Understanding Confidence and Competence in Team Discussions

Andrew Wang, CS ’19 (advised by Professor Danescu-Niculescu-Mizil)

March 5, 2018
1 Abstract

We aim to study group discussions from a computational social science perspective through an online game, StreetCrowd\textsuperscript{1}, that we have created. The game is a multiplayer version of the popular geography game geoguessr\textsuperscript{2}: players are placed in a random location in the world and attempt to locate themselves with the help of their teammates, using clues like words on streetposts and the types of vegetation they see. This game gives us a good example of goal-oriented group discussion with a clear metric of team performance, namely the distance of the team’s final guess from the true location they were placed in. Some questions we have asked in the past are: What linguistic or social network factors cause a team to be constructive or unconstructive? How do more or less confident players interact with their teammates differently? Besides clear applications to improving effectiveness of work groups, these questions are interesting from a social science perspective because team constructiveness is an example of “the whole being greater than the sum of its parts,” i.e. it must be understood not just by looking at the qualities of the individuals involved but by studying their arrangement, on a social network or interactional level.

We will explore questions about the relation between confidence and competence in teams and how these are affected by team discussion. In particular, we are now collecting data on players’ confidences before and after each team discussion, which lets us explore a rich set of questions about how team discussion affects confidence. Thus, we can try to understand what the composition of a team and the language used in a team discussion tell us about the change in the team’s confidence throughout the discussion, and whether this change is justified (i.e. whether it corresponds to the team’s actual correctness). We will also frame our inquiry in terms of a prediction task: based on an observed team discussion, can a machine learning model predict whether the team will become more or less confident, and whether it will become over- or underconfident? Both approaches will bring us closer to improving the productivity of team discussions.

\textsuperscript{1}streetcrowd.us
\textsuperscript{2}geoguessr.com
2 Biographical Sketch

I am a junior from Moraga, California studying computer science and philosophy. At Cornell, I’m an officer of OpenSourceCornell and an e-board member of the Cornell Undergraduate Research Board; I also study piano and have played in chamber music groups on campus. As a freshman, I was eager to study artificial intelligence, but algorithms, networks and philosophy classes helped me see a fascinating theoretical dimension to the social sciences and convinced me to change my course. Incidentally, I’m a huge fan of Bach and Beethoven, masters of polyphony, or the interaction between simultaneous musical voices; (social) interaction is surprisingly fundamental.
3 Statement of Purpose

3.1 Proposed Project

Team discussion is a ubiquitous, decentralized method of aggregating individual answers to a problem in order to create a higher-quality collective answer. Understanding team discussion will allow us to provide recommendations and create tools to enable more productive collaboration, which makes it an essential object of study.

We plan to explore the relation between confidence and competence in teams and how these influence and are influenced by team discussion, as understood through social network and natural language processing (NLP) techniques. We study this relationship by analyzing data from our online game Streetcrowd, which we have created. The game is a multiplayer version of the popular geography game geoguessr: players see a Google StreetView image of a random location in the world along with their teammates, and together attempt to determine their location by communicating through chat, using clues like words on streetposts and the types of vegetation they see. Each round of the game has two phases: in the “solo” phase, each team member individually guesses the location of the image without communicating with others; in the “team” phase, the team members can chat with each other in order to collectively decide on a final answer. Teams are scored by the distance between their final guess and the true location. After the solo phase, we measure confidence by asking each team member to estimate the correctness of his or her guess on a 4-level scale; we ask each individual the same question after the team phase, with respect to the team guess.

As a method of aggregating individual answers to a problem, team discussion could be seen in opposition to the “wisdom of the crowd” method of simply taking the average over many individuals’ answers, which is a surprisingly capable rival: it has been shown in stock markets, elections and quiz shows to produce estimates that can even beat those of individual experts [7]. We have previously shown that in Streetcrowd teams, team discussion can beat the average correctness of individual team members’ guesses, often exceeding the quality of the best individual answer in a team [5]. Indeed, this suggests that team discussion potentially overcomes the problem observed in [3] that revealing information about other individuals’ answers to a problem can undermine the wisdom of a group of people, at least in terms of producing correct answers.

