Who should be the captain this week? Leveraging inferred diversity-enhanced crowd wisdom for a Fantasy Premier League captain prediction

AAAI Press
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

Abstract

Participants in Fantasy Sports make a critical decision: selecting productive players for their fantasy team. The well-established Wisdom of Crowd effect can predict productive, rewarding players; popular, frequently selected players are potentially good choices. Previous *performance* data contributes to the identification of a subset of participants who collectively predict productive players. However, performance data may not always be available. Here we study the assembly of a small subset of the crowd *a priori* using another important aspect of crowd wisdom: semantic diversity. We infer diversity from participants' Twitter posts (tweets) that users voluntarily, and naturally provide as part of their reasoning. We propose the SmartCrowd framework to select a small, smart crowd using participants' Twitter posts. SmartCrowd includes three steps: 1) characterize participants using their social media posts with summary word vectors, 2) cluster participants based on these vectors, and 3) sample participants from these clusters, maximizing multiple diversity measures to form final diverse crowds. We evaluated our approach to diversity characterization for the Fantasy Premier League (FPL) captain prediction problem, in which participants predict a successful weekly captain among a set of soccer players. Empirical evaluation shows that SmartCrowd generates diverse crowds outperforming random crowds, 93% of individual participants, and crowds consisting of the top 10%, 20% experts using previous performance data. We provide converging evidence that social media based diversity supports the sampling of smarter crowds that collectively predict productive players. These results have implications for other domains, such as economics and geopolitical forecasting, that benefit from aggregated judgments.

Introduction

Fantasy sports is 7.22 billion dollar business [1], with revenues larger than World Wide Wrestling and Nascar racing combined. More than 4 million people play the Fantasy Premier League (FPL) in the UK and Ireland[2].

Several research studies explore the team construction that yields the maximum reward within budgetary constraints. We study the weekly captain selection task within the starting ll player team. Relative to normal players, participants receive twice the number of points that the captain scores. A good captain choice reflects numerous parameters such as recent injury, past injury, player record with the team, the player's leadership ability, etc. Crowd wisdom provides a cheap method for determining all relevant parameter values. A captain selected by the greatest number of participants reflects the union of news and variables that different participants may track. In fact, popular choice is one of the most widely used captain selection strategies in FPL[3]. But better strategies exist based on the predictions of a smaller subgroup of carefully chosen participants. Such strategies may include using performance data, i.e., participants' current judgments (Davis-Stober et al. 2015) and expertise determined from their previous judgment history (Goldstein, McAfee, and Suri 2014).

However, participant judgment history is not always available. Without such history, Surowiecki (Surowiecki 2005) suggests that collective wisdom applies when the selected crowd consists of independent users who potentially provide diverse opinions. Intuitively, a diverse crowd employs diverse perspectives in decision making. Their aggregated decision, or crowd agreement, is likely to be more accurate than one based on an individual or a set of similar (and possibly biased) perspectives. Several studies explore user perceived or self-categorized diversity to form such a crowd. Such diversity is not *directly* available for FPL participants.

However, Twitter provides relevant semantic data for inferring diversity. Participants regularly use Twitter to track game and player updates. The @OfficialFPL channel itself has half a million followers. Participants also use Twitter to seek player suggestions and broadcast their selection rationale. Such data naturally arise from playing the game, and freely distributed and available for crawling unlike other social network data such as Facebook. We mine FPL user tweets to infer crowd diversity based on topic and communication patterns. We had 2786 manually verified participants in our *participantset* who happen to be on Twitter and also play in the Fantasy Premier League. We collected $\sim$ 4M soccer related tweets from these participants to infer diversity. We use this *participantset* for our experiments to sample various types of crowds.

[1] https://tinyurl.com/y8vwahva

[2] https://tinyurl.com/ybjh9lyt

[3] https://tinyurl.com/y8rghx9u

Below, we demonstrate the benefit of inferred diversity in small, smart crowd selection for FPL captain prediction. Our approach to Fantasy Sports extends to other wisdom of the crowd enhanced prediction problems such as geopolitical forecasting and election prediction.

We propose a diverse crowd selection approach (SmartCrowd) based on social media posts (tweets). Each Twitter user is represented by the collection of their FPL tweets; user diversity appears in the topic and latent communication patterns between tweet collections. Word2vec (Mikolov et al. 2013b) summarizes a user's set of tweets, generating one equal-length summary word vector for each user.[4] We cluster these vectors to derive user clusters. We have tested multiple clustering strategies, such as cosine distance and Euclidean distance measures, single-view spectral clustering. The best strategy is multi-view clustering that synthesizes views based on both cosine distance and Euclidean distance. To compose the final crowd, we select optimal representatives from the clusters using a multi-view objective optimization method with both distance measures as the objectives.

