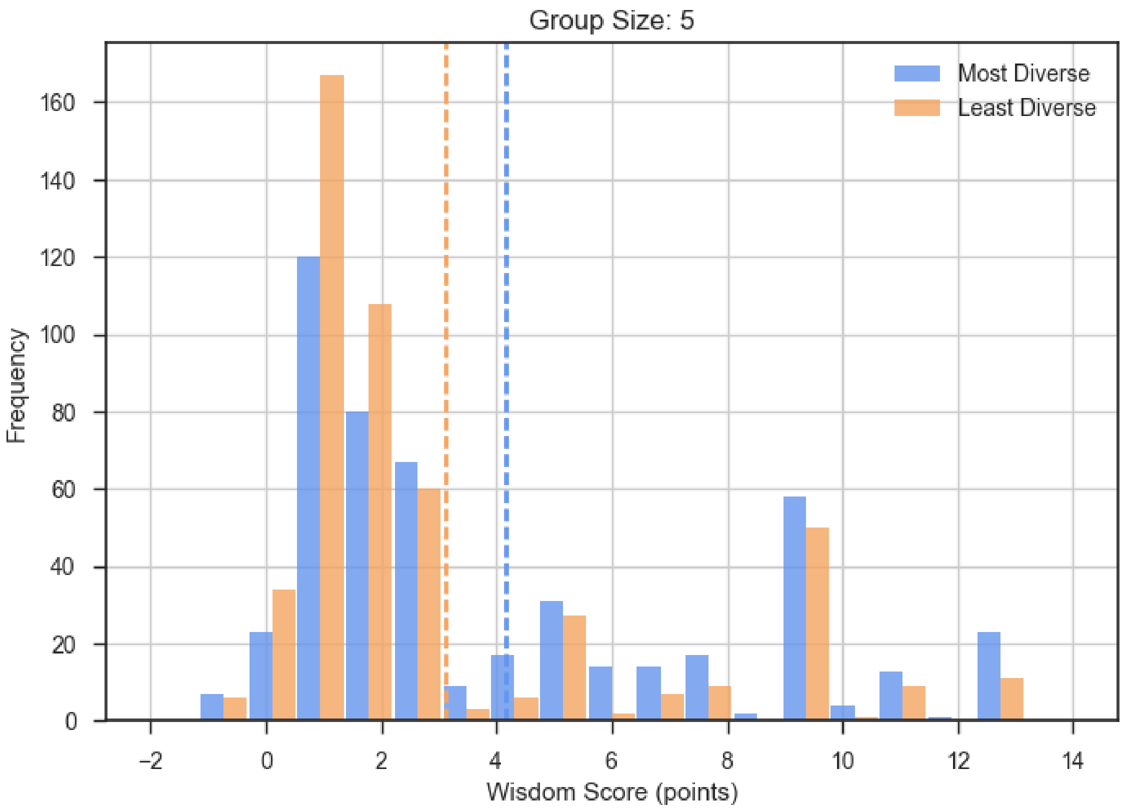


Figure 4.2: **Example of Wisdom Score Distributions for Group Size 5 (Most Diverse vs Least Diverse Groups)**. MD groups outscored LD groups by a statistically significant margin (p <0.005; Mann-Whitney test). Dotted lines are medians. Distributions for additional small group sizes are summarized in Figure 4.

detect a relationship between group diversity and wisdom score for small groups, because one would still expect that a unanimous pick by a diverse group would perform better on average than a unanimous pick by a less diverse group.

We first tested our main hypothesis that semantics-based diversity measures can predict WoC effects. Focusing initially on small groups (sizes 3-6), we compared the wisdom scores of the 500 most diverse (MD) groups to those of the 500 least diverse (LD) groups. MD and LD groups were selected by ranking all 5000 unique groups (of size n) according to their diversity and then choosing the top/bottom 500 groups, respectively. Figure 3 shows an example of the wisdom score distributions for MD and LD groups of size 5. Dotted lines indicate medians. We computed similar distributions for group sizes 3, 4 and 6. Figure 4

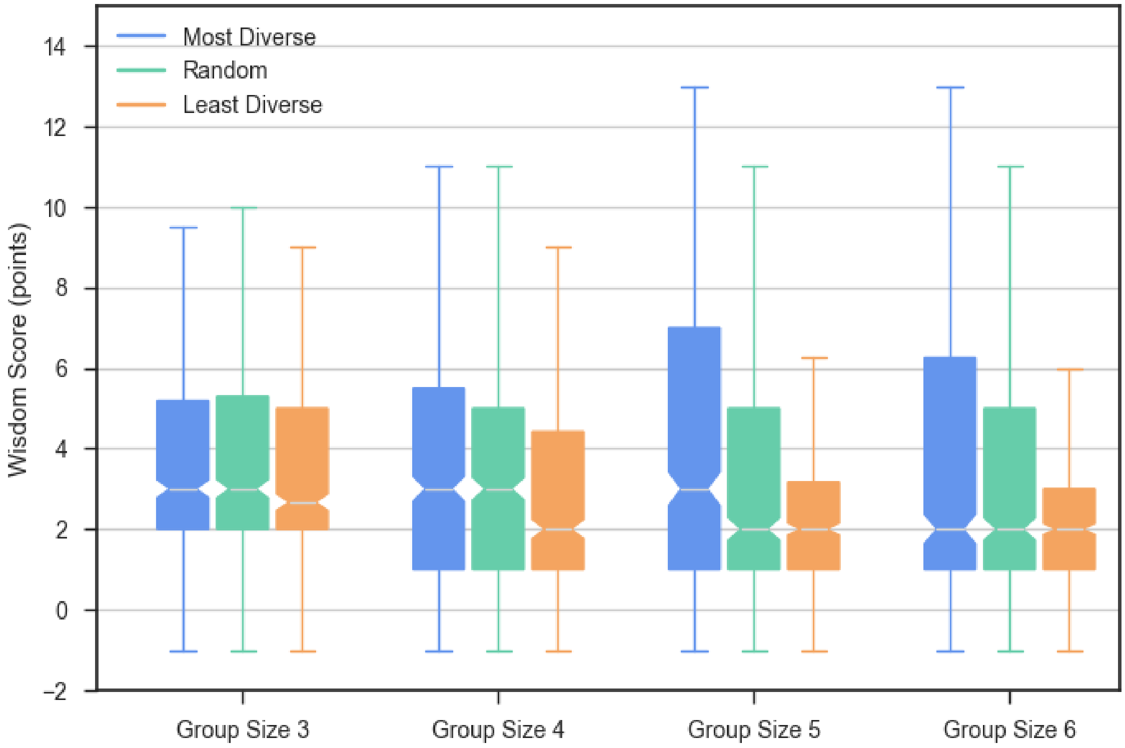


Figure 4.3: Summary of Distributions for Small Groups (Most Diverse vs Least Diverse vs Randomly Sampled). Distributions are summarized as box and whisker plots (box length indicates upper/lower quartiles, notch indicates median, whiskers indicate max/min values). Although differences in medians were small (see Table 2), MD groups outperformed LD groups for all small group sizes 3-6 ($p < 0.005$; Mann-Whitney test). Differences between MD and R groups were less pronounced (MD >R for group sizes 5 and 6; $p < 0.05$). See Table 2 for corresponding numerical data, including associated diversity values.

summarizes these distributions in the form of box and whisker plots. As expected, the MD groups outperformed the LD groups (p <0.005 for all group sizes; Mann-Whitney U test). For comparison, we also show the wisdom score distributions for 500 groups selected randomly (R) from the total 5000. The difference in wisdom scores between MD and R groups, and between LD and R groups, was generally less pronounced than the difference between MD and LD groups. This was expected, because randomly selected groups tended to have diversity values somewhere between those of MD and LD groups. According to our hypothesis, their group wisdom scores should therefore also fall somewhere between those of MD and LD groups.

Table 2 lists the median diversity values for each of the distributions shown in Figure 4, along with the corresponding group wisdom score data (median plus upper/lower quartiles). Although the wisdom score difference between LD and R groups was significant for all small group sizes (p<0.05), the difference between MD and R groups was statistically significant only for group sizes 5 and 6. We believe this is due to the fact that random selection provides some diversity “for free," and thus for MD groups to outscore R groups represents a high performance bar. Indeed, Table 2 shows that, on average, the median diversity of R groups was closer to that of MD groups than to LD groups.

We next extended our analysis to larger group sizes (10-20). Data for all groups – including wisdom scores and diversity values – are shown in Table 2. As for small groups, MD groups significantly outperformed LD groups for all group sizes 10-20. Figure 5 shows an example of the MD and LD wisdom score distributions for group size 16. The small difference in median values (dotted lines) highlights the fact that, while statistical differences between MD and LD wisdom scores were overall robust, effect sizes were generally small. Box plots comparing distributions (MD vs LD vs R) for three representative group sizes (13, 16, 19) are shown in Figure 6. Table 2 provides a more complete set of numerical data for all group sizes 10-20. These data show the same general relationship

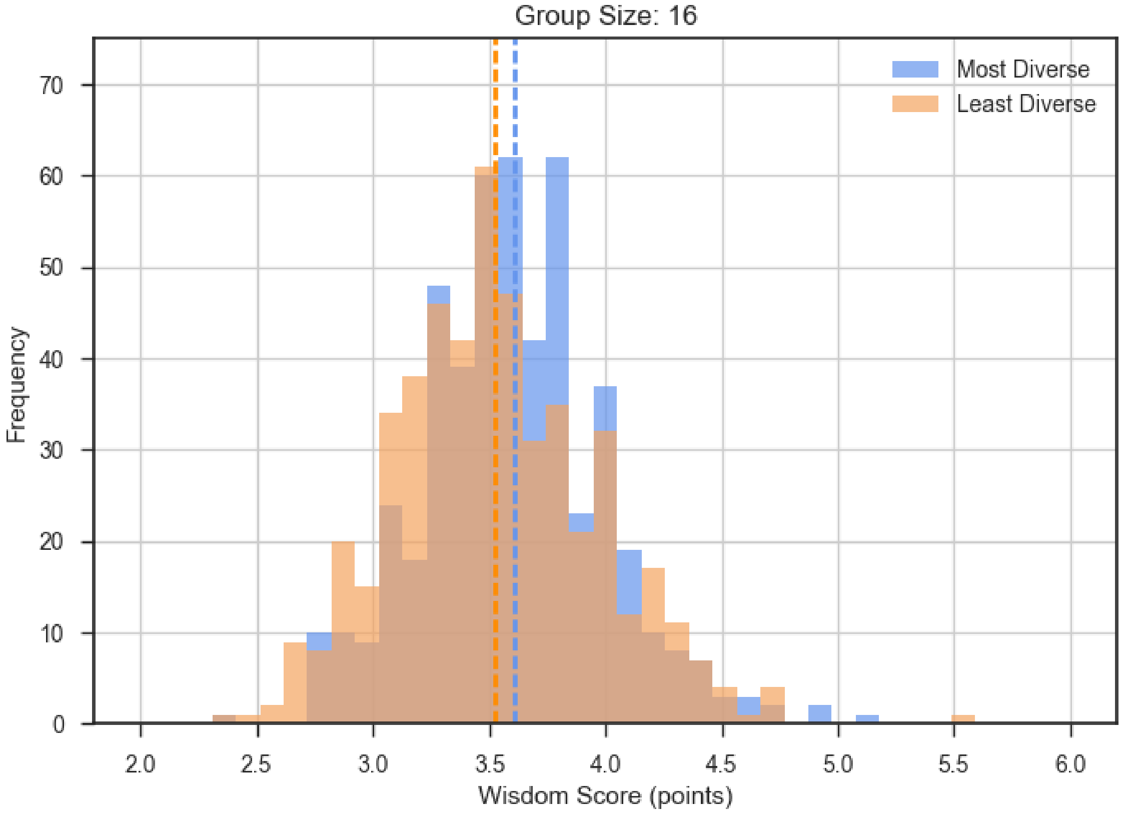


Figure 4.4: Example of Wisdom Score Distributions for Group Size 16 (Most Diverse vs Least Diverse Groups). MD groups significantly outscored LD groups (p <0.0005; Mann-Whitney test). Dotted lines are medians. Distributions for additional larger group sizes are summarized in Figure 6.

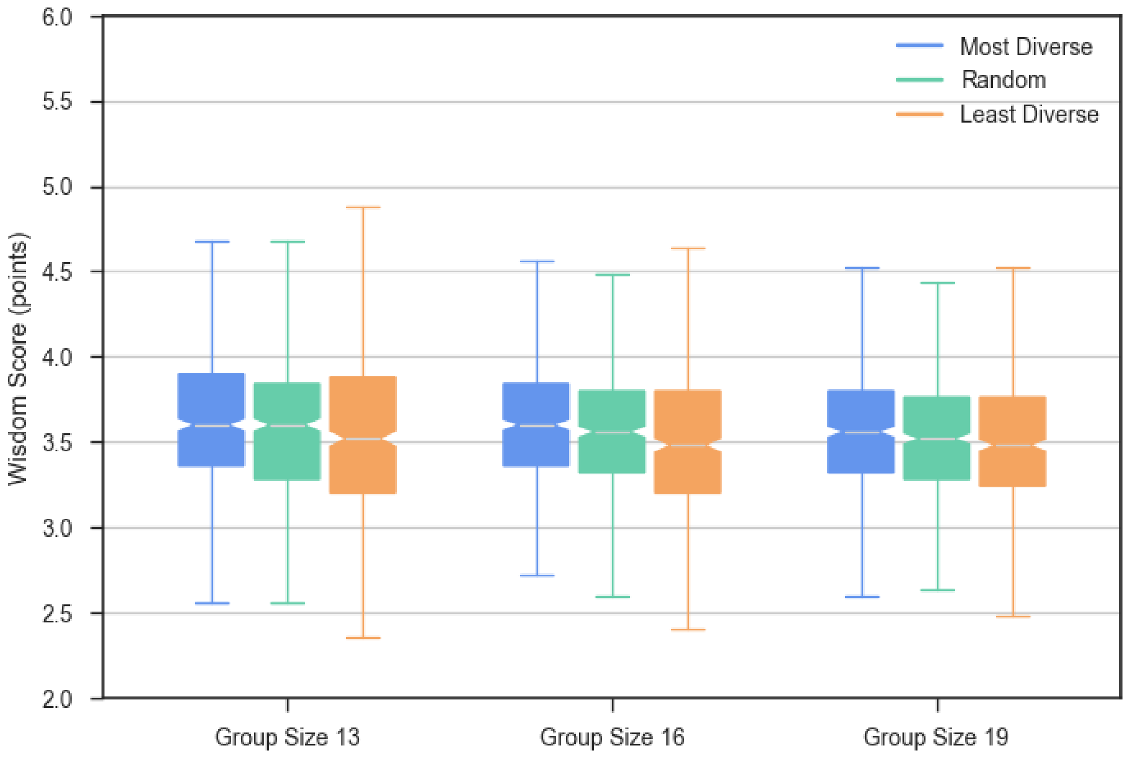


Figure 4.5: Summary of Distributions for Larger Groups (Most Diverse vs Least Diverse vs Randomly Sampled). MD groups outperformed LD groups for each of the group sizes shown. (p <0.005; Mann-Whitney test). Differences between MD and R groups were less pronounced (MD >R for each of the three group sizes shown; p <0.05). See Figure 4 for explanation of box plots. Data for additional group sizes are shown in Table 2.

Table 4.1: Wisdom Scores and Median Diversity Values for Various Group Sizes.