However, our preliminary results suggest that team discussion still suffers from another problem raised in [3]: imagine an intermediate setting between wisdom of the crowd and full team discussion, in which we still take the average of individual estimates, and in which individuals are still not allowed to communicate with each other, but in which each individual can see everyone else’s guesses across successive rounds of guessing. In this setting, seeing information about other people’s beliefs tends to make individuals more confident about their own, even more than is justified based on their level of correctness. Indeed, even in the full team discussion setting, we find that team members become systematically overconfident on average, i.e. that even though people have the ability to express doubts about other people’s beliefs, they underuse this ability. Furthermore, in team discussions, this effect seems to occur regardless of one’s actual level of correctness, and is asymmetric: underconfident team members get a confidence boost through team discussion, but overconfident team members remain at almost exactly the same level. The asymmetry is especially troublesome, since in a wisdom of the crowd setting, we could get an unbiased estimate of a group’s correctness by averaging their individual confidence estimates, but this result entails that a simple average will produce a biased estimate, and correcting this bias requires knowledge of individuals’ correctnesses (which is typically inaccessible in a real-world setting). To my knowledge, this asymmetry has not been articulated or explored in existing literature. Thus, we intend to explore the (mis)alignment of confidence and competence in teams and to build machine learning models that can improve team discussion by detecting when teams become more confident, and whether this confidence change is justified in terms of correctness.

3.2 Related Work

Our work follows a growing line of research studying the performance of work groups, or goal-oriented teams. For example, the social network structure of a team affects performance [1]; more successful team dialogs tend to be more aligned in linguistic structure [6]. We’ve also studied predictors of team constructiveness in Streetcrowd [5], as well as how the confidence of players going into a team discussion affects the discussion [2].

Another interesting line of work to draw from is that studying the wisdom of the crowd with social influence, the “intermediate setting” discussed above, which could be seen as a sort of “silent team discussion”
Figure 1: Individuals who play games in which the team members actually chat tend to be more overconfident than individuals in games without discussion. Note the asymmetry: individuals who were underconfident in the solo round become more (over)confident compared to teams with no discussion, while correctly-confident and overconfident individuals don’t seem to be affected by team discussion.

setting in which group members are exposed to others’ answers but are not allowed to communicate otherwise. Thus, [3] finds that social influence undermines the accuracy and confidence estimates of the wisdom of the crowd estimate, while [4] finds that individuals exposed to others’ answers and confidence estimates tend to gravitate toward both expert answers and answers shared by a large group of less certain people.

3.3 The Emerging Story (Preliminary Results)

3.3.1 Some First Statistics

Note that for each individual in each team game, we have four values: his or her solo-round correctness and confidence, and team-round correctness and confidence (all values measured on a scale from 1 to 4, with 4 being the most correct/confident). We define overconfidence to be one’s confidence minus one’s correctness; hence, each individual has a solo-round overconfidence and a team-round overconfidence. Negative overconfidence corresponds to underconfidence.

We have collected 1180 team games so far, with 3100 corresponding solo-round games. We find that team discussion makes people more confident on average, with the mean change in confidence from solo to team round being 0.3 (≠ 0 with $p < 0.001$). Indeed, this effect persists even when we take into account the fact that teams could have also improved in correctness: the mean change in overconfidence is 0.13 (≠ 0 with $p < 0.001$). Furthermore, the average team-round overconfidence is a modest 0.05 (≠ 0 with $p = 0.08$), but the team-round overconfidence among teams that actually chat is 0.07 (≠ 0 with $p = 0.02$). Indeed, the games in which people (for whatever reason) choose not to actually chat with each other in the team round serve as a good baseline for what result to expect in the absence of any constructive collaboration, and we will take advantage of this as a point of comparison.

3.3.2 Asymmetry in Confidence Alignment

At this point we might wonder if team discussion at least helps align its members’ confidence estimates closer to their true correctnesses. We can formulate two competing hypotheses:

1. Individuals in games with chats, on average, move closer to 0 overconfidence than individuals in games without chats.

2. Individuals in games with chats, on average, increase in confidence more than individuals in games without chats.
The first hypothesis states that individuals on average become better calibrated in their confidence estimates in games with chats. Note that it is not a priori obvious which hypothesis should obtain: since Streetcrowd teams seem to overcome the biasing effects of social influence (posited in [3]) in creating a team answer that is consistently more correct than the average team member’s correctness (as found in previous work), we might suspect that it can overcome the same problem with respect to confidence, since the meta-problem of judging one’s confidence is simply another estimation problem that is being subject to team discussion. Furthermore, we might suspect that in this freer setting in which team members are allowed to communicate with each other, underconfident people might be boosted by agreement from their teammates, while overconfident people will back down when faced with skeptical opposition. On the other hand, social intuition suggests that less confident people could be more easily swayed by more confident people, while the converse is less likely to be true.

