

Can We Just Start Over Again? Resetting Remote Team Dynamics

MARK E. WHITING, University of Pennsylvania, USA

MICHAEL S. BERNSTEIN, Stanford University, USA

Interactions defining teamwork today are heavily influenced by constraints and expectations found in in-person teams, however, remote collaboration provides the opportunity to try new ways to make teams work. One foundation of teamwork is persistent identity — we are who we were last time we worked together. Breaking with the expectation of in-person teams, we present a system that affords discontinuous identity using two-way pseudonym masking — enabling teams with new behaviors to arise from the same group of individuals. With this scaffold, a novel family of experiments, comparing the same group across multiple fresh starts, are possible. Further, interventions that involve choosing between versions of the same team are unlocked. We present an overview of experiments and interventions leveraging this system, and propose methods for its broader use in organizations enacting the future of work.

1 INTRODUCTION

Every team is now a remote team, but we need not make them mimic in-person teams. Remote teams can move beyond the limitations of in-person interactions [9]: we can help start teams on the right foot, and we can measure team health in ways that weren't possible before, we can even evaluate biases in team processes in-situ. Here we introduce a technical scaffold enabling team interactions heretofore impossible in person, and we will discuss its use in experiments and in interventions that improve team outcomes.

In in-person interactions we uncontrollably respond to signals of identity in our collaborators [13] — race, gender, status and numerous others are enacted through every interaction, even when we work hard to consciously subdue their effect. Furthermore, our interaction histories are directly tied to those identities too [16] — someone who stymies a team's efforts will not be quickly forgotten. We ask, what if these signals of identity could be temporarily masked?

As the saying goes, "*on the internet, nobody knows your a dog*" — Peter Steiner. In practice, online identity is mediated by the platform and this ability can be instrumented. Manipulating identity online has been used to show that individuals enact behavior of their self perceived identity [20], and that hiring committees make judgements based on signals as subtle as applicant names [10]. However, it also enables the design of experiments where team dynamics are reset and teams can be the unit of analysis — persistence of identity becomes an experimental construct — building on the notion of parallel worlds style experiments [15]. We have introduced a method for manipulating identity in teams through two-way pseudonym masking (Fig. 1) [17, 18], making it possible to convene the same group of people multiple times without them realizing that they have ever interacted before.

In this work we offer initial exploration of what can happen when identity is not presumed consistent within remote teams. We discuss experiments that suggest that team viability — teams willingness to continue interacting — is path dependent and explore how, through identity manipulation, teams can be reset so their most viable states are activated. We also show how this approach to thinking about remote teams enables analysis of the correlates of team viability and



Fig. 1. Mask colors indicate pseudonym identities in each round of an experiment; each round actually consists of the same people. In this depiction, four rounds initialize new parallel identities and one reconvenes identities from a previous round.

show that using this, its possible to predict team outcomes with only a short slice of their text chat interactions. Lastly, we discuss how being able to study teams in this way affords digging deeper into nuanced concepts such as identity born status, potentially unlocking a remote workplace that is less biased across a range of fronts.

2 THE PARALLEL TEAMS SCAFFOLD

In traditional teams studies, teams often receive an individual treatment and are compared with other differently treated teams. With that approach, its impossible to disaggregate the effect of the team membership and the treatment they experienced – only broader claims can be made and many hypotheses can't be empirically isolated. In an attempt to resolve this, we designed an experimental scaffold that allows for repeated interaction of the same team without their knowing it, and, on occasion, reconvening the same team with full knowledge that they are working with the same collaborators (Fig. 1) [17, 18]. This means teams can receive multiple treatments, and interactions can be analyzed within and between subjects as repeated measures. Further, by being able to selectively reconvene one of the earlier teams, more subtle aspects of the in-team situation can be assessed, and critically, interventions that reconvene treated teams are made possible.


The parallel teams scaffold uses a text chat interface, and shows an *adjective-animal* pseudonym for each team member (Fig. 2). To each participant, their own name appears consistent between rounds, however, to all other participants, new pseudonyms are generated each round, so they appear to be new collaborators. Mentions of team member pseudonyms and their common misspellings or abbreviations (e.g. gorila for newGorilla), are automatically replaced by our server to show the appropriate pseudonym for each participants perspective. This way, the identities of other participants can be reset at any round – masked – or can be reenacted by using pseudonyms from a previous round – unmasked.