Group	Most Diverse				Least Diverse				Random			
size	LQ	Median	UQ	Diversity	LQ	Median	UQ	Diversity	LQ	Median	UQ	Diversity
3	2.000	3.000	5.200	0.930	2.000	2.667	5.000	0.701	2.000	3.000	5.333	0.863
4	1.000	3.000	5.500	0.924	1.000	2.000	4.500	0.736	1.000	3.000	5.000	0.865
5	1.000	3.000	7.000	0.918	1.000	2.000	3.333	0.751	1.000	2.000	5.000	0.864
6	1.000	2.000	6.500	0.913	1.000	2.000	3.000	0.762	1.000	2.000	5.000	0.864
10	3.400	3.680	3.960	0.892	3.280	3.640	3.960	0.765	3.320	3.640	4.000	0.841
11	3.400	3.680	3.960	0.891	3.240	3.560	4.000	0.768	3.360	3.680	3.960	0.845
12	3.360	3.600	3.880	0.889	3.240	3.560	3.920	0.771	3.360	3.640	3.960	0.842
13	3.360	3.600	3.920	0.889	3.200	3.520	3.880	0.775	3.280	3.600	3.840	0.843
14	3.360	3.600	3.840	0.886	3.240	3.520	3.800	0.779	3.280	3.560	3.840	0.844
15	3.320	3.600	3.840	0.884	3.240	3.520	3.800	0.778	3.320	3.560	3.840	0.841
16	3.360	3.600	3.840	0.883	3.200	3.480	3.800	0.783	3.320	3.560	3.800	0.841
17	3.320	3.560	3.840	0.882	3.240	3.520	3.800	0.783	3.280	3.560	3.800	0.840
18	3.360	3.560	3.800	0.881	3.200	3.520	3.800	0.786	3.280	3.520	3.800	0.843
19	3.320	3.560	3.800	0.881	3.240	3.520	3.760	0.787	3.280	3.520	3.760	0.839
20	3.320	3.520	3.760	0.878	3.240	3.480	3.720	0.787	3.240	3.520	3.760	0.839

LQ = Lower Quartile Wisdom Score, UQ = Upper Quartile Wisdom Score, Median = Median Wisdom Score

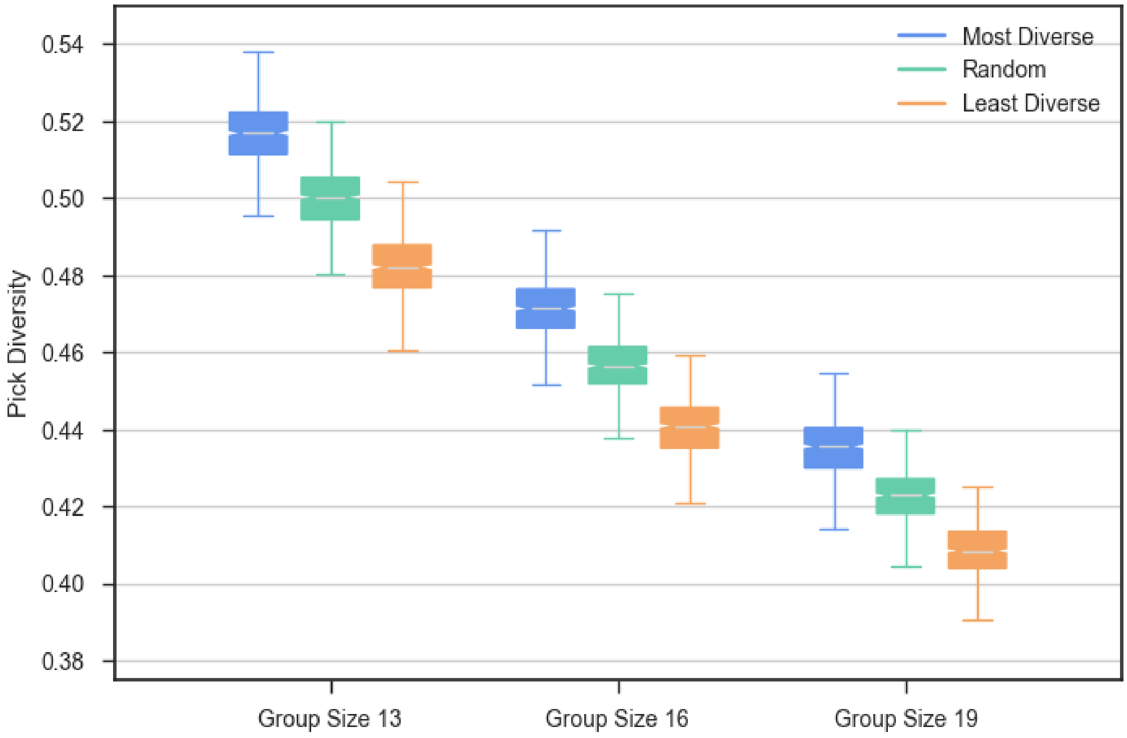


Figure 4.6: Box Plots Showing Differences in Pick Diversity for Most Diverse vs Least Diverse vs Randomly Sampled Groups. For each of the three group sizes shown, we observed a clear relationship between our semantic measure of group diversity and diversity of group members' judgments (captain picks).

between diversity and wisdom score for larger groups as for small groups, with MD groups significantly outscoring LD groups (p <0.05) for 10 out of 11 group sizes examined. As with small groups, differences between MD and R wisdom scores were more modest, with MD significantly outperforming R groups (p <0.05) for 5 out of 11 group sizes.

The preceding analyses suggest that a semantics-based diversity measure can be useful for selecting wiser groups. But is our measure *valid*? That is, does our diversity measure predict actual diversity of judgments (in this case, diversity of captain picks)? To test this, we analyzed pick diversity (as defined earlier) within our larger groups. (We focused on larger groups because our method for small group selection effectively eliminated any diversity among group members' picks.)

Note that, while the maximum pick diversity for each group size is 1 according to Equation 4, the minimum pick diversity varies as a function of group size. Thus it is more instructive to compare pick diversities within a particular group size than across group sizes. Figure 7 shows the relation between group diversity and pick diversity for three representative group sizes (13, 16, 19). For each group size, pick diversity differed substantially between MD and LD groups (p <0.001), with R groups showing intermediate pick diversity. This was true for all group sizes 10-12 (p <0.001).

One might argue that, because users are likely to tweet about their captain picks, any correlation between a tweet-based diversity measure and pick diversity is unsurprising. To estimate the frequency with which a user tended to tweet about his or her captain picks, we counted the number of times captain names – or to be more precise, the names of players whom the user happened to have captained during one or more game weeks – were mentioned in the user's tweets. For each of the 912 users, and for each word tweeted by each user, we checked whether the word matched the first or last name of any player captained by that user during weeks 1-25. The total number of words that were captain names (first or last) was 8,598 out of 1,294,294, or 0.7%. We also analyzed individual tweets and found that 7050 out of 86,938 total tweets (or 8%) contained captain names. Given that the large majorities of words and tweets did not refer specifically to captains, captain mentions alone are unlikely to explain the strong relationship between our tweet-based diversity measures and pick diversity.

4.3 Summary

These results demonstrate useful measures of crowd diversity inferred from linguistic analyses of crowd members' communications. By "useful" we mean that these measures can be exploited to create more effective ("wiser") crowds. We were able to extract such measures from a source, Twitter data, that has been criticized for its shallowness[1] and whose hallmark characteristic is a brevity constraint imposed on individual tweets (140 characters). Although statistically significant, the observed effect size was nevertheless small. A potential reasons for this is the absence of a crowd selection method. The current technique randomly sampled n crowds and selected the top 10% as diverse crowds. The next chapter investigates more sophisticated crowd selection.

[1] http://www.salon.com/2011/10/23/why_chomsky_is_wrong_about_twitter

5 Choices In Crowd Selection: Word embedding, clustering, and multi-view objective optimization

In the previous chapter, we discussed word vector generation and a diversity measure. In this chapter, we discuss a crowd selection approach using these word vectors.

5.1 Crowd selection methodology

We refer to the mechanism as SmartCrowd from hereon. To find diverse participants, our SmartCrowd approach first clusters similar participants according to their social media posts, concerning both topics and communication style. Participants within the same cluster are less diverse compared to those between clusters. We then approximate *diverse* crowds by sampling from different clusters. From a set of such crowds we selected those that maximize average pair-wise diversity measures.

As shown in Figure 5.1 our approach consists of three core components, further described below: social-media based participant representation (Process arrow P1), participant clustering (Process arrow P2), and diversity-based crowd selection (Process arrow P3).

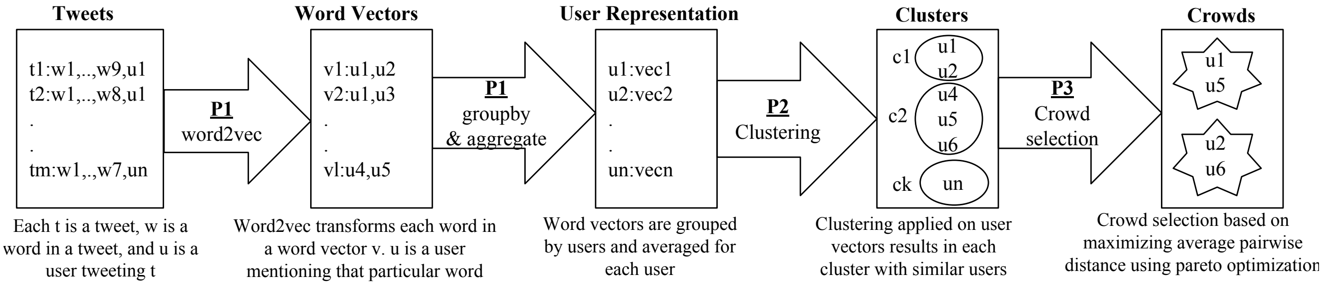

Tweets
t1:w1,..,w9,u1
t2:w1,..,w8,u1
.
.
tm:w1,.,w7,un
Each t is a tweet, w is a word in a tweet, and u is a user tweeting t
P1
word2vec
Word Vectors
v1:u1,u2
v2:u1,u3
.
.
vl:u4,u5
Word2vec transforms each word in a word vector v. u is a user mentioning that particular word
P1
groupby
& aggregate
User Representation
u1:vec1
u2:vec2
.
.
un:vecn
Word vectors are grouped by users and averaged for each user
P2
Clustering
Clusters
c1
u1
u2
c2
u4
u5
u6
ck
un
Clustering applied on user vectors results in each cluster with similar users
P3
Crowd
selection
Crowds
u1
u5
u2
u6
Crowd selection based on maximizing average pairwise distance using pareto optimization

Figure 5.1: Approach overview

P1: Social-Media Based Participant Representation. We generated a vector corresponding to each user as described in the previous chapter.

P2: Participant Clustering. Clustering the participants before crowd selection helps identify groups of similar users regarding topics and communication patterns. A group of users may be following the same teams, players, and use the same linguistic cues. We want to select an equal number of participants in a crowd, from each type of user set, to avoid oversampling users that follow one kind of signal. Such information may be captured by the multiple dimensions of word vector or multiple distance measures computing word vector similarities. For word2vec, cosine similarity describes the similarity between documents (e.g., a set of tweets) so that topic rather than document length determines similarity. Nevertheless, the summary vector already eliminates social media post size. Thus, Euclidean distance might also be appropriate. Related studies [56][61] show that both measures may work for some word-vector based applications. In the absence of a clear rule on which measure should be used for a particular application, we separately evaluate both measures.

We chose the spectral clustering algorithm because it shows exceptional performance in identifying clusters of irregular distributions[62]. Spectral clustering constructs an $n \times n$ similarity matrix A where n is the number of participants (users) in our application. To convert a distance matrix to a similarity matrix, we define an entry A_{ij} for a pair of participants (i, j) as,

$$W_{ij} = e^{-\frac{\delta(x_i, x_j)}{2\sigma^2}} \tag{5.1}$$

Here, x_i is a word2vec participant representation, δ can be Euclidean distance or Cosine distance (1-cosine similarity), and σ functions as a hyperparameter. We chose the standard σ value of 2.0. Using this matrix, spectral clustering returns a graph partition. As spectral clustering requires a given number of clusters, we use the well-known Silhouette Coefficient (SC) method [63] to find the optimal number of clusters. Thus, we run spectral clustering with different numbers of clusters, e.g., between [2, 30]. The SC is computed for each clustering result, and the maximum SC indicates the best clustering structure.

As the kind of diversity (similarity) that helps create a “good” clustering structure is unknown, we used multi-view clustering to synthesize views from multiple distance measures. For word2vec vectors, cosine similarity and Euclidean distance potentially capture different aspects of user clusters, albeit with modest divergence. Our experimental results confirm that multi-view clustering works substantially better than single-view spectral clustering with either Euclidean distance or cosine similarity for our application. It can also be applied for other types of word vectors with distance measures as a separate view or a view resulting from each dimension of a word vector.

P3: Diversity-based Crowd Composition. We considered two selection strategies from each cluster to compose a diverse crowd: random representative selection and average pairwise diversity-guided representative selection. Using random selection, we directly sample n participants at random from each cluster such that n is not larger than the minimum cluster size. Experimental results show that with a small number n, e.g., in [1, 3], the random representative selection method performs reasonably.

We may improve the selection strategy further by maximizing the desired diversity between representatives. The diversity of each generated crowd can be described as the average of pairwise distances between the selected representatives. Cluster-based representative selection already provides a good diversity measure, which can be further improved with the following method. We performed crowd selection based on maximizing both average pair-wise cosine distance and Euclidean distance using Pareto optimization. Here the Pareto front indicates a set of optimal crowds based on the two distance measures.

Algorithm 3 describes the crowd selection process that finds all of the crowds on the Pareto frontier. Let o_1 and o_2 represent two diversity measures. In each iteration, the

```
Input: Clusters $C = c1, c2, \ldots, ck$. $c1 = u1, u2, \ldots, up$. Representatives n
Output: a subset with $n$ participants $u$
P={}
for $i \leq I$ do
    Generate $s = \{p_1, p_2, \ldots, p_{n \times k}\}$ by selecting $n$ participants from each cluster
      at random
    if $\nexists z \in P such\ that$
      $((s.o1 < z.o1 \wedge s.o2 \leqslant z.o2) or (s.o1 \leqslant z.o1 \wedge s.o2 < z.o2))$ then
        $Q = \{z \in P | z.o1 < s.o1 \wedge z.o2 < s.o2\}$
        $P = (P \setminus Q) \bigcup \{s\}$
    end
    $i = i + 1$
end
```

Algorithm 1: Crowd selection from clusters

algorithm generates a crowd s by selecting n participants at random from each cluster to compare with the existing optimal solution. Comparison ensures that the generated crowd s is not strictly worse than existing crowds in P, such that either its o_1 or o_2 is better than one of the crowds in P. The process repeats for I iterations and results in set P that consists of crowds satisfying Pareto optimality.

Among all candidate crowds in P, a “knee point” reveals the best final crowd with conditions over the Pareto frontier [64]. Here, we do not select the best crowd from P but consider all the crowds in P as our final set of diverse crowds. We compute the wisdom score (described below) for each crowd of our final crowd set P. We then compare this set of wisdom scores to the set of wisdom scores of a different crowd selection strategy.

5.2 Results

We evaluated our SmartCrowd for the FPL captain prediction problem [3] for these goals: (1) how do the dominating factors in our approach, such as the chosen clustering algorithms and crowd composing methods, affect the performance of the selected crowds; (2) does our crowd selection strategy lead to wiser crowds, compared to other crowd selection methods; and (3) does the wisdom of the crowd effect depend on diversity.

5.2.1 Experiment designs

For participant clustering, we used spectral clustering with Euclidean distance or Cosine distance (1-cosine_similarity), and multi-view clustering, synthesizing the clustering structures on Euclidean distance and Cosine distance. We evaluated two representative participant selection strategies, 1) random sampling over clusters, and 2) average pairwise distance maximization based sampling. We maximized two average pairwise distance measures (Cosine and Euclidean distance) using Pareto optimization, as our multi-view clusters were generated using both measures. Pareto optimization resulted in 3-6 optimal crowds from our dataset each time we ran the experiment. We repeated the process several times, crowd formation with Pareto optimization (Algorithm 3), to obtain l crowds. In this thesis, we chose $l = 250$.

Wisdom Score: To compare crowds, we computed each crowd's "Wisdom Score" $G = \{U_1, U_2, \ldots, U_n\}$. We first extracted their captain picks for a week w_{index} as $C_{index} = \{c_1, c_2, \ldots, c_n\}$ where c_i is a captain picked by participant U_i in week w_{index}. Crowd wisdom is computed as,

$$WS = \frac{\sum_1^{25} Mod(C_{index})}{25} \tag{5.2}$$

Here, $Mod(C_{index})$ represents the points from the individual captain receiving the most votes from the crowd in the $index$ game week. In case of a non-unique mode - i.e., for a tie, we randomly selected one of these modes. A crowd's wisdom score was the average of its scores over all 25 game weeks considered in our analysis.

5.2.2 Data collection and implementation

We collected FPL related tweets using two FPL keywords, FPL and @OfficialFPL. As the tweets also contained their Twitter usernames, we matched these usernames to their FPL[1] usernames to extract their captain pick data. We manually verified 2786 such matches. We further collected their soccer related Twitter data by scraping their Twitter timeline (for a total 4,299,738 tweets)[2].