To help us unravel this distinction, we now note that taken alone, the overconfidence metric ignores potential boundary effects: for example, an individual with a correctness level of 4 cannot be overconfident by definition. If instead we perform a paired comparison in which we pair people from games in which people actually communicate via chat with people from games in which nobody writes in the chat, making sure the solo-correctnesses, team-correctnesses and solo-confidence levels match up within pairs, we can get a good idea of the effect that actual discussion has on individuals in a team, independent of boundary effects. Performing this pairing, we find that games with discussion have on average 0.18 overconfidence (\( \neq 0 \) with \( p < 0.001 \)), while games without discussion have on average -0.01 overconfidence (\( \neq 0 \) with \( p = 0.85 \)), suggesting that the latter retain the precision of the wisdom of the crowd estimate (since little information is exchanged in these games), while games with discussion suffer from some systematic bias.

The plot in Figure 1 performs this comparison for each possible level of solo-round overconfidence. Doing so, we find an asymmetry that supports hypothesis 2: underconfident individuals tend to become more (over)confident through team discussion, closer to the “correct” confidence level, while overconfident individuals are essentially unaffected by team discussion, compared to overconfident individuals in games without discussion. We might suspect that this effect is tied to the correctness of the individuals involved: for example, people with high confidence but very low correctness would be more willing to reduce their confidence levels. However, stratifying by correctness level (Figure 2), we find that this effect holds across correctness levels.

This finding is worrying for two reasons. First, it suggests that the skeptical voice is underrepresented, at least in Streetcrowd teams; a lack of skepticism could prevent team members from considering other ideas besides the first one on the table. Second, the clear dependence on overconfidence means that this phenomenon is tied to the correctness of the team members: this means that not only can we not simply average team members’ confidence judgments to get an unbiased estimate of the team’s correctness, but also that in order to correct for this systematic bias, we need to factor in each team member’s correctness, a quantity that is typically inaccessible in a real-world setting. This suggests that some kind of manipulation of the rules of the discussion is needed in order to remove this bias. To introduce a skeptical voice, a potential manipulation could be to introduce an adversary, a player whose goal is to thwart the team: if the teammates are alerted that a secret adversary could be among them, they may be less easily trusting of their teammates’ suggestions. To introduce skepticism without actually actively obstructing collaboration, we could even imagine taking this one step further and simply telling the team that there may be an adversary, without actually introducing one; one can imagine further variations on this theme.

### 3.3.3 Understanding the Mechanism

We can take a first stab at identifying the mechanism behind this phenomenon. We based hypothesis 2 on the intuition that individuals' confidence estimates will be swayed towards those of more influential teammates. Thus, we can pose these questions: given a team game, take the most influential teammate and the least influential teammate (where a teammate is more influential if the final team guess is closer to his solo-round guess). Which of the two teammates will move in confidence closer to the opposite teammate’s solo confidence (i.e., ideally, his confidence in the guess that he then proposes to the team)? Which of the two teammates will change more in confidence from solo to team round? We suspect that, independent of the correctness of the team, the more influential teammate will change confidence less, and will move less toward the other teammate’s solo confidence than vice versa. These suspicions are both confirmed by Figure
Figure 2: When we stratify by the correctness of the individual (i.e. in the top left chart, we only consider individuals with solo-round and team-round correctness of 1), we find the same asymmetry: overconfident individuals in chatting games are about as overconfident as individuals in non-chatting games, while underconfident individuals improve more in chatting games.
Figure 3: Less-influential players are swayed towards the confidence estimates of more-influential players, and change more in confidence too.

3. Combined with the fact that people with higher solo-round confidence tend to be more influential, this suggests a picture in which overconfident people lift their less confident teammates up, but less confident people have less sway over their more confident teammates. Thus we have a sense of where to begin studying the mechanism behind the asymmetry.

3.3.4 Looking across Games

Finally, we can take a first stab at understanding how feedback helps people calibrate their confidence levels, namely by tracking individuals as they play several games (receiving feedback on their performance after each game’s team round). We plot individuals’ confidences over time in Figure 4, finding that first-time players are especially susceptible to systematic overconfidence after team discussion, but they quickly calibrate after receiving their first round of feedback. This suggests that there is hope that manipulating the information displayed to team members could correct systematic overconfidence.

3.4 Conclusion

Our study of team discussions is motivated not just by theoretical interest but also by the promise of applications to improving collaboration. Given a growing line of existing work on team performance, our preliminary results and the many strands of inquiry they initiate, this project will easily admit a summer of deep exploration.
Figure 4: Track the solo and team overconfidences of individuals from their very first games onward. We find that first-time players are extremely susceptible to overconfidence after their first team discussion, but quickly learn to calibrate.

References