For evaluation, we collected captain picks for all participants, the points each captain earned over 25 weeks and participants' previous seasons' performance score. A participant's choice of a particular captain constitutes a "vote". For each virtual crowd, we computed two values: 1) the captain with the most participant votes and 2) the crowd's "wisdom score" (similar to Goldstein et al.) as the points earned by that captain. One crowd performs better than another other when it gets a higher wisdom score than the other crowd.

To show that diversity boosts the FPL captain prediction task, we compare diverse crowds generated from the *participantset* (2786 users) to "random" crowds of size 'n' randomly sampled from the entire *participantset*. We also compared diverse crowds to expert crowds where a crowd of size 'n' is sampled from the top k experts determined by a participants' previous seasons' performance, as used by Goldstein et al.

We empirically evaluate the following questions:

- (RQ1): Does semantic analysis of crowd members' communications(E.g. social media communication) inform crowd diversity?
- (RQ2): Can diversity based crowd selection outperform trivial, random crowd selection?
- (RQ3): Can crowd selected on the basis of inferred diversity outperform crowds selected on the basis of expertise?
- (RQ4): What is the relative advantage of diversity vs expertise criteria for crowd selection?
- (RQ5): Do the benefits of diversity based selection depend on the average expertise of the population from which the crowd is sampled?
- (RQ6): Do the benefits of diversity vary with crowd size?

[4]Other text summarization methods (e.g. TF-IDF and LDA (Blei, Ng, and Jordan 2003) for topic extraction and summarization) could apply.

Using the SmartCrowd approach, a diverse crowd (sampled from the *participantset*) outperforms random crowds (also sampled from the *participantset*) 85% of the time. Using past performance data, a diverse crowd outperforms expert crowds sampled from the top 20%, 10%, 5%, and slightly worse than the top 2% of the participants from the *participantset*. Diverse-expert crowds achieve the best wisdom score and diverse non-experts can replace experts without compromising performance. Our contributions include:

- Inferred crowd diversity from users' discussions on social media and a proposed a framework to capture such diversity and compose diverse crowds.
- An extensive experimental evaluation of our approach uses real FPL data and associated tweets and tests the word2vec method for summarizing tweets, different clustering algorithms, and several crowd composition methods to identify diverse crowds.
- Diversity enhanced crowd selection achieves a collective wisdom for FPL captain prediction.

Crowd formation and an accurate prediction reflect numerous computational choices, which we examine empirically. A multi-faceted experimental analysis confirms the superior judgment of a diverse crowd using social media data.

The rest of the paper is organized as follows. Section differentiates our work from other related work on crowd selection as well as the diversity-wisdom of crowd correlation found in other domains. Section provides background on word embedding, clustering, and multi-objective optimization. Section details our approach to crowd selection from participants' social media data. Section describes the setup, data collection, and the results of our experiments.

Related Work

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 diversity and accuracy of the collective judgment in crowd selection(Lorge et al. 1958). These experiments solicit participants to explicitly indicate diversity. Teng et al. asked participants to define their similarity to other members of groups (Ye and Robert Jr 2017). 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(RQ1-RQ3).

Other research does explore the correlations between content diversity and crowd wisdom. Hong et al. showed that opinion diversity in participant-generated content positively correlates with crowd performance(Hong et al. 2016). However, they did not explore a crowd selection strategy (RQ2-RQ3), and used cosine similarity between traditional word vectored representations to compute participant diversity. This word vectored representation neglects contextual similarity (Mikolov et al. 2013b) especially for short social media texts. Robert Jr. et al. explored diverse crowd formation for the generation of quality Wikipedia articles(Robert and

Romero 2015). They computed crowd diversity from the authors' stated topics of interest and showed that diversity could help to form smart crowds. However, they do not explore crowd selection based on such diversity (RQ2, RQ3). Moreover, they used explicit, stated participant topics instead of inferring diversity from raw social media posts. Several predictive analysis problems, such as the one discussed in this paper, do not provide explicit participant indications. For example, we do not have participants' FPL specific topic affinity listed on the FPL website. Instead, we use openly available social media data 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(Ren and Yan 2017). We employ Word2vec word vector generation to capture such latent communication patterns along with topic-specific words.