For evaluation purposes *only*, we collected 25 weeks of captain picks for 2015-16 FPL season for each participant. We also collected that captain's score based on his game performance from the same FPL portal. We further collected participant performance data for seven seasons (2009-2015), to compare with an expert-based crowd selection strategy that assumes the existence of historical performance data for defining expertise. [3]

5.2.3 Results and analysis

The results are organized according to the evaluation goals: (1) Factors affecting the SmartCrowd, to show how methods for clustering and proposed crowd composition method affect final SmartCrowd performance; (2) Comparison of different crowd selection methods. Based on the optimal SmartCrowd, we first compare the performance of SmartCrowd with a random crowd selection method, both of which do not depend on historical crowd performance data. Further, we show that the performance of SmartCrowd is comparable to expert crowds when expert participants can be selected using historical performance data; (3) Finally, we analyze the effect of diversity on crowd wisdom.

Factors Affecting the SmartCrowd

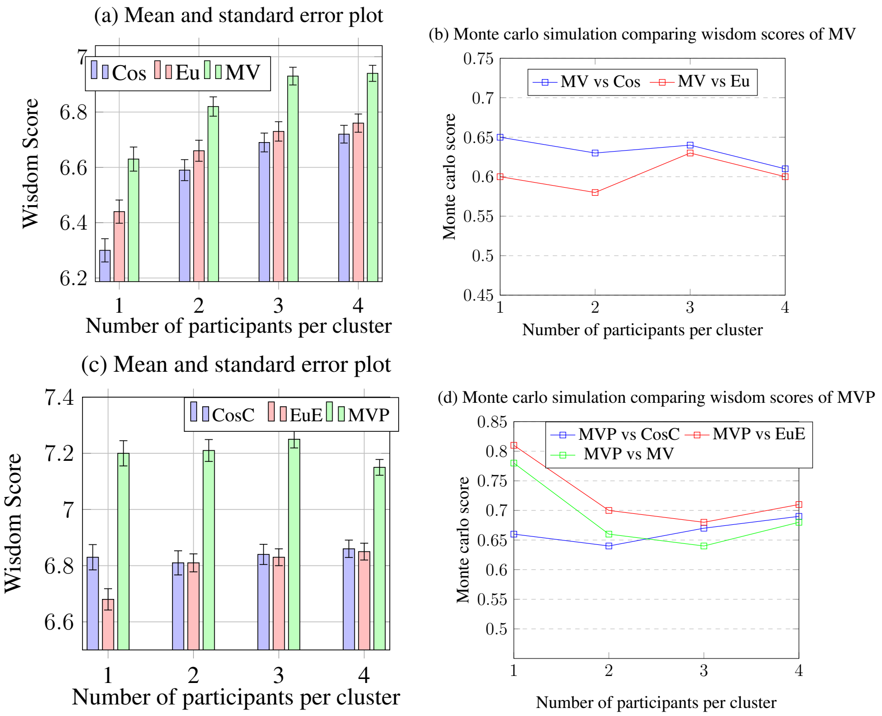
As described in Section 9.1, participant clustering, and diversity-based crowd composition are two key influences. Hence, we examined their effects on final crowd performance.

Participant clustering: The best number of clusters were 6(0.27), 7(0.23), and 7(0.45),

[1] fantasy.premierleague.com

[2] As the keyword list is not exhaustive, we may have more than $\sim$ 1M FPL tweets in our source dataset.

[3] As the dataset contains actual tweets and usernames, we have not uploaded the dataset. It will be made available from the corresponding author upon request.



(a) Mean and standard error plot

(b) Monte carlo simulation comparing wisdom scores of MV

(c) Mean and standard error plot

(d) Monte carlo simulation comparing wisdom scores of MVP

Figure 5.2: (a) and (b) compares crowds generated using Multi View clustering(MV), Cosine(Cos), and Euclidean(Euc) distance based clustering. (c) and (d) compare crowds generated by maximizing one distance measure (CosC, EucE) versus maximizing both distance measures (MVP). MVP crowds perform the best.

for Euclidean-spectral (spectral with Euclidean), Cosine-spectral(spectral with cosine), and Multi-view, respectively. The bracketed values indicate the corresponding maximum silhouette value. Note that Multi-view clustering produced the best clustering structure.

We used these clustering structures in subsequent analysis. We sampled crowds by selecting n participants from each cluster at random for a given clustering structure (Euclidean-spectral, Cosine-spectral, and Multi-view). We selected l such crowds from each clustering structure. Figure 5.2a,b shows the mean and standard error for the wisdom score for crowds generated from each clustering structure. Crowds from a multi-view clustering structure(MV) achieved the best average wisdom score. They also outperformed crowds generated from Cosine(Cos) and Euclidean clustering(Eu) structures, (T p-value < 0.05).

We also used Monte Carlo simulation to compare the wisdom score of a randomly selected crowd from set one to the wisdom score from a randomly selected crowd from set two. We repeated this 1000 times - each time counting whether the wisdom score of a set one crowd was higher than the wisdom score from a set two crowd. The ratio of the total counts to 1000 provides the Monte Carlo simulation score. A Monte Carlo score of $\sim$ 0.5 indicates that two sets of crowds are equally likely to beat each other. Monte Carlo score of $\sim$ 1.0 indicates that a crowd from set one almost always beats a crowd from set two. Figure 5.2b shows the Monte Carlo simulation score for comparing MV to Eu, and Cos. The Monte Carlo simulation score > 0.6 indicates that MV crowd is likely to outperform both Cos and Eu crowds.

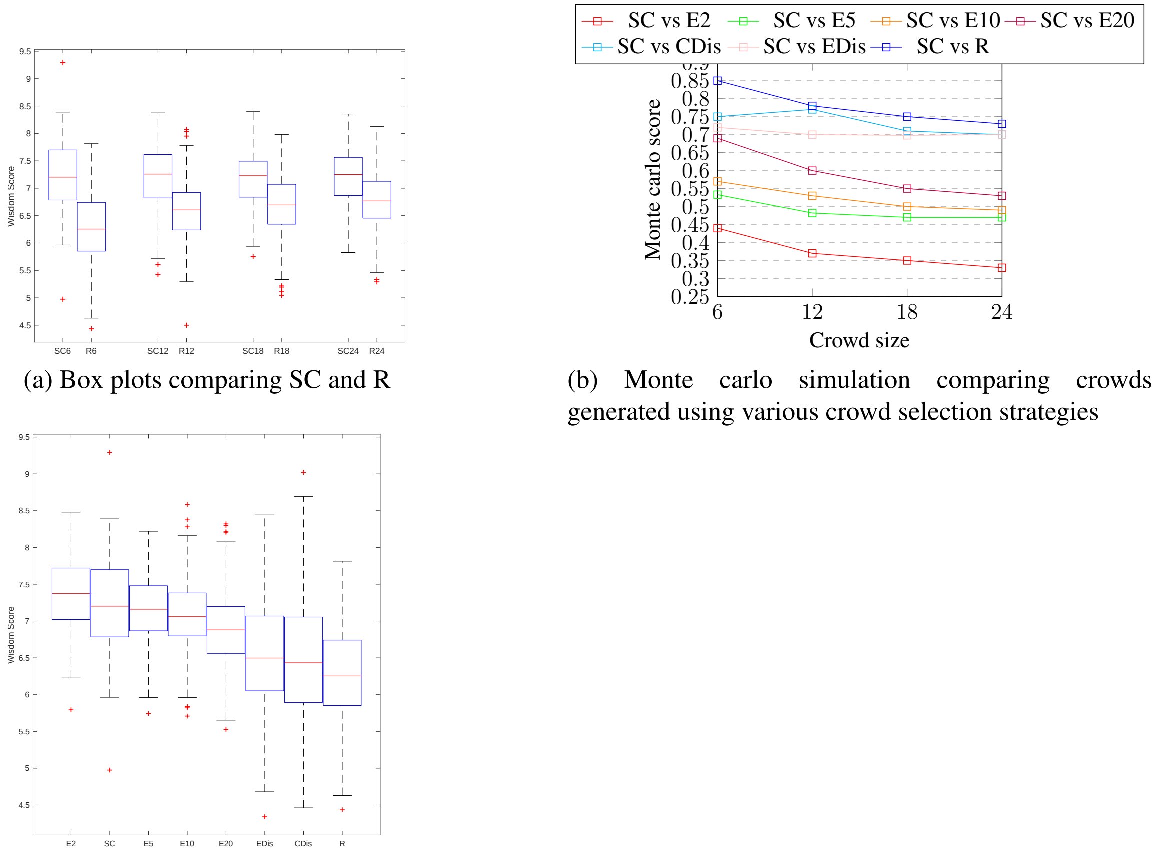
Diversity-based Crowd Composition Next, we evaluated a more sophisticated crowd composition method, i.e., the proposed Algorithm 3 for multi-view clustering. For single-view clustering, we had we separately maximized crowd selection based on Average pairwise Euclidean and Cosine distance respectively. Specifically, we sorted crowds generated by selecting n participants at random from each cluster for a given distance measure and selected the top l crowds. For multi-view clustering, we maximized *both* average pairwise

Euclidean and cosine distance for crowd selection. Figure 5.2c shows the average and standard error of l crowds' wisdom scores. Multi-view clustering combined with Pareto optimization based crowd selection generated crowds (MVP) that achieved the best wisdom score. These crowds also outperformed crowds generated using a single distance-based clustering and maximization (EuE and CosC) method (T-test p-value < 0.05). Figure 5.2d shows the Monte Carlo simulation scores comparing MVP crowds to EuE and CosC crowds. An MVP crowd was $\sim 80\%$ likely to have a higher wisdom score than EuE and CosC crowds. MVP crowds also outperformed MV crowds, i.e., crowds selected without distance measure maximization.

Comparison with Other Crowd Selection Strategies

Using the optimal settings obtained from the first set of experiments (Multi-view clustering and Pareto optimization based crowd selection), we compared SmartCrowd with other crowd selection methods. Without participants' prior performance knowledge, we considered randomly selected crowds as our baseline. Specifically, we generated random crowds by selecting $n \times k$ participants at random from all participants. Here, n indicates the number of representatives that we considered for SmartCrowd and k indicates the number of clusters in SmartCrowd. As we found 6 clusters in our SmartCrowd selection, we generated random crowds in multiples of 6, i.e., $6, 12, 18, 24$ corresponding to $n = \{1, 2, 3, 4\}$ representatives per cluster.

Figure 5.3a shows the box plot of wisdom scores for SmartCrowds and random crowds for different numbers of crowd participants. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the upper and lower quartile. Whiskers extend to the most extreme data points not considered outliers, and outliers appear as '+'. SmartCrowds (SC) have consistently larger wisdom scores than random crowds (R) for all crowd sizes. SC provided significantly higher wisdom scores than R (T p-value



(a) Box plots comparing SC and R

(b) Monte carlo simulation comparing crowds generated using various crowd selection strategies

(c) Box plots comparing crowds sampled using various crowd selection strategies

Figure 5.3: SmartCrowd(SC) crowds compared with Random(R), Expert(E), Euclidean(EDis), and Cosine(CDis) distance based crowds. (a) shows that SC performs significantly better than R. As shown in (b) and (c), SC performs better than R, EDis, and CDis. SC outperforms E20, E10, and E5 while almost equivalent to E2 and slightly worse than E1.

< 0.05). Figure 5.3b shows the Monte Carlo simulation score for comparing SC to R selection (SC vs R line). SC is 85% likely to beat a random crowd. The probability that SC outperforms a random crowd does decrease with increasing crowd size. Increasing crowd size in the random sample begins to approximate a better choice in aggregate. Thus, a smaller crowd size, e.g., one representative per cluster, to form a diverse crowd using SmartCrowd is sufficient.

Next, we compared SmartCrowd with expertise-based crowd selection. Expert crowds with a known performance history often perform very well [3]. For evaluation purposes, we sampled expert crowds only from the top performing participants. Figure 5.3c shows the box plots for expert crowds E2, E5, E10, and E20 generated from the top 2%, 5%, 10%, and 20% performance thresholds, respectively of crowd size six (one representative per cluster). E2 crowds do have the highest wisdom scores. SmartCrowd (for crowd size six) had comparable wisdom scores as E5. SmartCrowd significantly outperformed E10, E20 (T p-value <0.05). Moreover, the E2 expert crowd advantage is marginal and therefore comparable to SmartCrowd.

Figure 5.3b indicates Monte Carlo simulation scores comparing diverse crowds to various expert crowds. Monte Carlo scores of 0.7 and 0.58 for comparing SmartCrowd to E20 and E10 also show that a SmartCrowd crowd is quite likely to outperform E10 and E20 expert crowd for crowd size six. Increasing crowd size does not benefit SmartCrowd. Hence, we did not observe an improved wisdom score with increasing crowd size. Next, we compared the performance of a SmartCrowd to one assembled by maximizing either average pairwise cosine or Euclidean distance measure. We generated l random crowds and sorted them based on average pairwise Euclidean and Cosine distance, selecting the top 10% (l) crowds as representative. AvgE and AvgC in Figure 5.3b shows the resulting box plots of the wisdom scores. Figure 5.3a shows the Monte Carlo simulation scores comparing the SmartCrowd selection strategy to average pairwise Euclidean and Cosine distance-based crowd selection strategies. A Monte Carlo simulation score of 0.7 indicates that the SmartCrowds substantially outperformed these crowds.

Next, we compared various crowd formation strategies based on whether a crowd of size n outperforms an average individual. We ranked all 2786 participants using aggregated season scores, i.e., an average of all 25 weeks' captain scores. Figure 5.4 shows the percentile of participants that a crowd outperforms on average. We computed an average of the l crowds and computed the percentile of participant scores that it outperforms. On

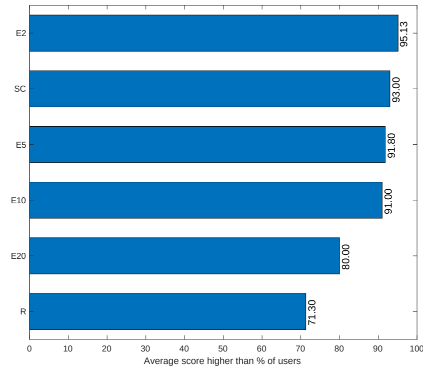
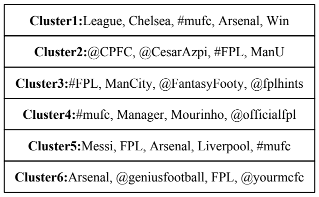


Figure 5.4: Wisdom of crowd effect



Cluster1: League, Chelsea, #mufc, Arsenal, Win
Cluster2:@CPFC, @CesarAzpi, #FPL, ManU
Cluster3:#FPL, ManCity, @FantasyFooty, @fplhints
Cluster4:#mufc, Manager, Mourinho, @officialfpl
Cluster5: Messi, FPL, Arsenal, Liverpool, #mufc
Cluster6: Arsenal, @geniusfootball, FPL, @yourmcfc

Figure 5.5: Cluster wise most frequent words

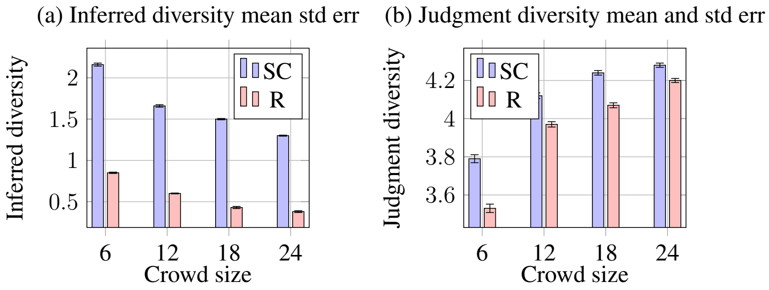
average a randomly generated crowd of size 6, achieves a better "wisdom score" than 72% of total participants. However, a diverse crowd of size 6 achieves a better "wisdom score" than 93% of the participants.

Diversity, Expertise, and Wisdom of Crowd Effect Analysis

Finally, we examined the diversity that SmartCrowd captures, including topic diversity, the effect of crowd size on diversity, and the relationship between social media-based diversity and other diversity measures.

Topic diversity. We computed the TF-IDF[4] score for each word in the tweets contributed by the participants in the same cluster, excluding stop words. We selected words with the highest TF-IDF scores to capture the most frequent topics discussed in each cluster as shown in Figure 5.5. Apparently, participants in different clusters show different interests in

[4] https://en.wikipedia.org/wiki/Tf-idf


(a) Inferred diversity mean std err

(b) Judgment diversity mean and std err

Figure 5.6: Inferred diversity (a) and Judgment diversity (b) comparison. SC has higher inferred and judgment diversity. Inferred diversity correlates with judgment diversity.

teams, players, and the informative FPL accounts. Thus, SmartCrowd finds representatives having diverse perspectives in soccer and FPL for captain choice.