Round activities are guided by instructions in the chat window. At the end of each round survey questions can be asked as part of an experiment. At the end of all the rounds participants are debriefed about the two-way pseudonym masking manipulation and a survey is given to check if it was effective. Teams for whom the manipulation was not effective can be filtered out before analysis, however, this generally happens in only about 3% of teams. The system is designed to enable recruiting and remunerating participants from online panels such as Amazon Mechanical Turk, or by providing participants direct links to an experiment construct. After an experiment, the system calculates and issues bonus payments for each participant adhering to a recommended pay rate of \$15 per hour [19].


In the the next several sections of this paper, we explore uses of the parallel teams scaffold in experiments and team interventions.

Round 2
Time left: 9:42

Members this round
conventionalHorse (you)
newGorilla
smallBear
youngRhino

 newGorilla 10:03
Hey, nice to meet you all!

 conventionalHorse 10:04
Hi – lets get started.

 youngRhino 10:04
OK, how about we list ideas?

Reply...

Fig. 2. In round 2 of a multi round experiment, three team members have started an activity. Visualizing conventionalHorse’s perspective, all other participants have new identities and appear to be new collaborators.

3 A WINDOW INTO TEAM VIABILITY

Over time many teams in any context lose hope for the long-term sustainability of the team. This is known as being a non-viable team [3, 8], or reaching team fracture [17]. Team viability, though less often the focus of research than team performance, is a critical antecedent to performance [2, 11], but also a correlate of psychological safety [4, 6]. A viable team is likely to perform well and be excited to tackle more problems afterward too, while teams that prioritize performance over all else can come away with internal strife [3].

With the parallel teams scaffold we first studied how consistent the team viability and team fracture scores were for a pair of the same teams across two conditions, when they know they’re the same – unmasked – and when they don’t – masked [17]. We found that the consistency of team viability depends on the kind of task being performed (Fig. 3), drawn McGrath’s task circumplex [12]. Tasks involving problem solving and resolving conflict tended to have the same high consistency masked or unmasked (73–80% consistency). However, a task involving generating ideas went from being comparably predictable when unmasked to being randomly consistent or inconsistent in the masked condition (48%).

This result suggests that team viability is path dependent in some cases, and can be substantially influenced by how the teams starts off. Further, this finding invites interventions targeting teams’ early interactions to scaffold long-lasting teams.

3.1 Predicting team viability

With the interaction data collected with the experiment scaffold just discussed, it became possible to compare the measures of team viability to teams behavior during their collaboration with a machine learning model [1]. Drawing on existing literature on team viability and performance we developed a set of features which could be evaluated from the chat transcripts alone. In other words, like the participants of the interaction, these features would have no external insight about the situation of the collaboration. Some features were computable using standard text analysis libraries, for more sophisticated and conceptual features, we employed Amazon Mechanical Turk workers to score the chat transcripts. We then trained machine learning models to classify if a team would end up in a higher or lower viability state and were able to reach as high as 92% classification accuracy. We also noted that a model trained using only the computationally derived features performed comparably well to the combined model.

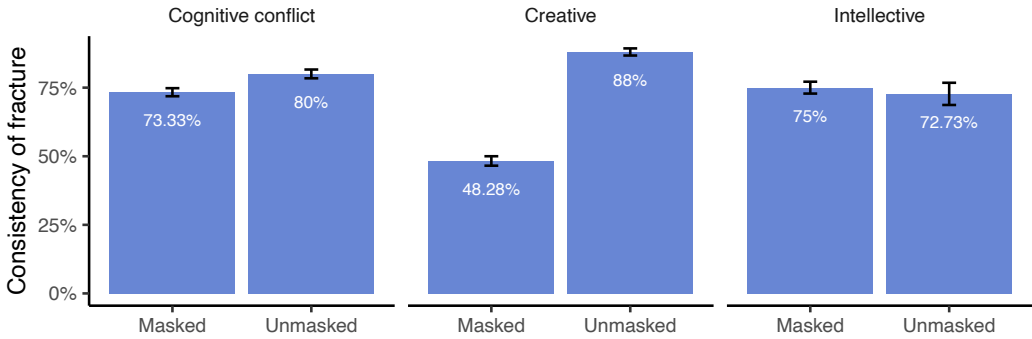


Fig. 3. Between an earlier and later round in the same experiment, unmasked teams were consistent in their fracture outcome 73–88% of the time across all tasks. When comparing masked rounds, teams were polarized by task in the consistency of their fracture outcomes. Teams performing cognitive conflict and intellective tasks were consistent at 73% — essentially the same consistency as unmasked teams. Teams in creative tasks were strongly inconsistent at 48% — essentially as consistent as a coin flip. Reproduced with permission [17].