Several studies explore crowd selection relying on current or historical judgment data(Olsson and Loveday 2015). Goldstein et al. proposed smart crowd generation in a domain similar to ours (FPL) but using the previous season's performance data(Goldstein, McAfee, and Suri 2014). Indeed, previous season performance data can indicate experts and a small crowd of such experts outperforms large crowds for the FPL captain prediction task. Section demonstrated surprisingly effective crowds without relying on such historical data (RQ3). As expertise contributes, we must examine the role of both diversity and expertise in crowd selection (RQ4, RQ5). Davis-Stober et al. also proposed crowd formation using performance data(Davis-Stober et al. 2015). They used current judgment data to select wise crowds optimally. Galesic et al. showed that a smart crowd exists in several prediction tasks, identified using their current judgments (Galesic, Barkoczi, and Katsikopoulos 2018). Nguyen et al. (Merayo, Nguyen, and others 2017) found that judgment diversity correlates with judgment accuracy in smart crowds. Using their method, inferred diversity based crowd selection should also have diverse judgment (RQ1).

Several research studies explore team selection for maximizing reward in a season-long fantasy tournament(Fry, Lundberg, and Ohlmann 2007)(Bergman and Imbrogno 2017)(Becker and Sun 2016). Some studies also explore the maximum number of wins a player will have in sports (Kaplan and Garstka 2001)(Clair and Letscher 2007). More recent studies explore team selection for daily fantasy sports(Hunter, Vielma, and Zaman 2016)(Haugh and Singal 2018). These successfully employ specific features of player data collected by a user. However, the collection of such broad data is challenging, e.g., each injury report of a player, player dynamics, player leadership skills, and 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 the existing approaches, our approach exploits crowd wisdom as a substitute for such specific information.

Background

Background on word vectors, clustering, and Pareto optimization follows.

Word vectors

A word vectored text representation improves and simplifies Natural Language Processing (NLP) applications such as search, language translation, and information extraction (Mikolov et al. 2013b)(Mikolov et al. 2013a). In this study, we intend to capture the topical and conversational diversity among these participants. As a word vector captures a context of a word where a context is identified by the surrounding words. Hence, it can capture the latent topic as well as the communication pattern of a user. Specifically, Given preceding words, such word vectors predict a probability distribution over the "next" word, given preceding words. 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 (Mikolov et al. 2013b). 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 (Mikolov, Yih, and Zweig 2013)(Mikolov et al. 2013a). Word2vec has been used to represent the similarity of social media posts, especially tweets, by averaging tweet word vectors (Zarrella et al. 2015).

Clustering

Data points in the same group (clusters) are more similar than those in different groups. Spectral clustering finds clusters using the eigenvectors of a similarity matrix. However, the chosen similarity measure affects clustering. Moreover, similarity determines diversity. In the absence of a diversity definition, we cannot justify a single, ideal similarity measure. Furthermore, multiple similarity measures provide complementary, potentially crucial information. Hence, we used multi-view clustering to find a clustering structure based on multiple similarity measures. The standard multi-view data clustering technique considers information from each view. Such approaches perform well for clustering real-world noisy multi-view data(Bickel and Scheffer 2004). We use a method proposed by Wang et al. that finds clusters using view agreement in the presence of noise (Wang et al. 2016), under the following key assumptions: 1) Features in each view are sufficient to discover most of the clustering structure, 2) The clustering structures agree between views, 3) An accurate underlying clustering assigns a point to the same cluster irrespective of the view.

These assumptions hold for our application, as we created views based on multiple similarity measures. We selected similarity measures without a significant divergence in resulting views that still represent complementary information. Hence, we expect the clustering structure to be in agreement across views, revealing a structure that indicates similar sets of participants.

Pareto Optimization

We compose a diverse crowd by picking one participant from each cluster. We used Pareto optimization to produce optimally diverse crowds from the clustering structure. Pareto optimization computes solutions that cannot be improved for any one objective without degrading at least one

Figure 1: Pareto Front Illustration

of the other objectives. Consider a scenario of choosing a product based on good quality and minimal price. Figure 1 shows the 2-dimensional solution space. The x-axis represents the inverse of price and the y-axis represents product quality with a higher number indicating higher quality. Pareto optimization selects a Pareto front as shown in Figure 1. Products shown in red satisfy Pareto optimum criteria and represent the Pareto front.

Approach

Our SmartCrowd approach first clusters similar participants according to their social media posts, concerning topics and communication style. 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.

Our approach (see Figure 2) consists of three core components: social-media based participant representation (Process arrow P1), participant clustering (Process arrow P2), and diversity-based crowd selection (Process arrow P3).

