A Novel Algorithm for Aggregating Crowdsourced Opinions

Ethan Prihar
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
ebprihar@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
nth@wpi.edu

ABSTRACT

Similar content has tremendous utility in classroom and online learning environments. For example, similar content can be used to combat cheating, track students' learning over time, and model students' latent knowledge. These different use cases for similar content all rely on different notions of similarity, which make it difficult to determine contents' similarities. Crowdsourcing is an effective way to identify similar content in a variety of situations by providing workers with guidelines on how to identify similar content for a particular use case. However, crowdsourced opinions are rarely homogeneous and therefore must be aggregated into what is most likely the truth. This work presents the Dynamically Weighted Majority Vote method. A novel algorithm that combines aggregating workers' crowdsourced opinions with estimating the reliability of each worker. This method was compared to the traditional majority vote method in both a simulation study and an empirical study, in which opinions on seventh grade mathematics problems' similarity were crowdsourced from middle school math teachers and college students. In both the simulation and the empirical study the Dynamically Weighted Majority Vote method outperformed the traditional majority vote method, suggesting that this method should be used instead of majority vote in future crowdsourcing endeavors.

Keywords

Crowdsourcing, Similarity, Community Detection, Hierarchical Clustering

1. INTRODUCTION

Within online learning platforms and intelligent tutoring systems there is a tremendous opportunity to utilize knowledge of content similarity. Similar problems can help prevent cheating during exams by randomly selecting from multiple similar problems when students receive the exam, measure students' learning gains by spreading out similar problems between assignments, and measure the effects of instructional interventions by comparing a student's scores on similar problems before and after the intervention. Similar instructional material can be used to offer students choices in which instructional material they receive, which has been shown to increase engagement and achievement [7]. While it is possible to implement these methods with general knowledge of content similarity, such as similarity in prerequisite knowledge or difficulty, if a more informed definition of content similarity is used, the success of these methods is likely to grow.

Although there is a lot of value in knowing what content is similar to other content, what content should be considered similar is highly dependent on use case. This makes it a challenge for content creators to define the similarity in the content, as they don't necessarily know what their content will be used for. While some content is obviously similar, for example, two mathematics problems that are identical except for the numbers used in the problems, in other situations it is much more difficult, especially when content is being aggregated from multiple sources that may not even use the same metrics for prerequisite knowledge or difficulty.

Crowdsourcing offers a way to derive which content is similar to other content for specific use cases. Crowdsourced opinions on similar content can be gathered each time a new use case for similar content arises. By informing the workers, whose opinions are being crowdsourced, of the specific use case and requirements for similarity, the methods that rely on content being similar are more likely to be successful. However, crowdsourcing opinions on similar content poses some challenges as well. Before an online learning platform or intelligent tutoring system uses crowdsourced assertions of similarity, steps must be taken to assess the trustworthiness of workers whose opinions are being crowdsourced and ensure the truthfulness of the final assertions of similarity.

In this work we present a novel algorithm that both measures the reliability of the workers whose opinions are being crowdsourced, and determines, from these individual's opinions, what content is most likely to be similar to other content. To evaluate this method, we first simulated a wide range of conditions in which assertions of similarity were made, and compared the performance of our algorithm to the traditional alternative. We then performed a case study where teachers and college students were told to identify middle-school mathematics problems that evaluated a simi-

lar skill set. The assertions of similarity collected from the case study were used to identify groups of similar problems and measure the reliability of each worker's assertions.

Ultimately, this work seeks to answer the following three research questions:

1. Can we exploit properties of community detection to more accurately form groups of content from crowdsourced opinions?
2. How does the resulting algorithm perform in a simulation study compared to the more traditional method?
3. How does the resulting algorithm perform in a case study using workers of various expertise to determine which mathematics problems are similar to each other?

2. BACKGROUND

2.1 Ensembling Crowdsourced Opinions

Identifying the truth from crowdsourced opinions is not a new problem. Most of the techniques employed to ensure the accuracy of crowdsourced opinions rely on ensuring that workers have sufficient knowledge of the subject matter. This can be done through testing workers before giving them tasks, tailoring tasks specific to their skill sets, recruiting high quality workers, and educating workers before assigning them tasks. This can also be done through encouragement with extrinsic motivators like money, promotions, or prizes, or intrinsic motivators like a sense of purpose, or by gamifying the crowdsourcing tasks [1].