Computationally derived features provide 2 interesting strengths: 1) they can be evaluated without leaking any sensitive conversation to human raters, and 2) they can be calculated in relative realtime on the chat interactions so far, instead of needing to wait for a human to review the transcripts and report back. Making use of this second advantage, we retrained models on time based snippets of the chat transcripts in our data to see if they could be useful in predicting teams’ future viability classification. We found that with only 70 seconds of a 10 minute interaction we could achieve 90% of the accuracy achieved with the full transcript in identifying the teams in the highest decile of viability (Fig. 4).

Being able to predict a teams’ future viability in realtime could be applied in many collaboration settings, however, understanding that team viability is path dependent makes it clear that doing this with real identities and in an in-person collaborative environment invites potential conflict and even abuse. These potential negative outcomes can be more easily navigated in a remote work setting because identity can be reset using an intervention like the one we’ve introduced.

4 GIVING TEAMS THEIR BEST START

Having found that how teams start dramatically influences how they end, we wondered if knowledge of how they start could be instrumented to search out the best start for a given group of people [18]. To study this, using the same experimental scaffold we allowed teams to collaborate several times in a row with their identities masked each time — several fresh starts in a row. We then reconvened them unmasked with the same identities as they had had in their previous highest viability team — reestablishing their best start. We found that their positive viability persisted and that this was significantly higher than their median viability in other rounds (Fig. 5a). We also showed that the lowest viability version of a team, when reconvened unmasked in the same way, sees improvement but that it remains significantly lower than the high viability pairing with the same individuals (Fig. 5b).

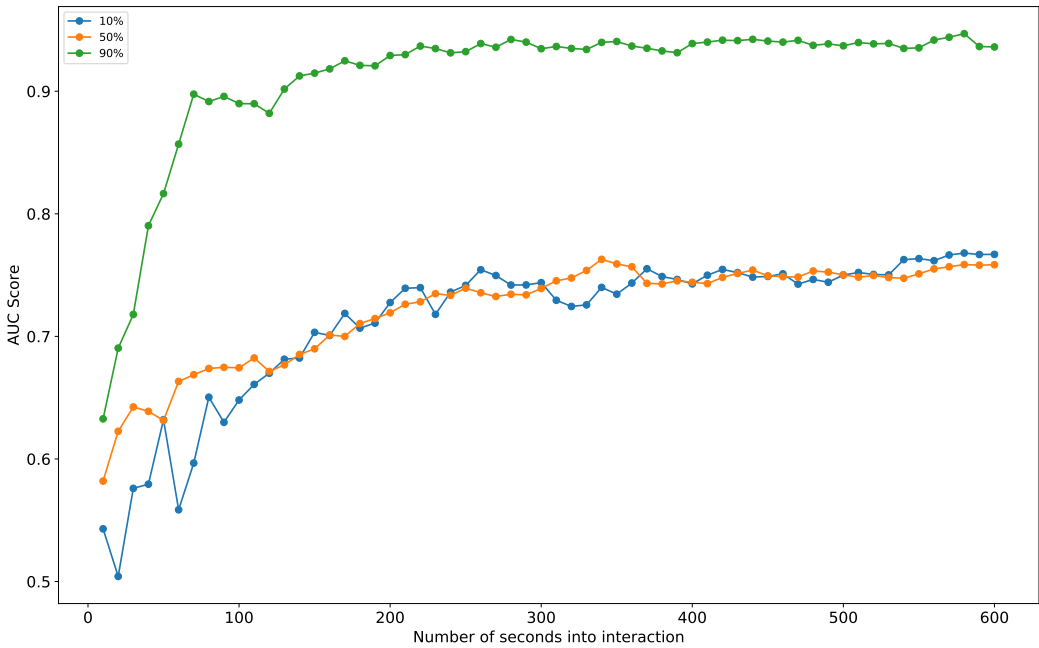
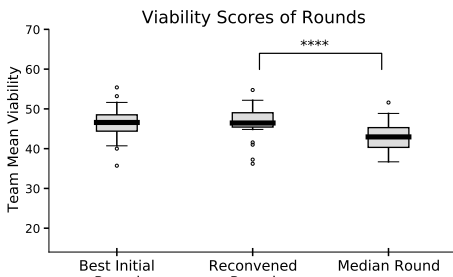
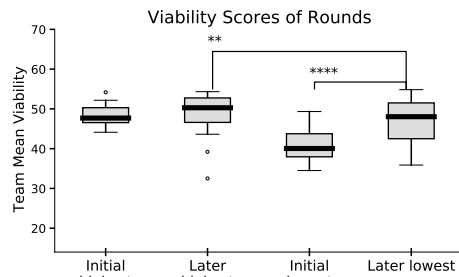