P1: Social-Media Based Participant Representation. Without using participants' history, we characterized them by summarizing their social media (Twitter) posts (tweets) with word2vec. As word2vec captures the context of a word as surrounding words, it can provide topic similarity as well as latent communication patterns. We used $\sim$ 4M participant tweets to train a word2vec model. The resulting model represents each word with a 300-dimension vector. Each participant's posts are then summarized into a 300-dimension vector (by averaging the individual word vectors in all posts). Consistent with (Bhatt et al. 2017)(Wijeratne et al. 2016) the resulting feature vector characterizes each participant, grouping words by participant's Twitter id and aggregating (averaging) vectors of these words (see *Tweets and* Word Vectors in Figure 2).

P2: Participant Clustering. Clustering the participants before crowd selection helps identify groups of similar users regarding topics and communication patterns. We want to avoid oversampling users following 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 (Zarrella et al. 2015)(Wijeratne et al. 2016) 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.

The spectral clustering algorithm shows exceptional performance in identifying clusters of irregular distributions(Ng, Jordan, and Weiss 2002). 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, $W_{ij} = e^{-\frac{\delta(x_i, x_j)}{2\sigma^2}}$. 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. We use the well-known Silhouette Coefficient (SC) method (Rousseeuw 1987) 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 randomly sample n participants from each cluster such that n is not larger than the minimum cluster size. With a small n, e.g., in [1, 3], random representative selection method performs reasonably.

We further improve the selection strategy 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 as follows. 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 1 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 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

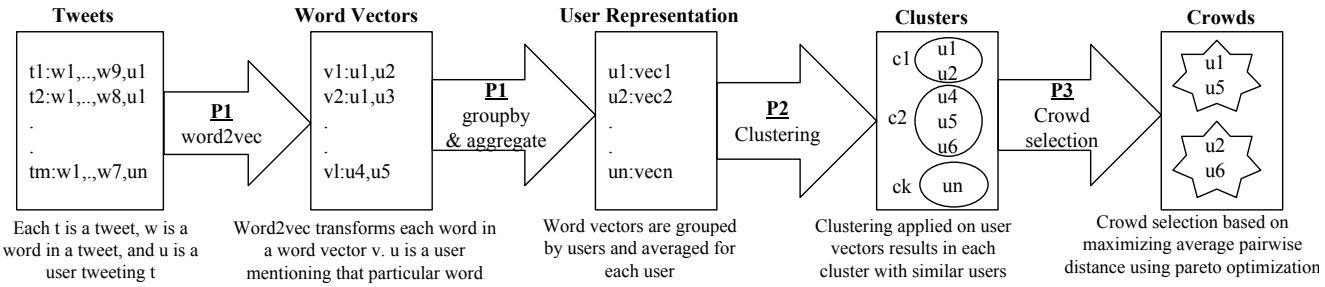


Figure 2: Approach overview

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

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 (Branke et al. 2004). Here, we do not select the best crowd from P but consider all crowds in P as our final set of diverse crowds. We compute the wisdom score (described below) for each crowd in our final crowd set P. We then compare these wisdom scores to the set of wisdom scores of a different crowd selection strategy.

Experiments

We evaluated SmartCrowd for the FPL captain prediction problem (Goldstein, McAfee, and Suri 2014).

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 using Pareto optimization over Cosine and Euclidean distance, as our multi-view clusters were generated using both measures. Pareto optimization consistently resulted in 3-6 optimal crowds in our dataset. We repeated crowd formation with Pareto optimization (Algorithm 1), to obtain l crowds. In this paper, 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, $WS = \frac{\sum_1^{25} Mod(C_{index})}{25}$. 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.

Data Collection and Implementation

We collected FPL related tweets using three FPL keywords, FPL, @OfficialFPL, and Fantasy Premier League. From these tweets, we extracted the names of the associated Twitter users. To obtain captain pick and previous season performance data, we matched these Twitter users using their name on Twitter and name on the official FPL website[5], on which registered users post their team lineup, including weekly captain picks. We expect 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. Hence, we found all matches with the same first and last name with at least ten tweets mentioning one of the FPL keywords, resulting in 70,440 matches. To further eliminated any non-unique names and their associated data, i.e., names appearing more than once on either Twitter or the FPL website, leaving 3829 user accounts. Finally, we manually verified 2786 matches based on their recent Twitter activity, location, and names on both Twitter and FPL website. During this step, we eliminated all the matches for which we couldn't verify the Twitter activity and/or location (including unavailable location information).

We collected their soccer related tweets by scraping Twitter user timelines (for a total 4,299,738 tweets). We found $\sim$ 1M tweets from this set of tweets having at least one of the three FPL keywords[6].