While there are many methods to encourage individuals whose opinions are being crowdsourced to be accurate, this work is focused on how to validate the quality of individuals' opinions after their task is complete. Current methods for accomplishing this place the burden of validation back onto the workers. Having workers rank the quality of other workers assertions is one method of validation. Another common method for validation is to have multiple workers perform the same task and merge the output of each worker, either as an average or as a majority vote [1].

There are also more advanced ways of algorithmically validating crowdsourced opinions. Item response theory and latent factor analysis based models have out-performed majority voting based validation methods on tasks related to identifying facial expressions and answering questions about geography [6, 10]. These models also determine the quality of individuals whose opinions are being crowdsourced, which can be used to refine the pool of individuals used for future crowdsourcing tasks [6, 10]. The novel algorithm in this work also aggregates crowdsourced opinions while evaluating the quality of each worker.

2.2 Community Detection

The field of community detection is focused around determining groups of similar items from a network of connected items. This has many applications throughout mathematics, physics, biology, computer science, and social sciences. Many things can represented as a network, for example, interstellar objects, neurons, city streets, and social media can all be represented as networks of interconnected items [3]. Finding similar educational content can be framed as a community detection problem by representing educational content as a network in which items are connected by topic, difficulty, language, prerequisite knowledge, or, in the case of this work, opinions on similarity. Structuring the task of identifying similar educational content as a community detection problem allows for the use of various well-established community detection algorithms, such as hierarchical clustering. In hierarchical clustering, each item begins in it's own cluster. Then, clusters are merged based on the merge strategy and distance between clusters [5]. Hierarchical clustering was used in both the simulation and empirical study.

3. METHODOLOGY

3.1 Dynamically Weighted Majority Vote

The Dynamically Weighted Majority Vote (DWMV) method is our alternative to the traditional majority vote method for combining multiple crowdsourced opinions on tasks with binary outputs. The DWMV method calculates the weighted majority opinion for each task, then determines the weight of each worker by how closely their opinion agreed with the majority opinion. The closeness of a worker's opinion to the majority opinion can be determined with any function for comparing two vectors that results in a value greater than or equal to zero. For example, accuracy or Dice coefficient[2]. DWMV initializes all workers' weights to be equal at the beginning of the algorithm, and iteratively updates these weights until the weighted majority vote does not change between iterations. Once the weighted majority vote remains constant from one iteration to the next, the weights of the workers can be interpreted as a measure of confidence in each worker, and the final weighted majority vote can be used downstream in the same way the traditional majority vote would have been used. Algorithm 1 formally defines the DWMV algorithm. In Algorithm 1, the function $s(x, y)$ determines the closeness of worker i's opinion, $(B_{ij}[A_{ij} = 1])_{j=1}^{t}$, to the majority opinion, $(u_j[A_{ij} = 1])_{j=1}^{t}$. The algorithm requires a matrix A of response indicators, in which $a_{ij} = 1$ if worker i completed task j, and $a_{ij} = 0$ otherwise, and a matrix B of worker's responses to tasks, in which b_{ij} contains the binary response of worker i to task j. In Algorithm 1, vector u contains the final weighted majority vote for each task, and vector c contains the final measure of confidence for each worker, based on the similarity between the weighted majority votes and the individual worker's responses.

3.2 Simulation Study

To determine if DWMV had a positive impact on forming groups from crowdsourced opinion, a simulation study was performed to compare the DWMV method to the traditional majority vote method in a variety of conditions. Figure 1 illustrates the simulation process. In the simulation study, hierarchical clustering was used to form groups from simulated workers' opinions of item similarity aggregated using both the majority vote method and the DWMV method. Table 1 lists the different initial parameters and their values used in the simulation. Five trials of every possible combination of the values in Table 1 were simulated for a total of 37,500 simulation runs.