Some words, e.g., join, team, etc. do not appear in the figure as these words do not explain the clusters. Moreover, some teams appear in more than one cluster. However, the TF-IDF scores of these words varied for each cluster. To capture this, we ran a Spearman's correlation analysis for each pair of six clusters. Thirteen of fifteen cluster pairs were negatively correlated, confirming cluster diversity.

Next, we confirmed whether SmartCrowd's outperformance truly results from diversity. Multi-view clustering creates clusters of different sizes. If small clusters contained mostly experts, we effectively assure at least n experts in our diverse crowds. SmartCrowd's outperformance could merely reflect expertise instead of diversity. To exclude this explanation, we eliminated the two smallest clusters of sizes four and seven and followed our crowd generation strategy based on Algorithm 3. We compared the resulting crowds without these clusters to crowds generated with all clusters(SmartCrowds). The resulting Monte Carlo simulation score $\sim$ 0.5 indicated that the two sets of crowds had similar performance. Therefore the eliminated crowds do not account for SmartCrowd outperformance.

Inferred diversity vs. judgment diversity. SmartCrowd selects diverse crowds by clustering similar participants represented by their word2vec vector. We refer this diversity

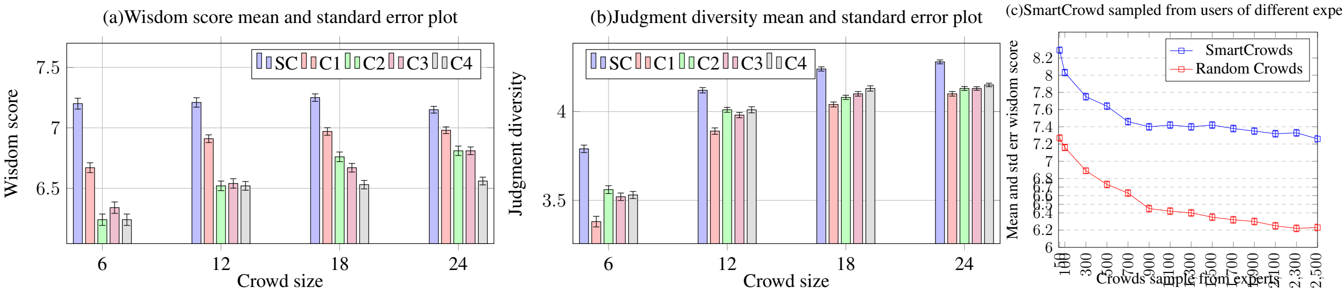


Figure 5.7: Inferred diversity comparison for SmartCrowd & Random crowd. SC has higher inferred diversity than R.

as "inferred". This is computed as the summation of average pairwise cosine and Euclidean distance of a group. Figure 5.6a compares crowds based on "inferred diversity", and shows that SC crowds are more diverse than Random crowds. Inferred diversity decreases with increasing crowd size as a newly added participant's social media is likely to be closer (regarding Euclidean and Cosine) to at least one existing participant.

We examined whether inferred diversity produces a set of participants with different judgments. We randomly sampled 10,000 participant pairs from a single cluster (selected at random) – " similar participants". The probability of a participant pair selecting different captain choices is

$$p_d = \frac{ND_{total}}{10000} \tag{5.3}$$

Here, ND_{total} is the number of times a participant pair differed in captain choice. We also generated another set of participant pairs, "diverse participants" by selecting two participants from different clusters and computed p_d. p_d for "similar participants" was 0.81 while p_d for "diverse participants" was 0.85. Hence, a crowd sampled from "inferred diversity" measure is also likely to demonstrate judgment diversity.

Further, we examined Merayo et al. 's. [21]"judgment diversity" measure to compare SmartCrowd with Random crowds. Accordingly, judgment diversity likely implies a less biased sample of participants, which provides a better-aggregated opinion. Merayo et al. 's judgment diversity measure is

$$D = \frac{\sum_{i,j} d(u_i, u_j)}{n(n-1)}, \tag{5.4}$$

where $d(u_i, u_j)$ is the difference between the performance scores of participants u_i and u_j (e.g., scores corresponding to their captain picks) and n is the total number of participants in the crowd. Using this metric, we investigated whether SmartCrowd generates crowds with better judgment diversity than a random crowd. We represented the judgment diversity of a crowd with the average of D over 25 weeks. Figure 5.6b confirms that SmartCrowd results in greater judgment diversity than a randomly selected crowd. Judgment diversity concerning captain score increases with increasing crowd size as participants chose a captain among 100+ soccer players. Hence, a new participant may choose a captain that is not already chosen by other members of the existing crowd.

The consistency between judgment diversity and inferred diversity is further confirmed with crowds formed by sampling only within a specific cluster. SmartCrowd samples crowds by selecting participants from each cluster. Hence, crowds formed by participants from the same cluster should have low diversity. We sampled crowds from each cluster by selecting n participants at random. Figure 5.7a compares the wisdom score of SmartCrowd and non-diverse crowds. C1, C2, C3, and C4 represent crowds sampled from cluster1, cluster2, cluster3, and cluster4 respectively. We ignored two clusters with less than ten users as we cannot generate l crowds of size ≥ 6 from these clusters. Crowds generated using SmartCrowd consistently outperformed crowds generated from one cluster regarding wisdom score. Figure 5.7b shows the average judgment diversity of crowds generated using SmartCrowd (SC) and (non-diverse) crowds generated from each cluster. SC also has the highest overall judgment diversity. Thus, in the absence of historical judgment data, our inferred diversity measure can serve as an effective proxy for judgment diversity and the attendant benefits to accuracy consistent with the findings of Merayo et al.[21]. Because the judgment diversity measure is a measure of variance, a larger variance is correlated with a larger mean and hence is expected to correlate with a better answer in aggregate. As a diverse crowd provides different judgments, it results in increased variance, and hence we expect a crowd to perform better than a non-diverse crowd. We examined whether diversity is meaningful in sampling crowds from users in different ranges of expertise. We generated crowds with the top-k experts, $k \in [50, 2500]$. Figure 5.7c shows the wisdom score achieved by crowds sampled using SmartCrowd(SC) and crowds sampled at Random(R). Crowds sampled using SmartCrowd benefit performance regardless of the expertise range. Moreover, the best performance results from diverse experts. In other words, one can effectively predict a captain despite the differing (and uncontrolled) expertise range inherent in Twitter data. Interestingly, crowds sampled from the top 50 and 100 experts achieve better wisdom score than any single user.

We also examined whether diversity can replace the expertise and the performance of hybrid (consisting of both experts and diverse non-experts) crowds. For this, we considered the top 100 users as experts and the rest of the users as non-experts. We formed crowds of size six from the top 100 expert users and kept on replacing n users with n non-expert but diverse users, $n \in [0, 6]$. We sampled the n users from the remaining r clusters. For example, if the initial set of expert users are already selected from the three out of six clusters, and we want to replace $n = 3$ users, then we select one user from each of the remaining three clusters. To select a user from a given cluster, we again maximize the two average pairwise distance measures. Figure 5.8b shows the results for replacing n experts with diverse non-experts. Diverse but non-expert participants can replace experts without trading performance. In fact, diverse participants replacing 1-2 experts results in better-performing crowds than all experts. All non-expert crowds do not perform better than all experts. Note that in this case, these crowds do not consist of any of the top 100 expert participants, unlike the experiments for comparing diverse crowds with expert crowds.

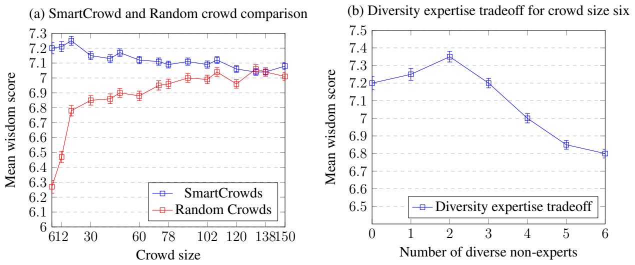


Figure 5.8: (a)SmartCrowd(SC) and Random crowd(R) wisdom score comparison. SC with 6 participants achieve wisdom score that is achieved by 100+ participants of R., (b) shows the effect of replacing experts with diverse non-experts.

Finally, we examined the effect of crowd size and wisdom score. Crowd size potentially affects prediction performance. Figure 5.8a plots the mean and standard error wisdom score for increasing crowd size. With increased crowd size, random crowd performance approaches a SmartCrowd, while SmartCrowd's performance slightly decreases. They achieve similar wisdom scores for crowd size at or above 108. However, even at a large crowd size, e.g., 150, random selection does not perform better than SmartCrowd with only 6-12 representatives. With *only six participants* SmartCrowd can judge as accurately as 100+ Randomly selected users.

5.3 Summary

This chapter demonstrated that social media data can be used to infer diversity, sampling diverse, and consequently smart, crowds for the selection of top performing FPL captains. A crowd sampled using the proposed technique is notably more accurate than a crowd sampled at random and comparable to crowds of the top 2-% experts. Hence, social media data provide an effective proxy for often unavailable historical expertise data. We clustered participants based on their social media content, and showed that multiple similarity measures improve clustering over a single similarity measure. Clustering users in this way allowed us to sample by diversity to improve FPL captain prediction. Average pairwise diversity maximization further improved crowd wisdom. We also showed that the performance was truly attributed to diversity and diverse non-experts can replace expert participants in a crowd without compromising performance. Hence, such a technique is crucial when one does not have an access to expert opinion.

6 Top-Down diversity and wisdom of crowd

The previous chapters showed that word2vec based diversity, that is, bottom-up diversity, can help identify diverse and subsequently smart crowds. This chapter investigates top-down diversity and its effect on diverse crowd selection and its wisdom. This chapter also investigates the explanation of diversity.

6.1 Top-down diversity and crowd selection

Figure 6.1 describes our approach to generating diverse crowds based on FPL captain selection strategies in social media posts. Each participant is categorized based on the number of tweets indicating the two captain selection strategies. Hence, classifying tweets indicating a captain selection strategy and participant categorization are the two key components of our diverse crowd generation. Crowd selection from these categories completes the crowd formation process. Subsequent knowledge graph analysis provides an explanation for calculated diversity.

6.1.1 Tweet classification

Consistent with common practice [65], we used machine learning-based text classification for identifying player selection strategy from tweets. We manually annotated 165 of tweets as Popular choice and 258 tweets as Differential choice tweets. We considered an equal number of non-popular and non-differential choice tweets in training the classifiers. The two classification categories (popular choice and differential choice) are not orthogonal; the same tweet may provide evidence for both popular as well as differential choice. Hence,

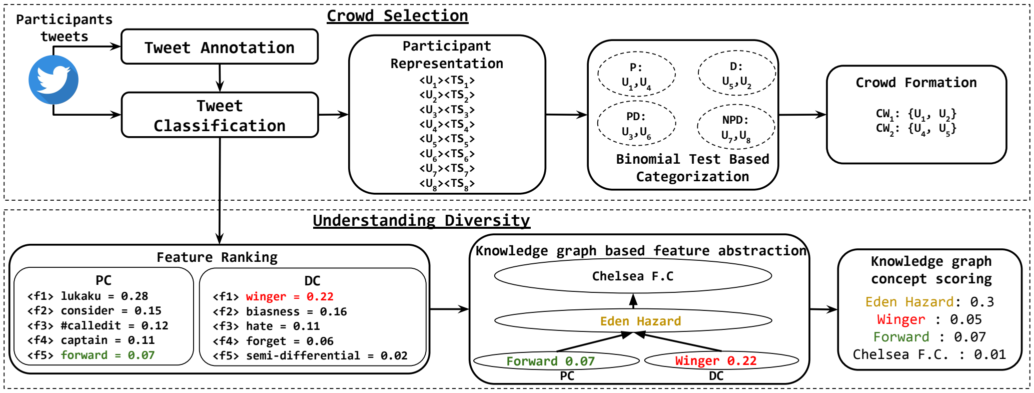


Figure 6.1: Proposed crowd selection and diversity explanation approach. Each participant is represented by his/her tweets and processed for identifying tweets referring to two solutions strategies. A knowledge graph-based feature abstraction identifies relevant concepts subsuming PC and DC keywords. $< f >$:feature, $< U_i >$: Participant, $< TS_i >$:Tweet Set per user, CW: Crowd

we trained two tweet classification models. Model 1 identifies whether a tweet belongs to popular choice and Model 2 determines whether a tweet belongs to differential choice.

We used a *Bag of Words* approach combined with term frequency and inverse document frequency (TF-IDF) for generating a feature vector for each tweet [66]. We considered uni-grams and bi-grams as features and found whether each feature is present in a tweet. These vectors are then processed for TF-IDF computation, and each feature is represented with its TF-IDF value instead of “1” or “0”. To avoid over-fitting in training based on these sparse tweet vectors, we used k-best feature selection. This feature selection technique uses Singular Value Decomposition, widely used in feature selection for text classification [67]. We trained two models using a Random Forest classifier with ten-fold cross-validation with a 70%, 30% train-validation split. We also reserved 10% of the labeled data as test data to determine classifier performance accuracy. We report the final accuracies for each model, i.e., popular choice, and differential choice classification, accuracies in Section **??**.

We processed the tweets for each participant using Model 1 to identify the number of tweets in popular choice and Model 2 to identify number of tweets in Differential choice. Formally, for each participant U_i, we have $\mathbf{U_i} = \{n_P, \overline{n}_P, n_D, \overline{n}_D\}$ (Same as TS_i in Figure 6.1). Here each n is a number where n_P is popular choice tweets, $\overline{n}_P$ non popular choice tweets, n_D Differential choice tweets, and $\overline{n}_P$ is non-differential choice tweets.

6.1.2 Binomial test based categorization

Each n in U_i may result from different distributions. Hence, we normalize each n by computing the Z-score for each n with respect to all participants. These Z-scores represent each participant relative to others in terms of n_P, $\overline{n}_P$, n_D, and $\overline{n}_D$. As each participant provides tweets suggesting both strategies or neither, we require a decision rule to categorize each participant according to the different amounts of data they provide. We do not consider these strategies as mutually exclusive and investigate whether a popular (or differential) choice best characterizes the participant or non-popular (or non-differential). We assign a participant as one of the following four *participant types*: 1. Popular Choice, 2. Differential Choice, 2. Popular and Differential, 3. Neither popular Nor differential. As the names suggest, Type 1 refers to participants most likely to exhibit popular choice tweets, Type 2 are participants most likely to exhibit Differential choice, Type 3 participants are ambiguous, providing substantial evidence for both, and Type 4 participants do not provide any evidence for either one of these strategies. We used a binomial test with a null hypothesis that the two strategies are equally likely to occur. Formally, a binomial test (BiTest) $\mathbf{B_q}$ for an event with q as the probability for an event to succeed and $\overline{q}$ as the probability of failure can be described as,

$$B_{q,\overline{q}} = \binom{N}{q} \cdot p_0^q (1 - p_0)^{\overline{q}} \tag{6.1}$$

Here, N is defined as the sum of q and $\overline{q}$. p_0 is the probability of occurrence of a success in each one of the N trials. In our study, we set $p_0 = 0.5$ and the binomial test was performed at 5% alpha level[1]. This determines whether q is likely to occur more than $\overline{q}$, 95% of the time.

[1] http://www.statisticshowto.com/what-is-an-alpha-level/

```
Input : $U_i = \{n_P, \overline{n}_P, n_D, \overline{n}_D\}$
Output : $T_1 \vee T_2 \vee T_3 \vee T_4$
$B_{n_P,\overline{n}_P} = BiTest(n_P, \overline{n}_P)$
$B_{\overline{n}_P,n_P} = BiTest(\overline{n}_P, n_P)$
$B_{n_D,\overline{n}_D} = BiTest(n_D, \overline{n}_D)$
$B_{\overline{n}_D,n_D} = BiTest(\overline{n}_D, n_D)$
$B_{\overline{n}_P,\overline{n}_D} = BiTest(\overline{n}_P, \overline{n}_D)$
$B_{\overline{n}_D,\overline{n}_P} = BiTest(\overline{n}_D, \overline{n}_P)$
if $B_{n_P,\overline{n}_P} > 0.05$ and $B_{n_D,\overline{n}_D} \leq 0.05$ and $B_{\overline{n}_P,\overline{n}_D} \leq 0.05$ then
  return $T_2$
else if $B_{n_P,\overline{n}_P} \leq 0.05$ and $B_{n_D,\overline{n}_D} > 0.05$ and $B_{\overline{n}_P,\overline{n}_D} \leq 0.05$ then
  return $T_1$
else if $B_{n_P,\overline{n}_P} \leq 0.05$ and $B_{n_D,\overline{n}_D} \leq 0.05$ then
  return $T_3$
else
  return $T_4$
end if
```

Algorithm 2: Binomial test based categorization of participants at 5% significance. Given the Z-scores of a participant U_i, the algorithm categorizes participants in one of following types: Type 1(T_1), Type 2(T_2), Type 3(T_3), Type 4(T_4)

The participant exemplifies Popular choice (Type 1) or Differential choice (Type 2) when the Binomial tests find significant evidence for only the corresponding type. Specifically, $\mathbf{B}(\mathbf{n_P}, \overline{\mathbf{n}}_\mathbf{P})$ should indicate that likelihood of n_P over $\overline{n}_P$ is more than 95%, and the possibility of other events is less than 95% to consider a participant as Type 1. Algorithm 3 formalizes this participant type assignment process. It starts with six binomial tests (lines 3-8). Based on the condition described above, it decides the participant type (line 9-17).