For evaluation purposes *only*, we collected 25 weeks of captain picks for previous 2015-16 FPL season for each participant. We also collected that captain's score based on

[5] fantasy.premierleague.com

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

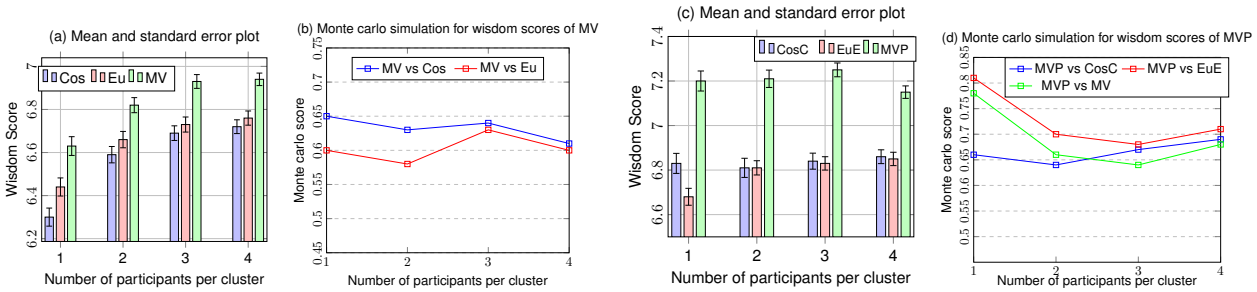


Figure 3: (a) and (b) compares crowds generated using Multi View clustering(MV), Cosine(Cos), and Euclidean(Euc) distance based clustering. (c) and (d) compare crowds generated by maximizing one distance measure (CosC, EucE) versus maximizing both distance measures (MVP). MVP crowds perform the best.

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 using historical performance data to define expertise. [7]

Results and Analysis

The results are organized according to the evaluation goals: (1) How methods for clustering and proposed crowd composition method affect final SmartCrowd performance (RQ1); (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 employ historical crowd performance data. We also show that the SmartCrowd performance is comparable to expert crowds when expert participants can be selected using historical performance data(RQ2, RQ3); (3) Finally, we examine diversity, expertise, and crowd size effects on crowd wisdom (RQ4-RQ6).

Factors Affecting SmartCrowd As described in Section , participant clustering and diversity-based crowd composition are two key influences.

Participant clustering: The best number of clusters were 6(0.27), 7(0.23), and 7(0.45), for Euclidean-spectral (spectral with Euclidean), Cosine-spectral(spectral with cosine), and Multi-view methods, respectively. The bracketed values indicate the corresponding maximum silhouette value. Multi-view clustering produced the best clustering structure.

We sampled crowds by selecting n participants from these clusters at random for a given clustering structure (Euclidean-spectral, Cosine-spectral, and Multi-view). We selected l such crowds from each clustering structure. Figure 3a,b shows the wisdom score statistics for crowds generated from each clustering structure. Crowds from a multi-view clustering structure(MV) achieved the best average wisdom score and outperformed crowds generated from Cosine(Cos) and Euclidean clustering(Eu) structure, (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 ~ 0.5 indicates that two sets of crowds are equally likely to beat each other. A score of ~ 1.0 indicates that a crowd from set one almost always beats a crowd from set two. Figure 3b 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., Algorithm 1 for multi-view clustering. Single-view clustering separately maximized crowd selection using average pairwise Euclidean and Cosine distance. Specifically, we had generated crowds 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 3c shows the statistics for l crowds' wisdom scores. Multi-view clustering combined with Pareto optimization based crowd selection generated crowds (MVP) with 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). Using Monte Carlo simulation, an MVP crowd was $\sim 80\%$ likely to have a higher wisdom score than EuE and CosC crowds (see Figure 3). MVP crowds also outperformed MV crowds, i.e., crowds selected without maximizing distance measure. Hence, the inferred diversity can inform diverse, smart crowd selection(RQ1).

Comparison with Other Crowd Selection Strategies Using the resulting optimal settings (Multi-view clustering and Pareto optimization based crowd selection), we compared SC to 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 considered for SC and k indicates the number of clusters in SC. As we found 6 clusters in our SC selection, we generated random crowds in multiples of 6, i.e., corresponding to $n = \{1, 2, 3, 4\}$ representatives per cluster.

[7] 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.

Figure 4a shows the box plot of wisdom scores for SCs and random crowds by crowd size. 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 '+'. SCs (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 < 0.05). Figure 4b 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 (RQ2). The probability that SC outperforms a random crowd does decrease with increasing crowd size. Thus using SC a smaller crowd size with just one representative per cluster is sufficient.