Algorithm 1 Dynamically Weighted Majority Vote

Require: $s(x, y)$: function for the similarity of two vectors
Require: w : number of workers
Require: t : number of tasks
Require: $A = (a_{wt})$: matrix of response indicators
Require: $B = (b_{wt})$: matrix of response values
  $v \leftarrow (0)_{j=1}^{t}$ ▷ initialize with values different from u
  $u \leftarrow (-1)_{j=1}^{t}$ ▷ initialize with values different from v
  $c \leftarrow (1)_{i=1}^{w}$ ▷ start with equal confidence in all workers
  while $u \neq v$ **do**
    $v \leftarrow u$
    $u \leftarrow \left(\begin{cases} 1, & \text{if } \frac{\sum_{i=1}^{w}(c \odot B \odot A)_{ij}}{\sum_{i=1}^{w}(c \odot A)_{ij}} \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \right)_{j=1}^{t}$
    $c \leftarrow \left(s\Big((u_j[A_{ij} = 1])_{j=1}^{t}, (B_{ij}[A_{ij} = 1])_{j=1}^{t} \Big) \right)_{i=1}^{w}$
  end while

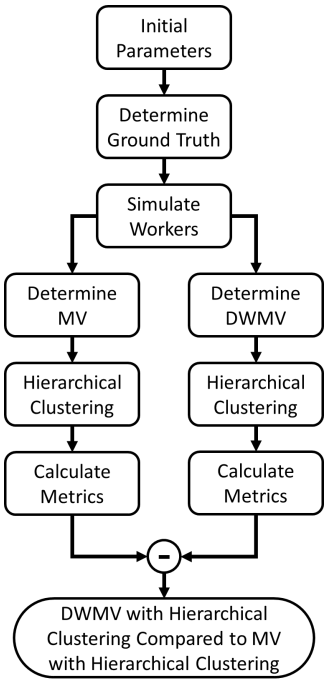

Figure 1: A flowchart of the simulation process, DWMV and majority vote were compared to each other through their use in community detection through hierarchical clustering.

Table 1: Simulation Parameters and Simulated Values

Parameter	Values
i	50, 100, 150, 200
g	5, 10, 15, 20, 25
w_{fp}	0.1, 0.2, 0.3, 0.4, 0.5
w_{fn}	0.1, 0.2, 0.3, 0.4, 0.5
p	20, 40, 60, 80, 100
d	0.25, 0.5, 0.75

The simulation began by randomly placing i items into g groups, where i and g are initial parameters of the simulation. Then the simulation crated ten workers. Each worker had a false positive rate and a false negative rate. These values were calculated separately to make the simulation more true to real life. In real life, it is not often that a worker would have an equal chance of incorrectly asserting that two items are or are not similar. The more likely case is that some workers think there is more similarity and other workers think there is less similarity between items than the actual similarity of items. The false positive and false negative rates of the workers were sampled separately for each worker from a uniform distribution in the range $[0, w_{fp}]$ and $[0, w_{fn}]$ respectively, where w_{fp} and w_{fn} are initial parameters of the simulation. Once the items were randomly placed in groups, and the error rates of the workers were randomly determined, a random p percent of all pairs of items were given to each worker, where p is an initial parameter of the simulation. Each worker then determined whether or not the items in each pair they received were similar to each other, taking into account their error rates.

Once all workers asserted whether or not each item pair they were given contained similar items, the majority vote and DWMV for the similarity of each item pair was calculated. The majority votes and DWMVs of item similarity were then used to form a network of item similarity, where each item is connected to every other item it was voted to be similar to. The majority vote network and DWMV network were both used to form groups through hierarchical clustering with Jaccard Index as the distance metric. Jaccard Index was used as the distance metric because Jaccard Index does not take into account true negatives [8]. Most items are not similar to each other, so a metric that takes into account true negatives would be over-inflated and not as informative in this context. After forming groups from the majority vote and DWMV similarity networks, the difference in accuracy, precision, and recall between the groups formed from the majority vote and DWMV similarity networks were used to determine if the DWMV method improved upon tradition majority vote.

3.3 Empirical Study: Similar Problems

In addition to a simulation, an empirical study was performed to compare DWMV to majority vote on a real crowdsorcing task. In this study, middle school mathematics teachers and college students were given 50 seventh grade mathematics problems from the Engage New York[1], Illustrative Mathematics[2], and Utah Middle School Math Project[3] curricula. Each worker was told to identify problems that evaluate similar mathematics skills. The workers' crowdsourced opinions of similarity were aggregated using both DWMV and majority vote, and then grouped using hierarchical clustering, with Jaccard Index as the distance metric with a threshold of 0.75. The resulting groups were then compared to a ground truth, provided by ASSISTments, an online learning platform [4], in the form of Common Core State Standards Mathematics Skill Codes[4], which each problem

[1]https://www.engageny.org/
[2]https://illustrativemathematics.org/
[3]http://utahmiddleschoolmath.org/
[4]http://www.corestandards.org/

was tagged with. These ground truth skill tags were determined by trained experts and the designers of the above stated curricula. The difference in accuracy, precision, and recall between groups formed with hierarchical clustering from DWMV and majority vote were again used to evaluate the quality of the DWMV algorithm.