Fig. 4. Evaluating classification accuracy with a logistic regression model using features extracted from subsets of 10 minute (600 second) chat interactions. Top decile, bottom decile and median split classifications were performed. In the top decile, 90% of the accuracy is attainable with the first 70 seconds of interaction. Reproduced with permission [1].



(a) Reconvening high viability teams



(b) Reconvening highest and lowest viability teams

Fig. 5. After repeated masked rounds the highest viability team (left) or the highest and lowest (right) were reconvened unmasked. Viability in unmasked rounds was at least as high as the round they were reconvening. Comparing the team with the lowest mean viability with its respective reconvened round, a substantial increase in mean viability is apparent. However, the reconvened low mean viability parallel worlds do not outperform either the initial or reconvened high viability ones. Reproduced with permission [18].

This implies that, if masking was used to initiate strong teams, it could then be applied again to reify the teams interaction histories and viability level, to help start that group of people out on the best foot possible.

While this kind of intermediation may be challenging in a in-person work world, in one where remote work is common, the path dependence of team viability and fracture offers an intervention that can be used to find the best version of a group of collaborators, and can reset the group if problematic norms emerge.

5 DEVELOPING STATUS IN TEAMS

In online interaction, individual identity is a dominant initial signal used to establish a status hierarchy among collaborators — aspects of identity such as gender, age, and race dramatically influence perceived status [13]. In addition to this, behavior in the group is the main alternative signal of status [5, 14]. By using 2-way pseudonym masking in a context where gender, age and race are not provided as signals, the impact of behavior born status enactment can be studied. In ongoing work, we are using this technique to understand if, without overt identity signals, high status is emergent from the team interactions or is predefined by behavioral tendencies of individuals. In other words, would the same leader (high-status) individual arise every time the group restarts? Our initial results show that teams who don't know they're working with the same people again are more likely to select new members as the highest status individual than teams where people see the same pseudonyms for their collaborators.

These initial results suggest that using masking can help remove latent power imbalances in remote team activities, but also open the door to a broader range of analysis and potential interventions for extant imbalances like racism.

6 CURRENT IMPLICATIONS

We've shown that teams capacity for viability is not limited to how they currently act — if identities could be reset, they might be more viable, and if not, with the right intervention, its possible to reestablish the old ones. However, instrumenting such interventions in traditional workplaces is challenging. In remote work settings there are more chances for using a tool like this while also minimizing potential harmful uses. One clear context is in the forming of teams and hiring processes where the expectation of having long term persistent identity are not present. Our results suggest identity based interventions could be informative for selecting for team viability in these contexts, but these tools may also make new kinds of team forming exercises possible.

Making predictions about teams' viability is also ripe for use in contexts where a substantial portion of their communication is conducted via chat or messages, as is the case in many remote companies today. Though this work is at an early stage, the approach could be used to help structure and refine team composition, e.g., selecting for viability, but also team processes, e.g., endorsing team interactions that improve the health of the team.

However, these opportunities also introduce new complexities to managing teams and as a consequence, we recommend careful application of these techniques. For example, viability is a group outcome, not one directly driven by individuals, as far as the theory suggests at this time. If a viability prediction tool consistently suggested that teams with a particular person on them had low viability, some might conclude that the individual in question was causing this outcome but the reality is generally more nuanced and interdependent on broader membership [7]. Further, to avoid targeted attacks that might identify particular individuals, the prediction tools we are releasing do not afford evaluation of one person conversations, and require a minimum amount of data before drawing conclusions [1]. Another kind of malicious use may train such a system on biased samples of team communication (e.g., all white male teams), meaning that teams with different communication habits would receive inaccurate assessments [21]. More experimentation is required to fully understand how these tools might deal with longer term team interactions, larger scale teams, and those doing high stakes work.