Next, we compared SC with crowd selection based on expertise. Expert crowds with a known performance history often perform very well as shown by Goldstein et al. For evaluation purposes, we sampled expert crowds only from top performing participants. Figure 4c 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. SC (for crowd size six) had comparable wisdom scores as E5. SC significantly outperformed E10, E20 (T p-value $<$0.05). Moreover, the E2 expert crowd advantage is marginal and therefore comparable to SC (RQ3).

Figure 4b indicates Monte Carlo simulation scores comparing diverse crowds to various expert crowds. Monte Carlo scores of 0.7 and 0.58 for comparing SC to E20 and E10 also show that a SC crowd is likely to outperform E10 and E20 expert crowd for crowd size six. Because increasing crowd size does not benefit SC we did not observe an improved wisdom score with increasing crowd size. Next, we compared the performance of an SC 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 4b shows the resulting wisdom score box plots. Figure 4a shows the Monte Carlo simulation scores comparing the SC 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.

We also 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 shows the percentile of participants that a crowd outperforms. We computed an average of the l crowds and computed the percentile of participant scores that it outperforms. On average a randomly generated crowd of size 6, achieves a better "wisdom score" than 72% of all 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 SC 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[8] score for each word in participants' tweets 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 (see left of Figure 6). Participants in different clusters show different interests in teams, players, and useful FPL accounts, i.e., with diverse perspectives on captain choice. Some words, e.g., join, team, etc. do not appear in the figure as these words do not explain the clusters. Also, 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 15 cluster pairs were negatively correlated, confirming cluster diversity.

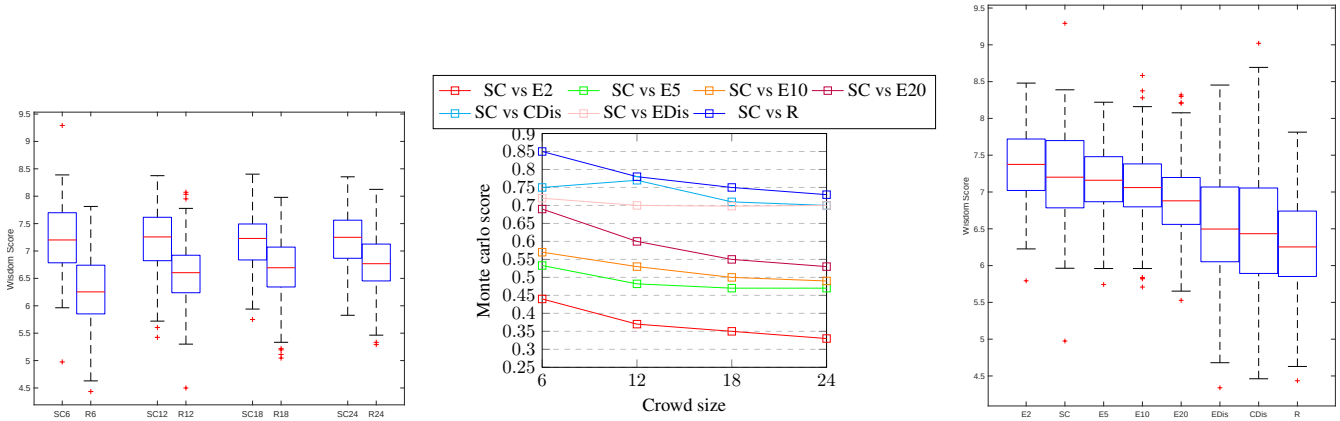
We further verified the sensitivity of the wisdom score and the wisdom of crowd effect for different sample sizes. We selected 20%, 40%, 60%, and 80% of the *participantset* and performed SC selection and Random crowd selection. Figure 6 (right) shows the results for these sample sizes. We notice that the effect can be observed with varying sample sizes, with relatively large standard error for smaller sample sizes. Hence, the effect is robust for different sample sizes.

Next, we confirmed whether SC's outperformance truly comes 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. SC's advantage 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 1. We compared the resulting crowds without these clusters to crowds generated with all SC clusters. The resulting Monte Carlo simulation score ~ 0.5 indicated that the two sets of crowds had similar performance. Therefore the eliminated crowds do not account for the SC advantage.

Inferred diversity vs. judgment diversity. SC selects diverse crowds by clustering similar participants represented by a word2vec vector. We refer this diversity as "inferred". Figure 7a 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 also 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 $p_d = \frac{ND_{total}}{10000}$. 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

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



(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 4: 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.