4. RESULTS

4.1 Simulation Study

To compare the DWMV method to the traditional majority vote method, the difference in accuracy, precision, and recall as a function of w_{fp}, w_{fn}, i, g, and p, as described in Section 3.2, were calculated. The first positive takeaway from the simulation is that DWMV was almost always more accurate than majority vote, regardless of the simulation parameters. Only when the simulation had more than twenty groups or the maximum false negative rate of workers was 20% or less did DWMV not reliably out perform majority vote, but it did not significantly underperform either. At most, DWMV was slightly less accurate than majority vote when workers had very low false negative rates. Interestingly this increase in performance was not shared by both precision and recall. While recall followed the trend of accuracy and showed almost entirely positive improvements from using DWMV over majority vote, precision did not.

Another interesting finding is that all three performance metrics increased as both the maximum false negative rate and fraction of links seen by workers increased. This implies that as workers answer more problems, and become worse at correctly identifying when items are similar, the benefit of using DWMV over majority vote increases.

Overall, t-tests [9] showed that using DWMV led to a statistically reliable ($p < 0.001$) 0.18% increase in accuracy, a statistically reliable 1.78% ($p < 0.001$) increase in recall, but no statistically reliable ($p = 0.28$) change in precision. While small, these reliable improvements in accuracy and recall over the traditional majority vote method are an indication of the potential positive effects of transitioning to using DWMV instead of majority vote when aggregating crowdsourced opinion.

There were also some interesting differences in how different types of error affected the weights of workers as determined by the DWMV method. Figure 2 shows the average and 95% confidence interval of the DWMV weights of workers as a function of the workers' false positive and false negative rates. The false positive rate of the workers seems to decrease their weight in the final weighted majority vote of the DWMV method much more quickly than their false negative rate. A potential cause of this is that, in the simulated groups of similar items, there were far more pairs of items that were not similar to each other than there were pairs of items that were similar. For example, to have an equal number of items that are similar and not similar to each other, each item would have to be similar to half the items. The only way to facilitate that in the context of this simulation would be to have only two equally sized groups of items. In the simulation there were always at least five groups, and up to 25 groups of similar items, which caused most problems to not being similar to each other. Therefore, when a worker had a large false positive rate, there were more opportunities for them to make a mistake compared to a worker with a large false negative rate. Additionally, the large number of dissimilar problem pairs compared to the number of similar problem pairs caused workers with very low false positive rates to have higher weights than workers with equally low false negative rates, because workers with low false positive rates, regardless of their false negative rates, had much fewer opportunities to make a mistake. These findings suggest that the distribution of correct responses in crowdsourcing tasks affects which type of worker error has a larger impact on workers' weights in the DWMV method.

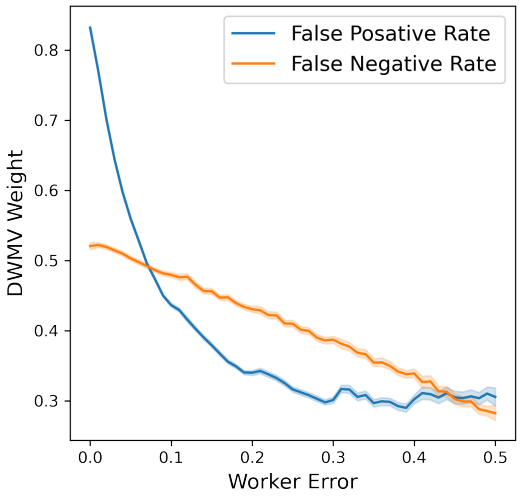

Figure 2: The average and 95% confidence interval of the DWMV weights of workers as a function of the workers' false positive and false negative rates.