6.1.3 Diverse crowd selection

Type 1 (Popular) and Type 2 (Differential) participants provide strong evidence of using the corresponding strategy in player selection. About half of the participants are Type 3 (Both) or Type 4 (Neither), who provide ambiguous or unclear strategy indicators that will likely muddy diversity. Hence, we avoided participants belonging to these two types in our diverse crowd formation and created our diverse crowd using clear Type 1 and Type 2 participants. We selected n participants from Type 1 and Type 2 to build our diverse crowd.

6.1.4 Understanding diversity

The diverse crowds that we generated in the previous step depend on tweet classification. To explain the kind of diversity such a tweet classification captures in Fantasy Premier League domain, we extracted the top most informative features (keywords) from our Random Forest Classifier[68]. We mapped these keywords to an English Premier League domain-specific knowledge graph extracted from DBpedia using a domain-specific knowledge graph extraction tool[19]. Such a knowledge graph provides a good representation of a corresponding domain [19][69]. The resulting hierarchy for the English Premier League in Figure 6.1 indicates that the top concept *Chelsea F.C.* subsumes *Eden Hazard* who has two attributes (subsumes) *Forward* and *Winger*. Hence, a *parent* concept subsuming multiple child concepts explains child concepts. We use the parents in this hierarchical structure to encompass the keywords identifying multiple strategies.

We seek the concept in a knowledge graph *subsuming* most of these keywords. As we had two classification models corresponding to two of the captain selection strategies, we obtain two lists of keywords from each model. Each keyword has an importance value between 0 and 1 from the Random Forest Classifier. We use these values to assign each concept in the domain-specific knowledge graph a weight, as shown in Figure 6.1. Specifically, we compute 3-hop parents of each concept and assign a score for each of parent concept as follows,

$$S = \frac{C_w}{P_l} \tag{6.2}$$

Here, C_w refers to the concept weight indicated by the keyword weight and P_l indicates a parent level of the current concept. For each concept C associated with the keyword, we get the parents of C and compute its corresponding S. If we find that the parent concept being processed as part of C is already identified as a parent for another concept, then we add this score to the existing score. Hence, a score associated with each parent concept indicates the number of keywords the particular parent concept subsumes. We repeat the same procedure for two hops, and three hop parents. As 1-hop parents are more relevant than 2-hop and 3-hop parents, the concept score is multiplied by the inverse of parent level P_l. As shown in Figure 6.1, we found *Forward* and *Winger* as important keywords to distinguish popular choice tweets and differential choice tweets respectively. In the English Premier League knowledge graph, we start with these two concepts and ascend the hierarchy (extracting *Eden Hazard*, and *Chelsea F.C.*) and compute S for each concept. As concepts with high scores can best explain multiple keywords and hence the player selection strategies, we score each concept in the knowledge graph and consider the top-N concepts for understanding diversity. This allows us to identify concepts that unify the categories with implicit contents that are not explicit in the tweets themselves.

6.2 Results

In this section, we describe the dataset, evaluation measure, and results.

6.2.1 Dataset

We collected FPL related tweets using the Twitter streaming API with two FPL related keywords, *FPL*, and *@OfficialFPL*. We determined captain pick data from the FPL portal[2] by matching Twitter usernames associated with these tweets to their FPL usernames. We used FPL captain pick data for the 2016-17 season.

We manually verified 3385 participant matches based on Twitter username and FPL username and collected their additional tweets by crawling their Twitter timelines. For each participant, we collected tweets ranging from 2014 to August 2016 (before the start of FPL 2016 season). We filtered these tweets using FPL related keywords to consider only

[2] fantasy.premierleague.com

relevant tweets in the participant representation. We obtained $\sim$1M participant tweets for the 3385 participants with 1282 median tweets and 1385 average tweets per participant. Hence, for each participant, we have a set of his/her FPL related tweets, and captain picks for 25 game weeks. [3]

6.2.2 Evaluation measure

Consistent with Goldstein et al.[3], we computed a crowd's wisdom score (**WS**) as the FPL score of a captain receiving the greatest number of votes from a crowd. For a crowd of participants, $G = \{U_1, U_2, \ldots, U_n\}$, we extracted their captain picks for a week w_{index} as $C_{index} = \{c_1, c_2, \ldots, c_n\}$ where c_i is a captain picked by participant U_i in week w_{index}. The wisdom score for a crowd is computed as,

$$WS = \frac{\sum_1^{25} Mod(C_{index})}{25} \tag{6.3}$$

Here, $Mod(C_{index})$ represents the corresponding real-world points of the individual captain receiving the most votes from the crowd in the $index$ game week. In case of a non-unique mode - i.e., for a tie, we selected one of these modes randomly. A crowd's wisdom score was the average of its weekly scores over the 25 game weeks considered in our analysis.

6.2.3 Results and analysis

We first show the results for tweet classification. We used ten-fold cross-validation for training and unseen labeled test data for validating the resulting classifier.

Table 6.1 shows the cross-validation results for tweet classification and Table 6.2 shows results for tweet classification on the unseen test data. We report the classifier's performance for both identifying popular (or differential) and non-popular (or non-differential)

[3] We have not provided unrestricted access to the dataset, as it contains actual tweets and usernames. However, the dataset is available from the corresponding author upon request.

Table 6.1: Tweet classification cross-validation results. Labels ‘0’ and ‘1’ indicate results for classifying a tweet to $\overline{P}(\overline{D})$ and $P(D)$, respectively. The classifier achieved 0.85 average F-score.

Strategies	Label	Precision	Recall	F-score
Popular	0	0.87	0.90	0.93
	1	0.97	0.66	0.76
Differential	0	0.97	0.67	0.79
	1	0.87	0.99	0.93

Table 6.2: Tweet classification results on the test dataset. Labels have the same meaning as in Table 6.1. The classifier achieved a 0.79 average F-score for labels ‘0’ and ‘1’.

Strategies	Label	Precision	Recall	F-score
Popular	0	0.86	0.92	0.89
	1	0.86	0.76	0.81
Differential	0	0.77	0.50	0.61
	1	0.78	0.92	0.84

judgments. The label ‘0’ indicates Popular choice (or Differential choice) tweets and label ‘1’ indicates non-popular choice (or non-differential choice) tweets. Rows with ‘1’ indicate the classifier’s performance for identifying Popular choice (or Differential choice) and Rows with ‘0’ the indicate classifier’s performance for identifying non-popular choice (or non-differential choice) tweets. For each participant, we computed the four U_i values with the count of tweets identified in ‘1’ and ‘0’ using Model 1 and Model 2 as follows: 1) n_P is the number of tweets identified by Model 1 in class ‘1’. 2) $\overline{n}_P$ is the number of tweets identified by Model 1 in class ‘0’. 3) n_D is the number of tweets identified by Model 2 in class ‘1’. 4) $\overline{n}_D$ is the number of tweets identified by Model 2 in class ‘0’. In other words, we ignored tweets for which the classifier was not able to decide which class (either ‘1’ or ‘0’) it belongs.

Model 1 achieved an 84.5% F-score for the Popular Choice tweet classification model and Model 2 achieved an 86% F-score for the Differential Choice tweet classification in cross-validation. A classification model with high training accuracy may also indicate over-fitting. To guard against over-fitting, we measured these results on test data that the classifier did not encounter while training. On test data, the models achieved adequate 85%

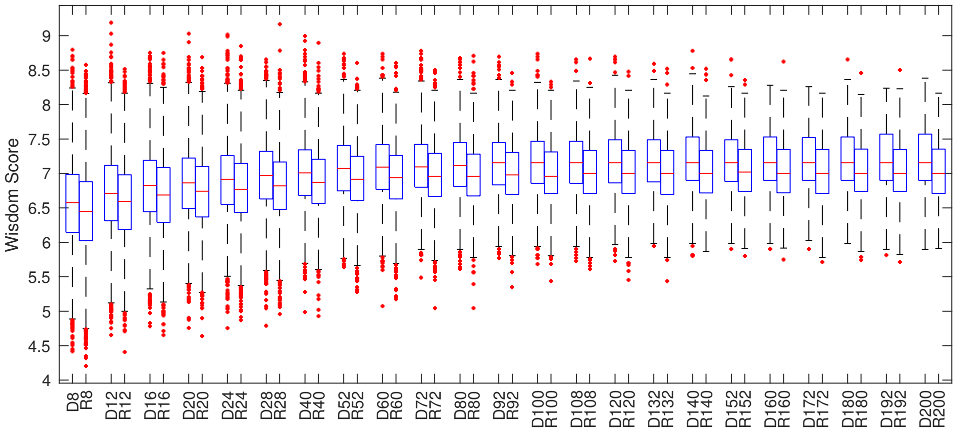


Figure 6.2: Box plots comparing wisdom scores of diverse crowds (D) and random crowds (R). Diverse crowds achieved a better wisdom score with a smaller standard deviation compared to random crowds.{D8,R8}: Diverse and Random crowd of size 8.

and 72.5% F-scores for Popular choice and Differential Choice, respectively which rules out over-fitting.

We used these classifier models to generate participant representations and select diverse crowds (see 9.1). Out of 3385 total participants, we had 895 participants identified as Type 1 (Popular choice) and 789 participants identified as Type 2 (Differential choice). Next, we evaluate diverse crowd formation using the *wisdom score* achieved by crowds selected from these types. As described in Section 6.1.3 we generated *diverse* crowds by randomly picking n participants from the two participant types. We generated l such crowds, where $l = 5000$ referred to as D (Diverse crowds). We compared these crowds with R (Random crowds), i.e., crowds generated by randomly selecting the same number of participants from the complete set of 3385 participants. Figure 6.2 shows box-plots for different crowd sizes. Here, crowd size is a multiple of 2 as we had two types of participant categories to generate diverse crowds. For each crowd, we computed its wisdom score based on the *Wisdom Score* formula resulting in two score lists (the diverse crowd score list and random crowd score list). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the upper and lower quartile, respectively for these lists. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. Diverse crowd lists always achieved

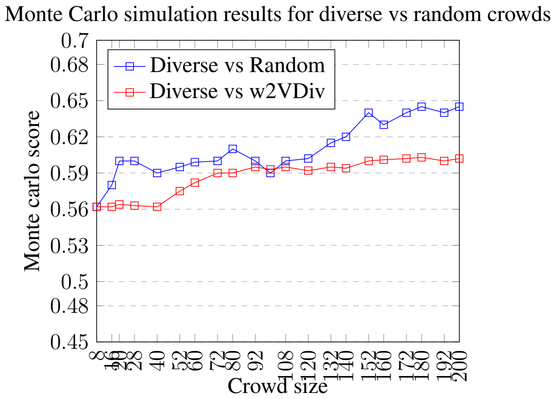


Figure 6.3: A diverse crowd consistently out performed Random crowds for group sizes 10 to 100. Our diverse crowds also outperformed crowds sampled using word2vec based diversity, w2VDiv.

a better median wisdom score than Random crowds for crowd sizes ranging from 8 to 200. Such an effect was also indicated by $p < 0.05$ for the t-test between Diverse crowds and Random crowds for each crowd size. We had a modest yet statistically significant effect for the diverse crowd out-performing Random crowds. As larger crowds tend to be more accurate than smaller crowds[2], we get better wisdom scores for large crowds than small crowds.

Figure 6.2 indicates that diverse crowds achieved a better median wisdom score than Random crowds. However, it is also of interest to know how likely a diverse crowd is to produce a better wisdom score than a Random crowd. We used Monte Carlo simulation for this purpose. Specifically, we randomly selected a single diverse crowd from the Diverse crowd set and a single random crowd from the Random crowd set. We then computed a *win* if the diverse crowd had a higher wisdom score than the random crowd. We repeated this for 1000 times and calculated a Monte Carlo simulation score as a ratio of the number of wins to 1000. A Monte Carlo simulation score of $\sim$ 0.5 indicates that the two sets of crowds are equally likely to beat each other. A Monte Carlo score of $\sim$ 1.0 suggests that a crowd from set one almost always beats a crowd from set two. Figure 6.3 shows the results for these Monte Carlo simulations.

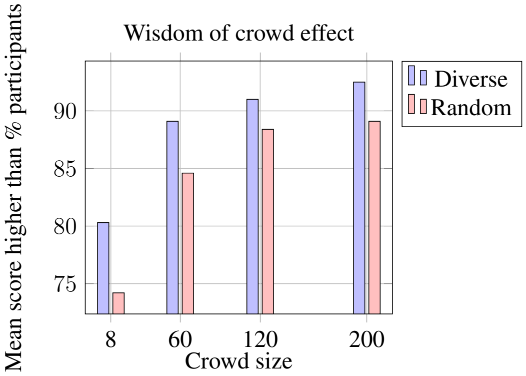


Figure 6.4: Wisdom of crowd effect. Diverse crowd of size 60 outperforms 89% of the individual participants.

We observed a ~ 0.63 Monte Carlo simulation score indicating that a Diverse crowd is 63% likely to outperform a Random crowd. We also compared our Diverse crowds generated from solution strategy to a diverse crowd formed using our previous word2vec based diverse crowd selection approach. For this analysis, we created word2vec representations of participants based on their tweets and generated $l \times 10 = 50,000$ random crowds. For each crowd, we computed an average pairwise cosine distance between participants of the crowd using their word2vec representations[70]. We selected the top 10% ($l = 5000$) of the crowds having the highest average pairwise distance, referred to as w2VDiv, and compared them with the Diverse crowds. A Monte Carlo simulation score of ~ 0.59 indicated that the proposed top-down, strategy-based method for assembling diverse crowds can assemble a better crowd than word2vec based diverse crowd selection.

Next, we evaluated our Diverse crowds regarding the *wisdom of crowd* effect. In other words, we measured the number of participants that a crowd, on an average, can outperform. Specifically, we computed the season score achieved by each participant (using the same formula as $\mathbf{WS}$) and found the number of participants that had a lower season score than an average wisdom score of a comparison crowd. Figure 6.4 plots the number of participants a crowd outperforms on an average.

On average, a random crowd of size 8 outperforms 74.2% of the individual participants

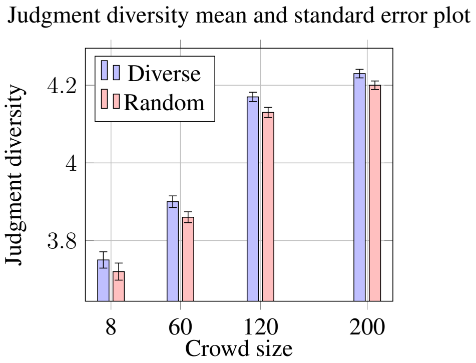


Figure 6.5: Judgment diversity comparison Diverse crowds and Random crowds. Diverse crowds had more judgment diversity than Random crowds.

while a diverse crowd of the same size outperforms 80.3% of the individual participants. On average, a diverse crowd of size 60 is better than almost 90% of the individual participants and approximates the performance of a random crowd more than three times as large.

We also examined whether diverse crowds produce diverse judgments. The intuition is that crowds producing diverse judgments likely imply a less biased sample of participants, which in turn likely yields a better-aggregated opinion. For this purpose, we used a *judgment diversity* measure proposed by Merayo et al.[21]. Formally the measure is defined as follow,

$$M = \frac{\sum_{i,j} d(u_i, u_j)}{n(n-1)} \tag{6.4}$$

Here, $d(u_i, u_j)$ is the difference between wisdom scores of participants u_i and u_j and n is the total number of participants in the crowd. Figure 9.3 shows the results for this metric for Diverse and Random crowds. We found that diverse crowds were more generally more diverse regarding the judgment diversity metric M. This suggests that a crowd of participants with a diverse player selection strategy ends up producing more diverse judgments than Random crowds. Our method assembles a crowd who can produce diverse and accurate judgment *using only their social media data*.