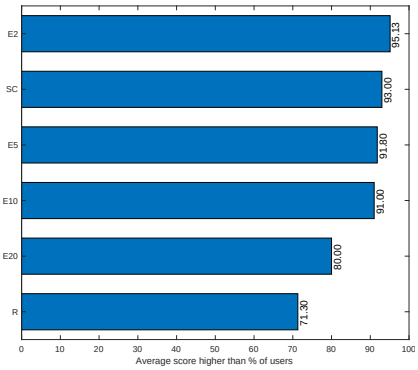


Figure 5: Wisdom of crowd effect

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

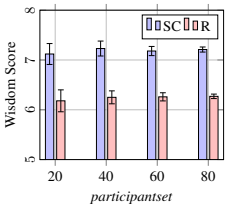


Figure 6: Most frequent words in cluster on left. Wisdom score for crowds generated from different size of *participantset* on right.

(a) Inferred diversity mean std err

(b) Judgment diversity mean and std err

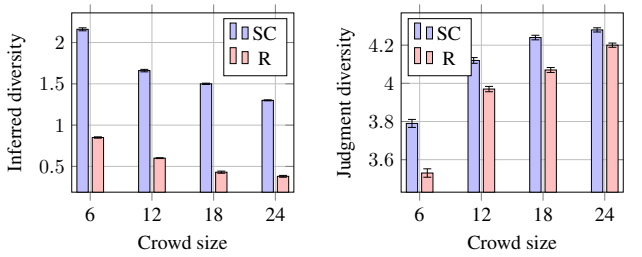


Figure 7: Inferred diversity (a) and Judgment diversity (b) comparison. SC has higher inferred and judgment diversity. Inferred diversity correlates with judgment diversity.

for "diverse participants" was 0.85. Hence, a crowd sampled from "inferred diversity" measure is also likely to demonstrate judgment diversity (RQ1).

Further, we examined Merayo et al. 's. "judgment diversity" measure to compare SC 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 $D = \frac{\sum_{i,j} d(u_i,u_j)}{n(n-1)}$, 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 crowd participants. Using this metric, we investigated whether SC 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 7b confirms that SC 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. SC 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 8a compares the wisdom score of SC 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. Using widsom scores, crowds generated using SC consistently outperformed crowds generated from one cluster. Figure 8b shows the average judgment diversity of crowds generated using SmartCrowd (SC) and (non-diverse) crowds generated from each cluster. SC

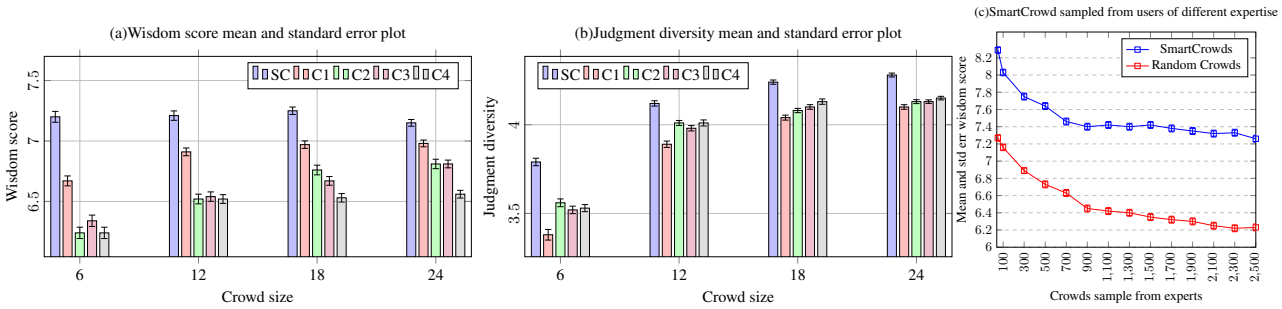


Figure 8: Inferred diversity comparison for SmartCrowd & Random crowd. SC has higher inferred diversity than R.

also has the highest overall judgment diversity. Thus, in the absence of historical judgment data, our inferred diversity measure serves as a proxy for judgment diversity with the attendant benefits to accuracy consistent with the findings of Merayo et al. 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 8c shows the wisdom score from crowds sampled using SmartCrowd(SC) and crowds sampled at Random(R). Crowds sampled using SC benefits 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 (RQ5).

We also examined whether diversity can replace expertise in the performance of hybrid (expert and diverse non-expert) crowds. 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 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 came from 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 9b shows the results for replacing n experts with diverse non-experts. Diverse, non-expert participants can replace experts without compromising 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 include any of the top 100 expert participants, unlike the experiments for comparing diverse crowds with expert crowds (RQ4).

Finally, we examined the effect of crowd size on wisdom score. Crowd size potentially affects prediction performance. Figure 9a plots the mean and standard error wisdom score for increasing crowd size. With increased crowd size, random crowd performance approaches SC, while SC'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 (RQ6).