4.2 Empirical Study: Similar Problems

In total, six teachers and four students completed the crowdsourcing task of grouping 50 seventh grade mathematics problems. Using each worker's assertions of similarity, the DWMV method and traditional majority vote were used to aggregate the opinions of the workers into a final network of similarity, which was then used to create groups of similar problems using hierarchical clustering. This is the exact same process that was used to form groups in the simulation study. Figure 3 shows the progressive iterations of DWMV. Iteration 1 shows the unweighted average of each workers assertions. The DWMV method's process of iterating between calculating a weight for each worker and calculating the weighted majority vote shifted the weighted average of workers assertions toward the ground truth similarity of problems. This convergence was present in the simulated example in Section 3.1 as well. The benefit of the DWMV method over traditional majority vote lies in this ability to converge towards ground truth. Figure 4 shows the weight of each worker as a function of their error rate. The cohort of middle school mathematics teachers performed much better overall than the cohort of college students. The average accuracy of the teachers was about 97% while the average ac-

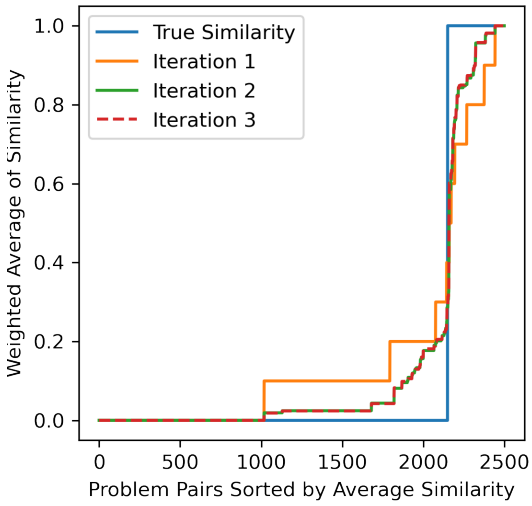

Figure 3: Progressive iterations of DWMV converging on empirical data.

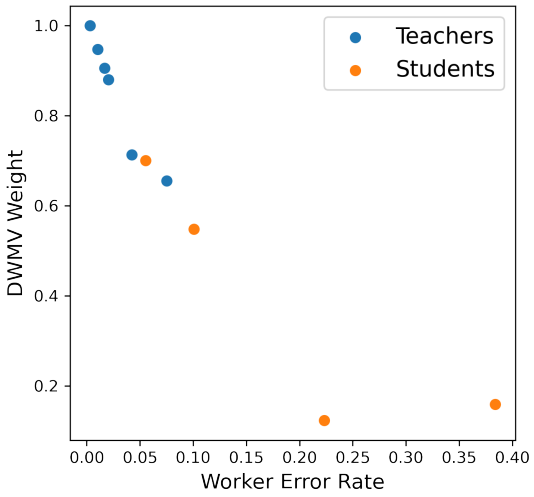

Figure 4: DWMV's confidence in each worker after the DWMV method converged.

curacy of the college students was only about 81%. Based on these weights, it is clear that the DWMV method valued the opinions of middle school mathematics teachers more than the opinions of college students, which is expected given the context and task. While, in this scenario, it might have been easy for a human in the loop to recognize that the teachers' opinions should be valued more, it will not always be the case that one group of workers is clearly more qualified than another group, and thus the DWMV method can help elucidate which workers are the most reliable.

Table 2 shows the difference in accuracy, precision, and recall between groups formed through hierarchical clustering from the assertions of similarity aggregated using DWMV and traditional majority vote. Similar to the simulation results, DWMV had the largest positive impact on recall, the second largest positive impact on accuracy, but no impact on precision. In this empirical study, both the traditional majority vote method and the DWMV method led to perfect precision, meaning all problems that were placed in groups together were similar to each other. However, traditional majority vote led to worse recall than DWMV. When tradition majority vote was used, three of the 50 problems were not placed in a group with any other problems, which is why the recall was so low. However, when DWMV was used, only one problem was not placed in a group of similar problems. This outlier problem, that neither traditional majority vote nor DWMV was able to correctly identify as similar to other problems in its group, had the following text:

> 22% of 65 is 14.3. What is 22.6% of 65? Round your answer to the nearest hundredths (second) decimal place.

Below are examples of problems in the same group as this problem, which were all correctly identified as similar to each other.

> Josiah and Tillery have new jobs at YumYum's Ice Cream Parlor. Josiah is Tillery's manager. In their first year, Josiah will be paid $14 per hour, and Tillery will be paid $7 per hour. They have been told that after every year with the company, they will each be given a raise of $2 per hour. Is the relationship between Josiah's pay and Tillery's pay rate proportional?