To explain this diversity, we used an English Premier League domain specific knowledge

Table 6.3: Highly ranked knowledge graph concepts subsuming both Popular choice and Differential choice keywords.

DBpedia Concept	**Score**
Eden_Hazard	0.34
Category:Chelsea_F.C._players	0.17
Romelu_Lukaku	0.17
Category:West_Bromwich_Albion_F.C._players	0.17
Category:Manchester_United_F.C._players	0.08
Midfielder(Winger)	0.05
Daniel_Sturridge	0.005

graph. We mapped the keywords identifying Popular choice and Differential choice strategies to the knowledge graph concepts using DBPedia lookup[4]. We then found concept scores using Equation 6.2. The concept that subsumes most of these keywords and are not far up in the concept hierarchy are ranked higher according Equation 6.2. We ranked each concept in the knowledge graph and sorted them in reverse order based on these scores.

Table 6.3 shows the resulting concept scores for a few concepts receiving high scores. We found *Eden Hazard* and *Romelu Lukaku*, two soccer players, subsuming both popular choice and differential choice features (keywords). These two players happened to be in the top 10 players scoring the most FPL points for the 2016-17 season. As our diverse crowd reflects both popular choice and differential choice strategies, they select a better player than Random crowds, albeit for different reasons. Hence, diversity in solution strategies leads to a better captain selection. Moreover, the concepts with high scores also help us interpret these two strategies. For FPL, we can determine whether diversity in solution strategy is related to the specific English Premier League teams or locations. We can also learn about teams whose players are chosen by both popular and differential choice. This information is helpful especially in deciding the kind of factors one should focus on in decision making so that the decision is not biased.

[4] https://wiki.dbpedia.org/lookup

6.3 Summary

Our strategy-based diversity framework can be used to interpret diversity in several domains, explaining the correlation between various domain features and collective intelligence. We also demonstrated that machine learning-based tweet classification methods work for classifying tweets by solution strategy. As the proposed approach only requires strategy characterization and training data, it applies to domains other than Fantasy Soccer.

The proposed diverse crowd selection achieved a statistically significant effect. Though of potential practical significance in some domains, the effect size was modest compared with simple random crowd selection strategy. One of the possible reasons is the limited number of strategies and the likely presence of additional strategic differences. Another explanation lies in the crowd selection strategies. The proposed methodology does not explore optimal crowd selection from the clusters. A crowd selection strategy that is consistent with the bionomial condition could potentially improve the consequent wisdom score the way it did for the word2vec based user characterization.

7 Future Work

while fantasy sports provide an ideal test bed for examining diversity-based approaches due to the availability of outcome measures, follow-on research should extend and validate these findings in other domains having more practical relevance, such as marketing, election prediction, and geopolitical forecasting.

Another facet of online communication not considered in the present study is sentiment. Twitter especially is a personality-driven medium featuring no shortage of affective content [71]. This sentiment content is likely to convey important predictive information about a user's future judgments.

These results suggest that, if deployed on more content-rich information sources – such as full-length blog postings, and articles – our methods may prove to be even more robust. Hence, future work may combine multiple data sources describing users, e.g., blog posts, different social media profiles, and prediction problem specific descriptions given by users.

The technique we have developed for captain selection can be extended to measure the wisdom of crowd effect in the choice of a whole team at the beginning of the season. The proposed approach can easily be extended to other Fantasy Sports.

The knowledge graph enhanced community detection improves community detection and characterization though it's just a scratch on the surface for using external knowledge in data-driven machine learning techniques. The current work opens up future direction for combining knowledge within the optimization to various machine learning and deep learning algorithms to achieve more accurate and explainable results.

8 Conclusions

A key contribution of this dissertation is *the demonstration of the successful use of social media data to perform wise crowd selection*. Traditional wisdom of crowd research studies either rely on the availability of performance data or require manual effort to compose diverse crowds. This dissertation shows that a crowd sampled using their openly available volunteered social media data allows wise crowd selection. Such a crowd outperforms random crowds and even crowds of experts. Further, there is a correlation between online social media conversations and individuals' judgments, with diversity in communications predicting sounder judgments. We also showed that random sampling may not always be the best sampling strategy especially for the applications that require diverse samples.

The proposed methodologies do not make domain-specific assumptions and hence can be extended to other domains. For instance, the bottom-up diversity quantification can be applied to geo-political forecasting domain. These techniques can be used to analyze wisdom of crowd effect for problems that involve the factors of skill and luck.

Fantasy Sports is a 7.2 billion dollar industry. Applying these techniques for a player prediction and a team prediction in Fantasy Sports can benefit individual users as well as Fantasy Sports organizers - to determine optimal price of the fantasy player.

The proposed knowledge graph based diversity explanation can be used to better explain machine learning algorithm feature importance. Specifically, the framework proposed in Chapter 6 can be used to explain random forest classification for a range of applications.

Another important contribution of this dissertation is knowledge graph enhanced community detection. This dissertation demonstrates that an optimization in machine learning that includes data and external knowledge can result in the improved results than considering only data. As demonstrated by the proposed community detection algorithm, appropriate inclusion of external knowledge in optimization can result in an explainable machine learning algorithm. The proposed community detection algorithm can serve as an excellent choice for network data exploration. In the context of wisdom of crowd, such a framework can be used to understand the kind of diversity that one needs to consider to make an accurate prediction. Accurate and explainable graph clustering is of an advantage to a number of domain specific applications. The proposed iterative optimization technique can be used for several domain specific leaning problems.

Finally, bottom-up diversity measures can identify a better preforming crowd than randomly selected crowds. However, contextual and independent crowds were found to have the most accurate prediction.

9 Knowledge-driven wisdom of crowd

The previous chapter described a top-down diversity measure, computed based on the solution strategies applied by individuals. These groups are often formed at highly abstract concept level. Moreover, the previous chapter did not consider an important criterion for the wisdom of crowd - influence. Twitter data contains influence information in terms of the follower as well as retweet relationship. A methodology considering both content and links can identify such a crowd. Hence, we need an algorithm that identifies contextually diverse crowd(s) that do not have individuals influencing each other. Moreover, we also need to explain the diversity measure. Traditional community detection and characterization algorithms fall short of this requirements as they do not consider the context. This chapter describes an approach that uses the context identified by a knowledge-graph and performs community detection and characterization to identify closely connected contextually similar groups of users. The proposed algorithm has applications beyond the wisdom of crowd analysis.

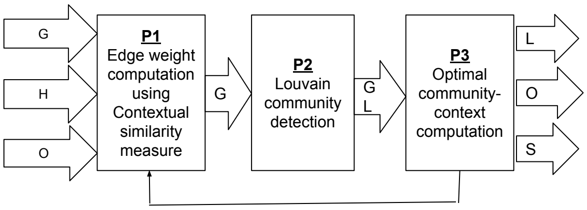

G
H
O
P1
Edge weight computation using Contextual similarity measure
G
P2
Louvain community detection
G
L
P3
Optimal community-context computation
L
O
S

Figure 9.1: Overview of the proposed approach. P1 computes contextual similarity between nodes and edge weights, inputs an updated graph to P2 which computes community labels (L). P3 computes community context O and concept weight vector S.

9.1 Approach

The proposed algorithm to generate community labels(communities) iteratively optimizes 1) community label assignment, keeping the community context constant and 2) community context assignment, keeping the community labels constant. We then recompute edge weights with the updated community context (O). Figure 9.1 summarizes this approach. Next, we describe the proposed contextual similarity measure (P1), community-context computation (P3), and the proposed way of integrating new node similarity values to find final community labels L and descriptions(O).

9.1.1 Contextual similarity measure

We describe our proposed similarity measure, $\phi(v_1, v_2, h_i, o_{ij})$, to compute a similarity score between nodes v_1 and v_2 in h_j with o_{ij}^{th} context. Here, i is the community to which the edge $v_1 - v_2$ belongs. Similarity is computed in the j^{th} domain-specific HKG. Note that similarity is computed in the context of o_{ij}, i.e., a hierarchy starting from o_{ij}. We extend this semantic similarity measure to compute similarity between two lists of concepts represented in a HKG. In a taxonomy with a given root node, similarity between two concepts can be computed [72], by finding the least common ancestor subsuming these concepts in the hierarchy. Similarity is the "informativeness" of that least common ancestor. More generic concepts provide less information. For example, in Figure 1.1, "USA" has less informativeness than "Ohio". Hence, the semantic similarity between "Cincinnati" and "Columbus" subsumed by Ohio is higher than "Columbus" and "Dallas" subsumed by USA. Informativeness, in its simplest form, is identified as $1 - \frac{\eta_i}{\eta_{root}}$ where η_i is number of concepts *subsumed* by i. Sanchez et al. proposed that inner HKG concepts should be evaluated separately from the leaves and revised informativeness formula as follows [72],

$$I_c = \left(2.0 - \frac{\sum_{l<c} \frac{1}{S_l}}{\sum_{l<root} \frac{1}{S_l}}\right) \tag{9.1}$$

Here, S_l refers to the number of concepts that subsumes l. The informativeness I of a concept c is summation of the subsumers over all leaves l such that $l < s$. We subtract the value from 2.0 as we want the values in (1.0, 2.0). In figure 1.2, $S_{Cincinnati} = 2$ and $S_{Columbus} = 2$ as they are subsumed by two concepts, “Cities in Ohio” and “Cities in USA”. Hence, $I_{CitiesinOhio} = 2 - \frac{\frac{1}{2}+\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}}$. The denominator has four terms corresponding to each one of the four leaves subsumed by “Cities in USA”(root).

As we have the nodes represented as a list of concepts, the existing similarity measure must find the least common ancestor of each pair of concepts from v_1 and v_2 and consider their informativeness score to compute semantic similarity. Instead, we compute the similarity between two lists. We extend each vertex list, v_1 and v_2, by recursively computing the subsuming “parents” of each concept $c \in v_i$ until o_{ij}. Along with each concept, we also compute its informativeness score. Consider an extended vertex list with concepts and informativeness score as v_{ext1} and v_{ext2}. Similarity is computed as the weighted Jacquard similarity [73] between v_{ext1} and v_{ext2}.

$$J(v_{ext1}, v_{ext2}) = \frac{\sum_l min(v^l_{ext1}, v^l_{ext2})}{\sum_l max(v^l_{ext1}, v^l_{ext2})} \tag{9.2}$$

Here, l represents vector dimensions. In our case, each one of these dimensions is a concept c and the value is its informativeness score. We chose weighted Jacquard similarity as it satisfies the following requirements. 1. v_1 and v_2 yield a low similarity value if they have fewer concepts in common. 2. v_1 and v_2 yield a low similarity value if the concepts are repeated a different number of times. If the concept c appears three times in v_{ext1} and four times in v_{ext2} then the numerator’s value for that concept will be less than the denominator leading to reduced similarity. 3. v_1 and v_2 yield a low similarity value if the concepts in common have less informativeness.

This similarity computation depends on o_{ij}, i.e., a concept of h_j^{th} knowledge graph representing community i. As an example, the similarity between $v_1 = \{Cincinnati\}$ and $v_2 = \{Columbus\}$ results in $v_{ext1} = \{(Cincinnati, 1.8), Ohio(1.6), USA(1.0), Columbus(0.0)\}$ and $v_{ext2} = \{Columbus(1.8), Ohio(1.6), USA(1.0), Cincinnati(0.0)\}$. The bracketed value is the informativeness score for each concept according to HKG in Figure 1.2. The weighted Jacquard between v_{ext1} and v_{ext2} results in similarity 0.419.

We used the Louvain algorithm to find community labels L for each node in the weighted graph. Next, we describe the process of finding an appropriate concept describing each community.

9.1.2 Optimal community context computation

In this subsection, we describe how we compute o_{ij}, an optimal context of $h_j \in H$ describing community $i \in C$. As described, context o_{ij} is essentially a hierarchy starting at concept o_{ij} in h_j. Hence, o_{ij} is represented by the concept $c \in h_j$ that is the most relevant concept for the community c. Such a concept is found based on two criteria: 1. appropriate generality (referred as purity) of a concept and 2. informativeness. Next, we describe the detailed procedure.

h_j hierarchies provide *real-world clustering knowledge*. As an example, in the context of “Cities in USA”, “Austin”, “Dallas”, and “Houston” forms the cluster “Cities in Texas”. In other words, as “Cities in Texas” subsumes three cities, it can represent and even validate these three cities being in one cluster. Each concept of h_j can potentially represent a community i based on node attribute values of nodes belonging to a community i. Our intuition for finding such a concept is as follows. *For any community, a concept can represent that community if it happens to subsume more concepts in a community than* *if the concepts of the community were distributed at random in a HKG*. As described above in Section 9.1.1, the use of $2 - informativeness$ can serve as a better approximation for “concepts distributed at random” than $\frac{\eta_i}{\eta_{root}}$. Hence, maximizing the following with respect to the concept of a knowledge graph can indicate the optimal context representing a community,

$$max_c \left(\eta_c - \eta_c \times \frac{\sum_{l<c} \frac{1}{S_c}}{\sum_{c<root} \frac{1}{S_l}} \right) \tag{9.3}$$

Here, η_i is the number of concepts (belonging to a community i) subsumed by c. We also minimize the number of concepts subsumed from neighboring communities. Considering this and rearranging terms, the final maximization term is:

$$o_{ij} = max_c \left((\eta_{n \in i} - \eta_{n \in \bar{i}}) \times I_c \right) \tag{9.4}$$

where $\eta_{n \in i}$ indicates the number of concepts in i subsumed by c and $\eta_{n \in \bar{i}}$ indicates the number of concepts in the neighboring communities of i subsumed by c. The first term corresponds to “purity” while the second term corresponds to the informativeness of c. In addition to the concept c maximizing the score, we also retain the actual score as s_{ij} which indicates the relative context importance of context j in community i.

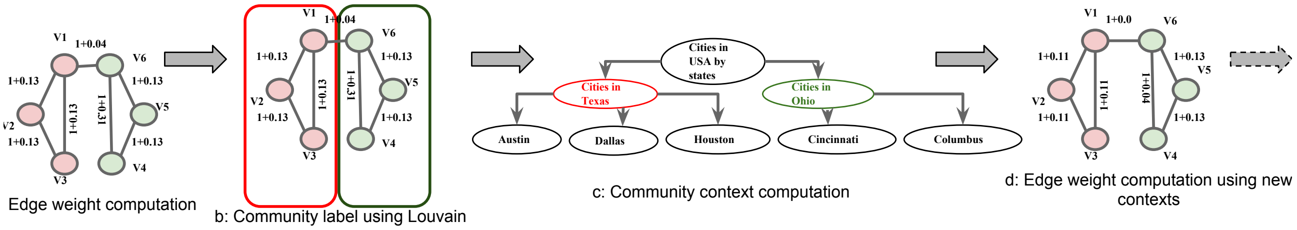
For the attribute list $T = \{c_1, c_2, \ldots, c_f\}$ and $\bar{T} = \{\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_f\}$ indicate the concepts of community i and neighboring communities in h_j respectively, Algorithm 3 finds the concept maximizing Equation 9.4. We pre-compute the hierarchical level, e.g., the root is set to ‘0’ and all the leaves are at level “tree height ” and the informativeness of each $c \in h_j$. We create a list with concepts at the lowest level and a score associated with each concept indicating the difference between the number of concepts each subsumes from T and number of concepts it subsumes from $\bar{T}$. Then, we compute a score for each concept and update the concept with the maximum score thus far and the maximum score.

Next hop "parents", i.e., concepts subsuming the current concept, are included in the list to investigate. The scores associated with the parent concepts are also attached as it indicates the number of concepts subsumed from T and $\bar{T}$. As a vertex may be represented with concepts other than leaves, there may be some concepts left in T and $\bar{T}$ that belong to higher level. They are added using add_list whenever the level that they belong to is processed. Because the root has no parents, the temp_list will eventually become empty. To avoid loops, we also condition on $level \geq 0$.

```
Input: $T$, $\bar{T}$, and $h_j$
Output: $c_{opt}$, $s_{max}$
$c_{opt} = root$, $s_{max} = root\_score$
Associate score with each concept. -1 for $\bar{T}$ and 1 for $T$
level = lowest_level()
list = add_list($T$, $\bar{T}$, level)
while list not empty and level $\geq$ 0 do
    temp_list = empty
    for $c \in$ list do
        $s_{cur}$ = score(c) $\times I_c$
        update_optimal($c_{opt}$, $s_{max}$, $c$, $s_{cur}$)
        for $p \in$ parents(c) do
            add_parent(temp_list, p, score(c))
        end
    end
    level = level - 1
    add_list($T$, $\bar{T}$, level)
    list = temp_list
end
```

Algorithm 3: Optimal community-context computation

One of the most important steps in the algorithm is add_parent. The concept maximizing the criteria must subsume at least one of the concept of i. Thus, we explore for a solution among hierarchical "parents" of any $c \in T$. We avoid adding a parent (stop looking for a solution in the path) if its informativeness score decreases so much so that even if it were to subsume rest of the remaining concepts, it could not get a higher score than max_score.