6.1 Research opportunities

This work has studied viability with an experimental scaffold that deliberately mediates how people perceive interactions. The scaffold takes advantage of design affordances available to online and remote interaction that are challenging or impossible to harness in the offline world. We note that moving forward, there are substantially more opportunities to study viability in contexts like these, but there are also exciting opportunities to study other areas too, such as group decision making, and perceived identity based injustices. Further, there are other advantageous affordances in the world of online interactions which can serve as starting points for new kinds of experiments and interventions that have not been possible until now. A particular area which we hope will become an active area of study in relation to this is better understanding how mediating identity may impact behavior in the long term.

7 CONCLUSION

Team viability is a key aspect of keeping teams functional and its more important than ever in an online world, where usual social cues are easily misunderstood. However, the online world also invites new forms of research, new kinds of interventions, and new early warnings about viability. In this work we have briefly outlined initial findings using an experimental scaffold to intermediate identity within online teams, we show that there are immediate ways to reach higher viability when establishing new online teams, and that there are ways to judge the viability of a team in relative realtime, inviting measures to course correct.

ACKNOWLEDGMENTS

We are grateful for the contributions of Melissa Valentine, numerous research assistants, Amazon Mechanical Turk worker participants and Upwork research engineers making this work possible. This line of work was supported by the Stanford Data Science Initiative, RISE Thailand Consortium, the Hasso Plattner Design Thinking Research Program, the Office of Naval Research (N00014-16-1-2894) and a National Science Foundation award IIS-1351131.

REFERENCES

- [1] 2020. (2020). Under review.
- [2] S. T. Bell. 2007. Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology* 92, 3 (2007), 595–615.
- [3] Suzanne T Bell and Brian J Marentette. 2011. Team viability for long-term and ongoing organizational teams. *Organizational Psychology Review* 1, 4 (2011), 275–292.
- [4] Bret Bradley, Bennett Postlethwaite, Anthony Klotz, Maria Hamdani, and Kenneth Brown. 2011. Reaping the Benefits of Task Conflict in Teams: The Critical Role of Team Psychological Safety Climate. *The Journal of applied psychology* 97 (07 2011), 151–8.
- [5] Shelley J. Correll, Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. It's the Conventional Thought That Counts: How Third-Order Inference Produces Status Advantage. *American Sociological Review* 82, 2 (2017), 297–327.
- [6] Amy Edmondson. 1999. Psychological safety and learning behavior in teams. *Administrative Science Quarterly* 44 (01 1999), 250–282.
- [7] J Richard Hackman. 2011. *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers.
- [8] J. Richard Hackman and Nancy Katz. 2010. Group Behavior and Performance. *Handbook of Social Psychology* (2010), 1208–1251.
- [9] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 119–125.
- [10] Eden B King, Saaid A Mendoza, Juan M Madera, Mikki R Hebl, and Jennifer L Knight. 2006. What's in a name? A multiracial investigation of the role of occupational stereotypes in selection decisions. *Journal of Applied Social Psychology* 36, 5 (2006), 1145–1159.

- [11] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *The Academy of Management Review* 26, 3 (2001), 356–376. <http://www.jstor.org/stable/259182>
- [12] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall, Englewood Cliffs, NJ.
- [13] Cecilia L Ridgeway, Elizabeth Heger Boyle, Kathy J Kuipers, and Dawn T Robinson. 1998. How do status beliefs develop? The role of resources and interactional experience. *American Sociological Review* (1998), 331–350.
- [14] Cecilia L Ridgeway and Shelley J Correll. 2006. Consensus and the creation of status beliefs. *Social Forces* 85, 1 (2006), 431–453.
- [15] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.
- [16] Daniel M Wegner. 1987. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*. Springer, New York, NY, 185–208.
- [17] Mark E Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S Bernstein. 2019. Did It Have To End This Way? Understanding the Consistency of Team Fracture. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
- [18] Mark E Whiting, Irena Gao, Michelle Xing, Junior Diarrassouba N’Godjigui, Tonya Nguyen, and Michael S Bernstein. 2020. Parallel Worlds: Repeated Initializations of the Same Team To Improve Team Viability. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 67 (May 2020), 23 pages. <https://doi.org/10.1145/3392877>
- [19] Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 197–206.
- [20] Nick Yee and Jeremy Bailenson. 2007. The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research* 33, 3 (2007), 271–290.
- [21] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it’s time to make it fair.