> To make a punch, Anna adds 8 ounces of apple juice for every 4 ounces of orange juice. If she uses 32 ounces of apple juice, which proportion can she use to find the number of ounces of orange juice x she should add to make the punch?

> A recent study claimed that in any given month, for every 5 text messages a boy sent or received, a girl sent or received 7 text messages. Is the relationship between the number of text messages sent or received by boys proportional to the number of text messages sent or received by girls?

Although all these problems are related to ratios and proportions, the other problems in the group with the outlier problem are longer word problems that do not explicitly use percentages. The teachers and students whose opinions were crowdsourced could have missed the connection due to the different wording in the problems, or they could believe that calculating percentages is a different skill than calculating proportions from word problems. Based on the differences between this single outlier problem and the other problems in its group, it is possible that the outlier problem was consciously excluded from its group and not simply an oversight.

The impact of using DWMV was larger in this empirical

Table 2: A comparison of majority vote to DWMV used to form groups of similar problems from crowdsourced assertions of similarity.

Metric	Majority Vote	DWMV	% Increase
Accuracy	0.987	0.997	1.054
Precision	1.000	1.000	0.000
Recall	0.903	0.977	8.228

study than it was in the simulation. In the simulation there was a larger than average improvement in accuracy and recall when the workers had very low false positive rates. Given that in this empirical study both sets of groups of similar problems had perfect precision, it is likely that the workers in this study had very low false positive rates, which likely contributed to why the positive impact of using DWMV instead of majority vote was larger in this empirical study than in the simulation as a whole. The results of this empirical study suggest that not only can DWMV out-perform traditional majority vote in simulations, but can also improve the recall and accuracy of groups of similar problems formed from crowdsourced opinions on content similarity in real-life scenarios as well.

5. CONCLUSION

Within online learning platforms and intelligent tutors, there is tremendous utility to knowing what content is similar to other content within the platform, but each application of similar content is likely to have different criteria for what is considered similar. Crowdsourcing opinions on the similarity of content is an accessible way for new applications to recognize similar content. However, crowdsourcing poses some difficulties, namely, how to identify reliable workers and properly aggregate opinions from multiple workers. This work has demonstrated the ability of the Dynamically Weighted Majority Vote method, a novel algorithm for aggregating crowdsourced opinion while rating workers, to accomplish those goals. DWMV has been shown, in both a simulation study and an empirical study, to lead to higher accuracy and recall that the traditional majority vote method on crowdsourcing tasks related to identifying similar content. In the simulation study, using DWMV before identifying groups of similar items through hierarchical clustering resulted in a statistically significant 0.18% increase in accuracy and a 1.78% increase in recall over using majority vote. The simulation study also revealed how the distribution of correct responses in the crowdsourcing tasks effects how the false positive and false negative rates of workers effects their weight in the DWMV method. In the empirical study, using DWMV before identifying groups of similar problems through hierarchical clustering resulted in about a 1% increase in accuracy and an 8% increase in recall over using majority vote, and provided perspective on the differences in accuracy between the expert middle school math teachers and the novice college students. Moving forward, when faced with the need to aggregate crowdsourced opinions, the learning science community can look to the DWMV method as an alternative to the traditional majority vote method. The DWMV method is a promising tool for increasing the reliability of crowdsourced opinion and, when paired with hierarchical clustering, identifying groups of similar content.

6. ACKNOWLEDGMENTS

We would like to thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; a) the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), b) the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), and c) the EIR. We also thank the Office of Naval Research (N00014-18-1-2768), Schmidt Futures, and an anonymous philanthropic foundation.

7. REFERENCES

[1] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

[2] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[3] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[4] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[5] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[6] P. Ruvolo, J. Whitehill, and J. R. Movellan. Exploiting commonality and interaction effects in crowdsourcing tasks using latent factor models. In *Neural Information Processing Systems. Workshop on Crowdsourcing: Theory, Algorithms and Applications*. Citeseer, 2013.

[7] D. M. Stenhoff, B. J. Davey, B. Lignugaris, et al. The effects of choice on assignment completion and percent correct by a high school student with a learning disability. *Education and treatment of Children*, 31(2):203–211, 2008.

[8] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[9] B. L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[10] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, and P. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22:2035–2043, 2009.