Edge weight computation

b: Community label using Louvain

c: Community context computation

d: Edge weight computation using new contexts

Figure 9.2: Demonstration on an example network. (a) Normalized edge weights are first computed using $\omega = 1.0$ and contextual similarity kernel with root node as the context identifying each community. (b) Community labels are computed using Louvain. (c) Optimal contexts "cities in texas' and "cities in ohio" computed for c1, and c2 respectively, (d) Normalized edge weights recomputed using new contexts. Note the modularity increase from (a) to (d).

9.1.3 Unified framework

Algorithm 4 describes the final algorithm and Figure 9.2 demonstrates the algorithm on an example network shown in Figure 1.1. We start by computing node pair similarities between all nodes for which $E_{ij} \neq 0$. We consider each edge ij as an edge from i's community to j's community. Hence, edge weight E_{ij} is computed with contexts for both communities L_i and L_j. Next, it computes community labels L by maximizing a modularity equation with respect to L. Note that $f(ij, L)$ is a function that determines whether i and j are in the same community based on their community labels. Specifically, $ij \in l$ iff $L_i = L_j = l$.

Modularity is an evaluative measure of community structure. Accordingly, *a part of graph (a group of nodes) is interesting if the number of edges within that group is higher than if the nodes were to assign into groups at random*, formally: $\sum_{i=1}^{k} (e_i - a_i^2)$. Here, e_i is the number of edges in a community i, and a_i is the expected number of edges in community i. Note that we used a similar idea in designing our optimal community context computation.

$\frac{k_i \times k_j}{4m^2}$ provides a better estimation of a_i^2 as the probability of an edge belonging within a community depends on the degree of nodes connected that edge [74]. Modularity maximization is one of the most widely used community detection technique. We used Louvain algorithm based modularity maximization as it has identified qualitatively robust community structure[20]. It is a greedy algorithm that processes each vertex at random and assigns a community label based on the one that can result in the maximum modularity gain. Details appear in [20].

Using the newly located community labels, we compute the optimal context representing each community $i \in K$. The process is repeated until maximum modularity is achieved or a max number of iterations. d_i in modularity $Q(E, L)$ is the degree of a node i, computed

```
Input: G=(V, E, A), $H = \{h_1, h_2, \dots, h_t\}, max\_iters, threshold$
Output: $L$, $O$, $S$
while mod < threshold or until max_iters do
    $w_{ij}(\omega, i, j, H, O, l) = \omega + \sum_{q=1}^{t} \phi(i, j, h_q, o_q, l)$
    $E_{ij} = w_{ij}(\omega, i, j, H, O, L_i) + w_{ij}(\omega, i, j, H, O, L_j)$
    $Q(E, L) = \frac{1}{2m} \sum_{l \epsilon K} \sum_{f(ij,L)} E_{ij} - \frac{d_i d_j}{4m^2}$
    $L = max_L Q(E, L)$
    for $i \in K$ do
        for $h_j \in H$ do
            $o_{ij} = max_c \left( (\eta_{n \in i} - \eta_{n \in \bar{i}}) \times I_c \right)$
        end
    end
end
```

Algorithm 4: Community detection and characterization algorithm

as the summation of all edges incident on i and m is the summation of all edge weights. ω is a hyper-parameter indicating the relative importance of edge to the node pair similarity computed using the contextual similarity. We iteratively update community label assignment and community-context vector o_i for each community i. Such an algorithm is likely to be stuck in local maxima. Thus, we repeated the process 10 times for each dataset, randomly selecting the vertex order to be processed by the Louvain algorithm. We consider the result that for which we achieved the maximum modularity value.

9.1.4 Algorithm complexity and convergence

Each iteration consists of the three steps (P1, P2, and P3) described in Figure 9.1. P1 and P2 process each edge resulting in $O(n)$ time complexity where n indicates the number of edges. P3 maps nodes from each community to a knowledge graph and computes an optimal context for each community resulting in the time complexity of $O(ck)$ for c communities and knowledge graph of k concepts. Hence, the time complexity of the algorithm is $O(n + kc)$ as the number of iterations $i << n$.

The Louvain algorithm (P2) optimizes Modularity to find a community structure. The algorithm could diverge if the optimal community context results in edge weights that could decrease Modularity. A generic community context will result in relatively less similarity and indicates that the communities should be computed only using ω that won't affect the Modularity value. A specific community context will change the edge weights to make the current community structure stronger. Hence, it is likely to increase the modularity value. Either way, the modularity value is not expected to decrease due to P3. We also found the algorithm converges to a satisfactory modularity value for all of datasets used in our experiments.

9.2 Results

9.3 Evaluation

The datasets, measures, comparison baselines and results follow. We refer to the proposed approach as "KDComm".

9.3.1 Datasets

We used four datasets to assess community detection accuracy and community structure characterization.

G+ ego network

This is a G+ user dataset with friends of a given user represented as nodes and friendship relationship represented as an edge[75]. Circles (communities) result from densely-connected sets of friends [74]. Each node has four features: job title, current place, university, and workplace. A user-pair(edge) is compared using knowledge graphs based on, *Category:Occupations*, *Category:Companies_by_country_and_industry*, *Category:Countries*, *Category:Universities_and_college*

Twitter

The Twitter dataset consisted of tweets about the configuration of a team for the Fantasy Premier League (FPL). We created a re-tweet network between these users based on information about their tweets. The re-tweet network between these users represents agreement. We used DBpedia spotlight [76] to identify soccer player mentions in these tweets. The final network consisted of users as nodes, re-tweet as edges, and FPL players mentioned by a user as node attributes.

These users have different types of teams where they select players of one position more than the others. These types include 1. Forwards, 2. Defenders, 3. Mid-fielders. As they discuss their players in their FPL related tweets, a dense re-tweet network between these users with community type characterization indicates a group of users interested in similar types of teams. Hence, given a network of these users, the task divides users into three circles — users with more "Forward" players in their team, more "Defender" players in their team, and more "Mid-fielder" players in their team. For KDComm, we generated three HKGs with following root nodes, Category:Association_football_defenders, Category:Association_f and Category:Association_football_midfielders.

We created ground truth circles using these users' actual team configurations available on the FPL website[70]. Users with more than the usual [1] number of players for any position is included in that circle [2].

DBLP

The DBLP dataset [77] is a co-author network, where each author is characterized by a set of keywords. Ground truth labels for authors are available for four categories: 1. Machine learning, 2. Data mining, 3. Databases, and 4. Information retrieval. We use a knowledge graph generated with root nodes *Category:Data_Mining*, *Category:Machine_Learning*, *Category:Databases*, and *Category:Category:Information_retrieval*.

Reddit

Each node in this dataset is a user, an edge indicates users are commenting/replying to the same post, and a node attribute is a set of comments made by that user. Each post has a “sub-reddit” that indicates the type of a post. The communities in this network can be evaluated using each user’s subreddits. Users belonging to the same community are likely to discuss the same sub-reddits [78] We considered the first four days of April 2015 to create this network[3]. We considered subreddits related to Economics and the NFL as they were the most discussed subreddits in the dataset. The domain-specific HKGs were extracted for *Category:Economics* and *Category:National_Football_League* as root nodes.

9.3.2 Evaluation measures

To evaluate community detection accuracy in G+, DBLP, and Twitter datasets, we used Yang et al. ’s community F-Measure and a Jacquard measure [79]. The evaluation

[1] http://www.soccer-training-guide.com/soccer-formations.html#.Wmk6GZM-eAI

[2] Please contact the corresponding author for the dataset.

[3] https://archive.org/details/2015_reddit_comments_corpus

function is,

$$\frac{1}{2\left|C^{*}\right|}\sum_{C_i^{*}\in C^{*}} C_j \overset{max}{\in} C\delta\left(C_i^{*},C_j\right)+\frac{1}{2\left|C\right|}\sum_{C_j\in C} C_i^{*} \overset{max}{\in} C^{*}\delta\left(C_i^{*},C_j\right) \tag{9.5}$$

Here, $\delta(C_i^{*},C_j)$ is a similarity measure, either Jacquard or F-score similarity (F-Measure). C is the community label set found by the algorithm and C^{*} is the ground truth community label set. For community detection evaluation in Reddit dataset, we used Hartman et al.'s rank entropy measure for a given community $R_e = \frac{-\sum_{j=1}^{L}\frac{n_{cj}}{n_c}log_2\frac{n_{cj}}{n_c}}{log_2 n_c}$. Here, j is a subreddit in a community c. n_{cj} is the number of times users of community i comment on subreddit j. n_c is total comments. A community c is likely to have a lower entropy value if the users of community c are commenting on a few subreddits most of the time.

9.3.3 Results and analysis

To evaluate KDComm, we use Liu et al.'s CPCD approach, which is superior to eight other community detection[42]. We also consider JCDC [16] which outperforms five other community detection approaches. Like CPCD, JCDC concerns edge weights based on

We used UNCut [43], which outperforms three other graph clustering approaches. We used Newman's community detection approach (referred to as SI) [15] that also uses attribute values in community structure detection and characterization. Finally, we also compared results with the Louvain algorithm, using only edge information. Evaluation results appear below for: 1. the similarity kernel. 2. community detection accuracy, 3. community structure characterization.

Contextual similarity measure evaluation

First we compare the proposed contextual similarity measure (referred as KGsim) with attribute value-based similarity. Two sets of user pairs ($n = 1000$) are created from four

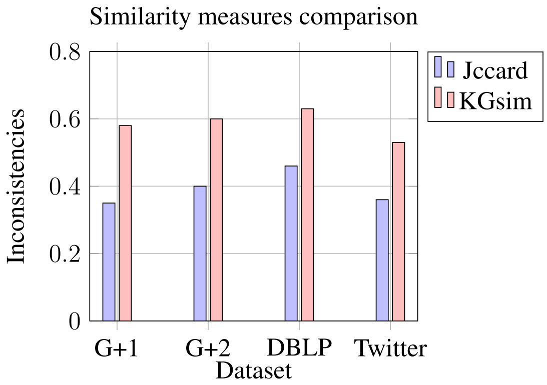


Figure 9.3: Similarity measures comparison. KGsim was able to assign appropriate edge weights to node pairs, resulting in lower inconsistencies corresponding to community labels.

datasets with ground truth community labels. IntraCommunitySet $= \{s_1, s_2, \ldots, s_n\}$ where each $s_i = \{(u_1, u_2)|u_1 and u_2 \in same community\}$. InterCommunitySet = $\{d_1, d_2, \ldots, d_n\}$ where each $d_i = \{(u_1, u_2)|u_1 and u_2 \in different communities\}$. We compared each $s_i \in$ IntraCommunitySet to all the $d_i \in$ InterCommunitySet resulting with n^2 comparisons. Ideally, each $s_i \in$ the IntraCommunitySet should be higher than all the $d_i \in$ the InterCommunitySet. The number of times that $s_i \in$ IntraCommunitySet is lower than $d_i \in$ InterCommunitySet is computed as number of "inconsistencies". We computed similarity using the proposed similarity measure and Jacquard similarity as Jacquard computes similarity using attribute values. Figure 9.3 plots the "inconsistencies" to the total comparison (n^2) ratio.

For the G+1 and G+2 datasets, we used the four features associated with each node as attribute values. For Twitter and DBLP, we used player names and author keywords respectively as attribute values. The proposed similarity measure (KGsim) had lower "inconsistencies" than Jacquard for all four datasets. Hence, KGsim can best assist edge re-weighting. We did not compute an appropriate context relevant to each community and used the "root" node as the context for each dataset.

Algorithm	DBLP		G+		Twitter		Reddit
	F	Jcc	F	Jcc	F	Jcc	R_e
Louvain	0.45	0.40	0.53	0.45	0.30	0.25	0.78
UNCut	0.57	0.51	0.5	0.42	0.35	0.30	0.75
CPCD	0.58	0.49	0.56	0.46	0.34	0.29	0.68
JCDC	0.54	0.5	0.58	0.48	0.33	0.28	0.62
SI	0.56	0.48	0.6	0.53	0.38	0.31	0.63
KDComm	0.66	0.59	0.71	0.60	0.47	0.39	0.48

Table 9.1: Community detection accuracy results. KDComm achieved the best F-score and Jacquard score for all three datasets.

Community detection accuracy

We compared community detection accuracy to other approaches. CPCD, SI and, UNcut used nominal node attribute values in the form of a 1/0 vector. We focused on the 100 most frequently used words of Reddit forums as attribute value vectors. For JCDC, we used the Jacquard similarity measure to compute similarities. Table 9.1 shows the results for the four datasets. The F-Measure and Jacquard scores reported for G+ are averaged over all the 20 ego networks. The proposed approach achieved better average scores for both measures (F-score and Jacquard) than all other approaches. For comparison with the G+ ego network dataset, we also performed a t-test between the set of F-scores received by KDComm and set of F-scores received by other approaches. A $p - value < 0.05$ also indicated superior performance of KDComm over all other baseline methods. Similarly, A $p - value < 0.05$ for Jacquard measure comparison confirms superior performance.

KDComm achieved the best F-score and Jacquard for the Twitter dataset, dividing users into three communities. As the Louvain algorithm found more than three communities, we merged communities based on community-context scores, merging users divided into two different "Defender" communities. KDComm also outperformed the other methods for the DBLP dataset as well, requiring similar community merging.

For the Reddit dataset, UNCut, CPCD, JCDC, and SI require a pre-determined number communities. We set the number of communities using KDComm. We report rank entropy

Dataset	JCDC		SI		KDComm	
	M_{11}	M_{22}	M_{11}	M_{22}	M_{11}	M_{22}
Twitter	0.168	0.154	0.41	0.285	0.6	0.7
G+1	0.56	0.381	0.36	0.263	0.7	0.8
G+2	0.482	0.58	0.7	0.536	0.6	0.75
DBLP	0.32	0.232	0.56	0.377	0.56	0.64

Table 9.2: Users within community characterization. M is a relevancy score matrix. KDComm found appropriate topics characterizing users within a community for all four datasets while JCDC found appropriate topics for two datasets.

averaged over all the communities. Lower entropy indicates a better community structure according to this measure [78]. KDComm achieved the lowest entropy among all methods.

Characterization of community structure

Next, we evaluated whether KDComm characterized users belonging to different communities with an appropriate community type. For each dataset, we considered users from two communities and evaluated whether KDComm, SI, and JCDC can find underlying two communities and compute an appropriate type of community-based node attributes. We considered attributes such that attribute type can identify community type. All the three methods(KDComm, SI, and JCDC) compute a "relevancy score" of each attribute type to each community, E.g., S for KDcomm. These "relevancy scores" for two attribute types and two communities can be represented as a 2x2 matrix, M. Each cell of this matrix indicates the relevancy score of attribute type to a community.

The relatively larger score for an attribute type indicates greater importance for that attribute type. All four datasets had community type and labels for nodes. We selected Twitter users from "Forwards" & "Defenders" communities, G+ users from "University" & "Workplace" communities and DBLP authors from "Data Mining" & "Machine Learning" communities. We considered two G+ ego networks (referred to as G+1 and G+2) for which we distinguished two ground truth communities based on "University" and "Workplace" attributes/contexts.

As the inputs were provided with two contexts/attribute types, a correct attribute type assignment is reflected by a higher score assigned to that attribute type relative to the other attribute type. As we used a normalized attribute/context score for each method, a score > 0.5 indicates a particular attribute type as the community type. We had attribute type 1, "forwards", "University", "Data Mining" as more relevant to Twitter, G+, and DBLP datasets' community one according to ground truth. We expect a context1 (T1) score higher than 0.5 for community 1 and a context2 (T2) score higher than 0.5 for community 2. Hence, we expect the relevancy score matrix M_{11} and M_{22} to be higher than 0.5. KDComm found the expected community-context scores for all the four datasets (see Table 9.2). Both JCDC and SI failed to find the expected community-context scores for at least two datasets.