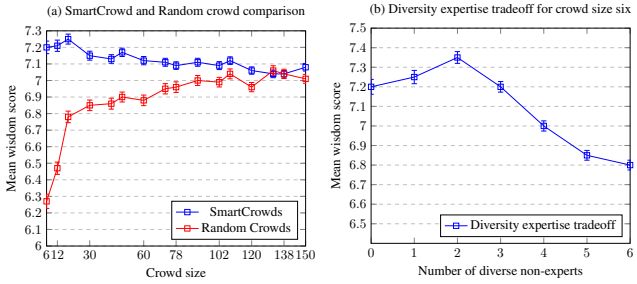


Figure 9: (a)SmartCrowd(SC) and Random crowd(R) wisdom score comparison. SC with 6 participants achieve wisdom score that is achieved by 100+ participants of R., (b) shows the effect of replacing experts with diverse non-experts.

Discussion and Conclusions

We have 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 significantly 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.

Unlike methods that rely on the explicit solicitation of auxiliar data, Twitter is already a popular medium for discussing FPL. Similarly, Reddit and sports blogs provide ac-

cessible data that is a natural facet of the task. To examine the sensitivity of our methods to data source, our future work will investigate the use of different data sources as well as different types of data such as links between users in Twitter, geo-location etc. To compensate for our relatively conservative approach to user identification, future work will also examine emerging, alternative methods to associate accounts on different social media with the same user (needs a citation!). The chief concern here is the potential benefit of a larger participantset.

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. We will explore other Fantasy Sports and even other domains with outcome measures to test the proposed method. Given our success, follow-on research shall extend and validate these findings in other domains such as marketing, election prediction, and geopolitical forecasting.

We are presently working with several geo-political social media corpora, where a crowd wisdom approach can help forecast the outcome of such events. The proposed technique is especially useful when opinions are not equally distributed across the corpus. Random sampling from such a corpus would merely replicate the existing bias. Hence, the proposed technique promises an unbiased (diverse) sample to predict the outcome of events substantially different from the FPL domain presented here.

References

Becker, A., and Sun, X. A. 2016. An analytical approach for fantasy football draft and lineup management. *Journal of Quantitative Analysis in Sports* 12(1):17–30.

Bergman, D., and Imbrogno, J. 2017. Surviving a national football league survivor pool. *Operations Research* 65(5):1343–1354.

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

Bickel, S., and Scheffer, T. 2004. Multi-view clustering. In *ICDM*, volume 4, 19–26.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

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

Clair, B., and Letscher, D. 2007. Optimal strategies for sports betting pools. *Operations Research* 55(6):1163–1177.

Davis-Stober, C. P.; Budescu, D. V.; Broomell, S. B.; and Dana, J. 2015. The composition of optimally wise crowds. *Decision Analysis* 12(3):130–143.

Fry, M. J.; Lundberg, A. W.; and Ohlmann, J. W. 2007. A player selection heuristic for a sports league draft. *Journal of Quantitative Analysis in Sports* 3(2).

Galesic, M.; Barkoczi, D.; and Katsikopoulos, K. 2018. Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision* 5(1):1.

Goldstein, D. G.; McAfee, R. P.; and Suri, S. 2014. The wisdom of smaller, smarter crowds. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 471–488. ACM.

Haugh, M. B., and Singal, R. 2018. How to play fantasy sports strategically (and win).

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

Hunter, D. S.; Vielma, J. P.; and Zaman, T. 2016. Picking winners using integer programming. *arXiv preprint arXiv:1604.01455*.

Kaplan, E. H., and Garstka, S. J. 2001. March madness and the office pool. *Management Science* 47(3):369–382.

Lorge, I.; Fox, D.; Davitz, J.; and Brenner, M. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological bulletin* 55(6):337.

Merayo, M. G.; Nguyen, N. T.; et al. 2017. Intelligent collective: The role of diversity and collective cardinality. In *Conference on Computational Collective Intelligence Technologies and Applications*, 83–92. Springer.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. 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*, 746–751.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856.

Olsson, H., and Loveday, J. 2015. A comparison of small crowd selection methods. In *CogSci*.

Ren, R., and Yan, B. 2017. 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*, 6342–6351. ACM.

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

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

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Wang, Y.; Zhang, W.; Wu, L.; Lin, X.; Fang, M.; and Pan, S. 2016. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. *arXiv preprint arXiv:1608.05560*.

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

Ye, T., and Robert Jr, L. P. 2017. 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*, 2344–2358. ACM.

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