9.3.4 Wisdom of crowds

Here we describe how we employed the proposed algorithm for the wisdom of crowd analysis. We considered the Twitter network and performed community detection considering two contexts,

- Soccer Positions: HKG of Defenders, Forward, and Mid-fielders as described in the Twitter dataset.
- Soccer Teams: HKG of soccer teams collected using the automatic hierarchical knowledge graph extraction framework[19].

We performed community detection and characterization using these two contexts. From the resulting communities, we formed 100 diverse crowds of size six by randomly picking two users from each type of community. Here, type of community refers to the type (specific soccer position or a team) for which the community had a maximum s_{ij} score. We explored both sets of community semantics: DiversePositions and DiverseTeams.

Crowd selection	Avg higher than % users
Random	76%
BottomUp	78%
DiversePositions	81%
DiverseTeams	86%

Table 9.3: Diversity based crowd selection and wisdom of crowd. DiverseTeams crowds outperform individual users 81% and 86% of the time depending upon community semantics.

Table 9.3 shows results for the number of individual users that a crowd outperformed on an average. Specifically, an average captain score of DiversePositions and DiverseTeams was compared with the captain score achieved by individual crowd members. We also created one more set of Random crowds, by selecting 100 crowds of six individuals at random. We found that a random crowd, on an average, performed better than 75% of the individual users. However, the DiversePositions crowd set outperformed 81% of the individual crowd members, and the DiverseTeams outperformed 86% of the individual crowd members. As the knowledge-driven community detection resulted in contextually diverse and independent crowds, these crowds achieved better scores than the crowds formed using bottom-up diversity measures introduced in previous chapters. Note that the crowd selection using multi-objective optimization wasn't performed.

9.3.5 School student communication network analysis

The proposed algorithm improves the state-of-the-art community detection and hence can be used for several applications. Here, we use this algorithm for the analysis of a network that resulted from high school students' Twitter conversation. We explore whether certain topics/contexts form a dense conversation community structure and contribute to the identification and characterization of insider-outsider [80], phenomena that contribute to harassment potential. We crawled for 388 high school students' tweets and had each student as a node, a mention or reply as an edge, and relevant domain-words from tweets as node attributes.

C(size)	Sports(Relevancy)	Music(Relevancy)
C1(47)	U.S. Women's soccer(0.36)	Bob Marley(0.64)
C2(40)	Cleveland Browns(0.45)	Keke Palmer(0.55)
C3(38)	American Football in Boston(0.39)	Machine Gun Kelly(0.61)

Table 9.4: Top 3 communities identified using Sports and Music contexts. Community description in Sports and Music contexts provided along with the normalized relevancy scores. Music context was found to be more relevant in creating community structure.

We explored two contexts, American Sports and American Music, to find whether they form modular conversational communities. First, we analyzed the conversation network without considering node attributes. The final modularity value of 0.32 does indicate a community structure based on edges alone. However, using a domain-specific knowledge graph created with "Category:Sports_in_the United_States", we also generated node (student) attributes as domain relevant concepts characterizing each node and performed the proposed community detection. We discovered community structure with improved modularity of 0.35. Similar processing with "Category:American_music" resulted in community structure with a higher modularity score of 0.38. Next, we used both contexts in community detection. We found a slightly better modularity score of 0.4. All of the modularity scores improve with node attributes, supporting the claim that the proposed algorithm favored American music (more informative context) and downplayed Sports (less informative context).

As described in Table 9.4, community-context relevancy scores also indicated that Music was more informative in finding community structure than sports. It also provides the most relevant contexts associated with four of the largest communities. To analyze the divergence from the edge-based community structure, we computed the F-measure defined in the evaluation. An F-score of 0.38 between edge-based community structure and community structure with both contexts suggested a divergence in assigning community labels to nodes in the presence of the contexts. Hence, contextual analysis has the potential to improve insider-outsider identification and characterization (with contexts identified for communities). Isolated nodes (student) suggest harassment potential [81]. Moreover, by characterizing context, the approach can also provide the foundation for predicting the harassment potential for a new node not considered in the original community detection.

9.4 Summary

This chapter presented an algorithm to incorporate hierarchical concepts about node attributes into community detection. Our core contributions include (1) a combined metric that describes concept informativeness in the hierarchy and concept purity in summarizing communities, which are used to guide the search for optimal concept generalization; (2) a node similarity measure that synthesizes multiple generalized concepts for community detection; and (3) a community detection algorithm that alternatively optimizes concept generalization and community structures. Our evaluation results showed that concept generalization can not only improve the quality of community detection, but also provides a meaning-oriented characterization of community structure. The results vary depending on the choice of domains and knowledge sources. The chapter also demonstrated that readily available and automatically extracted knowledge source can also have vital improvements on data-driven Machine Learning algorithms.

The chapter also demonstrated that the knowledge driven approach achieved was able to assemble a crowd that was better than 87% of the individuals and it can be applied for other network analysis tasks.

REFERENCES

[1] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[2] J. Surowiecki, *The wisdom of crowds.* Anchor, 2005.

[3] D. G. Goldstein, R. P. McAfee, and S. Suri, "The wisdom of smaller, smarter crowds," in *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 2014, pp. 471–488.

[4] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 46, pp. 16 385–16 389, 2004.

[5] C. P. Davis-Stober, D. V. Budescu, J. Dana, and S. B. Broomell, "When is a crowd wise?" *Decision*, vol. 1, no. 2, p. 79, 2014.

[6] H. V. D. Parunak and E. Downs, "Estimating diversity among forecaster models," *Ann Arbor*, vol. 1001, p. 48105, 2012.

[7] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181–207, 2003.

[8] E. Nevo, "Evolution of genome–phenome diversity under environmental stress," *Proceedings of the National Academy of Sciences*, vol. 98, no. 11, pp. 6233–6240, 2001.

[9] A. R. Solow and S. Polasky, “Measuring biological diversity,” *Environmental and Ecological Statistics*, vol. 1, no. 2, pp. 95–103, 1994.

[10] A. Stirling, “A general framework for analysing diversity in science, technology and society,” *Journal of the Royal Society Interface*, vol. 4, no. 15, pp. 707–719, 2007.

[11] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, “How social influence can undermine the wisdom of crowd effect,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9020–9025, 2011.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” *The semantic web*, pp. 722–735, 2007.

[15] M. E. Newman and A. Clauset, “Structure and inference in annotated networks,” *Nature communications*, vol. 7, 2016.

[16] Y. Zhang, E. Levina, J. Zhu *et al.*, “Community detection in networks with node features,” *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3153–3178, 2016.

[17] G. M. Namata, B. Staats, L. Getoor, and B. Shneiderman, “A dual-view approach to interactive network visualization,” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 939–942.

[18] P. Wang, J. Guo, and Y. Lan, “Modeling retail transaction data for personalized shopping recommendation,” in *Proceedings of the 23rd ACM international conference*

on conference on information and knowledge management. ACM, 2014, pp. 1979–1982.

[19] S. Lalithsena, P. Kapanipathi, and A. Sheth, “Harnessing relationships for domain-specific subgraph extraction: A recommendation use case,” in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 706–715.

[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and É. Lefebvre, “The louvain method for community detection in large networks,” *J of Statistical Mechanics: Theory and Experiment*, vol. 10, p. P10008, 2011.

[21] M. G. Merayo, N. T. Nguyen *et al.*, “Intelligent collective: The role of diversity and collective cardinality,” in *Conference on Computational Collective Intelligence Technologies and Applications*. Springer, 2017, pp. 83–92.

[22] I. Lorge, D. Fox, J. Davitz, and M. Brenner, “A survey of studies contrasting the quality of group performance and individual performance, 1920-1957.” *Psychological bulletin*, vol. 55, no. 6, p. 337, 1958.

[23] M. E. Shaw, *Group dynamics: The psychology of small group behavior*. McGraw-Hill College, 1981.

[24] L. P. Robert and D. M. Romero, “The influence of diversity and experience on the effects of crowd size,” *JAIST*, 2017.

[25] H. Olsson and J. Loveday, “A comparison of small crowd selection methods.” in *CogSci*, 2015.

[26] B. Mellers, E. Stone, P. Atanasov, N. Rohrbaugh, S. E. Metz, L. Ungar, M. M. Bishop, M. Horowitz, E. Merkle, and P. Tetlock, “The psychology of intelligence analysis: Drivers of prediction accuracy in world politics.” *Journal of experimental psychology: Applied*, vol. 21, no. 1, p. 1, 2015.

[27] T. Ye and L. P. Robert Jr, “Does collectivism inhibit individual creativity?: The effects of collectivism and perceived diversity on individual creativity and satisfaction in virtual ideation teams,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017, pp. 2344–2358.

[28] D. Loyd, C. Wang, K. Phillips, and R. Lount, “Social category diversity and pre-meeting elaboration,” *Organization Science*, pp. 1–16, 2013.

[29] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann *et al.*, “Life in the network: the coming age of computational social science,” *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[30] S. Pavoine, S. Ollier, and D. Pontier, “Measuring diversity from dissimilarities with rao’s quadratic entropy: Are any dissimilarities suitable?” *Theoretical population biology*, vol. 67, no. 4, pp. 231–239, 2005.

[31] H. Hong, Q. Du, G. Wang, W. Fan, and D. Xu, “Crowd wisdom: The impact of opinion diversity and participant independence on crowd performance,” 2016.

[32] L. Robert and D. M. Romero, “Crowd size, diversity and performance,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1379–1382.

[33] R. Ren and B. Yan, “Crowd diversity and performance in wikipedia: The mediating effects of task conflict and communication,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 6342–6351.

[34] M. J. Fry, A. W. Lundberg, and J. W. Ohlmann, “A player selection heuristic for a sports league draft,” *Journal of Quantitative Analysis in Sports*, vol. 3, no. 2, 2007.

[35] D. Bergman and J. Imbrogno, “Surviving a national football league survivor pool,” *Operations Research*, vol. 65, no. 5, pp. 1343–1354, 2017.

[36] A. Becker and X. A. Sun, “An analytical approach for fantasy football draft and lineup management,” *Journal of Quantitative Analysis in Sports*, vol. 12, no. 1, pp. 17–30, 2016.

[37] E. H. Kaplan and S. J. Garstka, “March madness and the office pool,” *Management Science*, vol. 47, no. 3, pp. 369–382, 2001.

[38] B. Clair and D. Letscher, “Optimal strategies for sports betting pools,” *Operations Research*, vol. 55, no. 6, pp. 1163–1177, 2007.

[39] D. S. Hunter, J. P. Vielma, and T. Zaman, “Picking winners using integer programming,” *arXiv preprint arXiv:1604.01455*, 2016.

[40] M. B. Haugh and R. Singal, “How to play fantasy sports strategically (and win),” 2018.

[41] C. Bothorel, J. D. Cruz, M. Magnani, and B. Micenkova, “Clustering attributed graphs: models, measures and methods,” *Network Science*, vol. 3, no. 3, pp. 408–444, 2015.

[42] L. Liu, L. Xu, Z. Wangy, and E. Chen, “Community detection based on structure and content: A content propagation perspective,” in *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 271–280.

[43] W. Ye, L. Zhou, X. Sun, C. Plant, and C. Böhm, “Attributed graph clustering with unimodal normalized cut,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 601–616.

[44] Y.-S. Cho, G. Ver Steeg, E. Ferrara, and A. Galstyan, “Latent space model for multi-modal social data,” in *Proceedings of the 25th International Conference on*

World Wide Web. International World Wide Web Conferences Steering Committee, 2016, pp. 447–458.

[45] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, and X. Li, "The author-topic-community model for author interest profiling and community discovery," *Knowledge and Information Systems*, vol. 44, no. 2, pp. 359–383, 2015.

[46] T. Ho and P. Do, "Discovering communities of users on social networks based on topic model combined with kohonen network," in *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*. IEEE, 2015, pp. 268–273.

[47] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," 2017. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14612

[48] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," 2016. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11964

[49] A. El Kassiri and F.-Z. Belouadha, "Towards a unified semantic model for online social networks analysis and interoperability," in *Intelligent Systems: Theories and Applications (SITA), 2015 10th International Conference on*. IEEE, 2015, pp. 1–6.

[50] S. Pool, F. Bonchi, and M. v. Leeuwen, "Description-driven community detection," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, p. 28, 2014.

[51] G. Erétéo, M. Buffa, F. Gandon, and O. Corby, "Analysis of a real online social network using semantic web frameworks," *The Semantic Web-ISWC 2009*, pp. 180–195, 2009.

[52] G. Palma, M.-E. Vidal, and L. Raschid, “Drug-target interaction prediction using semantic similarity and edge partitioning,” in *International Semantic Web Conference*. Springer, 2014, pp. 131–146.

[53] C. Wang, Y. Song, A. El-Kishky, D. Roth, M. Zhang, and J. Han, “Incorporating world knowledge to document clustering via heterogeneous information networks,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1215–1224.

[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[55] T. Mikolov, W.-t. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[56] G. Zarrella, J. Henderson, E. M. Merkhofer, and L. Strickhart, “Mitre: Seven systems for semantic similarity in tweets,” in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 12–17.

[57] F. Godin, B. Vandersmissen, A. Jalalvand, W. De Neve, and R. Van de Walle, “Alleviating manual feature engineering for part-of-speech tagging of twitter microposts using distributed word representations,” in *Workshop on Modern Machine Learning and Natural Language Processing, NIPS*, 2014.

[58] B. Hu, Z. Lu, H. Li, and Q. Chen, “Convolutional neural network architectures for matching natural language sentences,” in *Advances in neural information processing systems*, 2014, pp. 2042–2050.

[59] J. Weston, S. Chopra, and K. Adams, “# tagspace: Semantic embeddings from hashtags,” 2014.

[60] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, “Learning semantic similarity for very short texts,” in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1229–1234.

[61] S. Wijeratne, L. Balasuriya, D. Doran, and A. Sheth, “Word embeddings to enhance twitter gang member profile identification,” *arXiv preprint arXiv:1610.08597*, 2016.

[62] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in neural information processing systems*, 2002, pp. 849–856.

[63] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[64] J. Branke, K. Deb, H. Dierolf, and M. Osswald, “Finding knees in multi-objective optimization,” in *International conference on parallel problem solving from nature*. Springer, 2004, pp. 722–731.

[65] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, “A survey of text mining in social media: facebook and twitter perspectives,” *ASTESJ*, 2017.

[66] A. Bordes, Y.-L. Boureau, and J. Weston, “Learning end-to-end goal-oriented dialog,” *arXiv preprint arXiv:1605.07683*, 2016.

[67] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *JMLR*, 2003.

[68] D. Petkovic, R. Altman, M. Wong, and A. Vigil, “Improving the explainability of random forest classifier–user centered approach,” in *Pacific Symposium on Biocomputing*, 2018.

[69] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on twitter using a hierarchical knowledge base," in *European Semantic Web Conference*. Springer, 2014, pp. 99–113.

[70] S. Bhatt, B. Minnery, S. Nadella, B. Bullemer, V. Shalin, and A. Sheth, "Enhancing crowd wisdom using measures of diversity computed from social media data," in *Proceedings of the International Conference on Web Intelligence*. ACM, 2017, pp. 907–913.

[71] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm*, vol. 11, no. 538-541, p. 164, 2011.

[72] D. Sánchez and M. Batet, "A new model to compute the information content of concepts from taxonomic knowledge," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 8, no. 2, pp. 34–50, 2012.

[73] S. Ioffe, "Improved consistent sampling, weighted minhash and l1 sketching," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 246–255.

[74] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[75] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.

[76] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th international conference on semantic systems*. ACM, 2011, pp. 1–8.

[77] C. Jia, Y. Li, M. B. Carson, X. Wang, and J. Yu, "Node attribute-enhanced community detection in complex networks," *Scientific Reports*, vol. 7, 2017.

[78] R. Hartman, J. Faustino, D. Pinheiro, and R. Menezes, “Assessing the suitability of network community detection to available meta-data using rank stability,” in *Proceedings of the International Conference on Web Intelligence*. ACM, 2017, pp. 162–169.

[79] J. Yang, J. McAuley, and J. Leskovec, “Community detection in networks with node attributes,” in *Data Mining (ICDM), 2013 IEEE 13th international conference on*. IEEE, 2013, pp. 1151–1156.

[80] S. McKeown, R. Haji, and N. Ferguson, “Understanding peace and conflict through social identity theory,” *Contemporary Global Perspectives. Switzerland: Springer*, 2016.

[81] R. J. Hazler and S. A. Denham, “Social isolation of youth at risk: Conceptualizations and practical implications,” *Journal of Counseling & Development*, vol. 80, no. 4, pp. 403–409, 2002.